

Table of Contents

EXECUTIVE SUMMARY	2
INTRODUCTION	3
<i>Background</i>	3
<i>Aims</i>	4
<i>Technology</i>	5
<i>Structure</i>	6
DATA	7
<i>Data Selection</i>	7
<i>Exploratory Analysis</i>	15
METHODOLOGY	21
<i>Selection</i>	21
<i>Pre-Processing</i>	22
<i>Transformation</i>	26
ANALYSIS	41
<i>Data Mining</i>	41
RESULTS	43
<i>Correlation Matrix</i>	43
<i>Correlation Matrix with Significance Levels (p-value)</i>	43
<i>Correlogram</i>	44
<i>Rpart Decision Tree</i>	45
<i>Random Forest</i>	48
<i>Model Predictions</i>	49
<i>Best County Analysis</i>	50
<i>Best County Testing</i>	52
CONCLUSIONS	53
FURTHER DEVELOPMENT OR RESEARCH	54
REFERENCES	55
APPENDICES	57
PROJECT PLAN	57
REFLECTIVE JOURNALS	58
<i>Introduction</i>	59
<i>Journal</i>	60
OTHER MATERIALS USED.	68
APPENDIX 3	69
APPENDIX 4	70
APPENDIX 5	71
APPENDIX 6	74
APPENDIX 7	75
APPENDIX 8	75
APPENDIX 9	76
APPENDIX 10	77
APPENDIX 11	78
APPENDIX 12	79
APPENDIX 13	79
APPENDIX 14	80
APPENDIX 15	81
APPENDIX 16	82
APPENDIX 17	82
APPENDIX 18	83
APPENDIX 19	85
APPENDIX 20	86
APPENDIX 21	86
APPENDIX 22	87
APPENDIX 23	87
APPENDIX 24	88
PROJECT PROPOSAL – UPDATED DECEMBER 2020	89

Executive Summary

The Covid-19 pandemic has led to several big employers like Indeed, Siemens, Twitter, Salesforce & Spotify now allowing their employees to work remotely on a more permanent basis. As more companies recognise the benefits of large-scale remote working (both for the employee & employer), this list will grow as companies move from forced remote working to smarter working in a post-Covid-19 environment. Large numbers of employees now no longer must live in Dublin near their employer and are looking for alternative locations to call home.

The project aim was to identify which county in the Republic of Ireland would offer the best quality of life using variables that a typical family would consider. These variables include crime rate, classroom size, property prices / monthly rent costs, distance to an emergency department et cetera. These datasets are the most current available.

The project also compares the actual number of crimes versus the predicted number of crimes using decision trees and random forest algorithms for a sample of counties. The purpose of this was to determine if the eight independent variables are sufficient to predict crime rates. The random forest proved to be the most accurate but not accurate enough to be considered a good model.

The results from the analysis and associated charts/graphs show the top counties a user should consider for relocation. When using all nine variables in making the decision, the county with the best quality of life is Dublin. This prediction is not a true reflection of the best county, as including all nine variables is unrealistic. A more realistic scenario is that only selected variables would be needed for individual analysis. This option is explored in the results section using a 'R' Shiny application.

The 'R' Shiny application has been developed to allow users to pick the variables they would like to include in the analysis. The county with the best quality of life will be computed. Additionally, the application also provides summaries per county and per variable.

Introduction

Background

Before Covid-19, it was the norm to live in Dublin (or the commuter belt), near your Dublin based employer. E-working or remote working was a factor for some employers, with a 2018 Blueface report finding that 78% of Irish companies had a remote working policy in place. (78% of Irish businesses now have a Remote Working Policy in place Technology, news for Ireland, Ireland, Technology, 2020)

However, to look further into this, a research paper published in 2019 by the Department of Business, Enterprise, and Innovation (Department of Business, Enterprise, and Innovation, 2019) was reviewed. This research paper investigated what defines remote working and concluded it was either when employees worked from their homes or worked from a hub close to or within their local community. The report discusses a pilot survey that the Central Statistics Office (CSO) undertook in 2018. This pilot survey found 18% of respondents worked from home, mostly one or two days per week.

A Remote Work in Ireland Employee Survey was undertaken after this report addresses the lack of data around employee participation in remote working. However, the sample was considered skewed by a high response rate from the Finance and ICT sectors. Nevertheless, while the survey is not fully representative and likely overstates remote work, it does offer valuable insights. Some of these insights include:

- Remote working is more common in the private sector (63%) than in the public sector (28%).
- 48.5% of the respondents said they worked remotely.
- Working remotely every week (part of a working week) was most common at 51.1% compared to only 25.1% for the private sector & 10.1 % for the public sector who work remotely daily (every day)

The arrival of Covid-19 had a sudden and dramatic impact on remote working in Ireland. From March 16th, 2020, almost 100% of office-based work moved to remote working for both the private and public sector.

Many companies have now started to see benefits in allowing their staff to work remotely on a more permanent basis, and these include Indeed (Indeed to allow 'vast majority' of Irish employees to work from home forever, 2020) and Siemens (Kelly, 2020). They have both opted to let their office-based workforce work remotely on a more permanent basis, as an example.

This project was appealing as between April 2020 and August 2020, colleagues and family friends relocated out of Dublin without changing jobs. At least one of these moves was to get away from the high rents of Dublin, and Covid-19 provided the opportunity to keep their current employment and move to a better location. Interestingly, the decision to move developed quickly and how many other people are thinking like this, and where could they move?

Aims

In a post-Covid-19 environment, there will be a notable increase in the number of employees availing of remote working in Ireland. With this project, the aim was to identify, using the datasets selected, which county offers the best quality of life if a move from Dublin (or anywhere else) is an option for those staff choosing to work remotely. The project delivered by preparing the datasets, merging the data into a single data set, and then normalising the data to make it comparable. Without any weighting, the county with the best combination of variables was identified and then using only selected variables (this was to simulate a real-world scenario, where not every variable will be necessary to all readers).

In addition to this, the aim was to compare a selected variable from sample counties to the output from a Random Forest and decision tree algorithms. The crime data was chosen as the dependent variable that the models will predict. The crime variable was selected as of all the variables, there will be readers that would not have an interest in the other variables (e.g., short/medium term relocation would not be interested in property prices as they will be renting, couples with no children would not be interested in classroom sizes), but crime has the potential to affect all.

The report will then compare the model prediction to the actual data to indicate how the test counties are compared to the model prediction. The report will also compare the decision tree and Random Forest results to identify which algorithm offers the best prediction.

The report also includes a correlation heatmap to view correlations between the variables.

Technology

The primary language used for the analysis is the R programming language, with various libraries within R required. The project proposal described the R programming language for the pre-processing & transformation and then switching to Python for the analysis. Before beginning the project, some further research was conducted, including a comparison on how to do selected analysis in both languages using "Comparative Approaches to Using R and Python for Statistical Data Analysis" (Rui and Vera, 2017), and a paper "MatLab vs Python vs R ". (Colliau et al., 2017) After reviewing this a decision was made adopt the R programme language for the whole project. R Studio is the integrated development environment (IDE) of choice when using R, so this is the IDE that I will use. The R packages used are:

- Dplyr: Data manipulation.
- Ggplot2: Data visualisation.
- Readxl: Read Excel files.
- Tidy: To tidy messy data.
- Dbplyr: Interact with databases.
- Rvest: Web Scraper.
- Amelia: Missmap function to visualize missing data.
- Fuzzyjoin: join data frames in R.
- Hmisc: compute the significance levels.
- PerformanceAnalytics: to display a chart of a correlation matrix.
- Rconnect: Deploy and manage Shiny applications.
- Shiny: Interactive visualizations.
- Shinythemes: Alter the overall appearance of Shiny applications.
- Rpart: Decision Tree algorithm.
- randomForest: Random Forest algorithm.
- MLmetrics: Calculate MAPE (mean absolute percentage error).
- Corrplot: Graphical representation of a correlation matrix.
- SP: Spatial data classes and methods (required for maps).
- Rgeos: Interface to Geometry Engine(required for maps).
- Maptools: Tools for Handling Spatial Objects.
- SQLdf: Manipulate data frames using SQL.

Having reviewed several database options, the data was stored on an SQLite database, and the associated R package (RSQLite) was used with the R programming language. SQLite was chosen because its feature best suits the project. As a serverless application, it resides in a single file, making it easy to move and share. It is also ideal for projects with a low amount of requests, which this project will have.

For the data mining element of the project both a decision tree and Random Forest algorithms were implemented. The decision tree algorithm used is the Recursive Partitioning (rpart) package in 'R'. The most commonly used decision tree algorithm is the C5.0 algorithm, but this is used primarily for classification, whereas 'rpart' can be used either in regression or classification. Random forest was added to this analysis as it is particularly well-suited to a small sample size, which this analysis has.

For the data visualisations, the project includes the Shiny package in R. This package will allow interactive charts and graphs to allow end-users to interact with the data. In addition to the programming of the Shiny application, HTML and CSS languages were used to manipulate the presentation of the HTML pages that the Shiny Application generates.

For the administrative side of the project, products from the Microsoft Office suite were used. These include MS Project to develop and track the project timetable for the duration of the project, MS OneNote to keep a record of keynotes and data references identified as the project progressed, MS Word to produce this report and MS PowerPoint to create any slide packs required to deliver mid-term and end of year presentations. MS Excel was used to evaluate the datasets as they are acquired and as a second source when completing any evaluation of the outputs from R Studio.

Some of the datasets are sourced from the CSO, and their default format is PX. The PX format is a standard format for statistical files used by statistical offices around the world. A library in R (PC-Axis) can be used to read px files into R. PC-Axis was installed, and correct syntax followed for the library 'pxR'. However, when trying to merge this data with other data, errors occurred that were unresolvable. Following some research, an alternative method to using this data was identified and the decision to install the Px Win software package developed by the Swedish statistics office (Statistics Sweden, 2016) to read and convert the data to an alternative format. The converted files were saved to '.CSV' for convenience, which allowed R to read the data and merge it with other data without any issues.

Structure

The following is a brief overview of the document's structure and what is addressed in each section.

Methodology: An overview of the methodology applied to the project, plus summary details of each of the different stages that the project went through to get from raw datasets to completed analysis, including the transformation tasks required to get the data from raw to the required state.

Data Selection:	Provides a summary of each of the datasets used in the project, including their attributes, format of the raw dataset, and source of each dataset.
Transformation:	Details on the steps taken to transform each dataset from its raw state to the final state, ready to be used in the analysis.
Analysis:	An overview of the approaches included in analysing the data and why they were chosen. The report will also provide details on the essential features of the analysis.
Results:	Present the Interpretation and evaluation of the analysis.
Conclusions:	Describe the advantages/disadvantages, strengths and limitations of the project and its outcomes.
Future:	Details of further development or Research that could be provided with additional time and resources.

Data

This section is divided into two. The first section presents the raw data, including a description, its attributes, and its source. The second section will detail the exploratory data analysis.

Data Selection

This section provides a summary of each of the datasets used in the project. If a dataset is not used in the project or the dataset used is different from planned, the blockers encountered and why they were not overcome have been documented.

Name:	Recorded Crime Offences Under Reservation (Number) by Garda Station
Description:	This dataset is available in the '.PX' format. The dataset contains a count of offences for the years 2003 – 2019 by the type of offence, Garda Station location, and regional division.
Attributes:	The raw dataset attributes are:

File Format	CSV (initially .PX)
Data Format	Structured
Number of Columns	19 (All Chr)
Number of records (rows)	6,770

Source: This data is sourced from Irelands open data portal and is published by the CSO. (CJA07 - Recorded Crime Offences Under Reservation (Number) by Garda Station, Type of Offence and Year - data.gov.ie, 2020)

Name: Mainstream Primary Schools by Class Size

Description: This dataset is available in the '.PX' format. The dataset contains the number of children in every mainstream primary school class for all schools in the Republic of Ireland (the number of classes varies from 1 to 40 per school). The data is for the 2019 / 2020 school year and is summarised by the school.

Attributes: The raw dataset attributes are:

File Format	CSV (initially .PX)
Data Format	Structured
Number of Columns	49 (8 x Chr & 41 int)
Number of records (rows)	3,106

Source: This dataset is sourced from Irelands open data portal and is published by the CSO. (ED121 - Mainstream Primary Schools by Class Size, Teacher Size of School, Year and Statistic - data.gov.ie)

Name: E-Car charger Network

Description: The original source of this data was from Open Charge Map. (Open Charge Map - The global public registry of electric vehicle charging locations) Their dataset is Accessed via an API. However, the API is capped at 500 records, but there are many more chargers in the republic of Ireland. They do not have a premium API offering to obtain greater than 500 records, and the only filter option their API has is country.

Contact was made with ESB E-cars, who very nicely supplied their current list of all chargers on their network in Ireland. The list is an immaculate list with charger name, address and charging speed. The problem is that ESB only accounts for approximately 65% of the public charging network. ('EasyGo | Charging Network', 2021) The other prominent provider in Ireland is EasyGo. When contacted, they were helpful, pointing to their website, which has a map (and list) of their charging network and the other providers.

The data on their site is not downloadable, so the first attempt was to use a web scraper to extract the data using the 'rvest' package in R.

The next problem encountered was that the list of chargers displayed on the screen reflected the area visible on the accompanying map. The effect of this feature meant a web scraper could not be used to extract the data, as each time the site opens, it uses IP's address's current location to set the centre of the map, and if other areas of the country are required, the map must be moved manually. The second option was to zoom the map out to a province and copy the list displayed into an excel sheet. This approach worked, but the extracted data from the site needed some work to get it into a usable state (see pre-processing in the next section of the report).

Attributes: The raw dataset attributes are:

File Format	CSV
Data Format	Semi-Structured
Number of Columns	1 Chr
Number of records (rows)	2,826 (5 per charger location)

Source: This dataset is extracted from the website of a privately owned, public charging network operator. ('EasyGo | Charging Network', 2021)

Name: Traffic Collisions and Casualties by County

Description: This data is sourced from Irelands open data portal and is published by the Road Safety Authority. This dataset contains a single line summary per

county with the number of collisions (both fatal & injury) and the number of casualties (both fatal & injury).

Attributes: The raw dataset attributes are:

File Format	CSV
Data Format	Structured
Number of Columns	7 (1 x Chr & 6 x int)
Number of records (rows)	26 (1 per county)

Source: (ROA27 - Traffic Collisions and Casualties by County, Year and Statistic - data.gov.ie, 2020)

Name: Average Earnings

Description: This dataset was deemed unnecessary as the premise of the analysis is based on remote working, so earnings based on the county of residence are not valid for the analysis.

Name: Broadband Speed

Description: Not for the lack of trying, but this dataset has been excluded from the analysis as a source of this data was unavailable. Efforts were made to contact comparison sites to obtain data, but the best that could be obtained was a nationwide speed average. Contact with National Broadband Ireland (NBI) was also attempted, they are responsible for the roll-out of the national broadband plan, but no reply has been received at the time of writing. The alternative was to use a dataset over four years old, which a decision was made not to do as a lot has changed in the broadband domain in that period.

Name: Property Sale Prices

Description: The dataset contains the sale price of every property sale (Jan – Nov 2020) in the Republic of Ireland. This dataset was most recently updated on December 9th, 2020.

Attributes: The dataset attributes are:

File Format	CSV
Data Format	Structured
Number of Columns	9 (all Chr)
Number of records (rows)	40,353

Source: This data is published by the Property Services Regulatory Authority. (Property Services Regulatory Authority, 2020)

Name: Transport – Travel Times

Description: It was decided not to add this dataset to the analysis as the topic centres around remote working, and travel times would not be a factor.

Name: Weather

Description: This dataset was not pursued as it adds no value to the analysis, and no additional insight would be gained from using it.

Name: Tourist Attractions

Description: This dataset is available as an API GET call. The data set is loaded directly into a data frame in R from the API. The dataset contains the name and address of every tourist attraction registered with Failte Ireland.

Attributes: The dataset attributes are:

File Format	API / Link to source
Data Format	Structured
Number of Columns	8 (6 x Chr & 2 x num)
Number of records (rows)	3,324

Source: This dataset is sourced from Failte Ireland (Fáilte Ireland)

Name: Monthly Rental Costs

Description: This dataset reports on the average rent by location (e.g., town or area) and county for the republic of Ireland. This dataset was most recently updated on December 1st, 2020, with Q2 2020 data.

Attributes: The dataset attributes are:

File Format	CSV
Data Format	Structured
Number of Columns	7 (6 x chr & 1 x num)
Number of records (rows)	447

Source: This data is published by the Residential Tenancies Board (Residential Tenancies Board, 2020) and was accessed via the CSO website.

Name: Outpatient Waiting Lists

Description: The dataset contains the count of patients on an outpatient waiting list on different dates during 2020. The most recent data will be used in the analysis. The hospital summarises the data and contains a count by speciality, adult/child grouping, age profile & time band (waiting time in months).

Attributes: The dataset attributes are:

File Format	CSV
Data Format	Structured
Number of Columns	10 (all Chr)
Number of records (rows)	55,969

Source: This data is published by The National Treatment Purchase Fund (OP Waiting List by Group Hospital - OP Waiting List by Group Hospital 2020 - data.gov.ie, 2020) and was accessed via Irelands open data portal. This dataset was updated on August 27th, 2020.

Name: Air Quality

Description: A full nationwide dataset on this subject was unobtainable, so Air Quality has been removed from the analysis. The best data set I could locate was from the EPA, but this was only available for cities.

Name: Registered Pharmacies

Description: The dataset contains the count of pharmacies by county, with Dublin split into 25 (Co Dublin, Dublin 1 et cetera). The data was last updated on December 1st, 2020.

Attributes: The dataset attributes are:

File Format	CSV
Data Format	Structured
Number of Columns	2 (1 x Chr & 1 x int)
Number of records (rows)	50

Source: This dataset is published by the Pharmaceutical Society of Ireland and available Irelands open data portal. (PSI Registered Pharmacies - December 2020 - data.gov.ie, 2020)

Name: Population count and density by county

Description: This dataset is for use in conjunction with other datasets to create additional measures that can be used in the analysis.

Source: This dataset is sourced from Wikipedia ('List of Irish counties by population', 2020), with the population data confirmed by checking against the CSO dataset. (Population at Each Census 1841 to 2016, 2020)

Name: Average Distance to Emergency Hospitals at ED Level

Description: This data was used in the CSO report, "Measuring Distance to Everyday Services in Ireland". For emergency departments, the shortest-path analysis was performed on hospitals where adult emergency care is provided. It was last updated on February 17th, 2020.

This data set was added to the project after the outpatients waiting list data set had to be excluded following the transformation tasks, which identified no data for some counties in the Republic of Ireland.

Attributes: The dataset attributes are:

File Format	CSV
Data Format	Structured
Number of Columns	12 (4 x Chr & 8 x int)
Number of records (rows)	3,410

Source: It is available via the geohive site. (Ordnance Survey Ireland, 2020)

Exploratory Analysis

The exploratory data analysis (EDA) will allow me to glimpse the general characteristics of the dataset, which can be attained by generating descriptive statistics and data visualisations.

This analysis will be completed on the raw data sets to help summarise the data, find outliers/anomalies, and identify interesting patterns. This analysis will also help identify what actions need to be undertaken as part of the data transformation stage.

Crime Rate

The total number of reported crimes (n=564) averaged 393 (s=971) per garda station, per the 2019 data. The median number of crimes per Garda station is 63, indicating that the data set is positively skewed (a more significant number of low numbers). Figure 1 shows the spread by Garda Station and the high number of garda stations with low numbers.

Of the ten Garda Station with the highest numbers, seven are in Dublin, with Pearse Street been the highest (10,210 crimes).

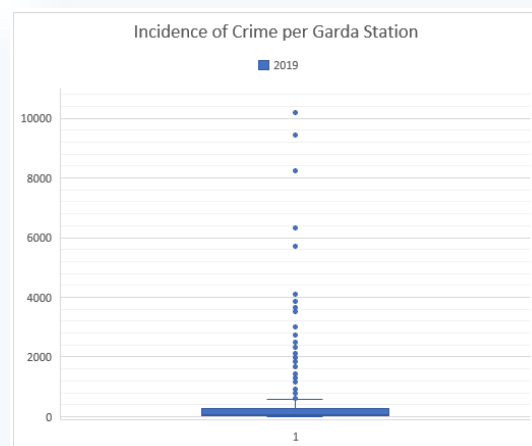


Figure 1

Classroom Size

Some exploratory analysis using a 'ggplot2' column chart in 'R' shows students per county (Appendix 14) using the totals column. All counties have some data, and there is no blank / missing data. A check using 'summary' in R to verify that every school had at least one classroom with students (*Figure 2*) was also completed.

The number of students (n=3,106) averaged 22 (s=5) per class in the current school year. The median number of students per class is 22, indicating that the data set has a symmetrical distribution. The histogram shown in Appendix 14 reflects this.

```
Class.1
Min.   : 1.00
1st Qu.:18.00
Median :23.00
Mean   :22.19
3rd Qu.:27.00
Max.   :40.00
```

Figure 2

E-car chargers

The exploratory analysis for this data set was completed post-pre-processing and transformation as the raw data was not in a condition that it could be analysed in any way. The number of electric car chargers (n=26) averaged 49 (s=64) per county. The median number of chargers per county is 49, indicating that the data set is positively skewed (greater number of low numbers).

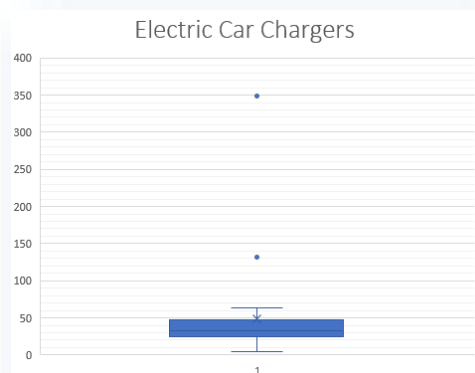


Figure 3

The bulk of the counties fall in the region of 24 chargers – 47 chargers (*Figure 3*), with only two outliers (Cork & Dublin). A map chart can be seen in Appendix 3.

Road Accident Casualties

Some exploratory analysis shows the number of casualties per county (*Appendix 4*). All counties have some data, and there is no blank / missing data. Dublin stands out as the county with a large number of casualties. The data will be normalised later to the number of casualties per 100k population.

The number of casualties (fatal & injured) (n=26) averaged 305 (s=428) per county. The range is large at 2,231, with one very high outlier causing this (Figure 4).

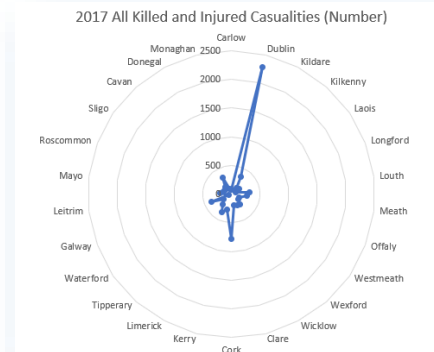


Figure 4

Property sale prices

Exploratory data analysis completed with a histogram (R ggplot2) identifies a small number of outlier sales over €1m (~1%). The outliers can be seen by looking at a histogram of all sales (Appendix 5: All Property Sales). A second histogram with a more focused perspective looking solely at the sales above €1m shows that a small number of the overall sales are for a value greater than 1 million euro (Appendix 5).

Additionally, sales at the other end of the scale were also investigated. There are a small number of sales below €50k also (Appendix 5). Using Google to research some of the below €20k sales (full address of the property is available on the register) and some appear to be residents buying out the local council's interest in a shared ownership scheme, I cannot find any apparent reason for the others.

A decision was made to remove sales below €20k and above €1m from the dataset as they will impact the average, but realistically very few properties will be sold in these two brackets.

The final nationwide population histogram shows a clear picture of the property sales in 2020 (Appendix 5) and using a facet wrap in ggplot to look at this by each county to make sure good content can be seen.

Before removing the sales below €20k and above €1m, the sold property price (n=40,353) averaged €310k (s=€1.1m). After removing the sales below €20k and above €1m, the sold property price (n=39,512) averaged €265k (s=€160k). Interestingly, the mode of this sales data is €150k (505 sales). The data set has a symmetrical distribution (skewness =1.3).

Attractions

```
> sapply(Attractions,function(x) sum(is.na(x)))
```

Name	Url	Telephone	Longitude	Latitude	AddressRegion	AddressLocality	AddressCountry
0	928	546	1	1	3	3	3

Figure 5

Exploratory data analysis completed with a bar chart (R ggplot2) to count the number of attractions by the county has identified that there are a small number of attractions where the address does not include a county (*Appendix 6*) and confirmed by counting NAs in R (*Figure 5*).

These will need to be resolved in the transformation of this dataset before moving on.

Rental Costs

Initial exploratory analysis shows that there are 47 locations with no average rental value (see *appendix 9 & figure 6*)

```
sapply(Initial_rent_DF,function(x) sum(is.na(x)))
```

i..Statistic	Quarter	Number.of.Bedrooms	Property.Type	Location	UNIT	VALUE
0	0	0	0	0	0	47

Figure 6

The monthly rental amount (n=447) averaged €1,193 (s=€484) across all locations and all property types. The median rental amount is €1,111, indicating that the data set has a symmetrical distribution (skewness = 0.52). The histogram shown in *Appendix 9* reflects this.

Outpatient waiting list

Initial review of data shows only one missing record (*Figure 7*) from this large dataset. This missing record is in the 'Time Bands' column, which I will investigate and update if required, as I will use this column to measure the average wait time.

```
> sapply(initial_df,function(x) sum(is.na(x)))
```

i..Archive_Date	Group	Hospital_HIPE	Hospital	Specialty_HIPE1	Speciality	Adult_Child	Age_Profile	Time_Bands	Total
0	0	0	0	0	0	0	0	1	0

Figure 7

The dataset has multiple records for each key date, and as only the most recent data will be used in the analysis, all other key dates will be removed from the dataset. In addition to this, it has been decided to split this data set into multiple data sets:

- Adult - Waiting time
- Adult - Count on waiting list
- Child - Waiting time
- Child - Count on waiting list

Sum of Total	Column Labels		
Row Labels	Adult	Child	Grand Total
0-3 Months	117513	15450	132963
3-6 Months	61087	9444	70531
6-9 Months	76848	13196	90044
9-12 Months	64641	9715	74356
12-15 Months	46939	7215	54154
15-18 Months	37676	6641	44317
18 Months +	121598	23033	144631
Grand Total	526302	84694	610996

Figure 8

An initial summary of the raw data shows that there are 611k patients on the outpatient waiting list on 27/08/2020, and ~14% of these are children (Figure 8). There will be further analysis done on this data set after transformation.

Registered Pharmacies

Exploratory data analysis completed with a bar chart (R ggplot2) to count the number of registered pharmacies by the county has identified that County Dublin data has been split into different areas. (Appendix 10). These will need to be resolved in the transformation of this dataset before moving on. There is also a “Grand Total” dimension that will also need to be removed.

Population

Census 2016 population data were used to derive new measures as well as normalise data sets. The total population is 4.7m, and a county by count breakdown can be seen in Figure 9.

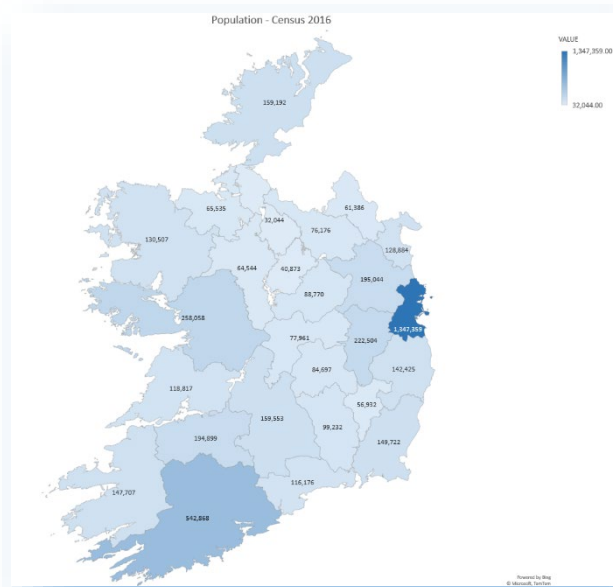


Figure 9

Distance to an Emergency Department

A review of the data in R shows that there are no missing values (Figure 10).

```
> #check for NA's
> sapply(ED_df,function(x) sum(is.na(x)))
```

FID	OBJECTID	CSOED_3409	EDNAME	COUNTY_31	County	Shape__Area	Shape__Length	DistanceRange	Distance_Rank	SHAPE_Length	SHAPE_Area
0	0	0	0	0	0	0	0	0	0	0	0

Figure 10

Data is by location, with 3,409 location across the 26 counties. Distance to an ED is in a range format, and this will need to be converted to a kilometre value in the transformation stage.

Location	Count
Less than 5km	397
5 - < 10km	263
10 - < 25km	918
25 - < 50km	1410
50km or more	421
Grand Total	3409

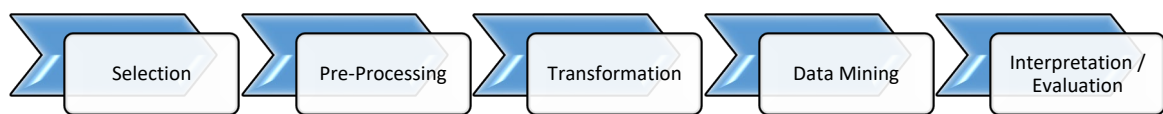
Figure 11

An initial count shows that the 25km – 50km range is the most common (Figure 11). Further analysis will be completed after the data is transformed.

Methodology

The methodology I used for this project is KDD (Knowledge, Discovery & Data Mining). This methodology is a core data analytics methodology and is best suited to this data analysis project. As KDD is an iterative process, outcomes can be refined, and conclusions can be enhanced as the data is transformed, allowing for more relevant results. This iterative process will benefit the project as each of the datasets will be evaluated initially and again as they are merged into a larger dataset and can be tweaked or modified to allow for better integration into the analysis.

The five stages of KDD are:



In addition to this, a published journal on Domain Knowledge in the initial stages of KDD (Szczyka et al., 2014) highlights the benefits of the overall analysis by increasing domain knowledge on the data subjects. With this in mind, there is time incorporated into the project data selection stage to study the data to be familiar with the data subjects before proceeding.

Selection

To identifying the datasets for the analysis, the first task was to identify the question that was to be answered.

“If moving from Dublin (or anywhere else),
which county would offer the best quality of life?”

The next step was to look at what data would help answer this question, focusing on datasets that a typical family would consider noteworthy. The resulting list of twenty datasets was further whittled down to fifteen after some high-level research. The remaining datasets (Figure 12) are, in my opinion, datasets that would add the best value to my analysis and would be of interest to the report readers.

Full details of each data set can be found in Data Selection (section 2) of this report.

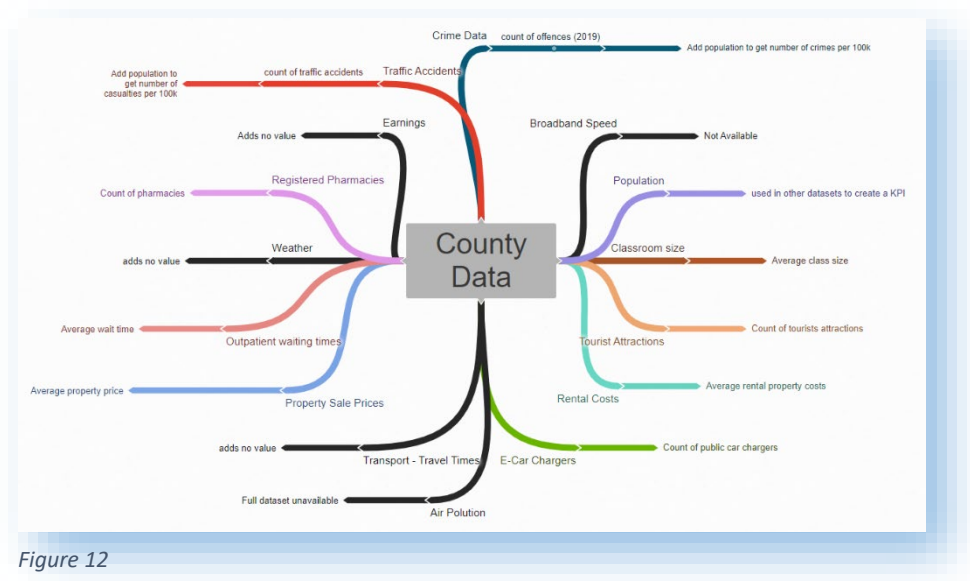


Figure 12

Pre-Processing

The pre-processing stage consisted of cleaning and preparing the raw data to obtain consistent, tidy data. Having read Hadley Wickham's paper on tidy data (Wickham, 2014) and applied the four main principles to my pre-processing:

1. Each attribute (variable) should be in its own column.
2. Each observation should be in a different row.
3. One table per topic
4. When using multiple tables, they should have a column in common (primary key)

As most of the data was clean when acquired, there was minimal cleansing, although some restructuring was required. An individual R file for each data set made it easier to manage and save the processed data to the database when complete.

This section provides details of the pre-processing work carried out on each dataset before beginning the transformation work.

The following data sets were sourced from the CSO via Irelands Open Data Portal (Data.gov.ie, no date) in the '.PX' format.

- Recorded Crime Offences Under Reservation (Number) by Garda Station
- Mainstream Primary Schools by Class Size

This format requires the PxWin software package developed by the Swedish statistics office (Statistics Sweden, 2016) to read and convert to alternative formats. The software has converted these files to '.CSV' for convenience.

Below is a high-level summary of the actions taken to tidy the data for each dataset:

Data Set	Details of pre-processing
Recorded Crime Offences	▪ Check for missing data using Missmap (see Appendix 8) & supply (sum NAs) ▪ Remove crime data for years 2003 – 2018 as only data for 2019 was used in the analysis. ▪ Some Garda divisions are spread over two counties, and there was a requirement to split these into their respective counties Details below (2)
Primary Schools by Class Size	▪ Check for missing data on key fields and at least one classroom for each school using Missmap (Appendix 14) & supply (sum NAs) ▪ Remove unrequired columns from the data set. • Roll Number • Academic year • School Address (excluding county) • Eircode • Local Authority
E-Car charger Network	▪ Please next paragraph (1) below for details on conversion from raw data ▪ Check for missing data using Missmap (Appendix 3) & supply (sum NAs)
Traffic Accident Casualties	▪ Check for missing data using Missmap (Appendix 4) & supply (sum NAs)
Property Sale Prices	▪ Check for missing data using Missmap & supply (sum NAs) (Appendix 5) ▪ The price variable was in a character format containing '€' & ',' symbols (e.g., €120,000.00). These needed to be removed using lapply and gsub functions in R. ▪ Variable converted to an integer. ▪ Unwanted columns were removed. • Date of Sale • Address • Postal Code • Property Description • Size Description

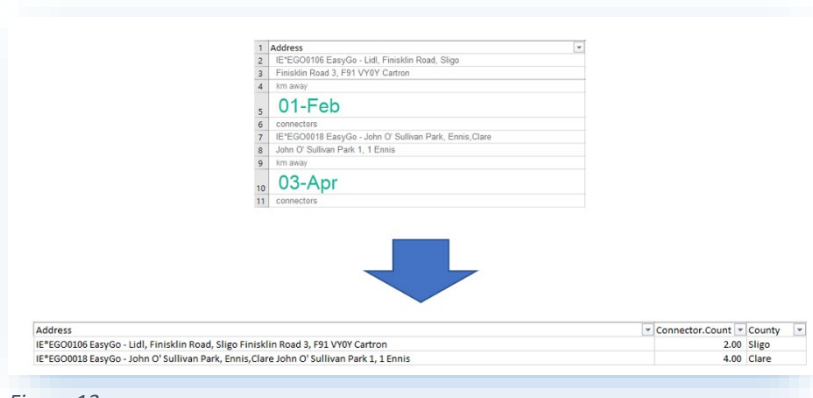
Data Set	Details of pre-processing
Tourist Attractions	▪ Check for missing data using Missmap & supply (sum NAs) (Appendix 6) • 3 x attractions with missing county details. • 'which(is.na) function used to find indexes of missing counties. • Google attraction to find the county of attraction. • Added county to the data set. ▪ Remove unwanted columns. • Name • URL • Telephone • Longitude & Latitude • Address Locality & Country
Monthly Rental Costs	▪ Check for missing data using Missmap & supply (sum NAs) (Appendix 9) • Found observations (47) with no average monthly rent cost
Outpatient Waiting Lists	▪ Check for missing data using Missmap & supply (sum NAs) (Appendix 11) • 1 Time band was found to be missing. • 'which(is.na) function used to find indexes of missing time band. • When viewed, this was for a key date other than the date I planned to use – ignored. ▪ Removed observations for all archive dates other than the most recent (27/08/2020)
Registered Pharmacies	▪ Check for missing data using Missmap & supply (sum NAs) (Appendix 10). ▪ Convert X variable (Count of pharmacies) to integer.
Population counts and density by county	▪ Remove the ',' symbol from the population variable and convert it to an integer. ▪ Check for missing data using Missmap & supply (sum NAs) (Appendix 12) ▪ Convert variables to an integer. ▪ Remove Northern Ireland counties from the data set. ▪ Rename columns.
Average Distance to Emergency Hospitals	▪ Check for missing data using Missmap & supply (sum NAs) (Appendix 13) ▪ Removed unwanted columns. • FID & ObjectID • ED Name • Shape Area & Length

(1) As mentioned in the above data selection section, the E-Car charging network data was troublesome to get. After contacting a couple of sources, the EasyGo website was identified as having a map (and list) of their charging network and the other providers in Ireland. However, extracting this data was not straight forward.

The first attempt to obtain the data was using the 'rvest' library in R to scrape the page, but this failed to read the list because the table has been dynamically generated based on the area of Ireland shown in the accompanying map. That was verified by using the CSS Selector Gadget Chrome addon.

The solution was to copy the displayed table in four sections (one for each province as this was the most significant area that the map could zoom out to). Each section was then concatenated together in MS Excel.

The resulting list in Excel was all in a single column, with five rows per charger location. A combination of lookups, concatenate, index & match and substitute formulae allowed the data to be converted into a usable table (Figure 13).



Address	Connector.Count	County
IE*EG00106 EasyGo - Lidl, Finisklin Road, Sligo	2.00	Sligo
IE*EG00018 EasyGo - John O' Sullivan Park, Ennis, Clare	4.00	Clare

Figure 13

(2) To complete the analysis, the data set had to be modified as the sum of crimes is by Garda Station, including the Garda station location and division. There are 115 garda stations (total = 564) that their county could not quickly be identified as their division is spread out over two counties.

Division

Laois/Offaly Division
 Cavan/Monaghan Division
 Roscommon/Longford Division
 Sligo/Leitrim Division
 Kilkenny/Carlow Division
 Laois/Offaly Division

An attempt was made to source a dataset/list of garda stations and their complete address, but the CSO dataset containing this has been discontinued. The garda website was also checked, but this uses a map feature or a search by division option. The data on Wikipedia was examined as a possible source but was found to be incomplete.

The solution was to create a list in 'R' that could be used to look up the details needed (sourced from Wikipedia & Google) and merged it with my dataset using an inner join function in 'R'. A new subset was then created, which uses the amended division (single county).

Transformation

For the project, transformation involved converting each dataset from its raw state (post-pre-processing) to an appropriate state to perform my analysis. This transformation included the alignment of each dataset to a universal unit of measure (County) and ensuring the required metric is correctly calculated before the datasets were merged. Additional measures were also calculated to normalise the data, plus each data set was ranked.

The tasks required were identified in the earlier data pre-processing stage and the exploratory analysis for each data set. After completion, each data set was saved to the SQLite database.

Crime Data

Garda.Station	Type.of.Offence	X2003	X2004	X2005	X2006	X2007	X2008	X2009	X2010	X2011	X2012	X2013	X2014	X2015	X2016	X2017	X2018	X2019
Abbeyfeale, Limerick Division	03 Attempts/threats to murder, assaults, harassments and r...	18	25	38	25	41	46	45	22	28	45	22	19	19	14	12	20	20
Abbeyfeale, Limerick Division	04 Dangerous or negligent acts	14	12	19	26	15	16	12	20	6	6	7	9	9	6	10	11	21
Abbeyfeale, Limerick Division	05 Kidnapping and related offences	0	0	1	0	0	0	1	0	0	0	0	0	0	0	1	0	0
Abbeyfeale, Limerick Division	06 Robbery, extortion and hijacking offences	0	0	1	0	0	3	2	0	0	0	1	0	0	1	1	0	0
Abbeyfeale, Limerick Division	07 Burglary and related offences	27	21	24	33	28	28	30	25	19	16	30	36	27	15	14	23	23

Figure 14

Figure 14 shows a snapshot of the raw data before the transformation. The transformation of the crime dataset into the required state, the following tasks were completed:

- Using aggregate, sum up the count of offences by Garda station.
- Using amended division column, identify which county each Garda station is located.
 - Create a vector that contains each county.
 - Use a 'For' loop to search through the Division column and use 'grepl' to compare counties' vector to find the appropriate county (Figure 15).

```

62 #create "County" column filled with NA's
63 workingCounty <- NA
64
65 # The data in a table
66 pop_df <- dbReadTable(db, "population")
67
68 #create vector with list of counties
69 county <- as.vector(pop_df$county)
70
71 #For loop to check "Garda.Station" column for "county" (using grepl)
72 for(i in 1:26){
73   workingCounty[i] <- ifelse(grepl(county[i], workingGarda.Station, ignore.case = T), county[i],
74                               ifelse(grepl("D.M.R.", workingGarda.Station, ignore.case = T), "Dublin", workingCounty))
75 }

```

Figure 15

- Using aggregate, sum up the count of offences by county.
- Data merged with population dataset using 'inner join' function.
- The number of crimes per 100k population calculated.
- Normalise function applied to crimes per 100k population (Appendix 17).
- The rank column created referencing normalised data.

Classroom size

Roll Number	Academic Year	Official Name	Address.1	Address.2	County	Eircode	Local Authority	Class.1	Class.2	Class.3	Class.4	Class.5	Class.6	Class.7	Class.8
00651R	2019	BORRIS MXD N S	Lower Main Street	Borris	Carlow	R95RH6F	Carlow County Council	21	18	26	30	31	31	28	30
00977B	2019	BALLYCONNELL N S	Ballyconnell	Tulow	Carlow	R93VA44	Carlow County Council	23	20	24	29	30	18	14	N/A
01116A	2019	BAILE AN CHUILINN N S	Balinkillen	Muine Bheag	Carlow	R21Y803	Carlow County Council	18	23	25	28	26	N/A	N/A	N/A
01215C	2019	NEWTOWN DUNLECKNEY MXD	Newtown	Muinebheag	Carlow	R21P714	Carlow County Council	29	27	30	32	28	N/A	N/A	N/A
01415K	2019	RATHOE NS	Rathoe	Co. Carlow	Carlow	R93X118	Carlow County Council	18	14	19	30	26	20	21	17

Figure 16

The transformation of the classroom data from raw (Figure 16 – before pre-processing) into the required state, the following tasks were completed:

- Calculate the average classroom size per school – using 'rowMeans' in R.
- Calculate the average classroom size per county - using 'aggregate' in R.
- Normalise function applied to average class size (Appendix 17).
- The rank column created referencing normalised data.

To validate the calculation performed in R, MS Excel was used to generate a pivot table (Figure 17) calculating the average classroom size per school and compared this to the R output for a sample population. This validation found no errors.

County	Province	Average Class Size
Mayo	Connacht	19.09
Roscommon	Connacht	19.98
Clare	Munster	20.08
Donegal	Ulster	20.17
Silgo	Connacht	20.50
Leitrim	Connacht	20.86
Kerry	Munster	20.94
Galway	Connacht	21.03
Longford	Leinster	21.08
Laois	Leinster	22.18
Tipperary	Munster	22.25
Limerick	Munster	22.40
Cavan	Ulster	22.41
Cork	Munster	22.45
Westmeath	Leinster	22.64
Monaghan	Ulster	22.84
Offaly	Leinster	22.90
Kilkenny	Leinster	23.12
Waterford	Munster	23.38
Wexford	Leinster	23.58
Dublin	Leinster	23.61
Wicklow	Leinster	23.69
Carlow	Leinster	23.80
Louth	Leinster	24.16
Kildare	Leinster	24.20
Meath	Leinster	24.65
Grand Total		22.23

Figure 17

Interestingly, all counties in Leinster except two (Longford & Laois) have an average classroom size more significant than the national average.

E-car chargers

The transformation of the electric car charger data into the necessary state required the following tasks were completed:

- Calculate the total number of car chargers per county - using 'aggregate' in R.
- Using the population data set, merge kilometres square size for each county
- Calculate the number of chargers per 100 km².
- Normalise function applied to chargers per 100 km² (Appendix 17).
- The rank column created referencing normalised data.

Traffic Accidents Casualties

County	X2017.Fatal.Collisions.Number	X2017.Injury.Collisions.Number	X2017.All.Fatal.and.Injury.Collisions.Number	X2017.Killed.Casualties.Number	X2017.Injured.Casualties.Number	X2017.All.Killed.and.Injured.Casualties.Number
Carlow	3	55	58	3	77	80
Dublin	24	1939	1963	24	2258	2282
Kildare	6	244	250	6	339	345
Kilkenny	5	92	97	5	137	142
Laois	2	111	113	2	169	171

Figure 18

The transformation of the traffic accidents data from raw (Figure 18) into the necessary state, the following tasks were completed:

- Remove all columns other than “All killed & injured casualties”.
- Merge with population data using the 'inner join' function.
- Calculate the number of casualties per 100k population.
- Normalise function applied to casualties per 100k (Appendix 17).
- The rank column created referencing normalised data.

A summary in 'R' after data transformation shows that the average casualties (includes killed and injured) per county is 305. When this is normalised, the average per county is 172 per 100k population (Figure 19).

```
> summary(casualties_df_final)
      County      Casualties      Population      Casualties_per_100k
Length:26      Min.   : 51.0      Min.   : 32044      Min.   :140.0
Class :character 1st Qu.: 143.0      1st Qu.: 76622      1st Qu.:148.5
Mode  :character Median : 199.5      Median : 123851     Median :163.0
              Mean  : 305.2      Mean  : 183149     Mean  :171.8
              3rd Qu.: 307.8      3rd Qu.: 159463     3rd Qu.:193.2
              Max.   :2282.0      Max.   :1347359     Max.   :245.0
```

Figure 19

Some further exploratory analysis after transformation using a 'ggplot2' bar chart, we can see which counties are above the average and below the average. Interestingly, Dublin is the county closest to the average (Appendix 4).

Property Sale prices

	Date of Sale, dd.mm.yyyy	Address	Postal Code	County	Price...	Not Full Market Price	VAT Exclusive	Description of Property	Property Size/Description
1	01/01/2020	APT 1 LEE HOUSE, STRAND LANE, THE QUAY, CARRICK ON ...		Tipperary	€120,000.00	No	No	Second-Hand Dwelling house /Apartment	
2	01/01/2020	ROSSACOSANE, TEMPLENOE, KENMARE		Kerry	€450,000.00	No	No	Second-Hand Dwelling house /Apartment	
3	02/01/2020	105 DONAGHMORE, THE ANCHORAGE, BETTYSTOWN		Meath	€125,000.00	No	No	Second-Hand Dwelling house /Apartment	
4	02/01/2020	109 Cedar Walk, Castlesods, Dublin Road		Carlow	€211,410.00	No	Yes	New Dwelling house /Apartment	
5	02/01/2020	11 OAK GLADE COURT, BLESSINGTON RD, NAAS		Kildare	€204,000.00	No	No	Second-Hand Dwelling house /Apartment	
6	02/01/2020	12 WHITECLIFF, WHITECHURCH RD, RATHFARNHAM DUBL...	Dublin 16	Dublin	€565,000.00	No	No	Second-Hand Dwelling house /Apartment	
7	02/01/2020	125 MILL TOWN HALL, MT ST ANNES, MILLTOWN	Dublin 6	Dublin	€588,500.00	No	No	Second-Hand Dwelling house /Apartment	
8	02/01/2020	12 BLOOMING, THE HARBOUR, SANDHOLME		Sligo	€18,000.00	No	No	Second-Hand Dwelling house /Apartment	

Figure 20

The transformation of the property sale price data (*Figure 20 shows a sample of raw data*) into the requisite state needed the following tasks to be performed.

The exploratory analysis identified some outliers that will be removed to make the data more useable, and these include all sales above €1m and all sales below €20k.

- Remove outliers identified in exploratory analysis. Figure 21 is a summary post removal of the outliers.
- Calculate the average sale price per county using the aggregate function.
- Format value to currency as a new column
- Remove decimal places.
- Normalise function applied to the average sale price (*Appendix 17*).
- The rank column created referencing normalised data.

Price	
Min.	: 20000
1st Qu.	: 154000
Median	: 238000
Mean	: 264945
3rd Qu.	: 339206
Max.	: 1000000

Figure 21

Following the transformation of this dataset, the final output from a ggplot2 bar chart can be seen in appendix 5: Average property sale price per county.

Attractions

	Name	Url	Telephone	Longitude	Latitude	AddressRegion	AddressLocality	AddressCountry
1785	'On The Nail' Literary Readings	http://onthenailreadings.blogspot.com	+353(0)872996409	-8.623180	52.66720	Limerick	Limerick City	Republic of Ireland
2962	'Star Wars' The Skellig Islands & Ring of Kerry - Railtours Ire...	http://railtoursireland.com/train-tour/star-wars-skellig-mich...	+353(0) 18560045	-6.250291	53.35100	Dublin	Dublin City	Republic of Ireland
1678	'A Rural Experience' Day Tours	http://www.arualexperience.com	+353(0)56727590	-7.074240	52.62980	Kilkenny	Gowran	Republic of Ireland
3071	11 Day Discover Ireland Tour - Vagabond Tours of Ireland	http://vagabondtoursofireland.com/tour/vacations-to-irelan...	+353(0) 14428559	-6.260205	53.34931	Dublin	Dublin City	Republic of Ireland
3072	12 Day Giant Irish Adventure Tour - Vagabond Tours of Irela...	http://vagabondtoursofireland.com/tour/irish-tours-12-day-...	+353(0) 14428559	-6.260205	53.34931	Dublin	Dublin City	Republic of Ireland
2717	12 Henrietta Street	http://thedrawingroom.ie/	+353(0)862746153	-6.270576	53.35246	Dublin	Dublin City	Republic of Ireland
2894	126 Artist-Run Gallery	http://www.126.ie	+353(0)864491366	-9.049473	53.27218	Galway	Galway City	Republic of Ireland
3284	14 Henrietta Street	http://facebook.com/14HenriettaStreetDublin	+353(0) 15240363	-6.270211	53.35251	Dublin	Dublin	Republic of Ireland
2834	1916 Freedom Tour	http://www.1916tour.ie	+353(0) 15311916	-6.250364	53.33914	Dublin	Dublin City	Republic of Ireland

Figure 22

The transformation of the attraction's dataset into the required state (Figure 22 – sample of the raw data set) required that the following tasks to be completed.

- The number of offences by county was calculated using the aggregate function.
- Using the population data set, merge kilometres square size for each county
- Calculate the number of attractions per 100 km².
- Normalise function applied to attractions per 100 km² (*Appendix 17*).
- The rank column created referencing normalised data.

Post transformation bar chart can be seen in Appendix 7: Count of Failte Ireland Attractions per County

Rent Costs

	T.Statistic	Quarter	Number.of.Bedrooms	Property.Type	Location	UNIT	VALUE
1	RTB Average Monthly Rent Report	2020Q2	All bedrooms	All property types	Carlow	Euro	821.37
2	RTB Average Monthly Rent Report	2020Q2	All bedrooms	All property types	Carlow Town	Euro	861.23
3	RTB Average Monthly Rent Report	2020Q2	All bedrooms	All property types	Graigucullen, Carlow	Euro	845.71
4	RTB Average Monthly Rent Report	2020Q2	All bedrooms	All property types	Tullow, Carlow	Euro	778.77
5	RTB Average Monthly Rent Report	2020Q2	All bedrooms	All property types	Cavan	Euro	637.08
6	RTB Average Monthly Rent Report	2020Q2	All bedrooms	All property types	Cavan Town	Euro	659.54

Figure 23

The transformation of the rental price dataset into the required state (*Figure 23 – sample of the raw data set*) required the following tasks to be completed. The exploratory analysis identified that 47 observations have no average rent. In reviewing the locations without a value, there are other locations in the affected counties with values (e.g., Waterford has 21 locations, including an overall average for Waterford on the report). The easiest solution is to remove all locations except the overall county average.

- Remove observations for all but overall county average values (*Figure 24*)

- Create a vector that contains each county.
- Use a 'For' loop to compare counties' vector to find the appropriate county.
- The new county column was then used to compare to the location column and remove any observations where the two columns did not match.

```
# The data in a table
pop_df <- dbReadTable(db, "population")

# create vector with a list of counties from population table
county <- as.vector(pop_df$county)

# Create new County column
Initial_rent_DF$county <- NA

for(i in 1:26){
  Initial_rent_DF$county <- ifelse(grep(county[i], Initial_rent_DF$location,
                                     ignore.case = T), county[i],
                                Initial_rent_DF$county)
}

# remove rows if county != location
final_rent_DF <- Initial_rent_DF[Initial_rent_DF$location == Initial_rent_DF$county,]
```

Figure 24

- Remove columns other than county and rent.
- Copy value column to new column “average” and convert to currency (for use in charts)
- Normalise function applied to average rent cost (*Appendix 17*).
- The rank column created referencing normalised data.

Following the data transformation, a ‘ggplot’ bar chart shows the average rental value per month per county, with a horizontal line added to show the nationwide average (*Appendix 9*). Dublin has the highest rental cost, with Laois the closest country to the average.

Outpatient Waiting Lists

	Archive_Date	Group	Hospital_HIPE	Hospital	Specialty_HIPE1	Specialty	Adult_Child	Age_Profile	Time_Bands	Total
1	30/01/2020	Children's Health Ireland	0	Children's Health Ireland	601	Paediatric ENT	Child	0-15	15-18 Months	604
2	30/01/2020	Children's Health Ireland	0	Children's Health Ireland	601	Paediatric ENT	Child	0-15	18 Months +	1420
3	30/01/2020	Children's Health Ireland	0	Children's Health Ireland	1302	Paediatric Neurology	Child	0-15	0-3 Months	346
4	30/01/2020	Children's Health Ireland	0	Children's Health Ireland	1302	Paediatric Neurology	Child	16-64	0-3 Months	8
5	30/01/2020	Children's Health Ireland	0	Children's Health Ireland	1302	Paediatric Neurology	Child	16-64	3-6 Months	4
6	30/01/2020	Children's Health Ireland	0	Children's Health Ireland	1302	Paediatric Neurology	Child	16-64	6-9 Months	7
7	30/01/2020	Children's Health Ireland	0	Children's Health Ireland	1302	Paediatric Neurology	Child	16-64	9-12 Months	4
8	30/01/2020	Children's Health Ireland	0	Children's Health Ireland	1302	Paediatric Neurology	Child	16-64	12-15 Months	3
9	30/01/2020	Children's Health Ireland	0	Children's Health Ireland	1402	Paediatric Neurosurgery	Child	16-64	15-18 Months	1
10	30/01/2020	Children's Health Ireland	0	Children's Health Ireland	1700	Ophthalmology	Child	0-15	0-3 Months	662
11	30/01/2020	Children's Health Ireland	0	Children's Health Ireland	1700	Ophthalmology	Child	16-64	0-3 Months	11
12	30/01/2020	Children's Health Ireland	0	Children's Health Ireland	1700	Ophthalmology	Child	16-64	3-6 Months	5
13	30/01/2020	Children's Health Ireland	0	Children's Health Ireland	1700	Ophthalmology	Child	16-64	6-9 Months	4
14	30/01/2020	Children's Health Ireland	0	Children's Health Ireland	1700	Ophthalmology	Child	16-64	9-12 Months	11

Figure 25

The transformation of the outpatients waiting list dataset into the necessary state (Figure 25 shows a sample of the raw data set) required the following tasks to be completed. The exploratory analysis identified no missing data that needs to be populated.

- Merge with hospital addresses dataset to identify the county of each hospital.
 - There are 44 hospitals included in the data set but there is no address or location data. A 'CSV' file with the county for each for each hospital was constructed and merged this with the main data using the 'inner-join' function in R.
- Add the number of months to the dataset.
 - A lookup dataset using the Time_Bands variable and the median number of months for each referenced period was developed (Figure 26).

```
#create new dataset to use in estimation of wait time
lookup <- data.frame(Time_Bands = c("0-3 Months", "3-6 Months",
                                   "6-9 Months", "9-12 Months",
                                   "12-15 Months", "15-18 Months",
                                   "18 Months +"),
                     Months = c(1.5, 4.5, 7.5, 10.5, 13.5, 16.5, 20),
                     stringsAsFactors = TRUE)
```

Figure 26

- Merge the lookup with the main data frame using the 'inner join' function in R.
- Create two subsets, one for the adult population and a second for the child population.
- Calculate the average wait time for both child and adult patients per county using the aggregate function.
- Calculate the total number of both child and adult patients on the waiting lists per county using the aggregate function.

The outcome for the outpatient dataset was that it had been split into four separate data sets.

- Average waiting time for a child patient (months)
- Average waiting time for a child patient (months)

- Number of adult patients on the queue
- Number of child patients on the queue

To check if it was suitable to use all four datasets in the project or if the means of both (waiting time & queue) were the same, a test on the two pairs of data sets was performed.

- The first set of tests was the waiting times' datasets.
 - Step 1
 - State the Hypothesis.
 - $H_0: \mu_{\text{Adult_Waiting_Time}} = \mu_{\text{Child_Waiting_Time}}$
 - We will only need to use one of the data sets.
 - $H_1: \mu_{\text{Adult_Waiting_Time}} \neq \mu_{\text{Child_Waiting_Time}}$
 - We can use both data sets.
 - Step 2
 - Choice of Formula
 - An f-test must first be completed to check if the two data sets have equal variance. That f-test result will determine which t-test is required. The resulting p-value of the f-test was 0.003579, which is less than the significance level of 0.05; therefore, the 2 data sets do not have equal variance (*Figure 27*).
 - Welch Two Sample t-test is required to test if the two data sets have equal means.

```
F test to compare two variances

data:  Child_wt_df$Months.Waiting and adult_wt_df$Months.Waiting
F = 5.8492, num df = 18, denom df = 19, p-value = 0.003579
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 2.297667 15.069994
sample estimates:
ratio of variances
 5.849188
```

Figure 27

- Step 3
 - Decision Rule
 - The null hypothesis will be accepted if the p-value is greater than or equal to the significance level of 0.05.
 - The null hypothesis will be rejected if the p-value is less than the significance level of 0.05.
- Step 4
 - Calculate t-test.

- The p-value of t-test is $p = 0.01867$ (Figure 28)

```
Welch Two Sample t-test
data: Child_wt_df$Months.Waiting and adult_wt_df$Months.Waiting
t = -2.5251, df = 23.729, p-value = 0.01867
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.0226518 -0.2026834
sample estimates:
mean of x mean of y
 8.387561  9.500228
```

Figure 28

- Step 5
 - Result & conclusion
 - P-value is 0.01867, which is less than the significance level of 0.05 – reject H_0 (null hypothesis)
- Step 6
 - Interpretation
 - There is a significant difference between the means of the two sets of data. Both datasets can be used in the project.
- The second set of tests is on the queue datasets.
 - Step 1
 - State the Hypothesis.
 - $H_0: \mu_{\text{Adult_Que}} = \mu_{\text{Child_Que}}$
Only one of the data sets is needed.
 - $H_1: \mu_{\text{Adult_Que}} \neq \mu_{\text{Child_Que}}$
Both data sets are required.
 - Step 2
 - Choice of Formula
 - An f-test must first be completed to check if the two data sets have equal variance. That f-test result will determine which t-test is required. The resulting p-value of the f-test was 0.0000001841, which is less than the significance level of 0.05; therefore, the 2 data sets do not have equal variance. (Figure 29).

```
F test to compare two variances
data: Adult_que_df$Patients.Waiting and Child_que_df$Patients.Waiting
F = 16.517, num df = 19, denom df = 18, p-value = 1.841e-07
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
  6.410996 42.048595
sample estimates:
ratio of variances
 16.51745
```

Figure 29

which is less than the significance level of 0.05; therefore, the 2 data sets do not have equal variance. (Figure 29).

- Welch Two Sample t-test is required to test if the two data sets have equal means.
- Step 3
 - Decision Rule
 - The null hypothesis will be accepted if the p-value is greater than or equal to the significance level of 0.05.
 - The null hypothesis will be rejected if the p-value is less than the significance level of 0.05.
- Step 4
 - Calculate t-test.
 - The p-value of t-test is $p = 0.0395$ (Figure 30)

```
Welch Two Sample t-test

data:  Adult_que_df$Patients.Waiting and Child_que_df$Patients.Waiting
t = 2.1928, df = 21.407, p-value = 0.0395
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1152.159 42562.883
sample estimates:
mean of x mean of y
26315.100 4457.579
```

Figure 30

- Step 5
 - Result & conclusion
 - P-value is 0.0395, which is less than the significance level of 0.05 – reject H_0 (null hypothesis)
- Step 6
 - Interpretation
 - There is a significant difference between the means of the two sets of data. Both datasets can be used in the project.

Some exploratory analysis using a plot in R (Appendix 11) identified a problem. There are only 20 counties on the list. A table of the data was used to confirm this. (Appendix 11) This data set cannot be used for the analysis, as it lacked an observation for every county.

Registered Pharmacies

The registered pharmacies dataset was very clean and almost already prepared when downloaded (Figure 31). The transformation of the dataset into the required state required the following tasks to be completed.

	County	X
49	Grand Total	1,890
33	Dublin 6W	10
7	Co. Galway	112
6	Co. Dublin	113
43	Dublin 16	14
45	Dublin 18	15
47	Dublin 22	15
12	Co. Leitrim	16

Figure 31

- Dublin is broken into areas (e.g., Dublin 8, Dublin 9, et cetera), and there was a requirement to join them into a single number for Dublin. In addition, the text description of the other counties did not match the other data sets (e.g., Co. Galway vs Galway).
 - Create a vector that contains each county name.
 - Use a 'For' loop to match the county column with a lookup vector of counties.
- Create a new data frame excluding the old county column & total row.
- Aggregate to sum up “Dublin” pharmacy numbers.
 - Rename columns.
- Using the population data set, merge kilometres square size for each county
- Calculate the number of pharmacies per 100 km².
- Normalise function applied to pharmacies per 100 km² (*Appendix 17*).
- The rank column created referencing normalised data.

Population

The population data set is used in conjunction with other data sets. The population data were used to calculate the per 100k of the population in:

- Count of crimes per 100k population
- Count of casualties per 100k population

The per 100km² (km²) measure was calculated by dividing density by the population to calculate the area squared, which was used in:

- Count of public electric car chargers per 100km²
- Count of Pharmacies per 100km²
- Count of Tourist Attractions per 100km²

The map chart shown in Appendix 12 presents the population per county.

Average Distance to Emergency Department

The average distance to an emergency department data set required the following tasks to be completed:

- There is a county column, but some counties are split (e.g., Cork county & Cork city). The county data from the population data set was used to correct this.
 - Create a vector that contains each county.

- Use a 'For' Loop to search through the county column and use 'grep' to compare to the vector of counties to find the appropriate county.
- The average distance to an ED variable is in the form of a range (e.g., 25 – 50km). A new variable was required using the midpoint of each range.
 - Created a dataset using the Distance Range variable and the median number of months for each referenced period(*Figure 32*).

```
#create new dataset to use in estimation of distance of ED
lookup <- data.frame(DistanceRange = c("25 - < 50km", "50km or more",
                                       "10 - < 25km", "Less than 5km",
                                       "5 - < 10km"),
                    Km = c(37.5, 62.5, 17.5, 2.5, 7.5),
                    stringsAsFactors = TRUE)
```

Figure 32

- A merge with the main data frame using the 'inner join' function in R was performed.
- Create a subset with only required columns.
- Used the aggregate function to calculate the average distance to ED.
- Normalise function applied to average distance to ED (*Appendix 17*).
- The rank column created referencing normalised data.

Combined Data Set

Before embarking on any data analysis or modelling, the nine transformed data sets must be merged into a single data set. The merging was complete one data set at a time. A copy of the crime data set was created firstly and renamed as the new combined data. The following steps were followed for each data set merge from here on.

1. Load following data set from the database.
2. Check column names to confirm naming convention. Rename any columns as needed.
 - County.
 - df_rank (e.g., class_rank).
 - df_normalise (e.g., class_normalise).
3. Perform 'inner join' with the combined data frame.
4. Check the new combined data frame.
5. The newly merged data frame has an issue with "Meath" and "Westmeath". The join creates an additional observation (*Figure 33*). To resolve this, the observation where "County.x" (County column from combined DF) = "Westmeath" and the "County.y" (County column from recently joined DF) = "Meath" (*Figure 34*) should be removed.
6. Check the new combined data frame.
7. Remove "County.y" column (County column from recently joined DF).
8. Rename "County.x" → "County".

With the new combined data frame prepared, two additional subsets were created. The first included just the normalised columns, and the second was with the key data columns (e.g., Crimes per 100k). All three data frames were saved to the database (*Appendix 15*).

County.x	Crime.Count	Crimes_per_100k	Crime_normalise	Crime_rank	County.y	Average_Class_Size	class_rank	class_normalise
Wicklow	5881	3988	0.3822930	17	Wicklow	23.68539	22	0.8258594
Wexford	4410	2945	0.1891091	3	Wexford	23.58386	20	0.8076104
Westmeath	4782	5386	0.6412299	24	Meath	24.65421	26	1.0000000
Westmeath	4782	5386	0.6412299	24	Westmeath	22.63863	15	0.6377100
Waterford	5904	5081	0.5847379	22	Waterford	23.37748	19	0.7705140
Tipperary	5129	3214	0.2389331	8	Tipperary	22.24654	11	0.5672340
Sligo	2139	3263	0.2480089	10	Sligo	20.50003	5	0.2533060
Roscommon	1242	1924	0.0000000	1	Roscommon	19.97738	2	0.1593635
Offaly	2534	3250	0.2456010	9	Offaly	22.90044	17	0.6847690
Monaghan	2511	4090	0.4011854	18	Monaghan	22.83871	16	0.6736724
Meath	5869	3009	0.2009631	5	Meath	24.65421	26	1.0000000

Figure 33

```
#westmeath has been added to the dataset following the merge twice
#(once using westmeath & once using meath) will remove match with meath
Combo_df <- Combo_df[!(Combo_df$County.x == "Westmeath" & Combo_df$County.y == "Meath"),]
```

Figure 34

With the data now combined, the normalised data set was summed by row. The principle applied here is that each variable was normalised based on whether it was more important to have a higher value or lower value. Taking crime as the first example, the lower the crime rate, the better, so the crime variable was normalised from low to high, as a comparison, the electric car chargers are the opposite, the more of them in a county, the better, so this variable was normalised from high to low. What that means for the normalised data set is that a summation per county will provide a total, and the county with the lowest total has the best combination of variables.

The results section below details the results, with a bar chart summary and a choropleth map. The map required a shapefile of the Irish counties sourced from “GADM maps and data”. (GADM, no date) The map is coloured using 26 shades of green. The hex codes for the different shades of green were sourced from “icolorpalette”. (icolorpalette-aurora-green, no date)

Shiny Application

As included in the project proposal, an 'R' Shiny application was developed as part of the project. The application has four pages, each with its purpose. The application was created using a three R files structure, these are:

1. app.R to launch the application.
2. server.R contains all the backend code and is where all the data processing is completed.
3. UI.R UI stands for user interface. The UI is the frontend of the application and is what the end-user will see.

The application uses the navbar framework throughout, and the "superhero" theme has been applied. Shiny has a built-in function called "themeSelector", which allows various themes to be selected during the design phase. The "superhero" theme was chosen as it made the colour on the charts and their white background stand out. The font associated with this theme was also easy to read on different size screens.

The application also benefits from the use of fluid pages (shiny & bootstrap feature). A fluid page layout consists of rows and columns. The rows ensure the different elements appear on the same line. Fluid pages also scale the different elements to fill all available browser width.

Link to application: https://mrk-kelly.shinyapps.io/Project_app/

Introduction Page

The introduction page (Figure 35) is the page displayed on loading and contains no reports or charts. The project poster is centred on the page, with a brief description of the project under the poster, plus a link to the NCI showcase website. The top of the page contains the navigation bar.

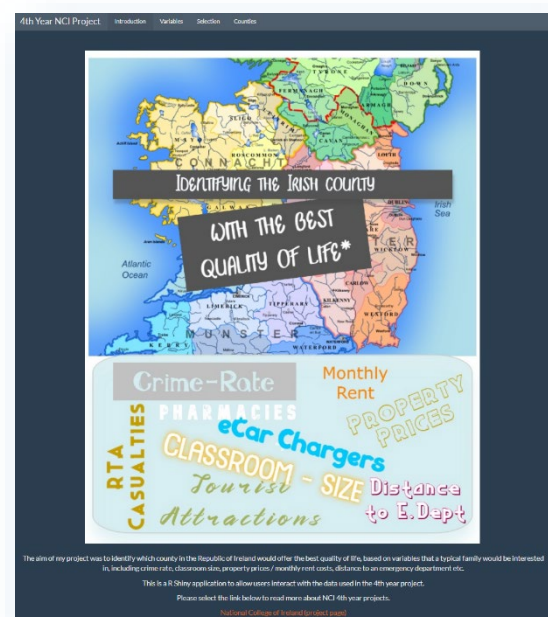


Figure 35

Variables Page

The purpose of this page is for the end-user to select one of the nine variables used in the analysis and be presented with a summary for that variable (Figure 36). The page includes:

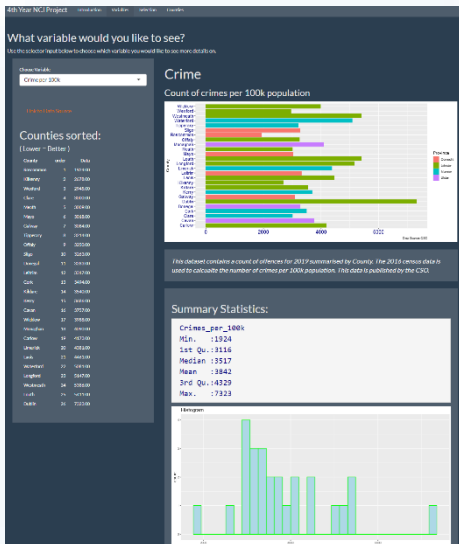


Figure 36

- Count of crimes per 100k population bar chart with each county (coloured by province)
- A brief description of the data set
- Summary statistics of the data set
- Histogram of the data set
- List of the 26 counties (including relevant value) sorted by rank.
- A link to the data source for that variable

Counties Page

This page allows the end-user to select one of the twenty-six counties and be presented with a summary for that county (Figure 37).

The page then displays the following information for the selected county.

- Population (per 2016 census)
- The measure for each variable (example - Average kilometres to ED:36.9)
- The rank of the selected county for each variable

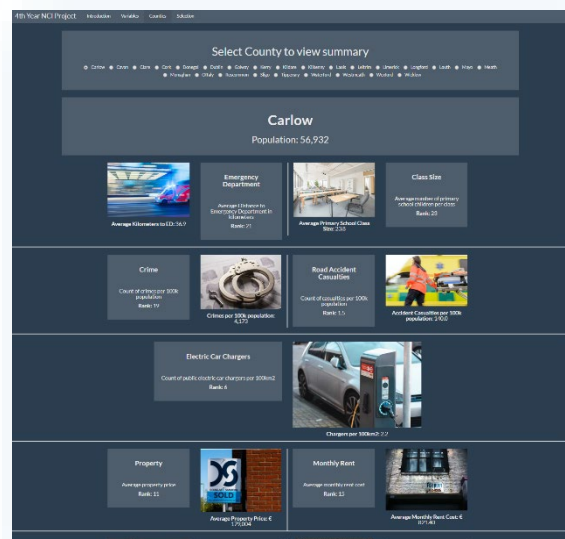


Figure 37

For the first two pages, the server takes the variable or county selection from the frontend and uses that to generate the plots, text, and tables, which are then returned to the UI and rendered on the page.

Variable Selection Page

The last page is titled “Selection” (Figure 38). This page allows the user to select the variables that they would like to include in the analysis. As a default, the ‘Crime Rates’ variable is included, with all others excluded on the opening of the page. As the user includes additional variables, the displayed information will update accordingly. The information displayed on the page is (based on variable selection):

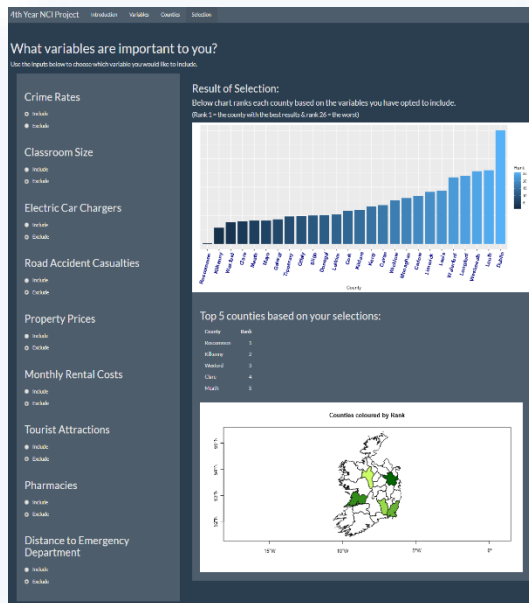


Figure 38

- Bar chart containing all counties sorted by rank.
- Top five counties listed in order.
- Map of Ireland with the top five counties coloured.

To calculate the rank based on the selection, the UI takes the user selection and converts it to a 1 or 0 (1 = include and 0 = exclude). The server uses this passed value for each variable to multiply the existing data frame and save it to a new data frame. The new data frame is then ranked and used to generate the charts and tables, which are then returned to the UI to display.

There were some issues with different elements of this page. Printing messages to the console at different stages helped identify the issues, but not for all issues. Two methods helped route out the problems quickly.

1. Standard error:

Using the `stderr` function in the ‘R’ code, printed the values in a data frame. This could be used before and after a calculation to identify if the variable has been calculated correctly.

Sample R code:

```
cat(file=stderr(), "new class", new$class.size)
```

Sample console output:

```
new class 0.8471233 0.5969276 0.1786574 0.6044277 0.1936798 0.8115157 0.3485003 0.3318327 0.9191299 0.7243513 0.5552807 0.3172474 0.5956174 0.3569611 0.9108935 0 1 0.6736724 0.684769 0.1593635 0.253306 0.567234 0.770
```

2. Export to CSV: When the standard error did not assist in identify the problem, exporting the data frames at different stages allowed further analysis and provided an understanding of what was happening with the calculations.

The other problem that arose during the development of this page was with the Map. The library initially used for the loading of the shapefile was the 'SP' library with the 'readShapeSpatial' function, and this was working fine when executing the map plot independently. However, this is now deprecated, and when executed as part of the Shiny application, it fell over. The alternative method to load the shapefile is the SF library and the 'st_read' function, which worked without error.

Analysis

Data Mining

In the data mining stage of the project, various tasks were performed. Before proceeding, the first step is to split the data into training and test data frames. This will allow the model to be trained using the training data set and then tested using the test data set. A decision was made to use an 80/20 split. This split profile stems from the Pareto Principle (80% of effects come from 20% of causes) and is a very common split in data analysis. (Dunford, Su and Tamang, 2014)

Using the 'normalised' data set, the creation of a correlation matrix (Pearson parametric correlation test) was completed. The matrix will be used to explore the relationship/dependency between the variables. The results section below contains the output table of correlation coefficients (rounded to two decimal places). (*Figure 35*)

A correlation matrix with significance levels (p-value) was also produced using the 'Hmisc' package in 'R'. The p-value is the probability that the correlation between the variables used occurred by chance or another way to put it, it's the probability that the null hypothesis is true. The hypotheses specification for this test is:

- Null hypothesis H_0 : The correlation is not statically significant between the two variables. There is not a significant linear relationship(correlation) between the two variables.
- Alternative hypothesis H_1 : The correlation is statically significant between the two variables. There is a significant linear relationship (correlation) between the two variables.

The results section below contains the output matrix with the significance levels (p-values) (*Figure 41*).

To assist in the presentation of the relationships the 'corrplot' package was used to create a graphical display of the correlation matrix, highlighting the most correlated variables in the data.

As well as gain an understanding of the combined data, the other data mining aim was to test if a model can predict the crime per 100k population (dependent variable) using the other variables as the independent variables. The Crime data was chosen as the dependent variable as some of the available variables will be less critical than others for readers (e.g., families with no children will not be interested in class size). However, everyone can be affected by crime and would have an interest in this.

Both a decision tree and Random Forest algorithms were implemented for this analysis, with the results compared to each other and the actuals. Random Forest was added to this analysis as they are particularly well-suited to a small sample size, which this analysis has.

The decision tree algorithm used is the Recursive Partitioning (rpart) package in 'R'. The most commonly used decision tree algorithm is the C5.0 algorithm, but this is used primarily for classification, whereas 'rpart' can be used either in regression or classification.

The decision tree was executed using three combinations of arguments (*for the function of each arguments, see Appendix 16*). Each tree will be trained using the training data first, then executed using the test data. The three combinations are:

Combination 1	
Method:	Anova
Minsplit	10
Minbucket	5
Maxdepth	20
Xval	10
Usesurrogate	2

Combination 2	
Method:	Anova
Minsplit	5
Minbucket	2
Maxdepth	20
Xval	5
Usesurrogate	2

Combination 3	
Method:	Anova
Minsplit	3
Minbucket	1
Maxdepth	20
Xval	3
Usesurrogate	2

For the Random Forest algorithm, the 'tuneRF' algorithm was used. The algorithm starts with the default value of mtry, and then searches for the optimal value (with respect to Out-

of-Bag error estimate) of mtry (number of random variables used in each tree) for the randomForest (*Figure 39*). (RDocumentation, no date)

```
Call:
randomForest(formula = Crimes_per_100k ~ ., data = train_data, mtry = 2)
Type of random forest: regression
Number of trees: 500
No. of variables tried at each split: 2

Mean of squared residuals: 1264060
% Var explained: 9.53
```

Figure 39

Results

Correlation Matrix

The results of the correlation matrix can be seen in Figure 40. As a general rule of thumb, the following are indicative of how to read the results:

- Coefficient between -0.3 and $+0.3$ = weak correlation.
- Coefficient less than -0.7 or greater than $+0.7$ = strong correlation.
- Coefficient between -0.3 and -0.7 or between $+0.3$ and $+0.7$ = moderate correlation.
- Coefficient values below zero indicate a negative relationship and above zero a positive relationship. A value equal to zero indicates no relationship.

Based on the above, the variables with a strong relationship (negative/positive) are:

- Attractions and electric car chargers are positively correlated, indicating that a higher number of attractions, there are a higher number of electric car chargers.
- Monthly Rent and Property Prices are positively correlated, which indicates that areas with high monthly rent also have high property prices.
- Electric car chargers and monthly rent are negatively correlated, which indicates that areas with high monthly rent are associated with fewer electric car chargers.
- Pharmacies and monthly rent are negatively correlated, which indicates that higher rent costs are associated with fewer pharmacies in the area.
- Pharmacies and attractions are positively correlated, which indicates that where there are a higher number of attractions, there is a higher number of pharmacies.

	Crime	Class.size	Car.Chargers	Road.Casualties	Property.Price	Monthly.Rent	Attractions	Pharmacies	Ed.Dept
Crime	1.00	0.39	-0.67	0.16	0.42	0.50	-0.63	-0.68	-0.52
Class.size	0.39	1.00	-0.27	-0.25	0.61	0.58	-0.20	-0.27	-0.51
Car.Chargers	-0.67	-0.27	1.00	-0.01	-0.64	-0.72	0.98	1.00	0.57
Road.Casualties	0.16	-0.25	-0.01	1.00	-0.36	-0.23	0.04	-0.01	0.11
Property.Price	0.42	0.61	-0.64	-0.36	1.00	0.96	-0.62	-0.64	-0.50
Monthly.Rent	0.50	0.58	-0.72	-0.23	0.96	1.00	-0.69	-0.72	-0.58
Attractions	-0.63	-0.20	0.98	0.04	-0.62	-0.69	1.00	0.98	0.53
Pharmacies	-0.68	-0.27	1.00	-0.01	-0.64	-0.72	0.98	1.00	0.57
Ed.Dept	-0.52	-0.51	0.57	0.11	-0.50	-0.58	0.53	0.57	1.00

Figure 40

Correlation Matrix with Significance Levels (p-value)

	Crime	Class.size	Car.Chargers	Road.Casualties	Property.Price	Monthly.Rent	Attractions	Pharmacies	Ed.Dept
Crime	NA	0.0482560455	1.600021e-04	0.40951513	3.419946e-02	1.055039e-02	0.0006016957	1.232497e-04	0.006159513
Class.size	0.0482560455	NA	1.899520e-01	0.21290311	9.976417e-04	1.941098e-03	0.3370463461	1.843918e-01	0.008122747
Car.Chargers	0.0001600021	0.1899519805	NA	0.96725074	4.692791e-04	4.138368e-05	0.0000000000	0.000000e+00	0.002769090
Road.Casualties	0.4095151281	0.2129031131	9.672507e-01	NA	7.567442e-02	2.637889e-01	0.8541339405	9.599905e-01	0.615892195
Property.Price	0.0341994551	0.0009976417	4.692791e-04	0.07567442	NA	3.330669e-15	0.0007587295	4.561056e-04	0.010380029
Monthly.Rent	0.0105503938	0.0019410981	4.138368e-05	0.26378888	3.330669e-15	NA	0.0001165823	3.655501e-05	0.002316038
Attractions	0.0006016957	0.3370463461	0.000000e+00	0.85413394	7.587295e-04	1.165823e-04	NA	0.000000e+00	0.005582887
Pharmacies	0.0001232497	0.1843917743	0.000000e+00	0.95999046	4.561056e-04	3.655501e-05	0.0000000000	NA	0.002538211
Ed.Dept	0.0061595125	0.0081227473	2.769090e-03	0.61589220	1.038003e-02	2.316038e-03	0.0055828868	2.538211e-03	NA

Figure 41

Figure 41 is the output showing the significance levels (p-values) from the normalised data set. If the p-value is less than the significance level ($\alpha = 0.05$), then the null hypothesis can be rejected. If the p-value is not less than the significance level ($\alpha = 0.05$), then the null hypothesis cannot be rejected. There are several variable pairings that the null hypothesis can be accepted for as there is enough evidence to conclude that they do not have a significant relationship. The following variables pairings do not have a significant relationship.

- Road casualties and crime.
- Electric car chargers and classroom size
- Road casualties and classroom size
- Tourist attractions and classroom size
- Pharmacies and classroom size
- Property prices and road casualties
- Monthly rent and road casualties
- Tourist attractions and road casualties
- Pharmacies and road casualties
- Distance to an emergency department and road casualties

For further understanding, a flatter matrix table that displays both the p-values and correlation coefficients for each variable pairing side by side can be seen in Appendix 19.

Correlogram

The correlogram shown in Figure 42 is a graphical display of the correlation matrix seen above, and its purpose is to highlight the most correlated variables. Positive correlations are displayed in blue and negative correlations in red colour. Colour intensity and the size of the circle are proportional to the correlation coefficients. The legend colour shows the correlation coefficients and the corresponding colours on the right side of the correlogram. (Sthda, 2019)

The correlogram reaffirms the earlier interpretations about our data, such as

- Attractions and electric car chargers are positively correlated, indicating that a higher number of attractions, there are a higher number of electric car chargers.
- Monthly Rent and Property Prices are positively correlated, which indicates that areas with high monthly rent also have high property prices.
- Electric car chargers and monthly rent are negatively correlated, which indicates that areas with high monthly rent are associated with fewer electric car chargers.
- Pharmacies and monthly rent are negatively correlated, which indicates that higher rent costs are associated with fewer pharmacies in the area.
- Pharmacies and attractions are positively correlated, which indicates that where there are a higher number of attractions, there is a higher number of pharmacies.

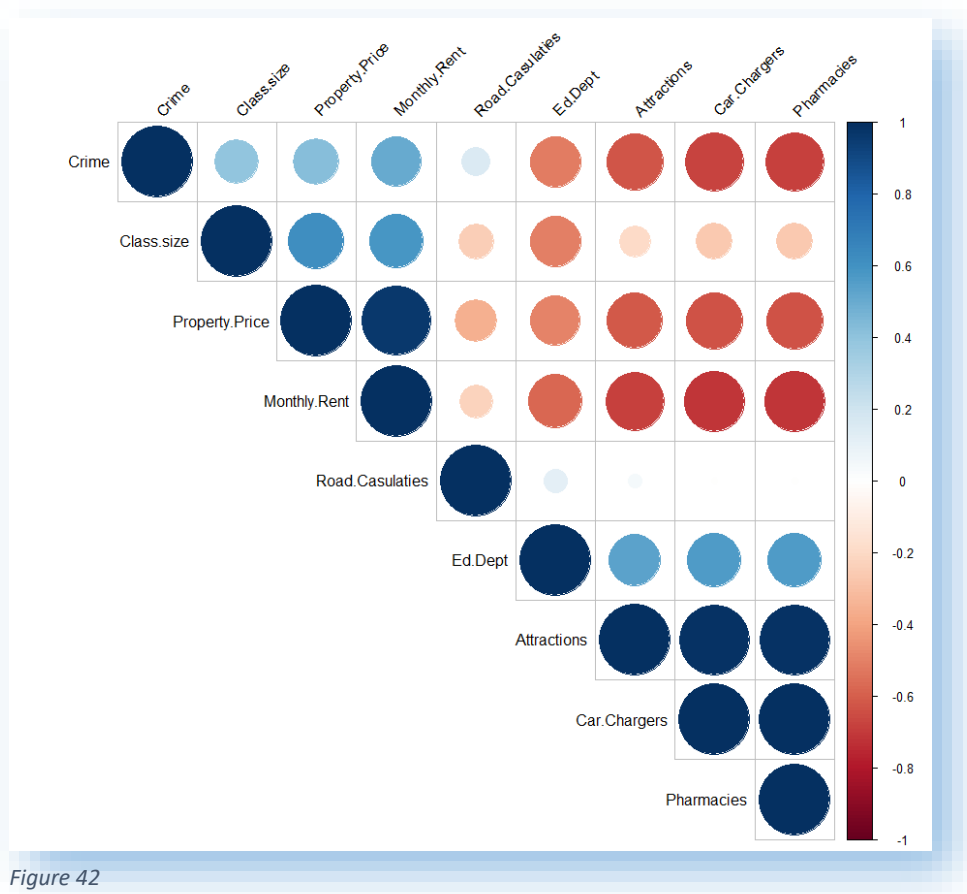


Figure 42

An alternative to the correlogram is the correlation heatmap (Appendix 20). The same data is presented using a seven-step colour scheme, and the resulting interpretations are the same as the above correlation matrix and correlogram.

Rpart Decision Tree

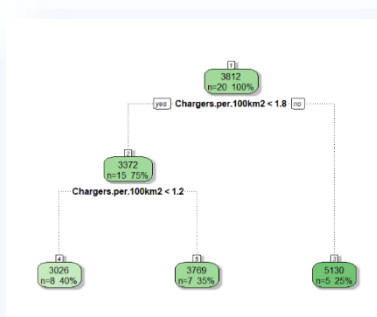


Figure 43

The 'rpart' decision tree was executed using three combinations of arguments (see above), with the following diagrams visualising the trees. All three combinations are using the Anova (regression) method, a max depth of 20 (maximum depth of any node of the final tree) and usesurrogate of '2' (if all surrogates are missing, then send the observation in the majority direction).

The first combination, with a minsplit of 10 (the minimum number of observations that must exist in a node), minbucket of 5 (the minimum number of observations in any terminal), and xval of 10 (number of cross-validations),

resulted in the tree seen in Figure 43 when using the training data.

This tree heavily relies on the chargers per 100km² variable to make its decision. As seen earlier in the report, the crime and car charger variables are negatively correlated, with a -0.67-correlation coefficient and a p-value of 1.600021e-04. The variable importance tables (Figure 44) confirm the importance of the Chargers per 100km² variable.

Variable importance				
Chargers.per.100km2	Average_Class_Size	Pharmacies_per_100km2	Average.Rent	Average_Price
27	16	16	14	14
Attractions.per.100km2	Ave.Distance_to_ED	Casualties_per_100k		
10	2	1		

Figure 44

The confusion matrix generated using the training data (Figure 45) has been summarised into a table and the predicted value calculated as a percentage of the actual values. The variance between the predicted and actual values (n=20) averaged 104.23% (s=21.82%).

train_data.Crimes_per_100k	Predicted_Values	variance
16	3018	3026 100.265%
22	3214	3026 94.151%
5	3283	3026 92.172%
12	3317	3026 91.227%
15	5411	5130 94.807%
9	3540	5130 144.915%
24	5386	5130 95.247%
6	7323	5130 70.053%
26	3988	5130 128.636%
4	3494	3769 107.871%
2	3757	3769 100.319%
7	3084	3026 98.119%
19	3250	3769 115.969%
10	2678	3026 112.995%
14	5147	3769 73.227%
17	3009	3769 125.258%
8	3686	3026 82.094%
11	4461	3769 84.488%
21	3263	3769 115.507%
20	1924	3026 157.277%

Figure 45

The second combination, with a minsplit of 5, minbucket of 2, and xval of 5, resulted in the tree seen in Figure 46 when using the training data.

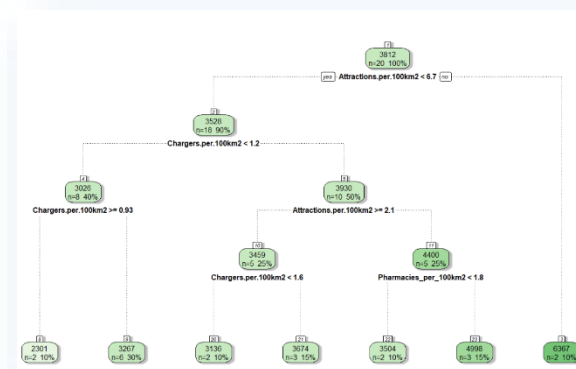


Figure 46

As shown in Figure 46, this version of the 'rpart' tree uses more variables in its decision making. The variable importance tables (Figure 47) confirms that three variables are considerably more important than the rest of the variables.

Figure 47

Variable importance				
Pharmacies_per_100km2	Chargers.per.100km2	Attractions.per.100km2	Average_Class_Size	Average.Rent
27	25	23	6	6
Average_Price	Ave_Distance_to_ED	Casualties_per_100k		
6	4	2		

The confusion matrix generated using the training data (Figure 48) has been summarised into a table and the predicted value calculated as a percentage of the actual values. The variance between the predicted and actual values (n=20) averaged 100.85% (s=9.32%).

train_data.Crimes_per_100k	Predicted_Values	variance
16	3018	3267 108.25%
22	3214	3267 101.65%
5	3283	3267 99.51%
12	3317	3267 98.49%
15	5411	6367 117.67%
9	3540	3674 103.79%
24	5386	4998 92.80%
6	7323	6367 86.95%
26	3988	3674 92.13%
4	3494	3674 105.15%
2	3757	3504 93.27%
7	3084	3267 105.93%
19	3250	3504 107.82%
10	2678	2301 85.92%
14	5147	4998 97.11%
17	3009	3136 104.22%
8	3686	3267 88.63%
11	4461	4998 112.04%
21	3263	3136 96.11%
20	1924	2301 119.59%

Figure 48

The third combination, with a minsplit of 3, minbucket of 1, and xval of 3, resulted in the tree seen in Figure 49 when using the training data.

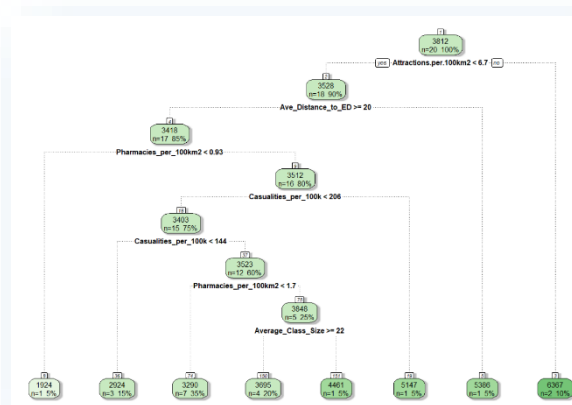


Figure 49

As shown in Figure 49, this version of the 'rpart' tree uses more variables than the second or first combinations in its decision making. The variable importance tables (Figure 50) confirms that three variables are considerably more important than the rest of the variables.

Variable importance				
Pharmacies_per_100km2	Chargers.per.100km2	Attractions.per.100km2	Ave_Distance_to_ED	Casualties_per_100k
31	26	25	8	6
Average_Class_Size	Average.Rent	Average_Price		
2	1	1		

Figure 50

The confusion matrix generated using the training data (Figure 51) has been summarised into a table and the predicted value calculated as a percentage of the actual values. The

variance between the predicted and actual values (n=20) averaged 100.45% (s=6.75%). Interestingly, there are six variables predicted to within 1% of their actual values.

Before comparing the results of the decision trees when performed on the test data, the following are the details on the random forest algorithm.

	train_data.Crimes_per_100k	Predicted_Values	variance
16	3018	3290	109.01%
22	3214	3290	102.36%
5	3283	3290	100.21%
12	3317	3290	99.19%
15	5411	6367	117.67%
9	3540	3695	104.38%
24	5386	5386	100.00%
6	7323	6367	86.95%
26	3988	3695	92.65%
2	3494	3695	105.75%
7	3757	3695	98.35%
19	3084	2924	94.81%
10	3250	3290	101.23%
14	2678	2924	109.19%
17	5147	5147	100.00%
8	3009	2924	97.18%
11	3686	3290	89.26%
21	4461	4461	100.00%
20	3263	3290	100.83%
	1924	1924	100.00%

Figure 51

Random Forest

As identified using the 'tuneRF' algorithm, the random forest was executed with a mtry of 2 and 'ntree' (number of trees used in the forest) of 500. When executed on the training data set, the percentage of variance explained is low, at 9.57%.

Looking at feature importance in a random forest, the variables with the highest importance are the variables that will have a significant impact on the outcome values. This measure is known as 'IncNodePurity' and is based on the Gini impurity index (used to calculating the splits in trees). The higher the 'IncNodePurity' value, the more critical the random forest model rates the average distance to an ED as the highest, followed by tourist attractions and pharmacies (Figure 52).

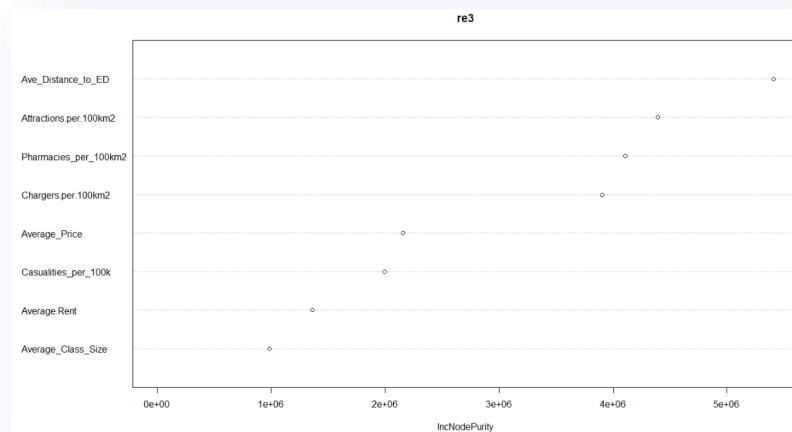


Figure 52

Model Predictions

		County1	County3	County13	County18	County23	County25
Actual Values		4173	3000	4381	4090	5081	2945
Predicted Values	rpart1	5130	3769	5130	3769	3769	5130
Predicted value as % of Actual	rpart1	123%	126%	117%	92%	74%	174%
Predicted Values	rpart2	6367	3136	3674	3504	3674	3674
Predicted value as % of Actual	rpart2	153%	105%	84%	86%	72%	125%
Predicted Values	rpart3	6367	3290	3695	3695	2924	3695
Predicted value as % of Actual	rpart3	153%	110%	84%	90%	58%	126%
Predicted Values	random forest	4224	3263	3953	3625	3582	3930
Predicted value as % of Actual	random forest	101%	109%	90%	89%	70%	133%

Figure 53

When executed with the test data, the output of all four models can be seen in Figure 53. When executed using the training data, 'rpart' combination two and three were similar, with the third combination having predicted several values correctly. This

level of prediction is not the case with the test data. All three 'rpart' models' predictions are spread out over an extensive range. This lack of accuracy is most likely due to overfitting. The confusion matrix and summary table can be seen for all four models in Appendix 21.

The metrics selected are Mean Absolute Percentage Error (MAPE) and Root Mean Square Error (RMSE) to evaluate the models best when comparing them to each other.

		MAPE	RMSE
Training Data	Combination 1	0.16671	845.67
	Combination 2	0.07760	397.47
	Combination 3	0.04613	340.22
Test Data	Combination 1	0.28921	1,201.86
	Combination 2	0.23337	1,168.08
	Combination 3	0.25913	1,337.88
	Random Forest	0.15679	784.05

Figure 54

MAPE is one of the most common measures used to predict error. The measure is calculated as a percentage and is indicative of how accurate a forecast model is. MAPE works best if there are no zeros or extreme values in the data.

RMSE is used to determine how focused the data is around the line of best fit in the model. It is the standard deviation of the prediction errors (residuals). The residuals measure how far from the regression line each data point is, and RMSE is a measure of how spread out these residuals are.

Both of these measures are commonly used in regression analysis to verify results and model accuracy.

Figure 54 details display the calculated values for the RMSE and MAPE. Focusing on the test data results for the MAPE first, the lower the error percentage, the better the model is. The lowest MAPE from the three, with 23.337%. This MAPE compares to 15.679% for the random forest, approximately a 32% decrease in the error rate.

Unlike MAPE, RMSE is not measured as a percentage. Instead, it is measured in the same units as the response variable. The lower the value of RMSE indicates a better fitting model. Again, looking at just the test data in Figure 54, the random forest model has the lowest RMSE value.

Both assessments indicate that the random forest is the better model for predicting the crime per 100k population variable. However, the prediction was not as accurate as it could be. The low rate of accuracy is most likely due to overfitting. Some possible solutions to this problem include making the model simpler by removing some of the variables, cross-validation or train the model with more data (not always possible – in this case the project would have to increase the granularity to do this)

Best County Analysis

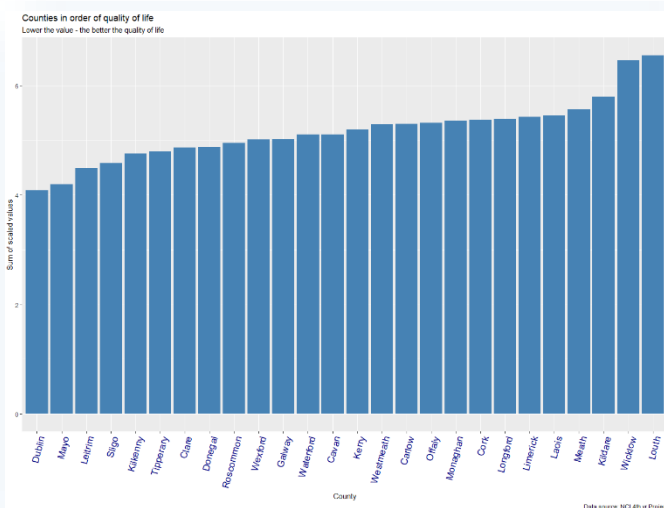


Figure 55

Using the normalised data frame and all possible variables and sorted in order, Figure 55 displays the counties in order (the lower the normalised total – the better the rank). The top county when using the full range of variables is Dublin. This prediction is not a true reflection of the best county, as including all nine variables is unrealistic. A more realistic scenario is that only selected variables would be needed for individual analysis. This option is covered in the next section.

An interesting observation that becomes more obvious when the results are mapped out (Figure 56) and coloured by rank (the darker the colour, the lower down the rank) is that Dublin may be at the top of the rankings. However, all the counties surrounding Dublin (commuter belt) are at the bottom of the rankings.

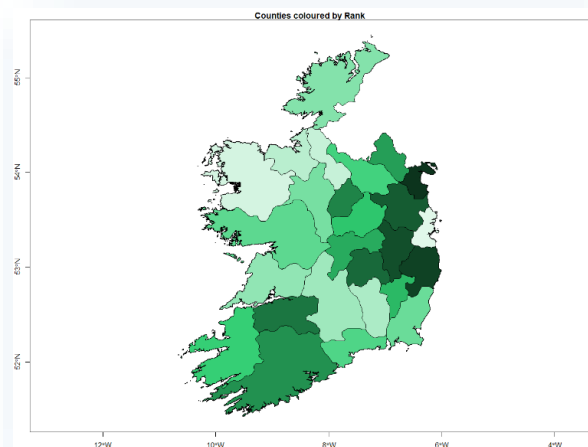


Figure 56

To further investigate the different possible outcomes based on user inputs, two types of end-user were identified. A selection was made based on the user story.

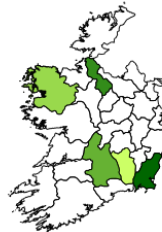
1. The first user story is for a young couple looking for a non-permanent move out of Dublin for approximately five years. They have no children and like to keep busy when not at work.

Selected variables

Variable	Include / Exclude	Reason
Crime	Include	No one likes to be a victim of crime
Classroom size	Exclude	No children
Electric Car Chargers	Exclude	Cannot afford one
Road Accident Casualties	Include	Drive a car
Property Prices	Exclude	Temporary move
Rental Costs	Include	Temporary move
Tourist Attractions	Include	Like to have things to do
Pharmacies	Exclude	Do not often visit as young
Distance to ED	Exclude	Do not often visit as young

Based on the above inputs for this hypothetical couple, the results from the Shiny application indicate the top five counties that they should consider moving have been identified. These are:

1. Kilkenny
2. Mayo
3. Tipperary
4. Leitrim
5. Wexford



The entire twenty-six counties in order can be seen in Appendix 22 or can be viewed using the Shiny application ([link](#)).

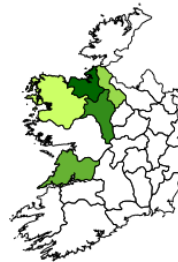
2. The second user story is a family looking for a permanent move out of Dublin. They have two children and like to keep them busy at the weekends. One of the children has asthma.

Selected variables

Variable	Include / Exclude	Reason
Crime	Include	No one likes to be a victim of crime
Classroom size	Include	Two children
Electric Car Chargers	Exclude	Cannot afford one
Road Accident Casualties	Include	Drive a car
Property Prices	Include	Permanent move
Rental Costs	Exclude	Permanent move
Tourist Attractions	Include	Like to have things to do with the children
Pharmacies	Include	To fill regular prescriptions
Distance to ED	Exclude	Do not visit often

Based on the above inputs for this hypothetical family, the results from the Shiny application indicate the top five counties that they should consider moving have been identified. These are:

1. Mayo
2. Leitrim
3. Clare
4. Roscommon
5. Sligo



The entire twenty-six counties in order can be seen in Appendix 23 or can be viewed using the Shiny application ([link](#)).

Best County Testing

To test the validity of these predictions, an MS Excel model was created which mirrored the functionality of the 'selection' page in shiny. Several selection combinations were applied to both models, with the resulting top five counties matching each time. A sample output can be seen in Appendix 24.

Conclusions

The data used in the project is readily available and regularly reported on, both academically and in the press. This project addressed that these data sets are all reported individually and with the question “If moving from Dublin (or anywhere else), which county would offer the best quality of life?” the project aimed to report on the data sets collectively. Selecting the best data to represent what the reader would be interested in was crucial in developing the analysis. Some data mining was included in the project to investigate if the chosen data could predict crime rates.

The data mining confirmed that yes, we could predict crime and that the random forest was more accurate than the tuned decision tree. However, the prediction was not as accurate as it could have been. The low rate of accuracy is most likely due to overfitting. The combined data set requires additional variables or for the existing data to go to a more granular level to solve this.

The other findings that are of interest are the correlation between some of the variables. Some of these were obvious, like monthly rent and property prices are positively correlated. This apparent correlation is a given in society. If the property costs more to purchase, it is going to cost more to rent that property. Some of the less apparent correlations, but not too unexpected, include the positive correlation between the count of attractions and the count of electric car chargers. This correlation makes some sense. Where there are going to be more people, there is a requirement for more electric car chargers. An example of a not so obvious correlation was between pharmacies per 100km² and the average monthly rent. They are negatively correlated, which indicates that the higher the average rent cost, the smaller the number of pharmacies.

The overall finding that Dublin comes out on top is both surprising and misleading. It is surprising because it is the lowest ranking for a number of the variables (Crime per 100k population, monthly rent, property prices). However, it is also the top-ranked for several variables (average distance to an ED, electric car chargers per 100km², pharmacies per 100km², tourist attractions per 100km²). This result is misleading as no end-users will choose to use all nine variables to make this decision.

That leads nicely into one of the advantages of the project. The Shiny application allows the user to see an overview of each county and use the selection screen to include only the variables they wish to include. The impact this will have is that the resulting recommended counties are specific to their interest and assist in making an informed decision using the aligned output.

Further Development or Research

With additional time and resources, the project would be further developed to include the following:

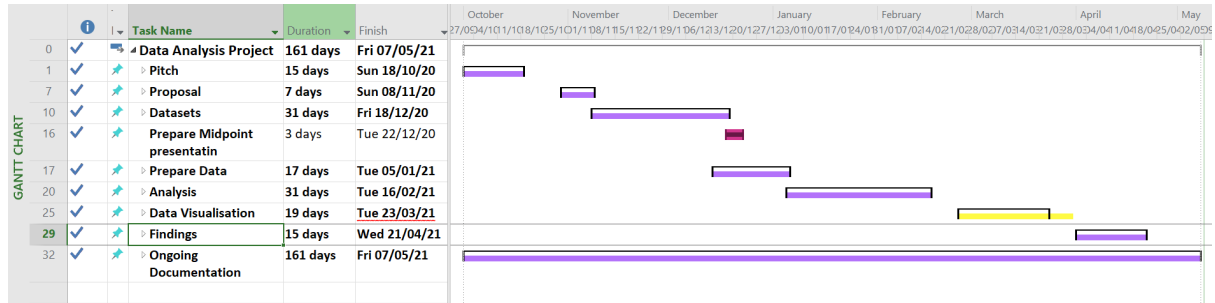
- Some of the data used in the project are updated regularly by the provider. Any additional development would include a plan to automate the loading of the data to the project were possible. This continuous refreshing of the data would keep the analysis current as further data is published would dynamically be updated.
- The raw crime data set has considerable granularity. The project in its current format only used the highest level of this data, the total number of crimes. The project would benefit from separating the crime data into a lower level, such as burglary or assault. This level of detail would be more relevant for users.
- One of the variables that was key to this analysis but could not be obtained was the broadband speed. To overcome the lack of available data, the solution would be to set up a nationwide survey to capture the broadband speed by location and provider. The inclusion of broadband speed data would be a valuable addition to the project as the foundation is around remote working.
- The current monthly rent data set is used as an average rental cost per county regardless of the property type. The additional development work here would be to split the data set by property type to give more options to the end-user.
- The Shiny application is beneficial in its current format. However, the selection page currently only allows the user to include or exclude a variable. For this, each variable is ascribed equal weighting. However, best practice would be to allow users to assign their own weighting to a variable. For example, both property prices and electric car chargers may be required by a user, but property prices may be more important to them than electric car chargers.
- It would be nice to capture users input into the analysis and ask them if there were any additional variable they would like included. Those variables could be investigated for future inclusion in the analysis.
- A factor that can only be supplied by a user is if they have family in a county. If possible, it would be good to allow users add a county (and weighting) for the location of family.

References

- ❖ Data.gov.ie. 2020. CJA07 - Recorded Crime Offences Under Reservation (Number) By Garda Station, Type of Offence and Year - Data.Gov.Ie. [online] Available at: <'> [Accessed 12 December 2020].
- ❖ Data.gov.ie. 2020. *OP Waiting List by Group Hospital - OP Waiting List by Group Hospital 2020 - Data.Gov.Ie.* [online] Available at: <<https://data.gov.ie/dataset/op-waiting-list-by-group-hospital/resource/2fe50521-ff05-4e2b-b3fb-d6a74cb726ce>> [Accessed 12 December 2020].
- ❖ Data.gov.ie. 2020. *ROA27 - Traffic Collisions and Casualties By County, Year and Statistic - Data.Gov.Ie.* [online] Available at: <<https://data.gov.ie/dataset/roa27-traffic-collisions-and-casualties-by-county-year-and-statistic>> [Accessed 13 December 2020].
- ❖ Data.gov.ie. 2020. *PSI Registered Pharmacies - December 2020 - Data.Gov.Ie.* [online] Available at: <https://data.gov.ie/dataset/https-www-thepsi-ie-libraries-monthly_statistics-pharmacies_-_website_statistics-sflb-ashx> [Accessed 14 December 2020].
- ❖ Data.cso.ie. 2020. *Population at Each Census 1841 to 2016.* [online] Available at: <<https://data.cso.ie/table/E2001>> [Accessed 14 December 2020].
- ❖ The Irish Times. 2020. Indeed to Allow 'Vast Majority' Of Irish Employees to Work from Home Forever. [online] Available at: <<https://www.irishtimes.com/business/work/indeed-to-allow-vast-majority-of-irish-employees-to-work-from-home-forever-1.4372417>> [Accessed 1 November 2020].
- ❖ Kelly, J., 2020. Siemens Says That 140,000 Of Its Employees Can Work from Anywhere. [online] Forbes. Available at: <<https://www.forbes.com/sites/jackkelly/2020/07/27/siemens-says-that-140000-of-its-employees-can-work-from-anywhere/>> [Accessed 1 November 2020].
- ❖ Businessworld.ie. 2020. 78% Of Irish Businesses Now Have A Remote Working Policy in Place Technology, News for Ireland, Ireland, Technology, [online] Available at: <<https://www.businessworld.ie/technology-news/78-of-irish-businesses-now-have-a-remote-working-policy-in-place-570184.html>> [Accessed 1 November 2020].
- ❖ Department of Business, Enterprise, and Innovation, 2019. Remote Work in Ireland. [online] Dublin: Enterprise Strategy, Competitiveness and Evaluation. Available at: <<https://dbei.gov.ie/en/Publications/Publication-files/Remote-Work-in-Ireland.pdf>> [Accessed 1 November].
- ❖ Szczuka, M. et al. (2014) 'Using Domain Knowledge in Initial Stages of KDD: Optimization of Compound Object Processing', *Fundamenta Informaticae*, 129, pp. 341–364. doi: 10.3233/FI-2014-975.
- ❖ ED121 - Mainstream Primary Schools by Class Size, Teacher Size of School, Year and Statistic - data.gov.ie. Available at: <https://data.gov.ie/dataset/ed121-mainstream-primary-schools-by-class-size-teacher-size-of-school-year-and-statistic> (Accessed: 20 December 2020).
- ❖ R Package pxR (no date). Available at: <https://pxr.r-forge.r-project.org/> (Accessed: 20 December 2020).
- ❖ Open Charge Map - The global public registry of electric vehicle charging locations (no date). Available at: <https://openchargemap.org/site> (Accessed: 21 December 2020).
- ❖ 'EasyGo | Charging Network' (2021). Available at: <https://easygo.ie/charging-network/> (Accessed: 10 February 2021).
- ❖ Authority, N. P. S. R. (2020) Residential Property Price Register, National Property Services Regulatory Authority. National Property Services Regulatory Authority. Available at:

Appendices

Project Plan



Project planned timeline was adhered to for most areas. As mentioned in the mid-term submission, I had planned to have all dataset work completed by December 18th, 2020. However, this body of work was not completed on time and ran over until approximately mid-January 2021. There was some over run time built into the larger stages of the project, so this did not have a large impact on the overall project timeline.

The data visualisation stage also ran over the allotted time. This was down to some issues with the Shiny application. The project had originally planned to be complete by mid-April 2021, so there was ample time available to absorb this slight overrun.

Reflective Journals

Data Analytics Project

Student name: **Mark Kelly**

Student Number: **17138311**

BSc (Honours) in Computing

Introduction

My Name is Mark Kelly, and I am a part time student completing the BSc (Honours) in Computing Data Analytics specialisation. Firstly, a little about myself. I am 45 years old and am currently working for the ESB. I have worked for the ESB for 23 years, having held a few roles throughout my employment. My current role is as an analyst for the Employee Relations division, which I have held for 2.5 years. I landed this role because of my domain expertise in HR and payroll reporting for ESB plus the fact I had already begun my 4-year degree. I have also been involved in several IT projects of various sizes in the last couple of years.

I have enjoyed my time in NCI so far and normally look forward to my time on campus. I would consider myself an enthusiastic learner and would estimate that I have missed only about half a dozen classes in the first 3 years. I have also maintained an average of over 70% in my results for each of the 3 years. Our class has shrunk quite a bit over that period, we started out with just over 50 in the class on the first evening with Leonie, to approximately 25 students today. For the first year, I was not the oldest in the class, but from second year on, I inherited that throne. Thanks to Covid-19, 4th year is now online, and I will be upfront in saying I do not like this style of learning. I much prefer to have the ability to walk up to the lecturer with my laptop when I have an issue or lean over to a classmate.

For 4th year, my plan is as disciplined as possible. I want to try clear as many of the CA's and assignments out of my way as early as possible, not wait for due dates. Most of the sizeable projects over the last 3 years have been completed in a team, and I have been very lucky with the team I joined as we all bring something different to the projects. A solo project of this size will be challenging, but hopefully my low hanging fruit first approach will enable me to have time to deal with the more challenging elements as they arise. I am lucky, in that Data Analytics was one of the two choices offered to the part time students, and I enjoy this kind of stuff, so that should at least make the project less of a hard task and more of an enjoyable challenge.

Finally, this journal must be uploaded monthly to Moodle, but I have added a reoccurring reminder to my calendar each week to stay on top of it.

Journal

Month: October 2020 (Week 2 – 5)

Week 2: (04/10/20 – 10/10/20):

WE had our first project class with Enda on the previous Saturday. Very little in the way of information regarding the project, on things like minimum requirement and essential elements to include. The presentation was more of an introduction to the project module and a briefing on ethics.

I have a project idea that I am working on and was hoping to firm up on the plans this week, hopefully next week's presentation will include some guidance. The project idea is very relevant in the current Covid-19 climate and will shed some light on how remote working has changed attitudes to work. I have put in a request with my employer for access to some data that I would need.

Week 3: (11/10/20 – 17/10/20):

Week 3 was not off to a good start, my request for access to the data I needed for my project was declined by my employer on GDPR grounds. Although disheartened, I was not overly surprised. I had a plan b, which is a little less complex, but hopefully we get more details on the requirements.

The second of the project classes with Enda was about the project pitch and this journal. Not a lot of project structure, and what little there was, seemed to be aimed at the Software Development side of the class. I decided to document my idea, using the pitch evaluation template for structure and send to Enda for some feedback.

In my rush for guidance, I forgot to include the fact that I wanted to use Machine Learning as a tool in the transformation of the data and the analysis of this data (I do not know how just yet). His reply was helpful in straightening my thoughts.

I set about putting my pitch idea down on paper, looking at the different data sources, existing reports that are similar and taking Enda's advice onboard. By Saturday, I had my pitch ready to go.

Week 4: (18/10/20 – 24/10/20):

Project pitch was submitted on the 18th as required. I received an email on the 21st, from my newly appointed supervisor. Ian has reviewed my project, just waiting on the second reviewer. I do not want to put too much time into my project this week, I am fairly time poor, and I think what time I have this week would be better spent on other projects and CA's, until I hear back about my project and then it will be full steam ahead on the proposal. I progressed my Introduction to A.I. CA, a two-player chess game, to almost completion this week.

Week 5: (25/10/20 – 31/10/20):

Not much in the way of activity this week either, waiting to hear from NCI / Ian on the faith of my project proposal. Ian has made contact and we have a scheduled call on Friday 30th @ 6pm to discuss. This is timely, as tomorrow's project class is about the proposal.

Had my call with Ian on Friday regarding my project. My project was approved with no amendments. He had some recommendations around my plan to use Tableau as a visualisation tool, he suggested I try use R Shiny instead as this would be more academic. We also discussed some timelines and the potential that after my analysis is complete, that my expectations may not be met. We have scheduled the next call for during reading week. Between now and then, I will write my proposal and start to gather potential datasets.

Month: November 2020 (Week 6 – 9)

Week 6: (01/11/20 – 07/11/20):

This week was all about getting the proposal completed along with the ethics form. We had two assignments due this week as well as a Business Data Analysis test, so I did not get as much work done on the datasets as I had hoped. I put together a timeline after speaking with Ian last week and put the timeline into MS Project this week. It looks very do-able if I can stay on top of the other module work. Also realised this week, that the next Saturday we have off from college is St Stephens Day 😞.

Week 7: (08/11/20 – 14/11/20):

This was reading week, so we had two big CA's due very early the following week which took up a lot of my time this week. I started to critically assess by dataset list to make sure the data I was looking for would add value to my analysis project. I started out with 15, and ruled out 3 very quickly as they added no real value or insight (in my opinion)

Week 8 (15/11/20 – 21/11/20):

Now that I have 12 datasets identified. My target was 10 and I included an additional 2 for good measure, this will give me some room to drop some datasets if I am having any difficulty.

I set about searching for my datasets and found a good portion of them before the week was out. So far, I have identified the following datasets (including sources and format)

- Crime Rates
- School Classroom sizes
- Traffic incidents
- House sale prices
- Rental property prices
- Tourist attractions
- Outpatient waiting lists.
- Registered Pharmacies
- Population

Week 9 (22/11/20 – 28/11/20):

This week I spent some hours trying to find my last remaining datasets before my meeting with Ian on Wednesday evening. Below is a summary of where I am at with these.

- Broadband speeds – I am looking for a dataset that will allow me to identify the average broadband speed, ideally by county or at a lower level that can be then summed up. This is

proving very difficult to find. Coreg (Communications Regulator) publish reports on cost and usage etc. but not on speed. Broadband comparison site, switcher.ie also issue some reports but the nearest they come is the max speed by provider. Finally, the National Broadband plan have made no data available.

- Electric Car chargers – As I write this journal, I have finally found a data source for this information. There is an API available from openchargemap.io which contains the data I need. There will be a good bit of data cleansing required as there is no structure to the address (example of one: “Lidl, Belgard”)
- Air pollution – There is some data available on this subject, but I have not found a dataset that meets my needs yet. The EPA has data for Dublin and some other sites around the country but not enough to give me a rating for each county.

I will continue my search until the end of this week, and after that I need to move on to the next task for my project.

The meeting went well with Ian, I appear to be on track with his expectations so far. I have asked Ian can we have a discussion at the next meeting re the best way to approach the next part of the project. I am ok with the exploratory data analysis, just need a bit of direction for what comes after that.

Month: December 2020 (Week 10 – 13)

Week 10: (29/11/20 – 05/12/20):

The task of acquiring the datasets from the relevant sources began this week. It is key that I have the data saved locally this week as I have built some time in to the timetable to study the data before I kick off with the exploratory analysis. All the datasets except one were obtained, the car-charge dataset is still giving me some problems. I had identified a source, but the free API is capped at 500 records per request. However, there are over 1,000 car chargers on the public network in the republic of Ireland. The provider does not appear to offer a premium API with a higher limit and the only parameter I am able to include in my API request is country. The API is designed primarily for use with Geographic Mapping. I have fallen back to contacting individual providers. ESB E-cars is the main public service provider, so I have put in a request with them for access to their directory of chargers.

Week 11: (06/12/20 – 12/12/20):

This week was spent getting familiar with the datasets and taking some notes on possible secondary measures that might be available when combining the data. I also identified the exploratory analysis that I want to complete for each dataset.

In addition to studying the data, I also read up on database solutions for use with R and the general guidance seems to point towards SQLite. I will set up my database next week and load each of the datasets after I have completed the transformation actions required.

Week 12: (13/12/20 – 19/12/20):

There was not much work completed on the project this week as we had 3 deliverables for other modules this week. This will put me behind on my exploratory analysis. My next opportunity to catch up will be over the Christmas break. I also need to start preparing for the mid-term submission.

I had a call with Ian this week also, we talked through the mid-term submission rubric plus the status of my datasets. I had some queries around the plan for the application of machine learning in my project and Ian was able to give me some good direction.

Week 13: (20/12/20 – 26/12/20):

This is the week of the mid-term submission. A lot of preparation needed this week. Luckily, I had included this time in my project plan, so I am not going to fall any further behind. I still plan to catch up with my exploratory analysis over the Christmas break. We also have three TABA's due in the first week of January, so it will be a busy Christmas.

*Month: January 2021 (Week 14 – 18)**Week 14 + 15: (27/12/20 – 09/01/21):*

The three TABA's were completed and submitted in this period.

Week 16: (10/01/21 – 16/01/21):

The plan for this week was to complete all exploratory analysis and this was complete. This was in the form of histograms, bar charts, scatter plots and checking for missing data. I mostly used ggplot2 and missingness map. This also allowed me to document what was needed in the transformation stage for each of my datasets. My SQLite database is also set and ready to go. Going forward, I will be writing my data frames to a table on the database instead of writing to CSV. This will also make the sharing of the population dataset easier. I have used the Htmltab library to extract the population details from a Wikipedia page (I checked the numbers first versus the CSO). The table also contains the list of counties, and their density, which I will use in the data transformations stage. Each dataset was also saved to my database.

Week 17: (17/01/21 – 23/01/21):

The transformation of the data from its raw form to tidy data is the task for this week. The first stumbling block was trying to find the cleanest way to search the address of each Garda station to extract the county. In the end, I used the county column from my population dataset to create a vector of counties, and then using a for loop and grepl, I was able to create a new column identifying the county. I went on to use this method for the same purpose in a few other datasets.

Most of the other datasets were fairly clean, with the property sale prices and the outpatient waiting list datasets requiring the most work. With the property sale price dataset, I needed to remove outliers from the dataset. Similarly, the list of average monthly rental prices was by location, so I needed to identify the county plus remove observations where no average was available. The other transformation challenge was with the outpatients waiting list. This dataset provided a count of patients, per hospital, per age bracket and per waiting time band. In addition to the planned

transformation, I have split this dataset into the following four separate datasets by county (location of hospital)

1. Average waiting time – child patient
2. Average waiting time – adult patient
3. Total number of patients waiting – adult patient.
4. Total number of patients waiting – child patient.

I will complete two t-tests on the datasets (wait time & patients waiting), next week and decide then if I need to use all four datasets or if 2 is suffice.

I also received a reply for ESB e-cars with a dataset this week, they provided a dataset of their public eCar charging network. Separately, EasyGo also replied to me with a link to a map page on their site which has all their public chargers, plus the ESB network and the Ionity network. There is a list feature on the site, this will allow me to view a list per province for all the above. I tried to use a web scraper to extract the data from this page, but this failed as the list is generated dynamically based on the visible area shown on the map and is not stored on the page. I was able to copy the list for each province into excel (5 cells per charge point – all in a single column), there was a good bit of excel work done to convert the data into a table. I will also load this data set to Kaggle when I have it tidy, as it might prove useful for someone else.

Week 18: (24/01/21 – 30/02/21):

We are back at class this week, but work continues. This week I have been adding a normalization calculation to my data. This is in two forms, the first is to make the data comparable per county. An example is the number of crimes per county, without the normalization, it would look like Leitrim has hardly any crime when compared Cork. However, if I merge the data set with the population data set and calculate the number of crimes per 100k population, Cork and Leitrim have a very similar crime rate. The next step in the normalization is to make each data set comparable to the other data sets which are measured in different units. To do this I will normalize the data to have values between 0 and 1 (the lower the number the better).

The final task for this week is to add the details of the transformation and charts from the exploratory analysis to the project report. I need to get in the habit of doing this as I work.

I have also reached out to Ian, for a meeting in the next few days to give him an update on the project.

Month: February 2021 (Week 19 – 22)

Week 19: (31/01/21 – 06/02/21):

I started this week off by going back to finish the transformation of my outpatient's data set. When I summarised the transformed data in a chart, it highlighted a problem that was not obvious when I had reviewed the data during my exploratory analysis. The outpatient list is grouped by hospital, which I then merged with a table containing each hospital address, however not every county has a hospital with an outpatient's waiting list. The result of this find was that I can no longer use this dataset in my analysis.

I set about looking for a replacement dataset, preferably something in a similar domain as this topic was of interest to most people that I had discussed my project with. I recalled reading an article when doing my initial search for data sets, about how the average amount of time that Irish people spend getting to an emergency department varies considerably depending on your address. I did some research and found the data set used to produce the article on the geohive website (Ordnance Survey Ireland). I spent the remainder of my free time this week getting this data set up to the same stage as my other data sets.

I also had my scheduled meeting with Ian on the Wednesday evening. We discussed my progress since our last meeting (Christmas week). He seemed happy with my progress so far and we agreed to meet weekly for this month and then review the schedule after that depending how far I have gotten with the project. I shared my plan with Ian for the following week, I had added some details of the data transformation to my report last week, but I had not been keeping the report up to date, so I am going to spend any free time next week on the report. We also discussed the results of my mid-term. I was happy with my grade and Ian just confirmed that he felt it was in line with the amount of effort I had put into the submission.

Week 20: (07/02/21 – 13/02/21):

This week's task as planned is to get the report up to date. I had been adding single line notes as I worked through the project, but I now needed to flesh that out and add the detail. I also want to change the layout of the report slightly. I think it will read better if the sections are in the same order as my methodology, so I am going to rearrange the sections.

I have completed the data selection, exploratory analysis, and data transformation stages of the project, so my aim is to get these three sections of the report completed and into a submittable state.

Wednesday's meeting with Ian went ahead as planned, and we discussed my plans for the week ahead. My timetable at mid-term submission was about 10 days / 2 weeks behind. I had put at least 1hr/2hrs per day into my project in the period between the TABA submission and when we returned to class and this had paid off as I am now back on target to complete the project as planned in early April.

Week 21: (14/02/21 – 20/02/21):

With some work done on the report last week, it was back to the data this week. With all the transformation now completed, and each of my data sets now tidy and saved into my database. It was now time to join my nine data sets into a single dataset.

I did this one at a time and completed some checks after each merge. I am using an inner join, and I need to check for consistency in the merge using the County column. The new combined data set has 26 rows (1 per county) and 35 columns.

I used a subset of the new combined data set (columns containing normalised (0 – 1) data only) to create a correlation matrix. This will allow me to visually see the strength of the relationships between my variables. I also used this to create a table showing me the correlation coefficients and p-values between each of my variables. This information will be used next week when I run my regression random forest.

Week 22: (21/02/21 – 27/02/21):

I have two tasks planned for this week, firstly I want to start get my head around the regression random forest. My plan is to pick a variable using the correlation matrix created last week that I want to try predicting. I will then split my data into training & test data (80:20) and see how it goes. This week, I want to only get it working so I will start with a basic version and then next week I will use different combination of variables and mtry values to see which works best.

My second task this week was to start the Shiny training webinars on the R-Bloggers site. At the time of writing, I am about 60% through the introduction webinars and I really like it. I found it hard to stop working on the samples as I really enjoyed using it and have even started on my first report using my data.

Wednesday's meeting with Ian went well. Work on the project will slow down for the next few weeks as the workload from the other models has upped a gear this week, with two large CA's announced this week. We have agreed to move our current weekly meeting to a fortnightly meeting.

*Month: March 2021 (Week 23 – 26)**Week 23: (28/02/21 – 06/03/21):*

There was not much work put into the project this week, what little free time I had was spent working on the Data & Web Mining CA.

Week 24: (07/03/21 – 13/03/21):

This week I got back to my modelling work in R. I have added three rpart decision trees with various arguments applied. I have also run a 'tuner' function on my dataset to identify the best mtry to use for my Random Forest.

I have done some research on ways to evaluate regression models and have settled on Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE). I have calculated these for each decision tree and the Random Forest. The results have been merged into a data frame and visualised to show the best option for my data.

Week 25: (14/03/21 – 20/03/21):

Most of any free time this week was taken up by my advance business data analysis CA that is due on the 29th. I wanted to concentrate on it this week to get it out of the way. For my project, the goal for this week was to draw up my plans for Shiny and what I hope to be able to achieve. My original plan was to be complete the data visualisation by March 23rd, this will now be complete around the April 3rd, putting me 10 days behind schedule. This is ok, as I had a planned finish date of April 12th (1 month early) to allow for so over runs and review of code and documentation.

Week 26: (21/03/21 – 27/03/21):

With the ABDA CA uploaded, it was on to Shiny this week, I started by creating three rough (but working) versions of the reports I wanted. This took a couple of days as some of the outputs were not what I was expecting. I added some “write to CSV” and “cat(file=stderr())” functions to my server.R file so that I could see what was happening with my variables as the code has been run. Several runs, later and I had mapped how my inputs were transforming my variables. Problem identified and resolved.

For the rest of this week, I want to turn the rough workings of my first Shiny report into the final production version of the report. Next week, I can work on the other 2 reports to complete them to the same standard.

*Month: April 2021 (Week 27 – 31)**Week 27: (28/03/21 – 03/04/21):*

Target task for this week was to complete my second and third Shiny reports. Once this was complete, I set about publishing the report to shinyapps.io. This provided harder than the tutorial or sample report that I had published a couple of months back. The published application was unable to read the CSV files that I had uploaded. After reading over the Q&A section, encoding was my issue. I resaved the R files with UTF-8 encoding and added the encode = “UTF-8” to the read.csv function, but no joy. The problem was solved thanks to article on RStudio Community which suggested converting the CSV to XLSX and use the Readxl library to load into R. This worked perfectly. I am happy with my Shiny report as it stands, it contains the three reports that I proposed. I am going to concentrate on getting my results & findings into the project report over the next couple of weeks. If I have some time at the end, I may revisit Shiny to add an additional report or two.

Week 28: (04/04/21 – 10/04/21):

With the Shiny report done, it was back to the report. I spent the week adding all the steps and outputs from the data mining process to the report. This is in a very rough format initially, but once it is all there, I can start to structure it better and get it into shape. I had a chat with my supervisor this week as well. He seems happy with my progress and is confident I will be finished with plenty of time to spare to review and tweak some elements if required. We have scheduled another catch up for the end of the month.

Week 29: (11/04/21 – 24/04/21):

There was very little free time in this 2-week period to do any work on the project, as both the Data & Web Mining and the Advanced Business Data Analysis TABA's have landed and required my full attention.

Week 31: (25/04/21 – 01/05/21):

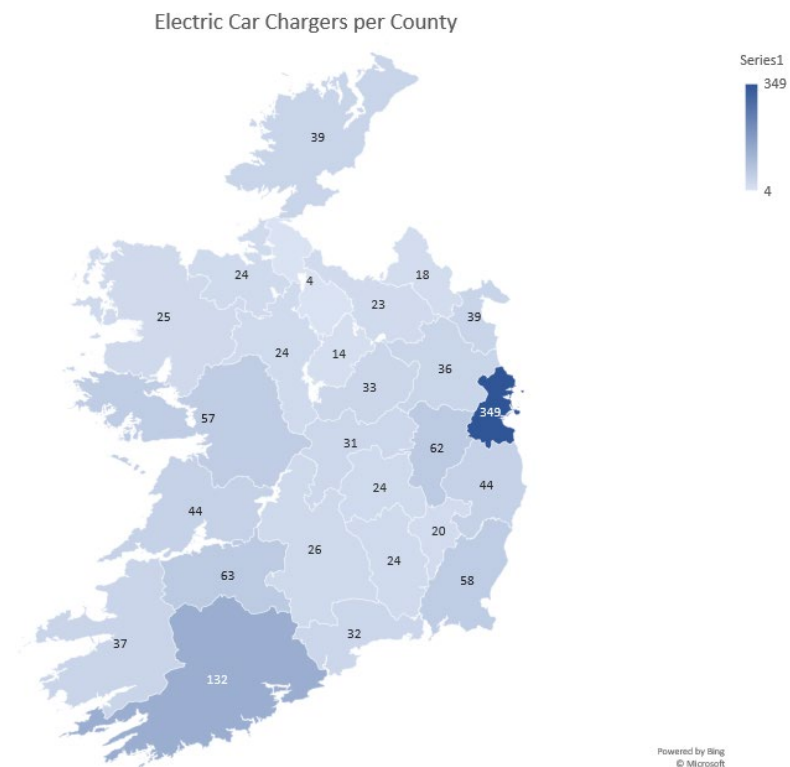
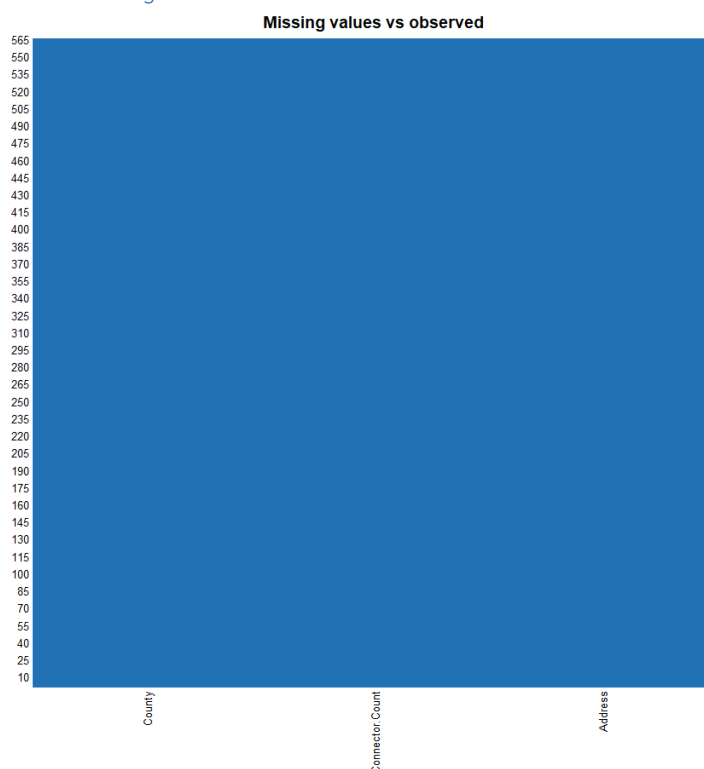
I finished the review of the two TABA's and went back to my report in the second half of the week. The two tasks left to do, are to tidy the results section of the report (everything is just added at the moment, no real structure) and I also need to amend the report as a whole to make it more formal. I have started on the formal conversion, its slower going than I thought it would be, but it is a bank holiday weekend and we are still in lockdown, so I anticipate this been completed over the weekend. All going well, the report will be completed with the next week.

Other materials used.

PX-Win:

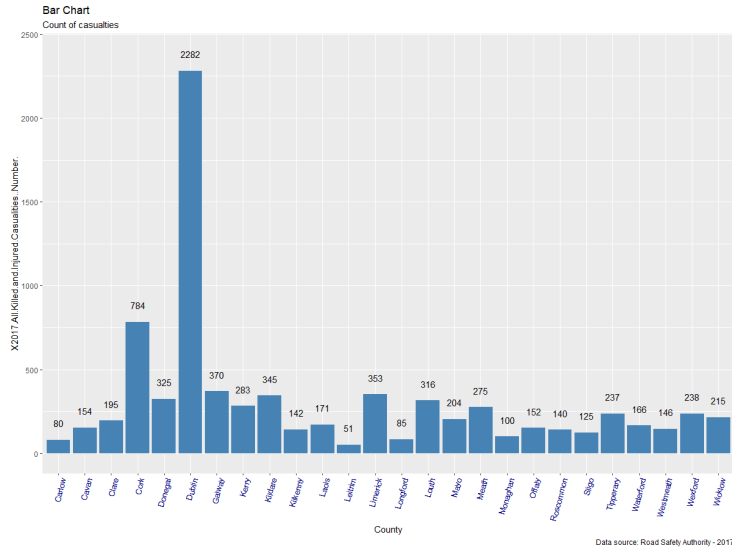
Software developed by the Swedish Central Statistics Office to read the .PX format used by the Irish Central Statistics Office

Appendix 3

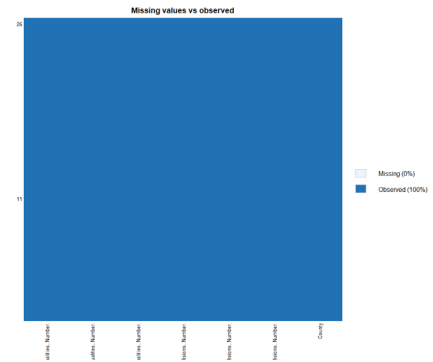
Map Chart – Electric Car Chargers by County*Missmap – Electric Car Chargers*

Appendix 4

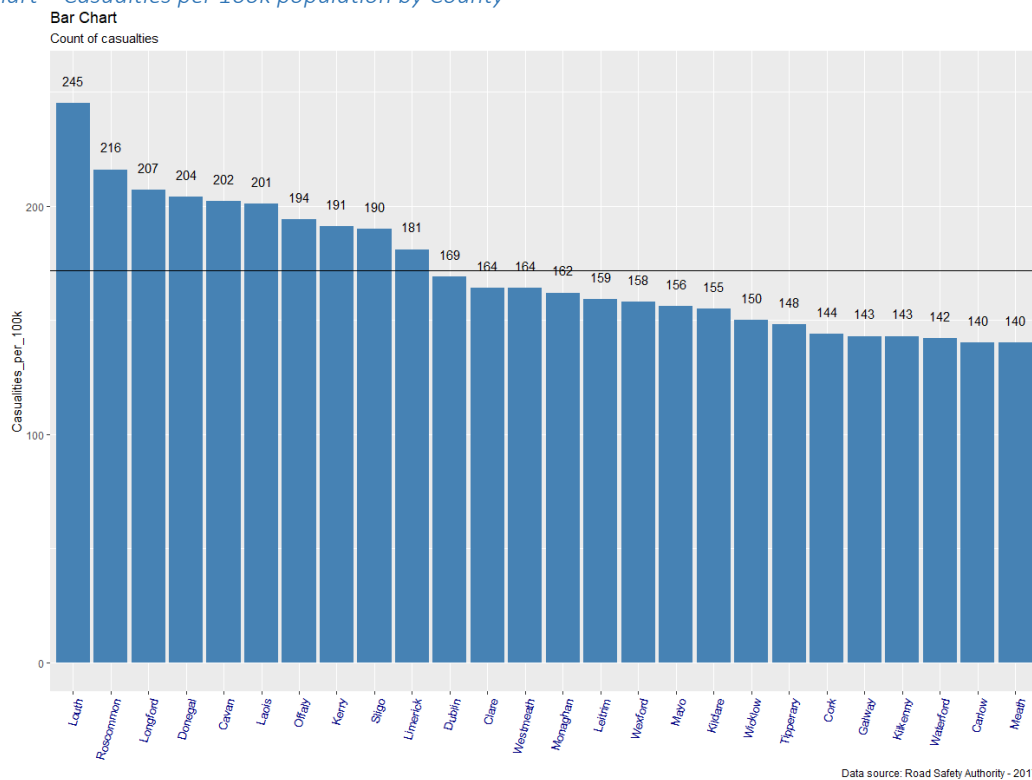
Bar Chart – Road Accident Casualties by County



Missmap - Casualties



Bar Chart – Casualties per 100k population by County



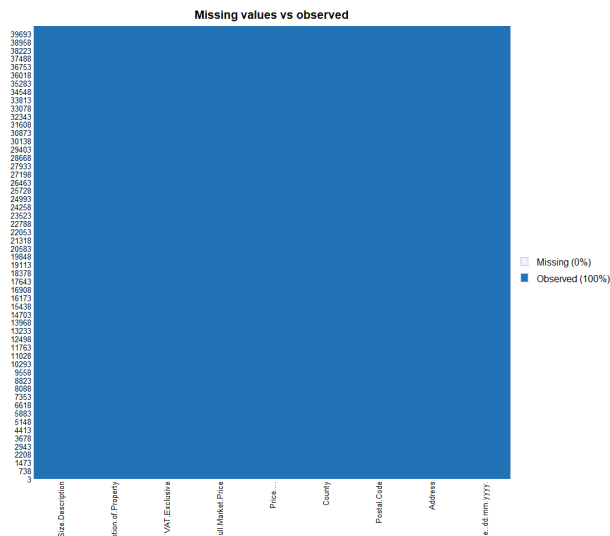
Appendix 5

sapply(DF,function(x) sum(is.na(x))) - Property Sales

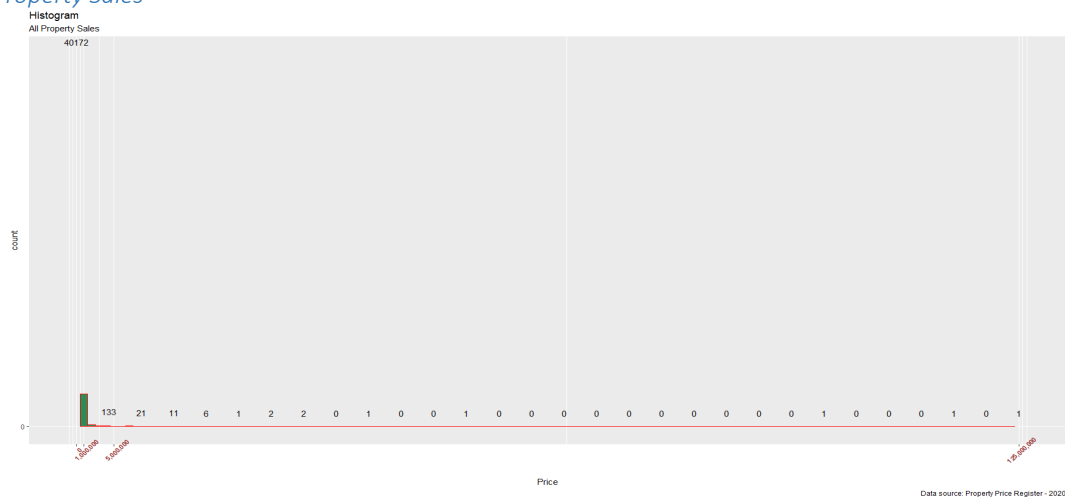
```
> sapply(PPR,function(x) sum(is.na(x)))
```

Date.of.Sale..dd.mm.yyyy.	Address	Postal.Code	County	Price...
0	0	0	0	0
Not.Full.Market.Price	VAT.Exclusive	Description.of.Property	Property.Size.Description	
0	0	0	0	

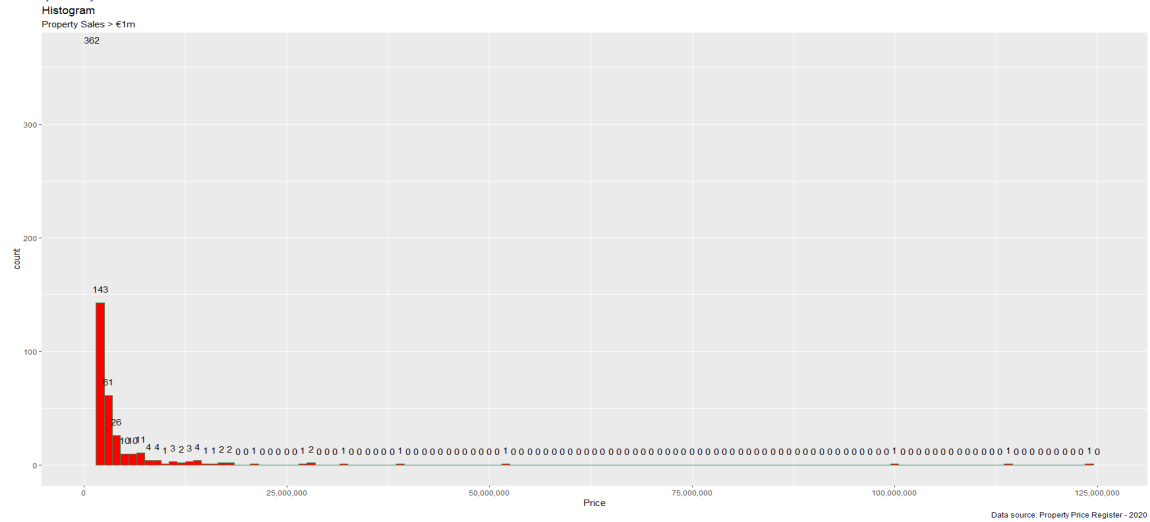
Missmap – Property Sales



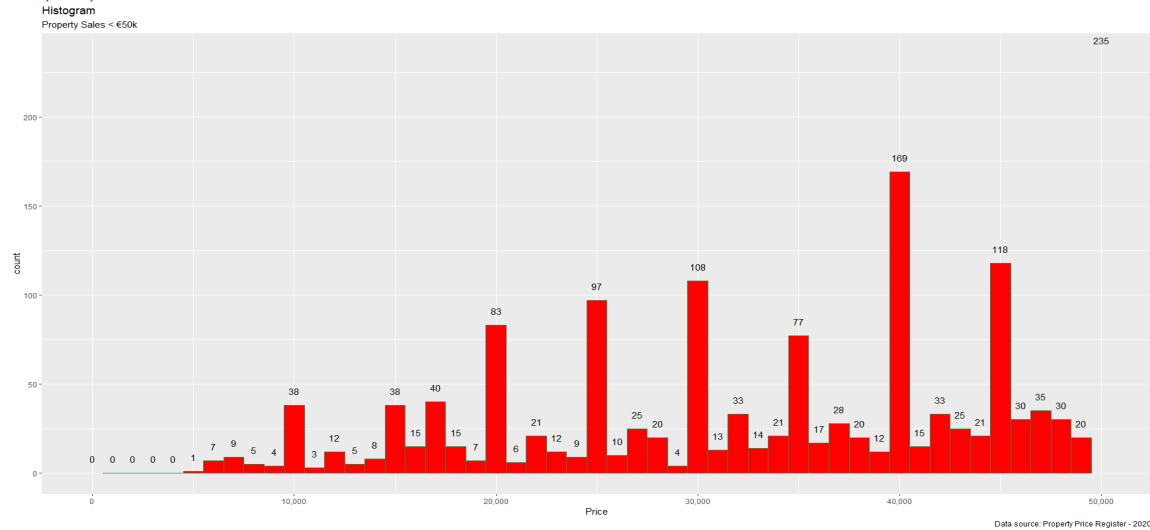
All Property Sales



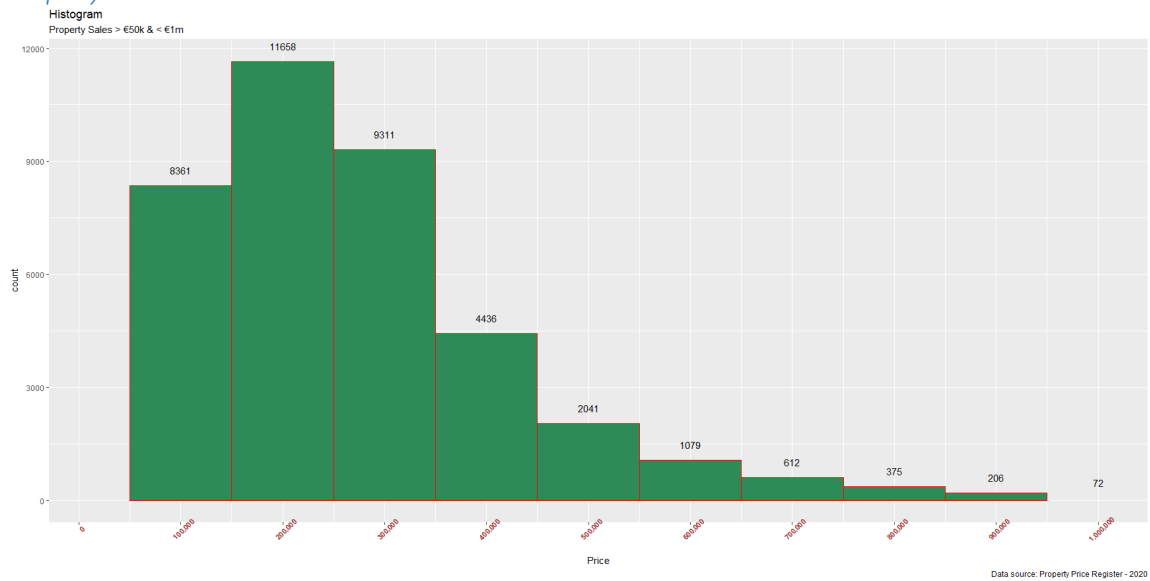
Property Sales > €1m



Property Sales < €50k



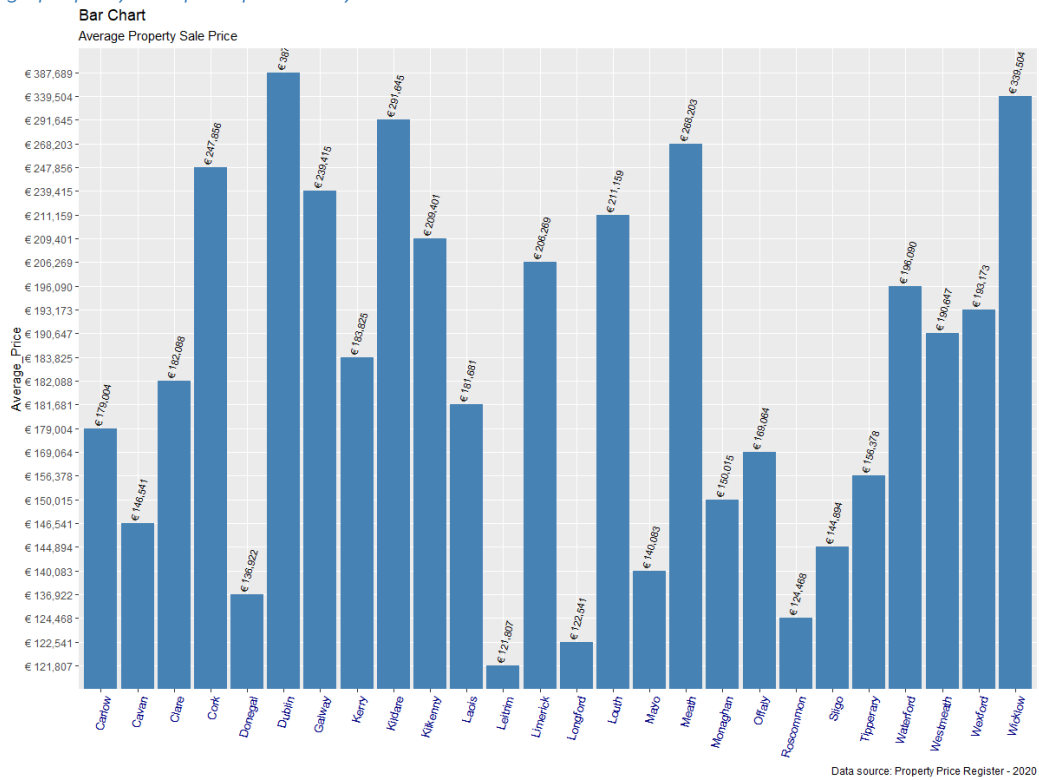
Property Sales > €50k & < €1m



Property Sales > €50k & < €1m by county



Average property sale price per county



Appendix 6

```
sapply(DF,function(x) sum(is.na(x))) - Attractions
```

```
> sapply(Attractions,function(x) sum(is.na(x)))
```

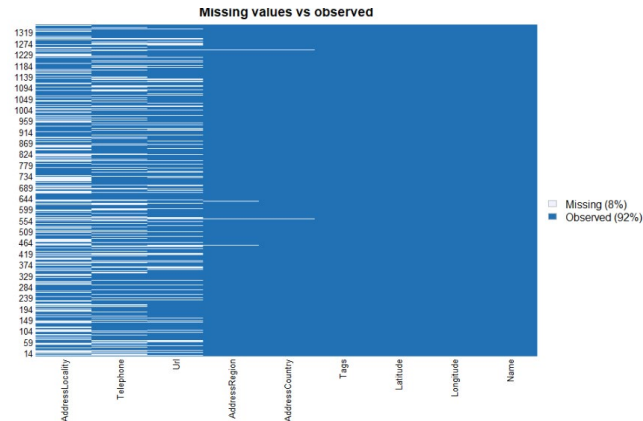
```

Name      0      Url      Telephone      Longitude
0      225      324      0

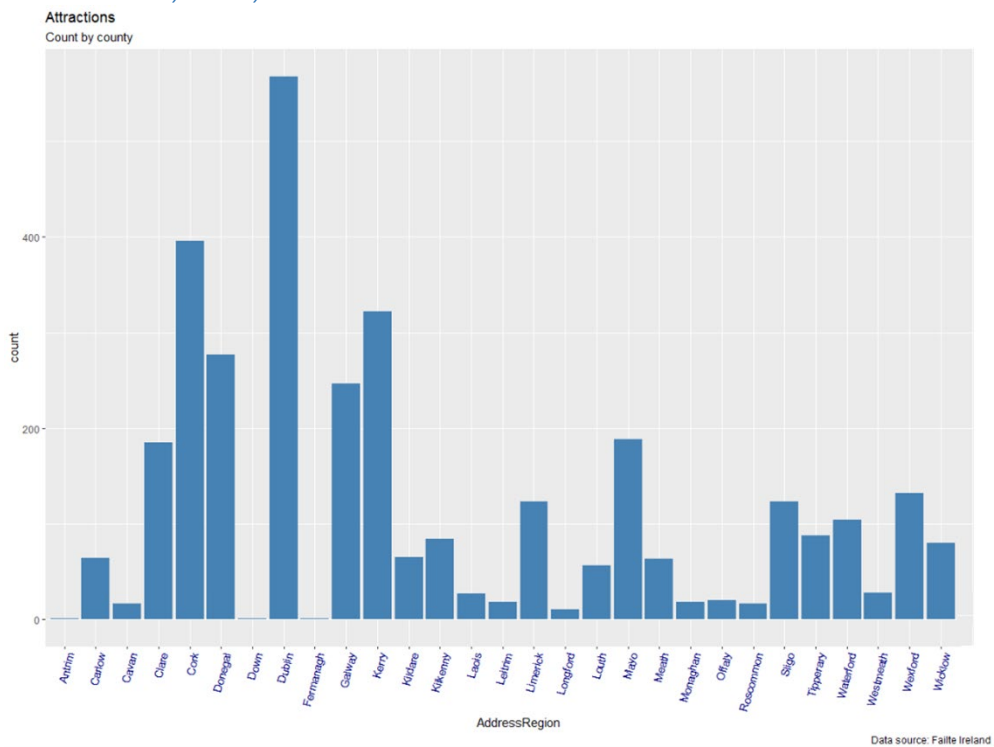
Latitude  AddressRegion  AddressLocality  AddressCountry  Tags
0           3           435           3           0

```

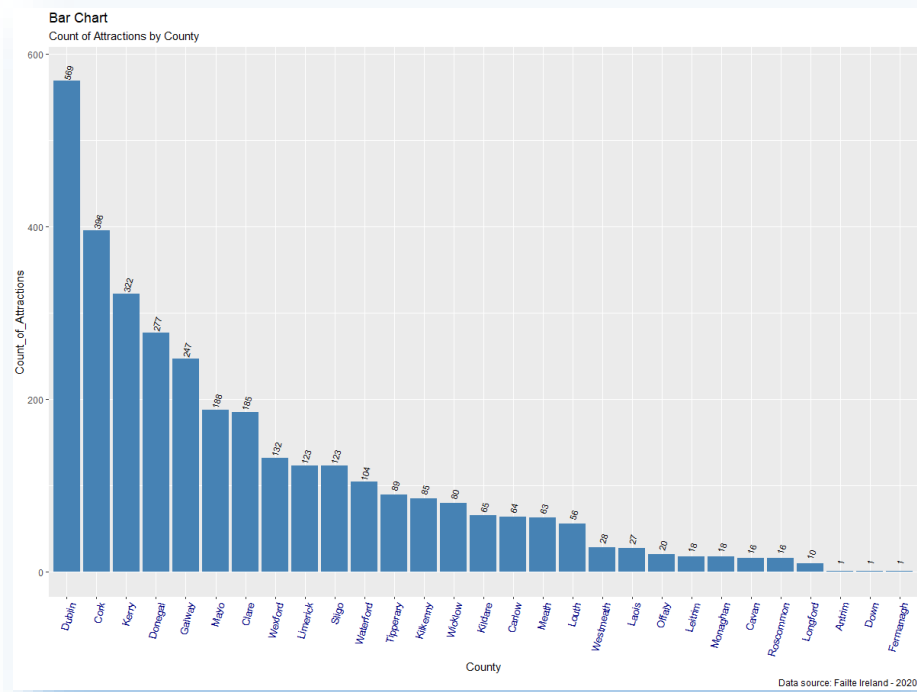
Attractions missmap



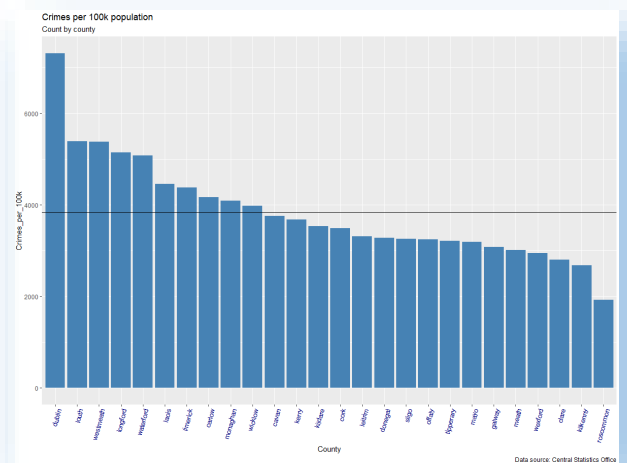
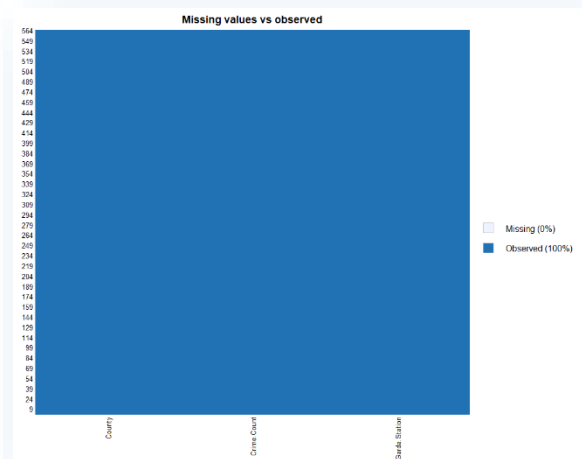
Bar Chart - Attractions by County



Appendix 7

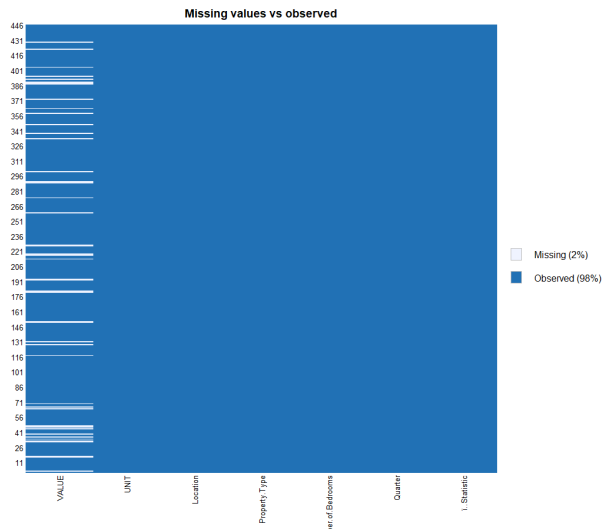


Appendix 8

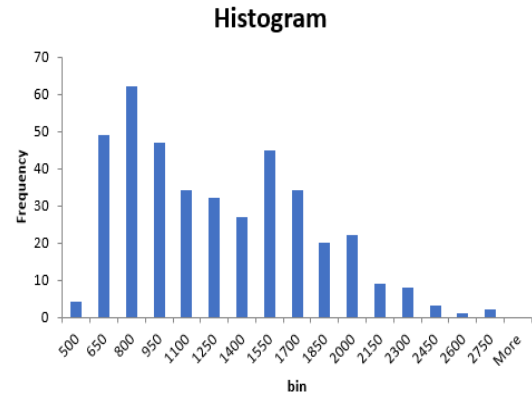


Appendix 9

Monthly Rental Costs Missmap



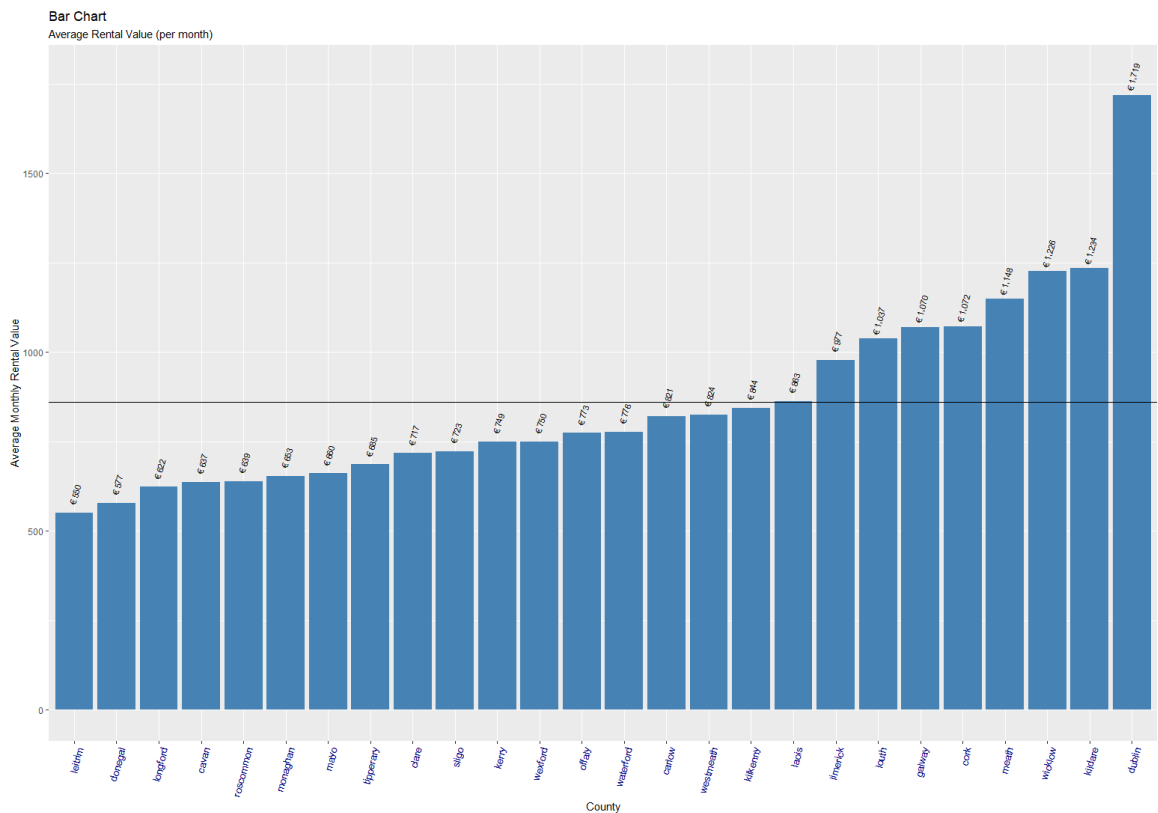
Histogram – Monthly Rental Costs



`sapply(DF,function(x) sum(is.na(x))) - Monthly Rental Costs`

```
> sapply(Initial_rent_DF,function(x) sum(is.na(x)))
i..Statistic      Quarter Number.of.Bedrooms  Property.Type      Location
0                0                0                0                0
UNIT              VALUE
0                47
```

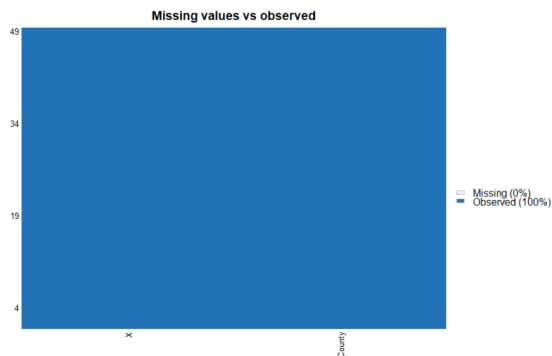
Average Rental value per County



Data source: Residential Tenancies Board - 2020

Appendix 10

Missmap - Pharmacies

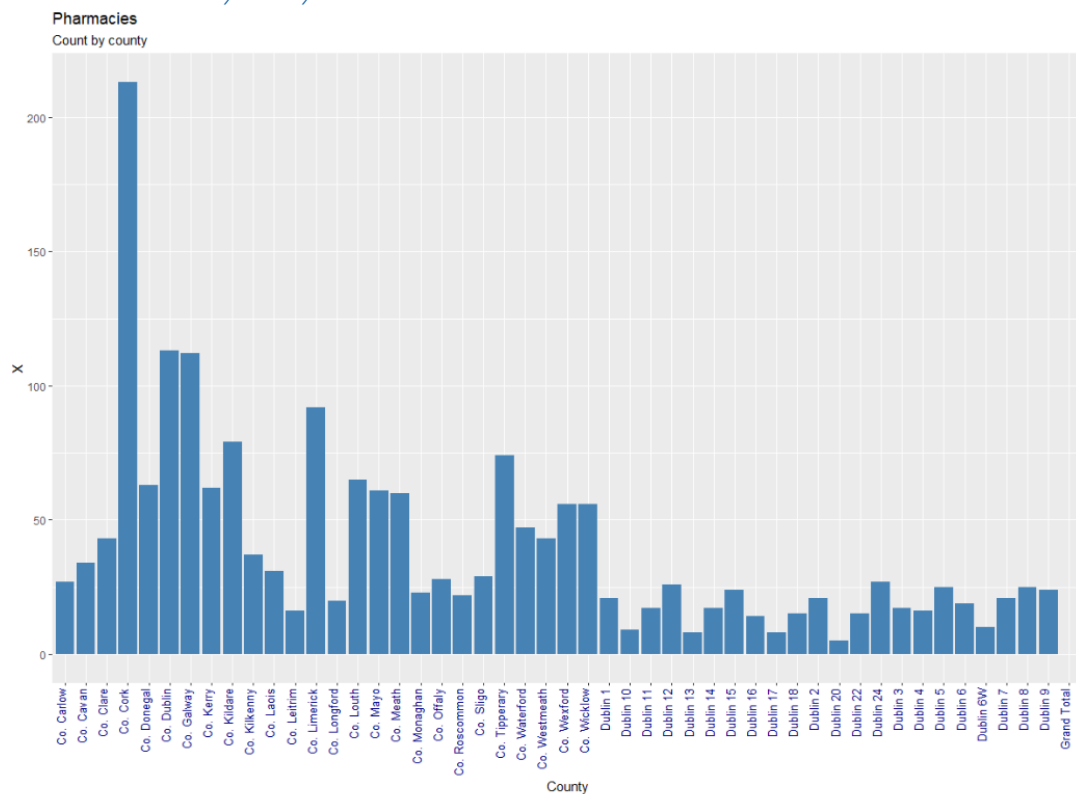


sapply(DF,function(x) sum(is.na(x))) - Pharmacies

```
> #check for NA's
> sapply(Pharma,function(x) sum(is.na(x)))
```

County	X
0	0

Registered Pharmacies by county

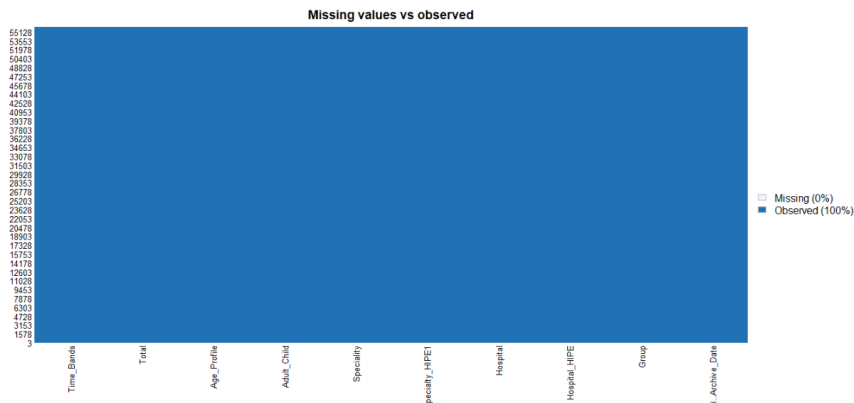


Appendix 11

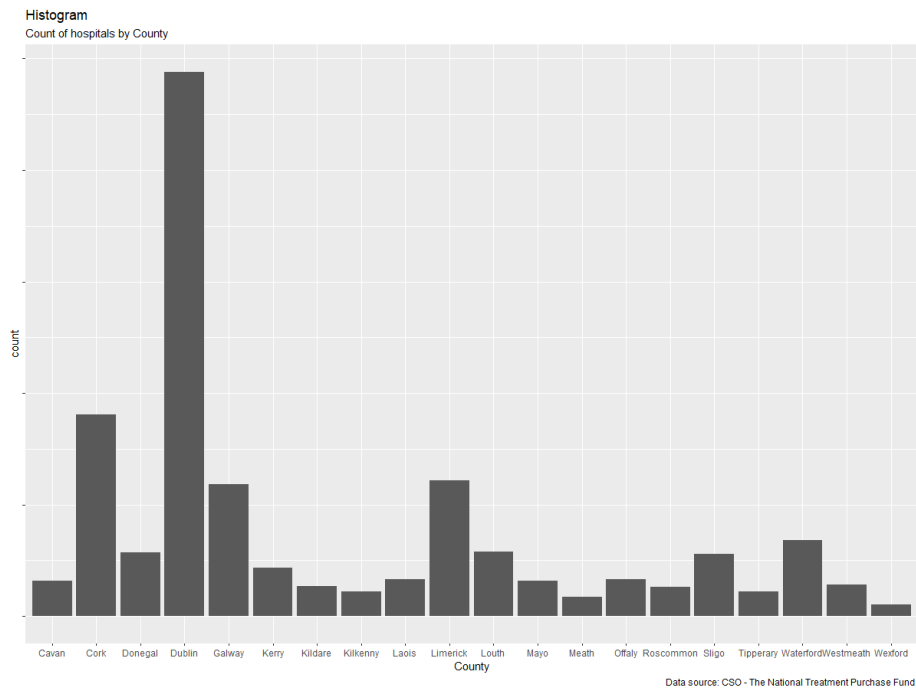
sapply(Df,function(x) sum(is.na(x))) - Outpatients

```
> sapply(initial_df,function(x) sum(is.na(x)))
Archive_Date      Group      Hospital_HIPE      Hospital      Specialty_HIPE1      Specialty
0                0                0                0                0                0
Adult_Child      Age_Profile      Time_Bands      Total
0                0                1                0
```

Outpatients Missmap



Hospitals with a waiting list by County - plot



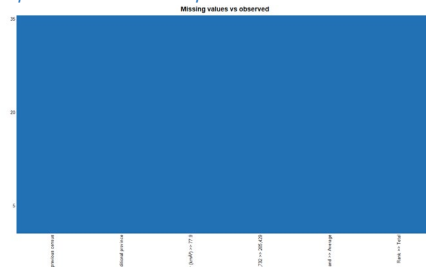
Hospitals with a waiting list by County - table

```
> table(workingFinal_df$County)
```

Cavan	Cork	Donegal	Dublin	Galway	Kerry	Kildare	Kilkenny
158	904	284	2440	592	218	133	110
Laois	Limerick	Louth	Mayo	Meath	Offaly	Roscommon	Sligo
164	608	289	160	88	164	131	280
Tipperary	Waterford	Westmeath	Wexford				
112	342	140	53				

Appendix 12

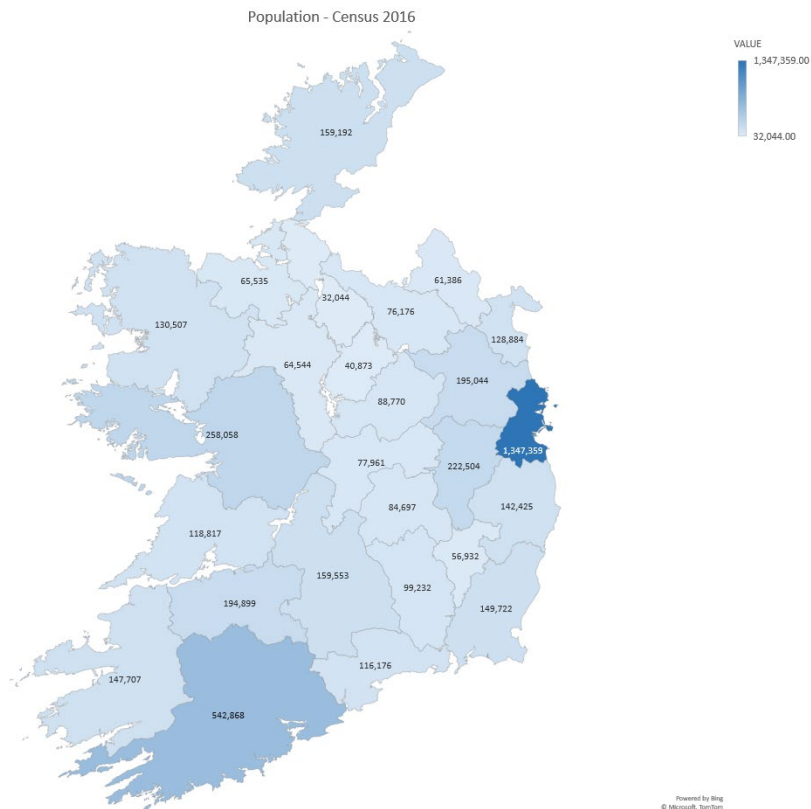
Population Missmap



sapply(DF,function(x) sum(is.na(x))) - Population

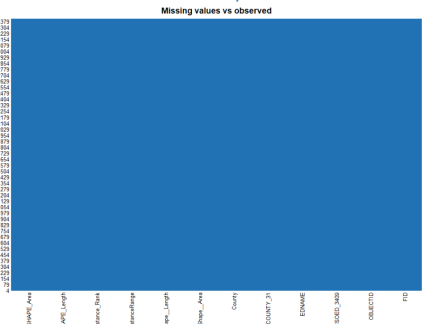
```
> sapply(Counties_df,function(x) sum(is.na(x)))
Rank >> Total County >> Island of Ireland >> Average
Population >> 6,573,732 >> 205,429 Density (kmA²) >> 77.9
Traditional province Change since previous census
0 0 0 0
```

Map Chart – population per County



Appendix 13

Distance to ED Missmap

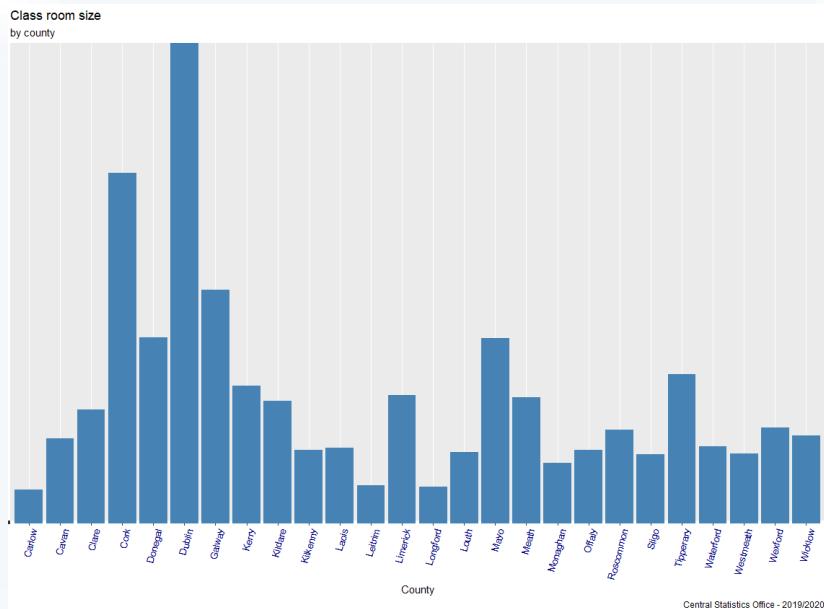


sapply(DF,function(x) sum(is.na(x))) – Distance to ED

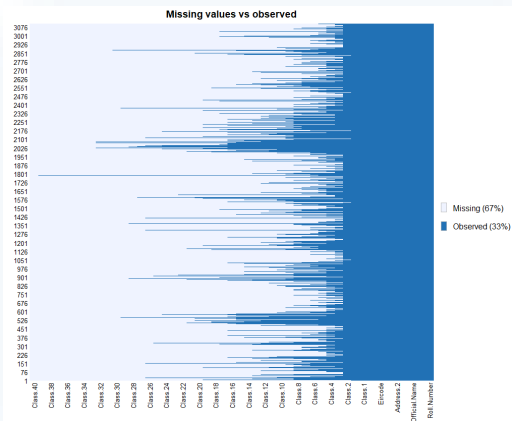
```
> #check for NA's
> sapply(ED_df,function(x) sum(is.na(x)))
FID OBJECTID CSOED_3409 EDNAME COUNTY_31 County Shape_Area
0 0 0 0 0 0 0
Shape_Length DistanceRange Distance_Rank SHAPE_Length SHAPE_Area
0 0 0 0 0 0
```

Appendix 14

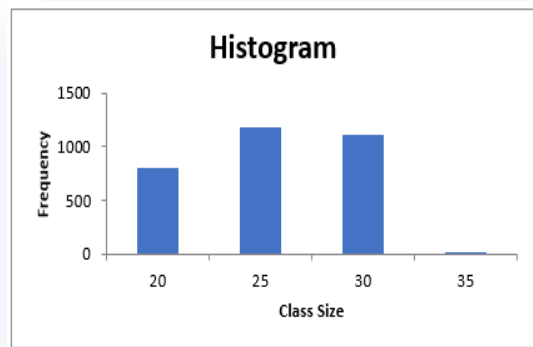
Bar Chart - Classroom size



Missmap – Classroom size



Histogram – classroom size



Appendix 15

Combined Data Frame – Normalised Data

County	Crime	Class.size	Car.Chargers	Road.Casualties	Property.Price	Monthly.Rent	Attractions	Pharmacies	Ed.Dept	normalise.sum
Carlow	0.4165586	0.8471233	0.9477392	0.00000000	0.215120709	0.23171723	0.8941369	0.9596607	0.7897607	5.301817
Cavan	0.3395073	0.5969276	0.9753213	0.59047619	0.093027924	0.07399546	0.9967372	0.9832234	0.4627592	5.111976
Clare	0.1992962	0.1786574	0.9731506	0.22857143	0.226720168	0.14255638	0.9234803	0.9928369	1.0000000	4.865269
Cork	0.2907946	0.6044277	0.9600772	0.03809524	0.474079891	0.44667722	0.9241695	0.9626767	0.6741892	5.375187
Donegal	0.2517133	0.1936798	0.9857886	0.60952381	0.056849007	0.02333005	0.9189926	0.9920829	0.8438001	4.875760
dublin	1.0000000	0.8115157	0.0000000	0.27619048	1.000000000	1.00000000	0.0000000	0.0000000	0.0000000	4.087706
Galway	0.2148546	0.3485003	0.9819682	0.02857143	0.442329733	0.44454619	0.9443434	0.9817154	0.6362341	5.023063
Kerry	0.3263567	0.3318327	0.9866530	0.48571429	0.233253254	0.17059352	0.9030864	0.9922714	0.7701107	5.199872
Kildare	0.2993147	0.9191299	0.9095369	0.14285714	0.638772427	0.58512559	0.9474886	0.9283695	0.4233721	5.793967
Kilkenny	0.1396555	0.7243513	0.9764056	0.02857143	0.329445867	0.25129017	0.9442396	0.9830349	0.3816571	4.758651
Laois	0.4699018	0.5552807	0.9697044	0.58095238	0.225191322	0.26753391	0.9845715	0.9822809	0.4239521	5.459369
Leitrim	0.2580107	0.3172474	1.0000000	0.18095238	0.000000000	0.00000000	0.9917556	0.9971725	0.7493948	4.494533
Limerick	0.4550843	0.5956174	0.9464603	0.39047619	0.317666074	0.36539860	0.9378302	0.9538172	0.4673126	5.429663
Longford	0.5969624	0.3569611	0.9729162	0.63809524	0.002760872	0.06188540	0.9953206	0.9819039	0.7874423	5.394248
Louth	0.6458603	0.9108935	0.8818319	1.00000000	0.336058898	0.41648338	0.8998432	0.8686145	0.5934498	6.553036
Mayo	0.2026301	1.0000000	0.9949228	0.15238095	0.068738549	0.09397064	0.9559129	0.9956645	0.7359794	4.200200
Meath	0.2009631	0.0000000	0.9658834	0.00000000	0.550603780	0.51143823	0.9662473	0.9679548	0.4059108	5.569001
Monaghan	0.4011854	0.6736724	0.9699872	0.20952381	0.106091911	0.08782575	0.9875873	0.9828464	0.9347613	5.353481
Offaly	0.2456010	0.6847690	0.9664448	0.51428571	0.177738084	0.19083401	0.9941857	0.9901979	0.5512329	5.315289
Roscommon	0.0000000	0.1593635	0.9819113	0.72380952	0.010009463	0.07580983	1.0000000	1.0000000	0.9983343	4.949238
Sligo	0.2480089	0.2533060	0.9725456	0.47619048	0.086831427	0.14765715	0.9025989	0.9868049	0.5096426	4.583586
Tipperary	0.2389331	0.5672340	0.9906049	0.07619048	0.130023992	0.11554624	0.9763627	0.9836004	0.7137032	4.792199
Waterford	0.5847379	0.7705140	0.9612050	0.01904762	0.279381636	0.19364971	0.9192606	0.9687088	0.4129364	5.109442
Westmeath	0.6412299	0.6377100	0.9591770	0.22857143	0.258912455	0.23478112	0.9853599	0.9721018	0.3743659	5.292209
Wexford	0.1891091	0.8076104	0.9417227	0.17142857	0.268412652	0.17079036	0.9192120	0.9715363	0.5798314	5.019653
Wicklow	0.3822930	0.8258594	0.9491118	0.09523810	0.818770997	0.57832171	0.9457682	0.9641847	0.8975955	6.457144

Combined Data Frame – Key Data

County.norm	Crimes_per_100k	Average_Class_Size	Chargers.per.100km2	Pharmacies_per_100km2	Attractions.per.100km2	Ave.Distance_to_ED	Casualties_per_100k	Average_Price	Average.Rent
Carlow	4173	23.80369	2.2148394	2.99	7.0874862	36.851852	140	179004.2	821.37
Cavan	3757	22.41174	1.1776754	1.74	0.8192524	23.792135	202	146541.9	637.08
Clare	3000	20.08472	1.2593017	1.23	5.2947911	45.248344	164	182088.3	717.19
Cork	3494	22.45347	1.7508953	2.83	5.2526860	32.236181	144	247856.9	1072.54
Donegal	3283	20.16830	0.7840772	1.27	5.5689586	39.010067	204	136922.6	577.88
dublin	7323	23.60559	37.8524946	53.90	61.7136659	5.310559	169	387689.8	1719.07
Galway	3084	21.02963	0.9277344	1.82	4.0201823	30.720339	143	239415.1	1070.05
Kerry	3686	20.93690	0.7515742	1.26	6.5407272	36.067073	191	183825.4	749.95
Kildare	3540	24.20429	3.6513545	4.65	3.8280330	22.219101	155	291645.8	1234.31
Kilkenny	2678	23.12065	1.1369019	1.75	4.0265277	20.553097	143	209401.3	844.24
Laois	4461	22.18004	1.3888889	1.79	1.5625000	22.242268	201	181681.8	863.22
Leitrim	3317	20.85576	0.2496879	1.00	1.1235955	35.239726	159	121807.4	550.62
Limerick	4381	22.40445	2.2629310	3.30	4.4181034	23.973988	181	206269.2	977.57
Longford	5147	21.07670	1.2681159	1.81	0.9057971	36.759259	207	122541.5	622.93
Louth	5411	24.15847	4.6931408	7.82	6.7388688	29.011628	245	211159.6	1037.26
Mayo	3018	19.09078	0.4406063	1.08	3.3133592	34.703947	156	140083.8	660.42
Meath	3009	24.65421	1.5325670	2.55	2.6819923	21.521739	140	268203.3	1148.21
Monaghan	4090	22.83871	1.3782542	1.76	1.3782542	42.642857	162	150015.4	653.24
Offaly	3250	22.90044	1.5114578	1.37	0.9751341	27.325581	194	169064.9	773.60
Roscommon	1924	19.97738	0.9298721	0.85	0.6199148	45.181818	216	124468.8	639.20
Sligo	3263	20.50003	1.2820513	1.55	6.5705128	25.664557	190	144894.4	723.15
Tipperary	3214	22.24654	0.6029685	1.72	2.0640074	33.814286	148	156378.5	685.63
Waterford	5081	23.37748	1.7084891	2.51	5.5525894	21.802326	142	196090.1	776.89
Westmeath	5386	22.63863	1.7847485	2.33	1.5143321	20.281905	164	190647.7	824.95
Wexford	2945	23.58386	2.4410774	2.36	5.5555556	28.467742	158	193173.6	750.18
Wicklow	3988	23.68539	2.1632252	2.75	3.9331367	41.158537	150	339504.2	1226.36

Appendix 16

Method:	Anova for a regression tree & class for a classification tree
minsplit	The minimum number of observations that must exist in a node in order for a split to be attempted.
minbucket	The minimum number of observations in any terminal <leaf> node. If only one of minbucket or minsplit is specified, the code either sets minsplit to minbucket*3 or minbucket to minsplit/3, as appropriate.
maxdepth	Set the maximum depth of any node of the final tree, with the root node counted as depth 0. Values greater than 30 rpart will give nonsense results on 32-bit machines.
xval	number of cross-validations.
usesurrogate	0 means display only; an observation with a missing value for the primary split rule is not sent further down the tree. 1 means use surrogates, in order, to split subjects missing the primary variable; if all surrogates are missing the observation is not split. For value 2, if all surrogates are missing, then send the observation in the majority direction. A value of 0 corresponds to the action of tree, and 2 to the recommendations of Breiman et.al (1984).

Appendix 17

Normalise function – Lower the value the better

```
#####normalize#####
#lower the better

normalise <- function (x){
  return((x-min(x))/(max(x)-min(x)))
}
```

Normalise function – Higher the value the better

```
#####normalize#####
#higher the better

normalise_high <- function (x){
  return((x-max(x))/(min(x)-max(x)))
}
```

Appendix 18

Rpart – Combination 1 – test data

```
Call:
rpart(formula = Crimes_per_100k ~ ., data = train_data, method = "anova",
      control = rpart.control(minsplit = 10, minbucket = 5, maxdepth = 20,
                             xval = 10, usesurrogate = 2))
n= 20
```

```
      CP nsplit rel error      xerror      xstd
1 0.41437982      0 1.0000000 1.0996217 0.4736183
2 0.07379329      1 0.5856202 0.9400938 0.3826779
3 0.01000000      2 0.5118269 1.0868985 0.4410593
```

```
Variable importance
Chargers.per.100km2      Average_Class_Size      Pharmacies_per_100km2      Average.Rent      Average_Price
      27              16              16              14              14
Attractions.per.100km2      Ave_Distance_to_ED      Casualties_per_100k
      10              2              1
```

Node number 1: 20 observations, complexity param=0.4143798

mean=3811.65, MSE=1397263

left son=2 (15 obs) right son=3 (5 obs)

Primary splits:

```
Chargers.per.100km2 < 1.767822 to the left, improve=0.4143798, (0 missing)
Pharmacies_per_100km2 < 1.77 to the left, improve=0.3238175, (0 missing)
Average_Class_Size < 21.05317 to the left, improve=0.2050859, (0 missing)
Casualties_per_100k < 161.5 to the left, improve=0.1780529, (0 missing)
Attractions.per.100km2 < 5.410822 to the left, improve=0.1457184, (0 missing)
```

Surrogate splits:

```
Average_Class_Size < 23.36312 to the left, agree=0.90, adj=0.6, (0 split)
Pharmacies_per_100km2 < 2.075 to the left, agree=0.90, adj=0.6, (0 split)
Average_Price < 279924.5 to the left, agree=0.90, adj=0.6, (0 split)
Average.Rent < 1187.285 to the left, agree=0.90, adj=0.6, (0 split)
Attractions.per.100km2 < 6.654691 to the left, agree=0.85, adj=0.4, (0 split)
```

Node number 2: 15 observations, complexity param=0.07379329

mean=3372.333, MSE=505286.2

left son=4 (8 obs) right son=5 (7 obs)

Primary splits:

```
Chargers.per.100km2 < 1.157289 to the left, improve=0.27207970, (0 missing)
Pharmacies_per_100km2 < 1.77 to the left, improve=0.21549940, (0 missing)
Average_Class_Size < 20.89633 to the left, improve=0.16742500, (0 missing)
Casualties_per_100k < 190.5 to the left, improve=0.12780370, (0 missing)
Attractions.per.100km2 < 1.813254 to the right, improve=0.09642075, (0 missing)
```

Surrogate splits:

```
Average_Class_Size < 21.05317 to the left, agree=0.800, adj=0.571, (0 split)
Pharmacies_per_100km2 < 1.32 to the left, agree=0.800, adj=0.571, (0 split)
Ave_Distance_to_ED < 29.02296 to the right, agree=0.800, adj=0.571, (0 split)
Attractions.per.100km2 < 1.049365 to the right, agree=0.667, adj=0.286, (0 split)
Casualties_per_100k < 174.5 to the left, agree=0.667, adj=0.286, (0 split)
```

Node number 3: 5 observations

mean=5129.6, MSE=1757202

Node number 4: 8 observations

mean=3025.5, MSE=245076

Node number 5: 7 observations

mean=3768.714, MSE=508073.3

Rpart – Combination 3 – test data

```

Call:
rpart(formula = Crimes_per_100k ~ ., data = train_data, method = "anova",
      control = rpart.control(minsplit = 5, minbucket = 2, maxdepth = 20,
                             xval = 5, usesurrogate = 2))
n= 20

      CP nsplit rel error   xerror   xstd
1 0.51925449    0 1.0000000 1.066889 0.4520127
2 0.12997096    1 0.4807455 1.321649 0.3883542
3 0.08759666    2 0.3507745 1.765042 0.5321901
4 0.05008844    4 0.1755812 1.765042 0.5321901
5 0.01242905    5 0.1254928 1.694200 0.5241488
6 0.01000000    6 0.1130637 1.660574 0.5235175

Variable importance
Pharmacies_per_100km2  27  Chargers.per.100km2  25  Attractions.per.100km2  23  Average_Class_Size  6  Average.Rent  6
Average_Price  6  Ave_Distance_to_ED  4  Casualties_per_100k  2

Node number 1: 20 observations, complexity param=0.5192545
mean=3811.65, MSE=1397263
left son=2 (18 obs) right son=3 (2 obs)
Primary splits:
Attractions.per.100km2 < 6.654691 to the left, improve=0.5192545, (0 missing)
Pharmacies_per_100km2 < 6.235 to the left, improve=0.5192545, (0 missing)
Chargers.per.100km2 < 4.172248 to the left, improve=0.5192545, (0 missing)
Ave_Distance_to_ED < 20.4075 to the right, improve=0.5141868, (0 missing)
Average_Price < 315575 to the left, improve=0.2703527, (0 missing)
Surrogate splits:
Chargers.per.100km2 < 4.172248 to the left, agree=1, adj=1, (0 split)
Pharmacies_per_100km2 < 6.235 to the left, agree=1, adj=1, (0 split)

Node number 2: 18 observations, complexity param=0.129971
mean=3527.722, MSE=644815.6
left son=4 (8 obs) right son=5 (10 obs)
Primary splits:
Chargers.per.100km2 < 1.157289 to the left, improve=0.3129293, (0 missing)
Pharmacies_per_100km2 < 1.77 to the left, improve=0.2929228, (0 missing)
Average_Class_Size < 20.07284 to the left, improve=0.2164692, (0 missing)
Attractions.per.100km2 < 1.813254 to the right, improve=0.1307536, (0 missing)
Casualties_per_100k < 143.5 to the left, improve=0.1131744, (0 missing)
Surrogate splits:
Average_Class_Size < 21.05317 to the left, agree=0.833, adj=0.625, (0 split)
Pharmacies_per_100km2 < 1.32 to the left, agree=0.833, adj=0.625, (0 split)
Ave_Distance_to_ED < 29.02296 to the right, agree=0.778, adj=0.500, (0 split)
Average_Price < 142489.1 to the left, agree=0.722, adj=0.375, (0 split)
Average.Rent < 704.39 to the left, agree=0.722, adj=0.375, (0 split)

Node number 3: 2 observations, complexity param=0.05008844
mean=6367, MSE=913936

Node number 4: 8 observations, complexity param=0.05008844
mean=3025.5, MSE=245076
left son=8 (2 obs) right son=9 (6 obs)
Primary splits:
Chargers.per.100km2 < 0.9288033 to the right, improve=0.7139285, (0 missing)
Average_Class_Size < 20.07284 to the left, improve=0.4181971, (0 missing)
Attractions.per.100km2 < 4.797743 to the left, improve=0.2865519, (0 missing)
Ave_Distance_to_ED < 37.53857 to the right, improve=0.2422160, (0 missing)
Casualties_per_100k < 197.5 to the right, improve=0.2422160, (0 missing)

Node number 5: 10 observations, complexity param=0.08759666
mean=3929.5, MSE=601400.2
left son=10 (5 obs) right son=11 (5 obs)
Primary splits:
Attractions.per.100km2 < 2.122246 to the right, improve=0.3684044, (0 missing)
Average_Class_Size < 22.76954 to the right, improve=0.2583388, (0 missing)
Pharmacies_per_100km2 < 2.44 to the right, improve=0.1971766, (0 missing)
Casualties_per_100k < 159.5 to the left, improve=0.1971766, (0 missing)
Average_Price < 219252.3 to the right, improve=0.1971766, (0 missing)
Surrogate splits:
Pharmacies_per_100km2 < 2.44 to the right, agree=0.9, adj=0.8, (0 split)
Casualties_per_100k < 159.5 to the left, agree=0.9, adj=0.8, (0 split)
Average_Price < 219252.3 to the right, agree=0.9, adj=0.8, (0 split)
Average.Rent < 967.88 to the right, agree=0.9, adj=0.8, (0 split)
Average_Class_Size < 23.29291 to the right, agree=0.8, adj=0.6, (0 split)

Node number 8: 2 observations, complexity param=0.01242905
mean=2301, MSE=142129

Node number 9: 6 observations, complexity param=0.01242905
mean=3267, MSE=46102.67

Node number 10: 5 observations, complexity param=0.01242905
mean=3458.8, MSE=105708.6
left son=20 (2 obs) right son=21 (3 obs)
Primary splits:
Chargers.per.100km2 < 1.641731 to the left, improve=0.6571517, (0 missing)
Pharmacies_per_100km2 < 2.65 to the left, improve=0.6571517, (0 missing)
Average_Price < 279924.5 to the left, improve=0.5874456, (0 missing)
Average.Rent < 1187.285 to the left, improve=0.5874456, (0 missing)
Ave_Distance_to_ED < 28.95037 to the left, improve=0.5022415, (0 missing)
Surrogate splits:
Pharmacies_per_100km2 < 2.65 to the left, agree=1.0, adj=1.0, (0 split)
Ave_Distance_to_ED < 28.95037 to the left, agree=0.8, adj=0.5, (0 split)
Average_Price < 279924.5 to the left, agree=0.8, adj=0.5, (0 split)
Average.Rent < 1187.285 to the left, agree=0.8, adj=0.5, (0 split)

```

```

Node number 11: 5 observations,      complexity param=0.08759666
mean=4400.2, MSE=653975
left son=22 (2 obs) right son=23 (3 obs)
Primary splits:
  Pharmacies_per_100km2 < 1.765      to the left,  improve=0.8196755, (0 missing)
  Ave_Distance_to_ED    < 23.0172   to the right, improve=0.2791574, (0 missing)
  Attractions.per.100km2 < 1.244733  to the left,  improve=0.2791574, (0 missing)
  Average_Price         < 175373.3  to the left,  improve=0.2791574, (0 missing)
  Average.Rent          < 799.275   to the left,  improve=0.2791574, (0 missing)
Surrogate splits:
  Average_Class_Size    < 22.29589  to the right, agree=0.8, adj=0.5, (0 split)
  Attractions.per.100km2 < 1.244733  to the left,  agree=0.8, adj=0.5, (0 split)
  Ave_Distance_to_ED    < 23.0172   to the right, agree=0.8, adj=0.5, (0 split)
  Average_Price         < 175373.3  to the left,  agree=0.8, adj=0.5, (0 split)
  Average.Rent          < 799.275   to the left,  agree=0.8, adj=0.5, (0 split)

Node number 20: 2 observations
mean=3136, MSE=16129

Node number 21: 3 observations
mean=3674, MSE=49650.67

Node number 22: 2 observations
mean=3503.5, MSE=64262.25

Node number 23: 3 observations
mean=4998, MSE=153704.7

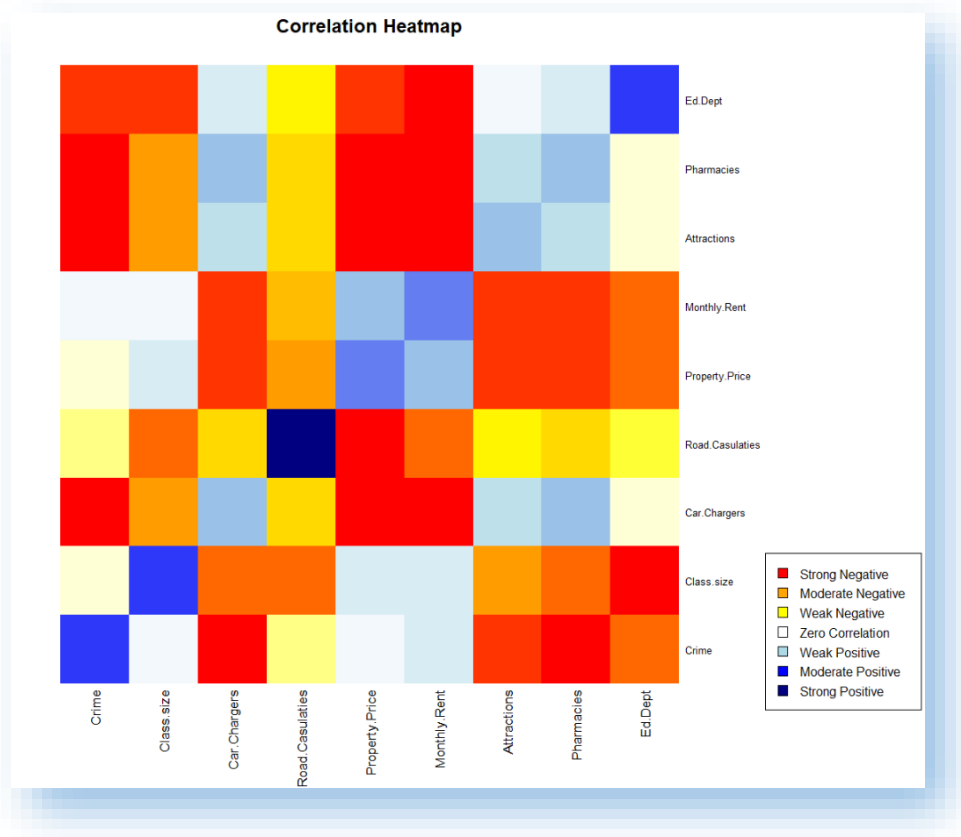
```

Appendix 19

A flatter matrix table that displays both the p-values and correlation coefficients

	row	column	cor	p
1	Crime	Class.size	0.39100175	4.825605e-02
2	Crime	Car.Chargers	-0.67399458	1.600021e-04
3	Class.size	Car.Chargers	-0.26546847	1.899520e-01
4	Crime	Road.Casualties	0.16888554	4.095151e-01
5	Class.size	Road.Casualties	-0.25272680	2.129031e-01
6	Car.Chargers	Road.Casualties	-0.00846823	9.672507e-01
7	Crime	Property.Price	0.41670748	3.419946e-02
8	Class.size	Property.Price	0.60745653	9.976417e-04
9	Car.Chargers	Property.Price	-0.63677193	4.692791e-04
10	Road.Casualties	Property.Price	-0.35439392	7.567442e-02
11	Crime	Monthly.Rent	0.49270883	1.055039e-02
12	Class.size	Monthly.Rent	0.57898905	1.941098e-03
13	Car.Chargers	Monthly.Rent	-0.71440876	4.138368e-05
14	Road.Casualties	Monthly.Rent	-0.22745445	2.637889e-01
15	Property.Price	Monthly.Rent	0.96322342	3.330669e-15
16	Crime	Attractions	-0.62742925	6.016957e-04
17	Class.size	Attractions	-0.19608220	3.370463e-01
18	Car.Chargers	Attractions	0.98223859	0.000000e+00
19	Road.Casualties	Attractions	0.03790666	8.541339e-01
20	Property.Price	Attractions	-0.61843553	7.587295e-04
21	Monthly.Rent	Attractions	-0.68404247	1.165823e-04
22	Crime	Pharmacies	-0.68230488	1.232497e-04
23	Class.size	Pharmacies	-0.26871283	1.843918e-01
24	Car.Chargers	Pharmacies	0.99866905	0.000000e+00
25	Road.Casualties	Pharmacies	-0.01034692	9.599905e-01
26	Property.Price	Pharmacies	-0.63782319	4.561056e-04
27	Monthly.Rent	Pharmacies	-0.71780852	3.655501e-05
28	Attractions	Pharmacies	0.98197812	0.000000e+00
29	Crime	Ed.Dept	-0.52263111	6.159513e-03
30	Class.size	Ed.Dept	-0.50757609	8.122747e-03
31	Car.Chargers	Ed.Dept	0.56266679	2.769090e-03
32	Road.Casualties	Ed.Dept	0.10319976	6.158922e-01
33	Property.Price	Ed.Dept	-0.49365340	1.038003e-02
34	Monthly.Rent	Ed.Dept	-0.57097954	2.316038e-03
35	Attractions	Ed.Dept	0.52782203	5.582887e-03
36	Pharmacies	Ed.Dept	0.56674440	2.538211e-03

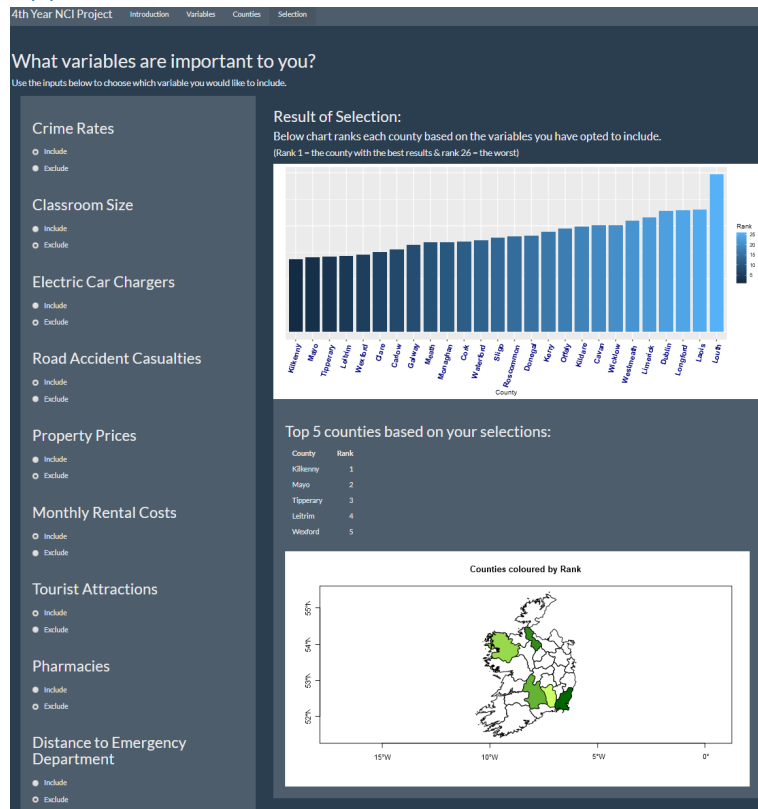
Appendix 20



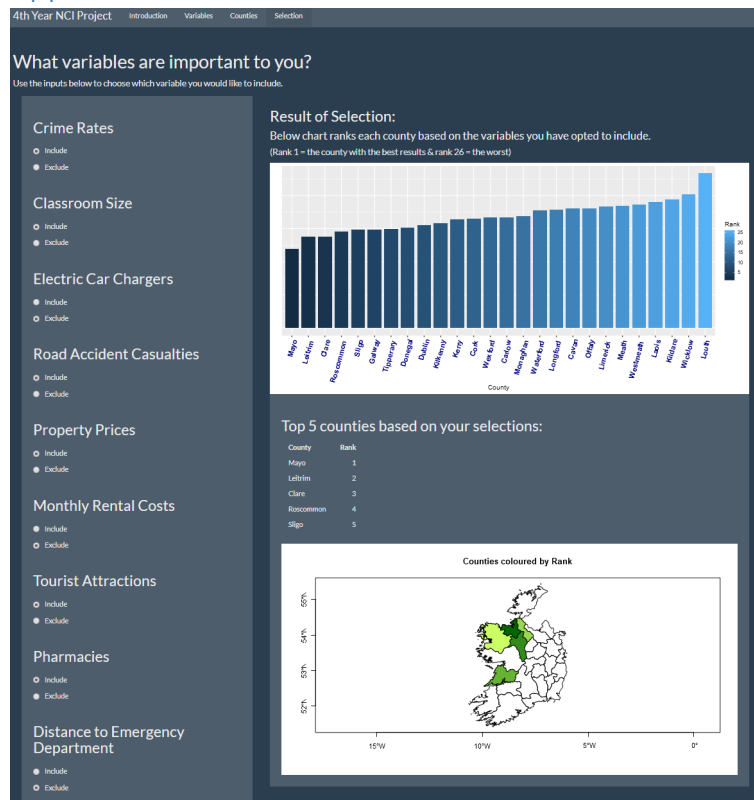
Appendix 21

Rpart Combination	Confusion matrix	Summary
1	pred_test1 Pred:3769 Pred:5130 Actual:2945 0 1 Actual:3000 1 0 Actual:4090 1 0 Actual:4173 0 1 Actual:4381 0 1 Actual:5081 1 0	test_data.Crimes_per_100k Predicted_values variance 1 4173 5130 122.9% 3 3000 3769 125.6% 13 4381 5130 117.1% 18 4090 3769 92.2% 23 5081 3769 74.2% 25 2945 5130 174.2%
2	pred_test2 Pred:3136 Pred:3504 Pred:3674 Pred:6367 Actual:2945 0 0 1 0 Actual:3000 1 0 0 0 Actual:4090 0 1 0 0 Actual:4173 0 0 0 1 Actual:4381 0 0 1 0 Actual:5081 0 0 1 0	test_data.Crimes_per_100k Predicted_values variance 1 4173 6367 152.6% 3 3000 3136 104.5% 13 4381 3674 83.9% 18 4090 3504 85.7% 23 5081 3674 72.3% 25 2945 3674 124.8%
3	pred_test3 Pred:2924 Pred:3290 Pred:3695 Pred:6367 Actual:2945 0 0 1 0 Actual:3000 0 1 0 0 Actual:4090 0 0 1 0 Actual:4173 0 0 0 1 Actual:4381 0 0 1 0 Actual:5081 1 0 0 0	test_data.Crimes_per_100k Predicted_values variance 1 4173 6367 152.6% 3 3000 3290 109.7% 13 4381 3695 84.3% 18 4090 3695 90.3% 23 5081 2924 57.5% 25 2945 3695 125.5%
Random Forest	> predict_rf 1 3 13 18 23 25 4224 3263 3953 3625 3582 3930	test_data.Crimes_per_100k Predicted_values variance 1 4173 6367 152.6% 3 3000 3290 109.7% 13 4381 3695 84.3% 18 4090 3695 90.3% 23 5081 2924 57.5% 25 2945 3695 125.5%

Appendix 22



Appendix 23



Appendix 24

Shiny selection test

Shiny selection

Crime Rates
☒ Include
☐ Exclude

Classroom Size
☒ Include
☐ Exclude

Electric Car Chargers
☐ Include
☐ Exclude

Road Accident Casualties
☒ Include
☐ Exclude

Property Prices
☐ Include
☐ Exclude

Monthly Rental Costs
☒ Include
☐ Exclude

Tourist Attractions
☒ Include
☐ Exclude

Pharmacies
☐ Include
☐ Exclude

Distance to Emergency Department
☐ Include
☐ Exclude

MS Excel selection

Crime ☒ Include

Class.size ☒ Include

Car.Chargers ☐ Include

Road.Casualties ☒ Include

Property.Price ☐ Include

Monthly.Rent ☒ Include

Attractions ☒ Include

Pharmacies ☐ Include

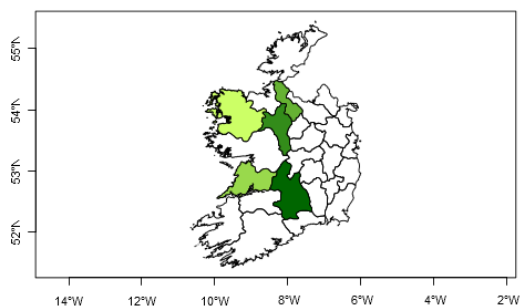
Ed.Dept ☐ Include

Shiny Output

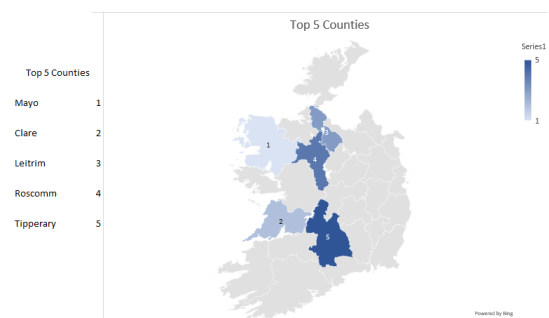
Top 5 counties based on your selections:

County	Rank
Mayo	1
Clare	2
Leitrim	3
Roscommon	4
Tipperary	5

Counties coloured by Rank



MS Excel Output



Project Proposal – updated December 2020

Objectives

For my 4th year project, I will be analysing multiple datasets from multiple sources and focusing on what this data tells me about each county in the Republic of Ireland. The area of focus of the project, centres around the change in corporate attitudes to remote working. Employees no longer need to live in Dublin to be near a Dublin based employer, so I will aim to identify which county will offer the best quality of life based on the analysis of my data sets.

There are several objectives required to complete this project successfully. Each objective will relate to a stage in my plan to take this project from an idea to an output that provides valuable insight.

The first objective will be to acquire the necessary datasets required to complete my project. These datasets will be acquired from publicly available sources.

- Central Statistics office
- The National Treatment Purchase Fund (via Data.Gov.ie)
- Property Price Register
- Residential Tenancy Board
- Failte Ireland

My second objective will be to clean and transform these datasets into a format that will easily allow me to interact with them. From what I have identified so far, the datasets I am interested in are not at the same level, for example, class size is reported on at school level and crime stats are reported on by Garda station or Region level. I will need to be able to analyse these at the same level. I will also need to identify the best database to use as part of my project.

My third, and most important objective will be the analysing of all the data to predict the county with the best standard of living. This objective will be the bulk of the work for my project, and it is the main aim of the project.

The fourth and final objective, will be to provide an insightful report based on my analysis as well as some interactive data visualizations to allow a reader of the report to interact with the data to help gain a better insight into the project.

All the above is planned to be complete one month before the due date in May 2021. This will give me some room for slippage in case I come up against a task that I find difficult. I have tried to include some time for the other modules where I know what is required.

Background

Prior to Covid-19, it was the norm to live in Dublin (or the commuter belt), near your Dublin based employer. E-working or remote working was a factor for some employers, with a 2018 Blueface report finding that 78% of Irish companies already had a remote working policy in place (78% of Irish businesses now have a Remote Working Policy in place Technology, news for Ireland, Ireland, Technology, 2020).

However, to look further into this, I reviewed a research paper published in 2019 by the Department of Business, Enterprise, and Innovation (Department of Business, Enterprise, and Innovation, 2019). This research paper investigated the definition of remote working and found that this was either when employees worked from their homes or where an employee works from a hub close to or within their local community. The report goes on to discuss a pilot survey that the Central Statistics Office undertook in 2018. The results of this pilot survey found that 18% of respondents worked from home, mostly one or two days per week.

A Remote Work in Ireland Employee Survey was undertaken after this report to address the lack of data around employee participation in remote working. However, the sample is skewed by a high response rate from the Finance and ICT sectors. Nevertheless, while the survey is not fully representative and likely overstates the take up of remote work, it does offer useful insights. Some of these insights include:

- Remote working is more common in the private sector (63%) compared to the public sector (28%)
- 48.5% of the respondents said they worked remotely.
- Working remotely on a weekly basis (part of a working week) was most common at 51.1% compared to only 25.1% for private sector & 10.1 % for the public sector who work remotely daily (every day)

The arrival of Covid-19 had a sudden and dramatic impact on remote working in Ireland. From March 16th, 2020, almost 100% of office-based work moved to remote working for both the private and public sector.

Many companies have now started to see benefits in allowing their staff to work remotely on a more permanent basis, these include Indeed (Indeed to allow 'vast majority' of Irish employees to work from home forever, 2020) and Siemens (Kelly, 2020) as an example.

In a post Covid-19 environment, there will be a large increase in the number of employees availing of remote working in Ireland. With this project I hope to identify, using the datasets I have chosen, which county will offer the best quality of life if a move from Dublin is an option for those staff choosing to work remotely.

The reason I choose to do this project, and why it appealed to me was that in the period between April 2020 and August 2020, I know of three colleagues and two-family friends that relocated out of Dublin without the need for changing jobs. At least one of these moves was to get away from the high rents of Dublin, and Covid-19 provided the opportunity to keep their current employment and move to a better location. I found it interesting that the decision to move was made so easily and thought, how many other people are thinking like this, and where will they move to?

Technical Approach

My approach to the project will be to break the project up into 4 distinct parts. These are:

1. Gather datasets.
 - a. Identify the best source.
 - b. Studying the data
2. Clean and transform datasets as required.
 - a. Ensure data is relatable.
 - b. Engineer features
3. Perform analysis.
 - a. Model data using R.
4. Create report and visualisations.
 - a. Data visualisations will be created using Shiny in R.

Research:

I have carried out research into possible topics for my datasets before beginning my project. There are many options for reliable data available to me and the first task will be figure out what will add the most to my analysis. After that I can then research how to obtain the required datasets.

For the next stage, I will need to research database options to store my datasets. Options to consider are MySQL, SQLite, and PostgreSQL. This research is not complete at the time of writing.

I will also need to conduct research on data manipulation, data cleaning, data visualization using Shiny and machine learning.

The aim of all the research will be to take on the project with as much planning as possible been done in advance. This will also help make the project an enjoyable experience.

Requirements:

- Acquire required datasets.
- Transform acquired datasets.
- Analyse data.
- Summarise findings.
- Present findings in a report
- Make dashboard & interactive reports available.

Implementation:

The implementation of my project will be done using R and associated libraries to analyses my datasets. The datasets will be stored in a database which I will interact with using R.

The data visualizations will be created using Shiny in R. As I am only starting to use R in semester 1, I will add more content on the implementation here as I learn more.

In addition to R, I plan to use Python and the Scikit-learn (sklearn) library to run a Random Forest algorithm to create an importance matrix. I will also use sklearn to create a rank / score for each county based on my data.

Special Resources Required

R Cookbook – Paul Teetor:

This book uses practical recipes to perform data analysis with R. I will use this for inspiration and troubleshooting when developing my analysis.

RStudio:

This is the IDE for R. As most of project will involve R, this is my IDE of choice.

Visual Studio Code:

If I opt to use Python for my machine learning, this is the IDE that I prefer to use for Python.

Contains documentation, examples and tutorials on the implementation of scikit-learn.

YouTube:

You can never rule out a YouTube video for help when you're stuck.

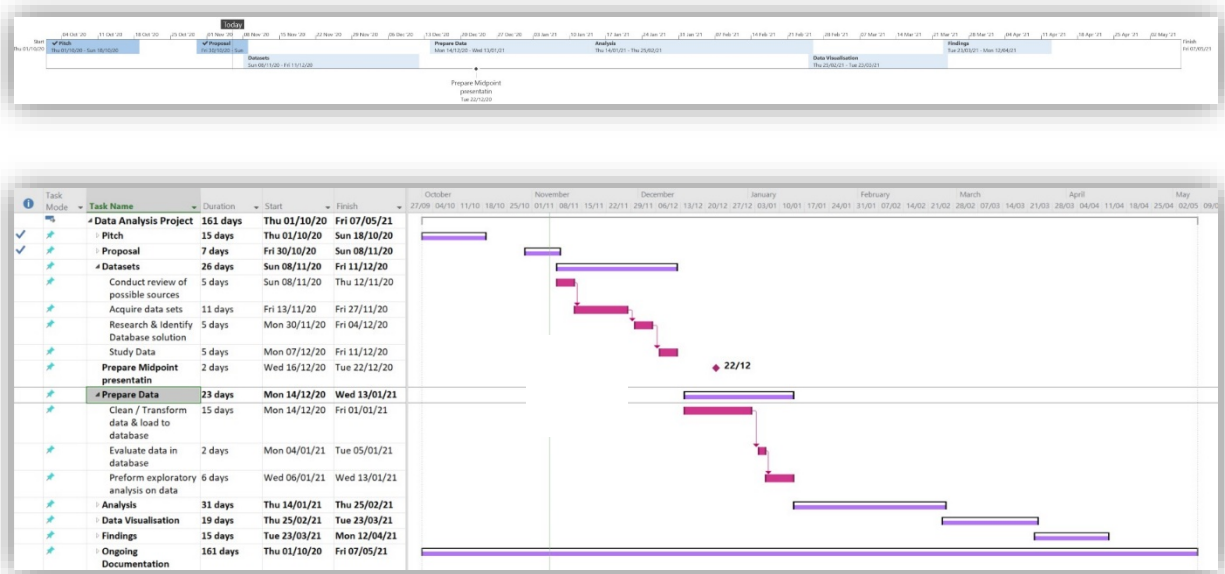
Microsoft Excel:

This will be used in the evaluation of the datasets and to possible to identify issues with csv files if they arise.

Microsoft Project:

To create project plan and track progress through project duration.

Project Plan



Technical Details

I aim to use the R programming language as the primary language for my 4th year project. R is a programming language used in the development of statistical software, data analysis and associated graphics. R also has use of thousands of packages (libraries) available to perform functionality that does not come as standard.

Packages that I have identified for possible use in R as part of this project:

- Dplyr: used for data manipulation.
- Ggplot2: Used in data visualisation.
- Htmltab: assemble data frames from HTML tables.
- Htttr: tools for working with URL's (API's)
- Jsonlite: JSON Parser
- RandomForest: Classification and Regression
- Readxl: Read Excel files
- rJSON: JSON for R
- Shiny: Web Application framework for R
- Tidyr: Tidy messy data
- Dbplyr: used to interact with databases.
- RSQLite: SQLite interface for R

In addition to R, I plan to use Python and the Scikit-learn (sklearn) library to run a Random Forest algorithm to create an importance matrix. I will also use sklearn to create a rank / score for each county based on my data which will be used to determine my results.

Evaluation

Evaluation will be done on each of the datasets as I acquire them. A follow-up evaluation will be carried on the datasets after transformation to make sure I have everything that I expected and that I have not lost, scrambled, or corrupted any data during the transformation process. These evaluations will include setting a specific data quality for the key fields needed for my analysis.

Following the development of the data visualisation, I will acquire some testers to review and test these.

I will also be meeting with my supervisor on a regular basis and showing the progress my project is making. They will help me make sure the project is traveling in the right direction, including the review of any of my evaluation work.