National College of Ireland

Technology Management
Data Analytics
2020/21
Ryan Johnston
X17437624
X17437624@student.ncirl.ie

Analysis of Crime in North Dublin
Technical Report

Contents

Executive Summary

Crime is an ongoing issue globally. This analysis aims to assist law enforcement or any other relevant parties, in the analysis of crime in Dublin, focusing on 40 key areas with relevant study to background motives of crime. North Dublin is be analysed as a key area, compared directly with South Dublin uncovering trends and incident prevention. Inner city areas have proven to have more incidents. The top 5 areas of crime are theft, public order, offences against government, damage to property and environment & drug controlled offences. This report focuses on external factors of crime including *income and poverty,* and *educational attainment.* These are two key areas that may influence crime.

The North of Dublin has a mean value of 198 whereas the south has a mean value of 214 for the total value of crimes all time. The top 5 stations in North Dublin are Store Street, Bridewell, Blanchardstown, Finglas and Coolock. The top 5 stations in South Dublin are Pearse Street, Tallaght, Kevin Street, Dundrum & Dun Laoighre. These are listed in order from highest to lowest for the total count of crimes all time.

1.0 Introduction

1.1. Background

It was chosen to undertake this exact analysis as crime is evident on a daily basis living on the North-side of Dublin. Digging deep into the analytics of the types of crime, as-well as the locations will give an insight as to where the largest volume of crime is occurring and the most common type. A question that can be asked to determine the accuracy representation of locations is: Is there only one Garda Station for multiple locations? For example Howth Garda Station is the only one for quite a distance, will this influence more crime? If people know that there is no Garda Station close to the area of an incident this may influence crime as the response time for the Garda might be longer than the response time for incidents with a Garda Station in closer proximity. Another important note is that if some areas have their own Garda stations, while others have one station for a larger district, is the crime higher in these areas or lower? These are potential findings of this analysis. A further in depth background analysis can be found in the *Section 9. Appendices* of this document.

Note: The crime dataset used for this analysis is from 2003-2019 which covers a 17 year period. When the period 'all time' is mentioned it reflects the years of the data, not all time crime dated outside of this period.

1.2. Aims

During this analysis a comparison of the North & South of Dublin, both with 20 locations each will be analysed to look at types of crime over a period of time per location. These 40 locations are 20 Garda stations for North Dublin and 20 for South Dublin.

- Assess background motives of crime that may have relevance to the study. For example, what factors may cause a higher volume of crime e.g. poverty, educational attainment etc. Trying to find a link between those to crime using data sets.
- To calculate the top 5 areas of crime in North and South Dublin
- Calculate the mean of north vs south to determine which area crime is occurring more and interpret the results.
- Calculate the top 5 types of crime of all time based on the overall value of each.
- Create visualisations that simplify the values for the viewer allowing for more efficient understanding and interpretation.
- Create a map visualisation showing locations of each garda station with interactivity.
- Construct a Shiny Interactive dashboard containing a map and data table that updates values based on the user selection of a slider bar.
- Carrying out statistical data analysis tests that apply to the data and topic.

1.3. Technology

R will be the main programming language for data manipulation due to its powerful visualisations and functionality. Python was considered however R was superior for the types of visualisation needed for this project. Libraries were used in R to help with reading, manipulating, styling and outputting the data.

These libraries include but are not limited to:

- *Leaflet* and *Shiny* for creating interactive maps and a dashboard,
- *TidyVerse* to help with data manipulation and querying,
- *Ggplot2* for powerful visualisations such as bar plots, line plots and point plots. Also *GGThemes* for styling these plots.
- *Flextable* for styling data table outputs.

R is just more suited for the purpose of this analysis whereas Python has a wider range use for software development and automation which is not required for this analysis.

RStudio is the choice of IDE as it is constructed specifically for the functionality of the R programming language. All processes within the KDD have been performed in R Studio.

Excel was used for the reformatting of many datasets to help with creating visualisations and performing tests. It was also used alongside Leaflet in R Studio to add the longitude & latitude coordinates to a map visualisation containing pin-points. These points were the locations of each Garda Station. This process will be further discussed in the *Data* section of this report.

Tableau was used for some visualisation of the data to use a wider range of tools rather than just sticking to R.

GitHub was used for version control in case of any fatal errors. A repository was created and **Git** was used to push the code onto this repository. Commits were made regularly as changes were made to the project and the repository was updated frequently.

SPSS was be used for the statistical side of the project, for performing statistical tests.

1.4. Structure

The following sections of this report have been summarised into the below headings. These can be used to help navigate to the preferred section of information based on the content in each heading summarised below.

2.0. Data

- Exploring for datasets within certain topics.
- Datasets that were chosen.
- How the data was retrieved.
- Where it was retrieved from.
- How it has been used.

3.0 Methodology

- Methodology which was used for the analysis.
- Why it was used.
- Steps within the methodology and the processes within each.

4.0 Analysis

- Software used.
- How the objectives were met.
- Statistical tests used.

5.0 Results

- Output of objectives that were mentioned at the start, supported with the use of visualisations.

6.0 Conclusions

- Conclusion as to whether the selected datasets were efficient.
- Completeness of the analysis/project.

7.0 Further Development

- What would be done differently with more time and experience.

8.0 References

- List of references that were used within the project report.

9.0 Appendices

- Project proposal.
- Project plan.
- Monthly journals.
- Other materials used.

2.0 Data

2.1 Obtaining the Data

After investigating the possible motives of crime, it was apparent that income and education could have a direct relationship on the volume of crime values. It was decided that having a dataset containing types of crime, areas and years could work well with an income dataset containing years and types of income and an education dataset that had the educational attainment for age groups also containing years for comparison.

The primary dataset used was sourced on the Central Statistics Office of Ireland and includes different types of crime offences. Some of the columns of this dataset include the type of crime, the total number of that crime in a given year and the Garda station the crime was reported to. This dataset includes the year 2003 up to 2019. This was the period which was decided to try and match the other datasets too. The crime dataset contains crime values for the whole Republic of Ireland. This dataset was downloaded directly from the CSO as a comma separated values file. *(Sam Scriven, 2020)*

Two secondary datasets were also been downloaded from the CSO.

First was the Income and Poverty Rates dataset which includes household income and poverty rates with variables such as the mean disposable income of a household and three different age ranges that contribute to this dataset. This income dataset period is from 2004-2019, close to that of the primary dataset. *(Tricia Brew, 2020)*

Educational attainment was another dataset that was used and contains 8 age ranges with a column showing the percentage of attainment for each type of education. The education dataset contains the year 2009-2020. This has caused some limitation as to what can be concluded due to the year range not being as wide as the other two datasets. However it was utilised, as it still contains a wide range of years that can give an insight. These two datasets were used alongside the primary dataset to analyse whether external factors have an impact on the frequency of crime. *(Sarah Crilly, 2020)*

A supplementary dataset was also constructed to aid the crimes dataset.

A location dataset was created using excel to support the use of a map Visualisation in R. Google Maps was used to get the latitude and longitude locations of each Garda Station in North and South Dublin & they were put into excel and saved as a comma separated values file. This was then loaded into R as a data frame. Leaflet contains native American state borders as locations when used alongside tMap which is another library in R for constructing maps. This is non-existent for the locations which were needed for this project so the only option was to plot each location manually. The exact latitude and longitude locations of the corresponding Garda Station for each location was and gives insight to the density of Garda stations in a proximity.

2.2 Exploratory Data Analysis

Before choosing the primary dataset, it was observed on the source website and potential findings were questioned. What can this dataset prove? if it can't prove anything by itself, how can it be supported to help prove a point. The primary dataset can reveal valid answers itself however it will reveal more insights when supported by other datasets, which is why the others were chosen. If the other datasets have no impact on crime after the analysis is concluded, that is still an important piece of information to take from this project.

The data was read into R Studio and each column is correct in the type of data, e.g. 'char' for words, 'int' for numbers etc. Upon analysing the crime data after reading it in, no null values were present. In the Income dataset, some NA's were present but only for income of people aged 0-17 which seemed accurate to progress with as it's such a wide range of ages and only people from age 16 would typically work meaning there is a redundancy of ages that can have an effect on the accuracy. Each column of the dataset was understood and self-explanatory. The only thing that needed to be done was to filter it to the locations needed for the analysis.

All data used is labelled 'under reservation' on the Central Statistics Office website indicating that it is not final. However these datasets were the only ones that could be of any use for the topic chosen. It is a valid assumption to assume the data is correct based off my knowledge of the areas in which each Garda station is located in, and weighing up whether or not it is deemed dangerous, the data seemed to line up to what I expected in this case. The areas that would be slated in the media through violent crimes had a higher rate of crime which was expected. The more well off areas had less crime activity which concluded that the data was accurate to work with.

Exploratory data analysis findings are included in the results section as visualisations. These visualisations will answer questions that had been thought of prior to analysing the data. The exploratory data analysis process was ongoing for the duration of the project as new findings were discovered as time went on.

Exploratory data analysis was conducted in R as follows;

1. **Crime Dataset**

Fig.1. *Glimpse* function - crime

```
Rows: 115,056
Columns: 6
$ Statistic       <chr> "Recorded Crime Offences Under Reservation", "Recorded Crime Offences Under Re…
$ Year            <int> 2003, 2003, 2003, 2003, 2003, 2003, 2003, 2003, 2003, 2003, 2003, 2003, 2003, …
$ Garda_Station   <chr> "35301 Abbeyfeale, Limerick Division", "35301 Abbeyfeale, Limerick Division", …
$ Type_Of_Offence <chr> "Attempts/threats to murder, assaults, harassments and related offences", "Dan…
$ Unit            <chr> "Number", "Number", "Number", "Number", "Number", "Number", "Number", "Number"…
$ Value           <int> 18, 14, 0, 0, 27, 20, 0, 2, 2, 13, 17, 5, 14, 23, 0, 0, 29, 41, 1, 11, 2, 18, …
```

The *glimpse* function in R quickly gives insight in to the type of variables and the data within the variables. The characters and integer types are correctly formatted. This also shows the total number of rows and columns in the dataset at the start.

Fig.2. *Status* function - crime

	variable	q_zeros	p_zeros	q_na	p_na	q_inf	p_inf	type	unique
Statistic	Statistic	0	0.0000000	0	0	0	0	character	1
Year	Year	0	0.0000000	0	0	0	0	integer	17
Garda_Station	Garda_Station	0	0.0000000	0	0	0	0	character	564
Type_Of_Offence	Type_Of_Offence	0	0.0000000	0	0	0	0	character	12
Unit	Unit	0	0.0000000	0	0	0	0	character	1
Value	Value	28720	0.2496176	0	0	0	0	integer	1178

This was used to check for the percentage of NA's. In the *q_na* column it is clear that there are no NA's present in this data. Unique values can also be observed in the *unique* column.

2. Education Dataset

Fig.3. *Glimpse* function - education

```
Rows: 2,376
Columns: 7
$ Statistic       <chr> "Persons Aged 15-64", "Persons Aged 15-64", "Persons Aged 15-64", "Persons Age…
$ Quarter         <chr> "2009Q2", "2009Q2", "2009Q2", "2009Q2", "2009Q2", "2009Q2", "2009Q2", "2009Q2"…
$ Age_Group       <chr> "15 - 19 years", "15 - 19 years", "15 - 19 years", "15 - 19 years", "15 - 19 y…
$ Sex             <chr> "Both sexes", "Both sexes", "Both sexes", "Both sexes", "Both sexes", "Both se…
$ Education_Level <chr> "Primary", "Lower secondary", "Upper secondary", "Post leaving cert", "Third l…
$ Unit            <chr> "%", "%", "%", "%", "%", "%", "%", "%", "%", "%", "%", "%", "%", "%", "%", "%"…
$ Value           <int> 2, 8, 52, 11, 27, 3, 7, 16, 2, 19, 51, 28, 2, 0, 0, 0, 0, 0, 18, 51, 28, 2, 1,…
```

All fields are correct here except 'Quarter'. This should be an int. The reason it is not an int is because of the 'Q2' at the end of each value. This will need to be changed by removing the 'Q2' and parsing the type to 'int'. The rows and columns can also be observed in the top left.

Fig.4. *Status* function - education

	variable	q_zeros	p_zeros	q_na	p_na	q_inf	p_inf	type	unique
Statistic	Statistic	0	0.00000000	0	0	0	0	character	1
Year	Year	0	0.00000000	0	0	0	0	character	12
Age_Group	Age_Group	0	0.00000000	0	0	0	0	character	8
Sex	Sex	0	0.00000000	0	0	0	0	character	3
Education_Level	Education_Level	0	0.00000000	0	0	0	0	character	9
Unit	Unit	0	0.00000000	0	0	0	0	character	1
Value	Value	171	0.06597222	0	0	0	0	integer	64

Similar to the crime dataset, there are no NA's present in this dataset. There are far less unique values here which is expected because the dataset is much smaller than the crime one.

3. Income Dataset

Fig.5. *Glimpse* function - income

```
Rows: 528
Columns: 5
$ Statistic <chr> "Median Real Household Disposable Income", "Median Real Household Disposable Income…
$ Year      <int> 2004, 2004, 2004, 2005, 2005, 2005, 2006, 2006, 2006, 2007, 2007, 2007, 2008, 2008,…
$ Age_Range <chr> "0 - 17 years", "18 - 64 years", "65 years and over", "0 - 17 years", "18 - 64 year…
$ Unit      <chr> "Euro", "Euro", "Euro", "Euro", "Euro", "Euro", "Euro", "Euro", "Euro", "Euro", "Eu…
$ Value     <dbl> NA, 43948, 18461, NA, 44557, 19759, NA, 44977, 21371, NA, 46901, 22865, NA, 47645, …
```

A new observation to make is that the value column in this case is a double. This is due to the fact that the value is money and money is measured as a double because it contains a decimal point typically. This is fine and will work similarly to an int so it is the correct type. All of the rest are correct also. The rows and columns can be seen up the top left.

Fig.6. *Status* function

	variable	q_zeros	p_zeros	q_na	p_na	q_inf	p_inf	type	unique
Statistic	Statistic	0	0	0	0.0000000	0	0	character	11
Year	Year	0	0	0	0.0000000	0	0	integer	16
Age_Range	Age_Range	0	0	0	0.0000000	0	0	character	3
Unit	Unit	0	0	0	0.0000000	0	0	character	2
Value	Value	0	0	64	0.1212121	0	0	numeric	409

An important observation to make here is that there are 64 NA's in this case. This is due to the fact that people aged 0-17 were not recorded for household income. These will be left as NA as it makes no sense to assign them 0 because we can't assume they made no money, having the value as Not Available is the best applicable solution in this case as it won't skew any results.

These were basic functions in R to give quick insight for exploratory data analysis. It can be concluded that each of the datasets are ready and there is no major problems.

3.0 Methodology

Knowledge Discovery in Databases methodology is what was chosen to follow for this analysis. This was chosen as it best fits this analysis. The CRISP-DM is more focused on the business side of things and driving business rules into the implementation which was not a requirement for this analysis.

Selection – The data was selected after brainstorming some ideas which were of interest. Using the Irish Government Database website proved to be useless for this topic and so the Central Statistics Office website was more suited. Looking through the datasets it was apparent that some of them were similar with a different additional variable, for example the dataset that was chosen contains Garda Stations for each year and what was reported to each one, others had similar data but without the information of where it happened or was reported to. Having the Garda Station location in the dataset was a great way to conduct a more specific analysis, focusing on areas, which further allowed me to then introduce other datasets. A background to crime was explored highlighting possible motives causing people to commit crime, and datasets were picked based on knowledge gained from this exploration.

Also within the selection, the full primary dataset was not required as this analysis is focusing on the Dublin region, so the Dublin Garda stations were selected, 20 from the North & 20 from the south to use as a comparison. The original dataset had the listings of Garda Stations all over Ireland so this greatly lessened the new number of observations.

Pre-Processing – The very first thing was to check for nulls which was done in the exploratory data analysis. The crime and education datasets contained no NA's. The *income* dataset contained 'N/A's' and it was decided to leave the values as N/A, as no assumptions could be made for the reasoning behind why they were not available. Assigning N/A's with values such as the mean of the column would have skewed the accuracy and gave different results so it was decided to leave them completely unvalued. Columns were renamed and removed as necessary to include only what was relevant for the analysis. There were no major outliers in these datasets and outliers would have been welcomed to the study as it focuses on crime statistics which can be very high or very low depending on the area. Extra data frames were made to help analyse certain areas and produce visualisations.

Other typical tasks that occurred in pre-processing were; sub-setting and filtering, math calculation such as computing the mean of a range of data, parsing character type fields to numeric, removing stop words (Q2 at the end of Quarter column in the income dataset) and aggregation of columns for visualisation purposes. These were all done to help construct new columns, rows or datasets that would be of benefit to the analysis for interpretation through visualisations.

Transformation – The location dataset was created using excel and contained the latitude and longitude locations of each Garda station. This was joined to the all-time crime values dataset which was used to create a map visualisation showing the total crimes between a selected period. The location dataset was joined through a left-join to two data frames; South location and North location. These datasets were constructed from the main crime dataset except they only had the Garda station location, year and value of crimes. This was done for the creation of a shiny dashboard which contains contain the years, locations and values. The original *crime* dataset was reduced from 115,056 observations to a total of 8160, 4080 for North Dublin & 4080 for South Dublin. This was done as the 8160 were the only rows that corresponded to the purpose of the analysis.

Data Mining

K-Means clustering was done on all datasets to group the data by similar mean values. The similar mean values is how the variables were grouped together into a cluster. A *K*-value was selected which outputted the total number of groups to be displayed on the cluster visualisation. This differed for each dataset. Clustering was done on all datasets to see the different groups within the data easily. Factoextra and Mclust were two libraries installed to allow for cluster visualisation. *(Agarwal, Nagpal, Sehgal, 2013).*

- **Crime value clustering** was done separately for the North & South of Dublin to group the data together to show areas of high, medium and low crime values. The variables in this case were the Garda Stations.
- **Income clustering** was done to group the average household disposable income by year. The years were the values in this case.
- **Education clustering** was done to group the different types of education by the most attained over the range of years 2009-2019. Each variable in this case were the types of education.

Linear modelling was done to try relationships in the crime dataset in regards to income. Income was the dependent variable and each type of crime were independent variables. The crime types were renamed for the process of linear modelling to simplify the task by including less character dense strings. The dataset for linear modelling was constructed in excel by joining the income dataset to the crime dataset and writing it to a comma separated values file. This needed to be done as one variable 'Type_Of_Offence' which contained all 12 types of crime needed to be abolished and each crime type needed to be its own variable. This was too complex to be done in R Studio and would have taken more time than it needed to as it wasn't a huge step for the success of the project.

Two models were made, a full model and a half model. The full model contained all the types of crime and the half model only contained those that showed as "significant" with the asterisk beside them in the model output. These models were done to assess the effect of each type of crime on income to see if there was any relevance between them. This was one of many ways which correlation & regression were checked between the datasets.

Evaluation – Upon completing the first 4 processes of the Knowledge Discovery in Databases model, somewhat of an evaluation was made. R and Tableau were then used to create visualisations from the data which can evaluate questions within the initial aims of the project. Visualisations from the data frames constructed in R can be used to interpret information about the data at ease and efficiently. Some visualisations contain multiple datasets however most are only constructed using the data within its own dataset. The graphs that were produced from one dataset such as crime can be compared to a graph of education to look at the yearly statistics to determine whether high educational attainment meant lower crime. This is a quick way to evaluate relationships and patterns in the data. The main evaluation of this analysis will be made from the interpretation of visualisation output. Visualisations are presented within the *Results* section of this report.

4.0 Analysis

Data analysis was an ongoing process that was used for the whole duration of the project. Creating new data frames and joining current ones were made possible through data analysis, and interpreting the results of the data.

4.1 Visualisation

Visualisation was the main analysis method used for this project. Visualisation enables efficient interpretation of data like no other method. It is aesthetically pleasing and easy to interpret by the general public which is why it was used. The crime dataset only contained one column which had values in it. The rest were just columns that aided the values in terms of categorisation being the type of crime and periods being the year it occurred. Although there was only one column, there was a lot of information that could be retrieved from it. The data frame was exported and reformatted so that each crime had its own column. This allowed for a deeper analysis into the top types of crime. A *leaflet* map visualisation was created which displayed crime values for the 40 locations in Dublin which were colour coded through pin-point markers plotted on the map. Leaflet uses the Google Maps API to construct the map that's displayed. The map contains hover labels to show the location name and an on click popup label to show the total number of reported crimes for the selected location. This is a powerful visualisation to represent the volume of crimes per Garda Station in North and South Dublin. *(Farmer and Wasser, 2020)*

A Shiny Dashboard was also implemented which contains a slider bar that enables user interaction. The dashboard was constructed similarly to the map with the hover and pop-up functionality however it works differently. The data within the pop up changes based on the users selection of the year on the slider bar. Below the map is a data table which will be filtered simultaneously with the map through the slider input but also allows for querying through a search bar to retrieve data on a specific search. This is a powerful visualisation which offers the greatest insight of the project, it is an innovative aspect that can be used for real world examples by data scientists within related fields such as criminology.

Tableau was used to create some visualisations to add some variety to the look of the visualisations for the analysis. Tableau allowed for easier visualisations to be created. Tableau was initially going to be used for all visualisations but it was more efficient to use R for them as the data was constantly being manipulated in R and would require writing to csv files every time to use the data in Tableau.

4.2 Statistical Tests

Statistical tests were only run on the crime dataset. This is because the crime dataset had the most valuable information for the analysis.

A test for normalisation was the first thing that needed to be done. This was done to give insight to the type of data being used and the tests that could be run on it. Testing for normality was done in R using the *ggqqplot* function and using the *Shapiro-Wilk* test. Firstly, Q-Q plots were constructed using a sample set of columns selected at random. The data within these columns is the number of crimes so it wouldn't make sense to test them all due to the similarity. This is why a random sample was selected. The Q-Q plots were constructed which showed the plots of data against the normality line. It was evident that from observation of the Q-Q plots, the data was not normally distributed.

Accuracy of these plots was questioned so it was then decided to do Shapiro Wilk tests. The Shapiro-Wilk test defines the normality based on the P-Value statistic. If P >= 0.05 the data is normally distributed. *(Technik, 2019)*

After finding the data was not normally distributed. Non parametric tests were explored and a Kruskal-Wallis H test was conducted in SPSS. The Kruskal-Wallis H is a one way ANOVA test which is rank based. This test is used to analyse significant differences between variables. This test had one grouping variable which was *year* and 12 independent variables which were types of crime labelled *Crime.1-Crime.12.* The Kruskal-Wallis H test outputs a *H* value which is the mean rank of the data from the whole observed row. An *Asymp. Sig.* value is also a result outputted from this test and represents the significance between each value in the variable. This can be used to show how significant the change of that particular crime was by year. If the *Asymp. Sig.* value > 0.05, it can be interpreted that there was a significant difference for the values in the variables based on the different years. This can be used to show if a crime is a reoccurring issue or if it isn't frequent. *(Kang and Kang, 2017)*

Spearman's correlation was also done in SPSS to check for correlation between *income* and types of crime that may be influenced by an increase or decrease the income value. The income value in this test is *mean household disposable income* which is typically generated by multiple individuals in a household. This test was done to check the relationship between the selected types of crime and income to give insight as to whether these types of crime activity are triggered by income related statistics.

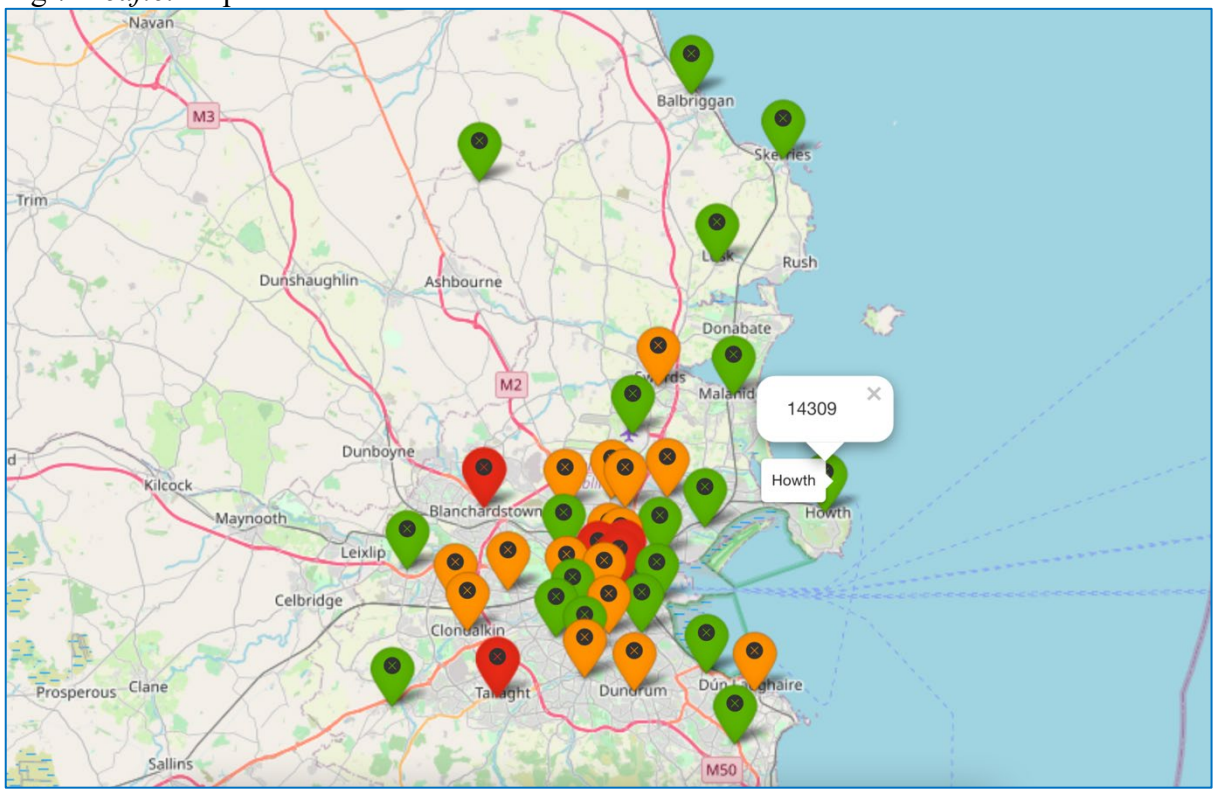
A small amount of statistical tests were performed due to a limitation with the data. The crime dataset didn't contain much data that could be tested. However, this data set was the best crime dataset for Ireland that could be found and it was important for the topic to keep the analysis on Irish crime data. A workaround for a crime data analysis would be to use American data as it is constructed differently and contains a lot more information than what was included in this dataset. The limitation of this dataset was observed at the statistic testing stage.

5.0 Results

This section includes the visualisation output and explanation for each to give greater understanding to the findings of this analysis. This section will be broken down into 3 **Visualisation, Data Mining** and **Testing.**

5.1 Visualisation

Fig.7. *Leaflet* map visualisation – total crime values



The above map was constructed using R to show the total number of crimes committed in each area between the year 2003-2019 which was the years focused on in this analysis. The markers on the map were created using the latitude & longitude locations of each garda station in the study. The colour of the markers symbolises the "danger' of the areas. These colours were programmed to output as following.

Green – Number of crimes <= 5 per day
Orange – Number of crimes >5 <=10
Red - Number of crimes > 10

These numbers were decided based off the total numbers of crimes for areas which could contain more crime such as Garda Stations in the City Centre. These Garda Stations would deal with a lot of theft related crimes as the city centre is a shopping hub.

It's apparent that Blanchardstown and Tallaght have high values, this may be due to the fact that they are responsible for crimes within a wider range of areas than some of the other stations. However this assumption is not entirely accurate as it is evident that Howth is considered to be "safe" based on the colour of the pin-point map marker. This may be to do with social class though, for example the housing in Howth is more expensive which may lead to a different more upper class population of residence. Howth is also a terrain dense area as it is constructed on a hill which is another factor that may cause the crimes to be low due to the fact that there is a limited amount of houses. These are factors that need to be taken into consideration when observing a map like this. The colours were just used to highlight the total number, not to suggest areas of complete safety.

Fig.8. *Datatable* containing Location, Latitude, Longitude and Total Values

Show 25 entries Search:

	Location	Latitude	Longitude	Total
1	Bridewell	53.35005	-6.275874	103752
2	Fitzgibbon	53.357847	-6.255418	39786
3	Mountjoy	53.360484	-6.266825	43011
4	Store	53.350401	-6.252243	157464
5	Balbriggan	53.614394	-6.190959	20962
6	Garristown	53.567034	-6.383895	1167
7	Lusk	53.523408	-6.167363	10562
8	Skerries	53.579616	-6.106996	5777
9	Ballymun	53.394395	-6.263822	33956
10	Dublin Airport	53.429677	-6.244087	8137
11	Santry	53.389765	-6.251572	39703
12	Coolock	53.395849	-6.213343	50225
13	Malahide	53.45112	-6.151918	18856
14	Swords	53.4562	-6.221127	44410
15	Clontarf	53.363376	-6.220033	30160

Showing 1 to 25 of 40 entries Previous 1 2 Next

The data table above represents the structure of the dataset used to create the *Leaflet* map visualisation above. The coordinates for each location represent the exact location of the Garda Station. The total value is the total number of crimes which is displayed as a pop-up on the map when a location is clicked and the location value represents the location of each marker on the map in which the name is displayed upon hovering over the marker.

Fig.9. *Barplot* - North Dublin total crime values

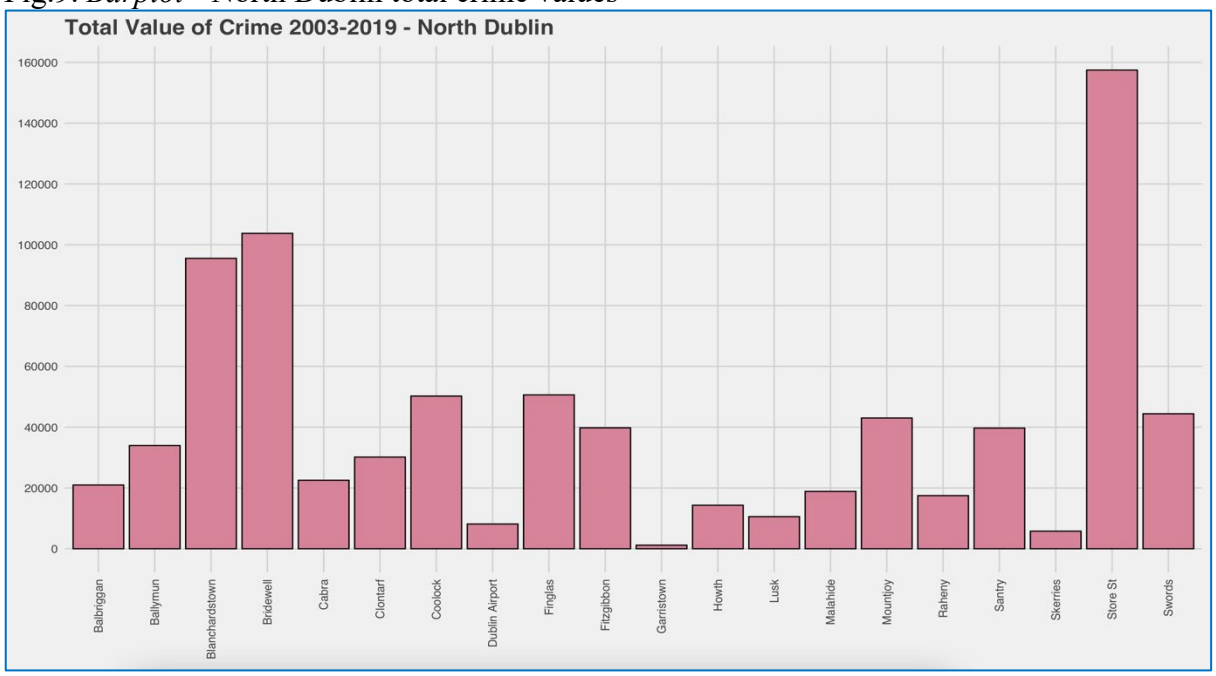
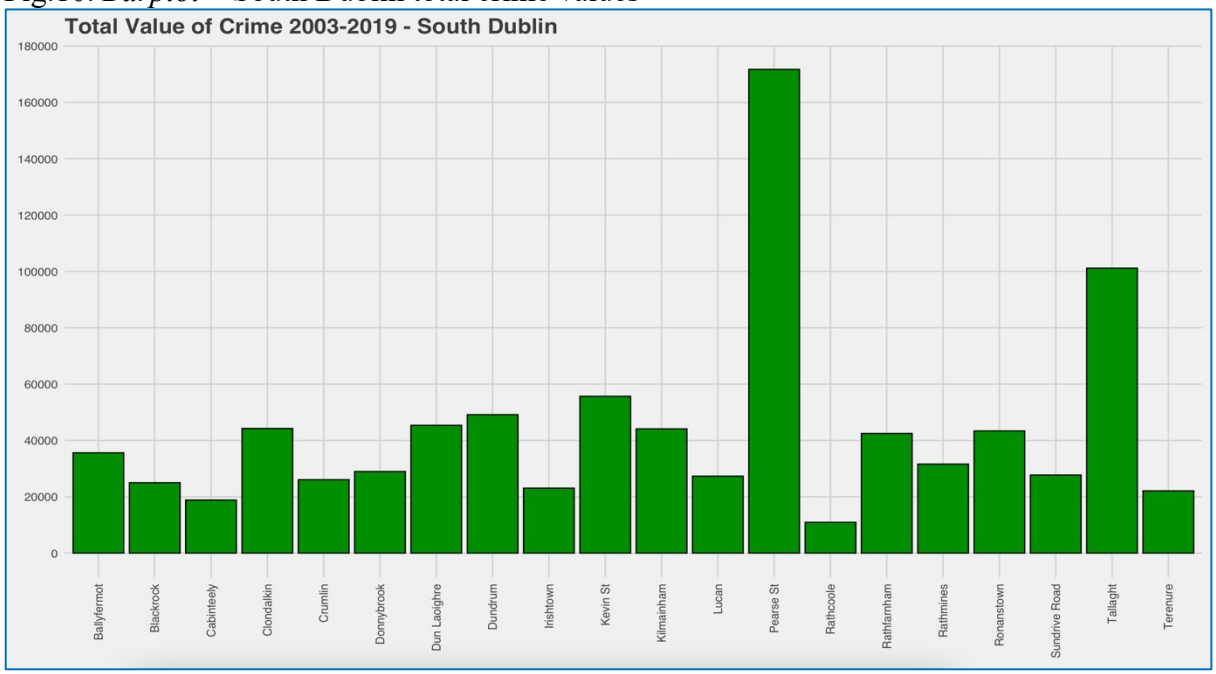
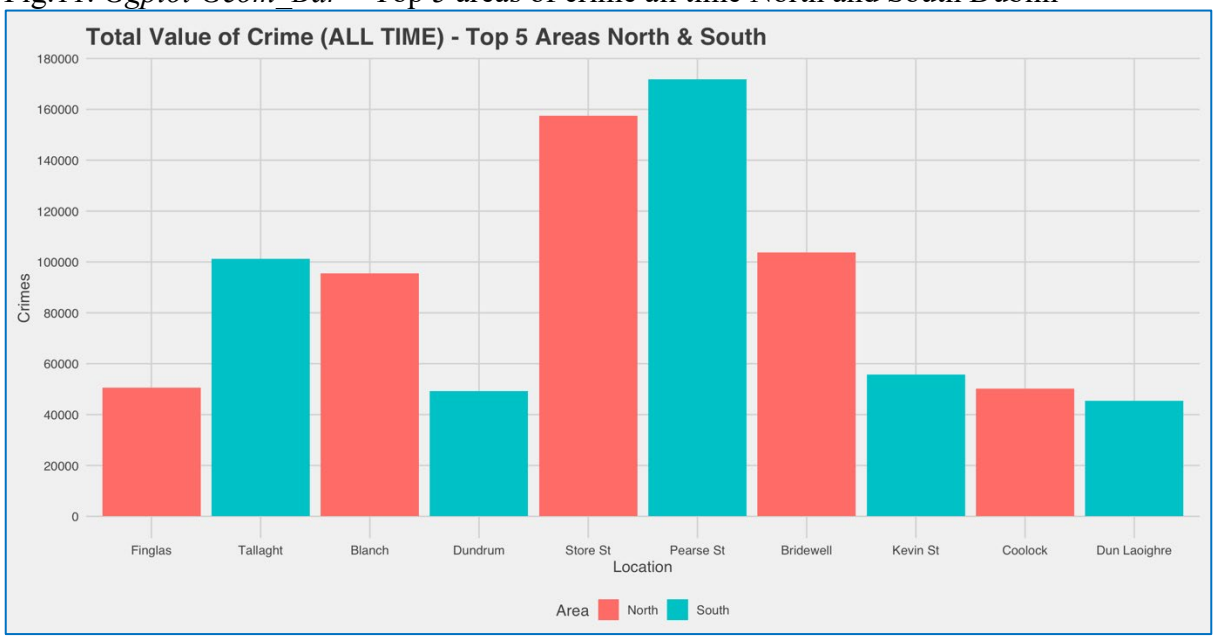


Fig.10. *Barplot* – South Dublin total crime values



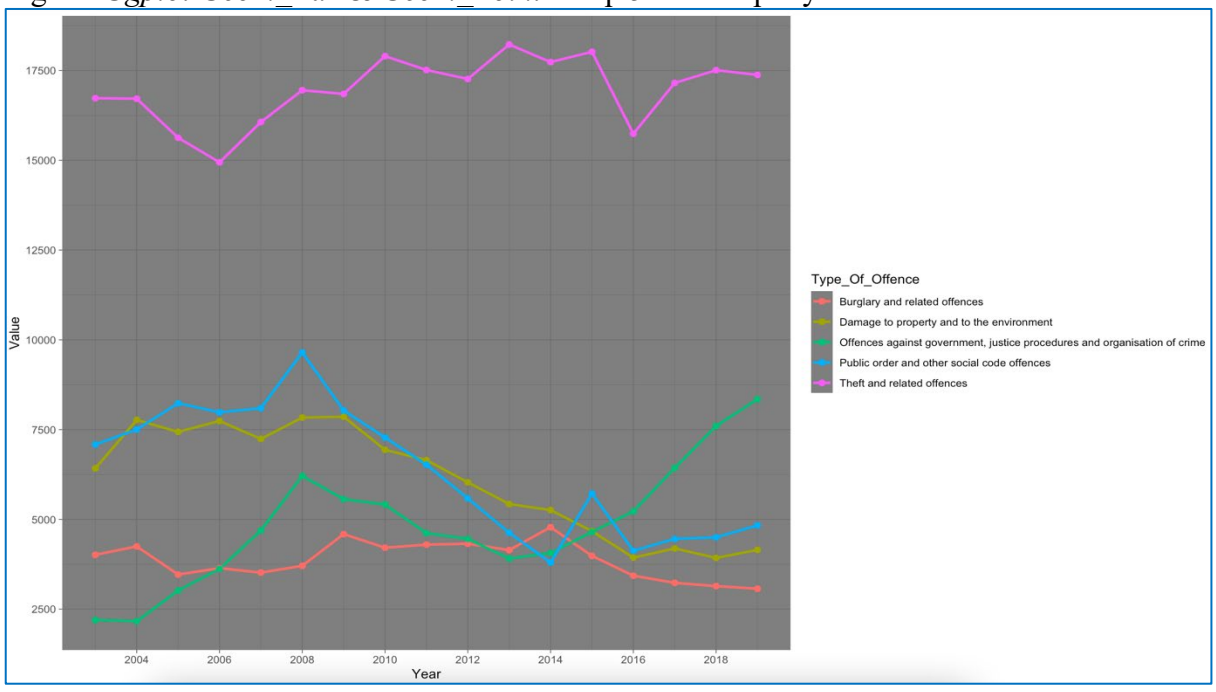
The two graphs above were constructed using the *barplot* function in R. These are made using the same data that constructed the total crime values map explained previously. The location with the least amount of crimes is Garristown which is located in North Dublin and has 1167 total crimes. The value that is the closest match to this in South Dublin is Rathcoole which has a total of 10946 crimes. This is a major difference in the lowest value. The two highest values are Store Street in North Dublin and Pearse Street in South Dublin which are both responsible for shopping districts such as Henry Street and Grafton Street. It was expected that inner city Garda Stations would have the highest crime values. There is a lot more variance in the North Dublin chart than South Dublin due to areas like Garristown, Skerries and Lusk which are all located in the same direction close to the border of County Dublin.

Fig.11. *Ggplot Geom_Bar* – Top 5 areas of crime all time North and South Dublin



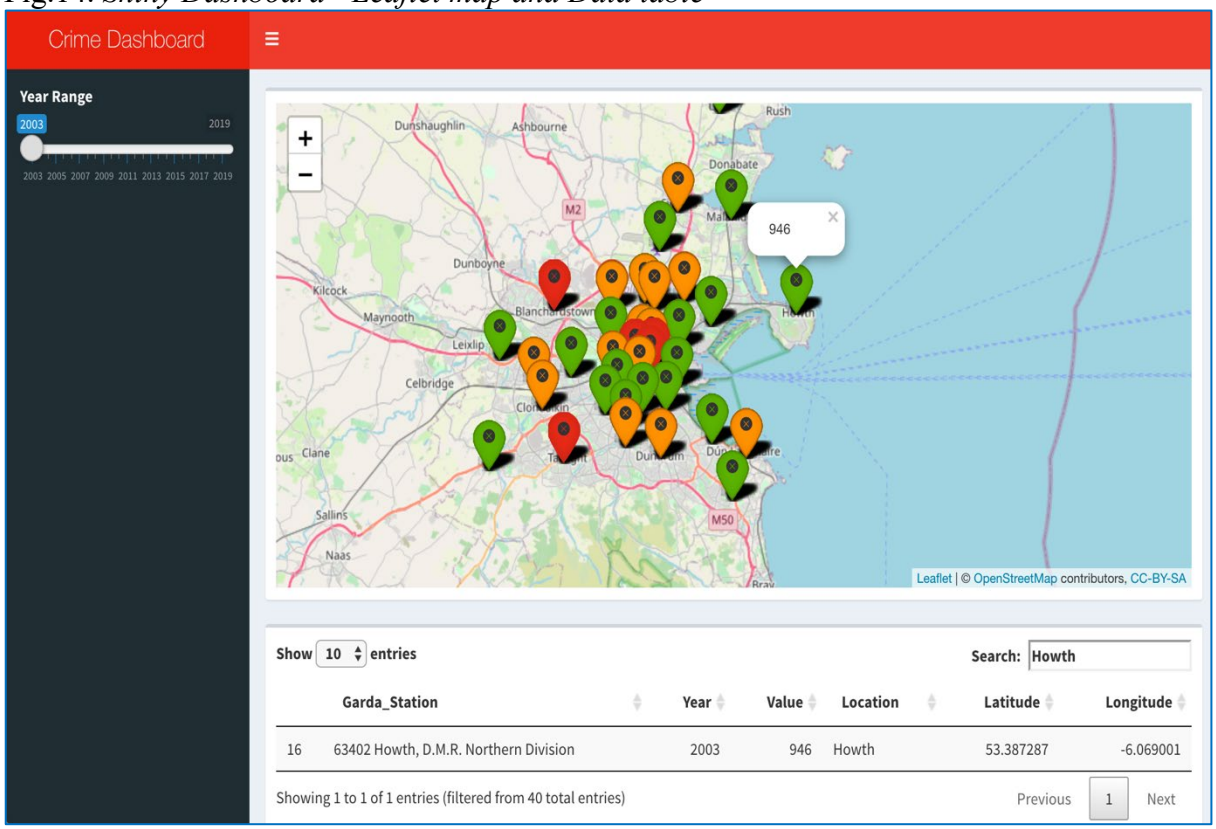
The graph above was created to show the top 5 locations for each area for comparison between the crime values. The North areas definitely outweigh those in the South for the total number of crimes between the top 5 areas. A total of 4 areas out of the 10 are located within the city centre. These areas are: Store Street, Pearse Street, Bridewell and Kevin Street. A high volume of crime is expected here as the population density in the city centre is much more than in any other location listed.

Fig.12. *Ggplot Geom_Bar & Geom_Point* - Top 5 Crimes per year



The visualisation above was constructed to show the top 5 types of crime, and there values per year. It is very evident that the highest number of crimes was *Theft & related offences* which gives an explanation as to why the crimes are so high for areas in the city centre. Theft is a big problem in Dublin and the shopping districts in the city centre are a the area targeted for this time of crime. All types of crime has increased over the years except offences against government which is represented by the green line and has increased since 2013

Fig.14. *Shiny Dashboard - Leaflet map and Data table*

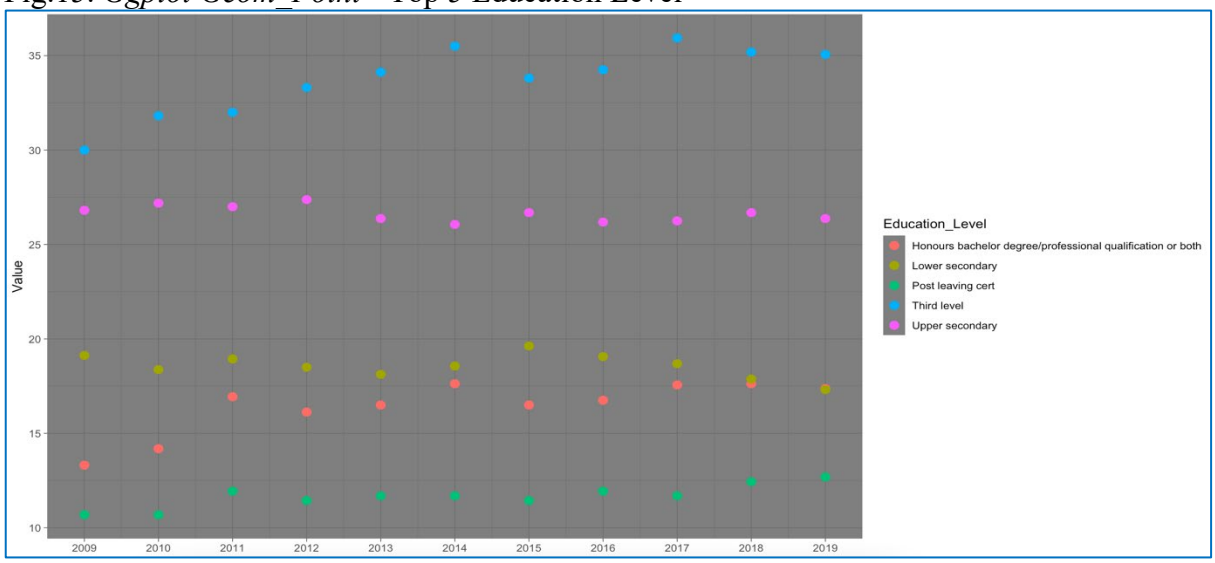


The visualisation above is a powerful interactive upgrade to the initial all time crime values represented on a map. This shiny dashboard was created to include a map and data table with a slider bar to control the data which is shown.

This dashboard features interactivity and reactivity as the slider bar is changed to a different year or a search is made in the data table search box as demonstrated. The year slider bar can be moved to filter out the data to only the selected year. It can be seen in the photo that the year is 2003 and “Howth” is searched in the search bar. Howth is selected on the map and a pop-up is displayed showing that there were a total of 946 crimes for the year 2003, this is matched in the data table. The values on the data table update with the slider bar and the values within each map marker. The marker colours also updates to match the values for the year selected. The marker colouring function is the same as what was used before except the values needed to be made to a fraction of the size of the ones used on the yearly dataset.

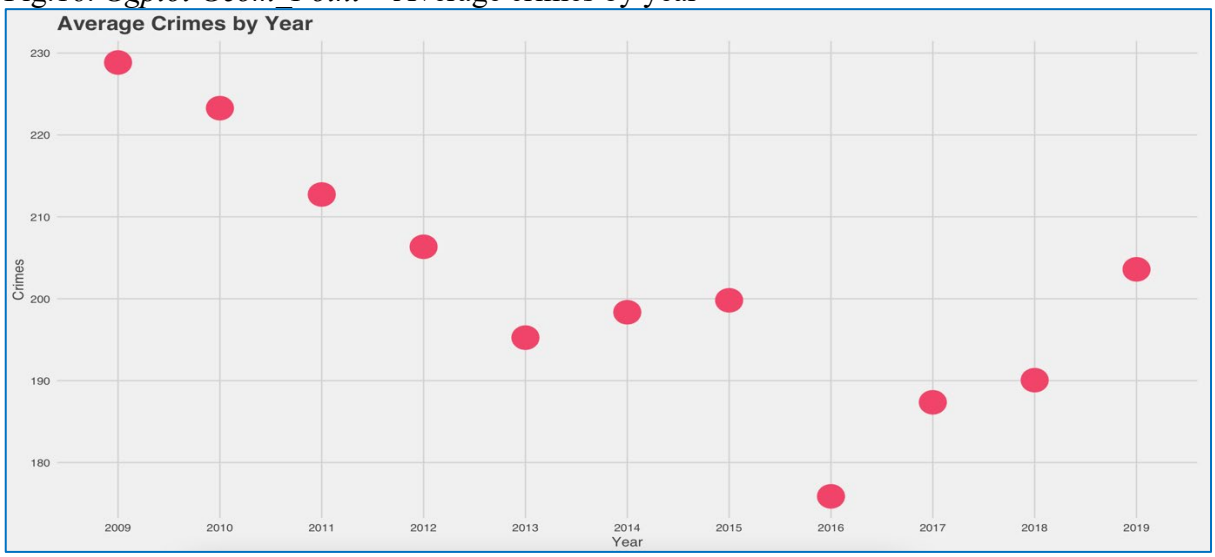
Information retrieval is simple with the use of this dashboard. A user can gain great insight into the amount of crime occurring in a particular year if they wish to combine the information to a separate study. This visualisation could be recommended for real world use of law officials and data scientists in the Garda. It can be used as a road map to investigate the areas of higher crime.

Fig.15. *Ggplot Geom_Point* - Top 5 Education Level



The graph above shows the top 5 types of education represented as percentage values for each year. This was constructed to be compared against average crimes which is displayed below.

Fig.16. *Ggplot Geom_Point* – Average crimes by year

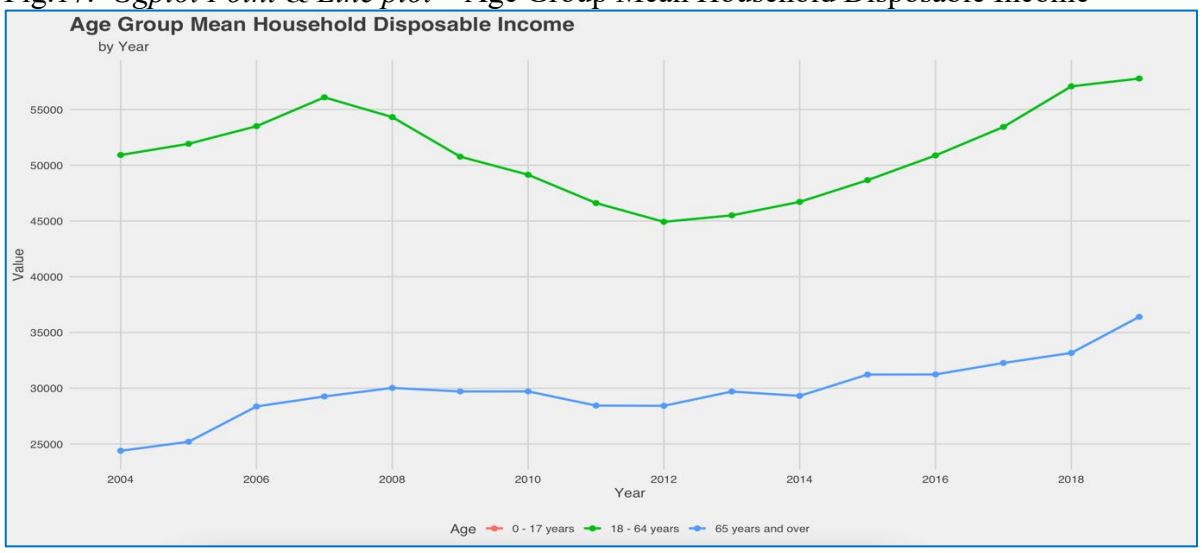


The plot above was constructed to show the average number of crimes by year matching the years of the education dataset.

The education dataset was limited to starting at the year 2009. This caused a limitation as to what could be proved from this. The average crimes dataset was targeted to only contain the same years as the education and these plots were compared.

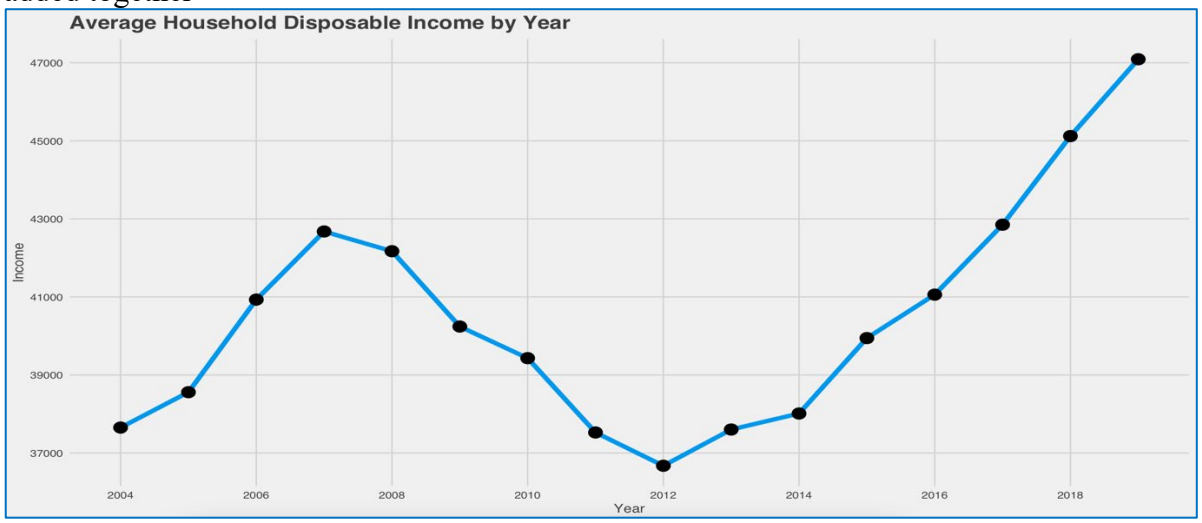
2009 contained the highest average of crimes and the lowest percent of Third Level education attainment. The percentage of honours degree for this year was also at its lowest. This can indicate that due to people not having qualifications to secure themselves a high paying job, they could turn to crime to fill the gap in their revenue. This could be drug offences from trying to sell drugs, theft and burglary offences and even fraud. 2016 were when crimes were at an all-time low, the change in educational attainment was not significant enough to justify that crime was a supplement to not securing a job. The two points made above contradict each other and in reality no assumptions can be made as there appears to be no link between the data sets when compared this way.

Fig.17. *Ggplot Point & Line plot* – Age Group Mean Household Disposable Income



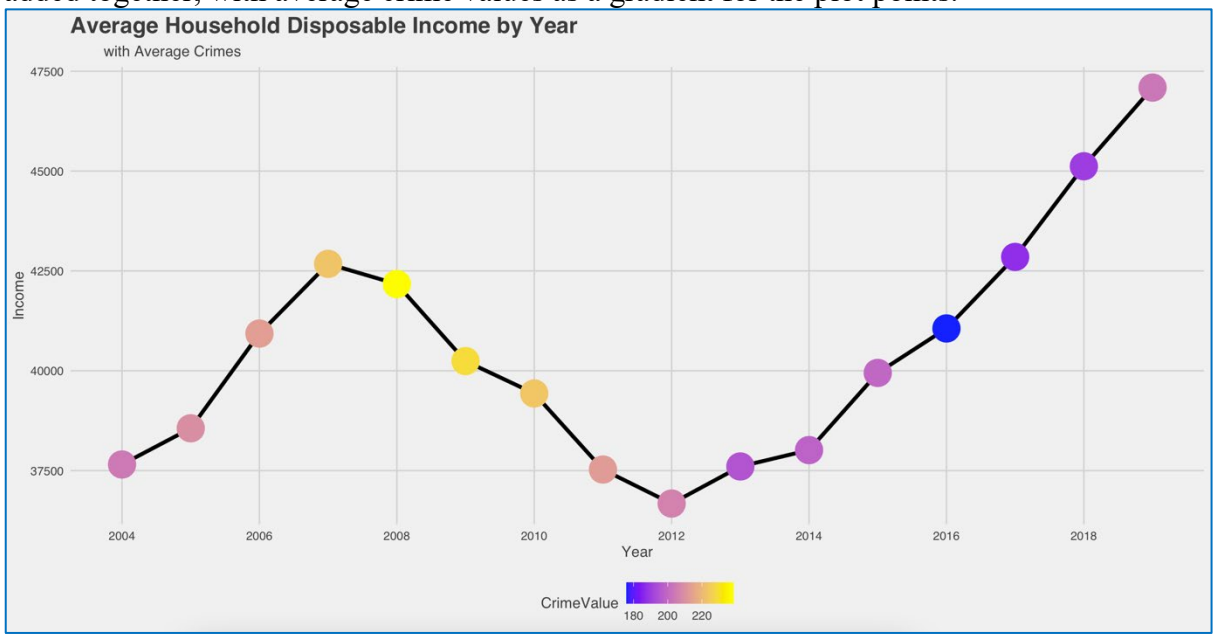
The graph above shows the mean disposable income for each age group which is indicated in the key at the bottom. The income dataset starts from the year 2004 which is one year later than the crime dataset. The age group of 0-17 years contains no values as the values for this variable was NA. As mentioned previously in the report there was nothing that could be done with these NA's that wouldn't cause the data to be skewed and hinder accuracy so they were left as was.The green line represents the age group of 18-64 years of age which is the age which most people will spend working, the minimum is approx. €45000 and the max is approx. €62500. It was expected that the value for this would be the highest of the two groups. The blue line represents those 65 years of age and over which and due to the value displayed, this revenue may be generated from pension schemes which is what causes it to be so much lower than the working group The minimum value for this group is approx. €24800 and the maximum is approx. €33000. Both values are almost half that of the 18-64 age group. It is evident on the graph that the year 2019 was the peak year for income for both age groups.

Fig.18. *Ggplot Point & Line* - Average Mean Household Disposable Income – age groups added together



The plot above shows the average household disposable income for the two groups combined mentioned above. This is the sum of disposable income for both age groups. The graph shows an incline until 2007, then a big dip to 2012 then a peaking incline up to 2019. Where the total disposable income exceeds €47,000.

Fig.19. *Ggplot Point & Line* – Average Mean Household Disposable Income for age groups added together, with average crime values as a gradient for the plot points.

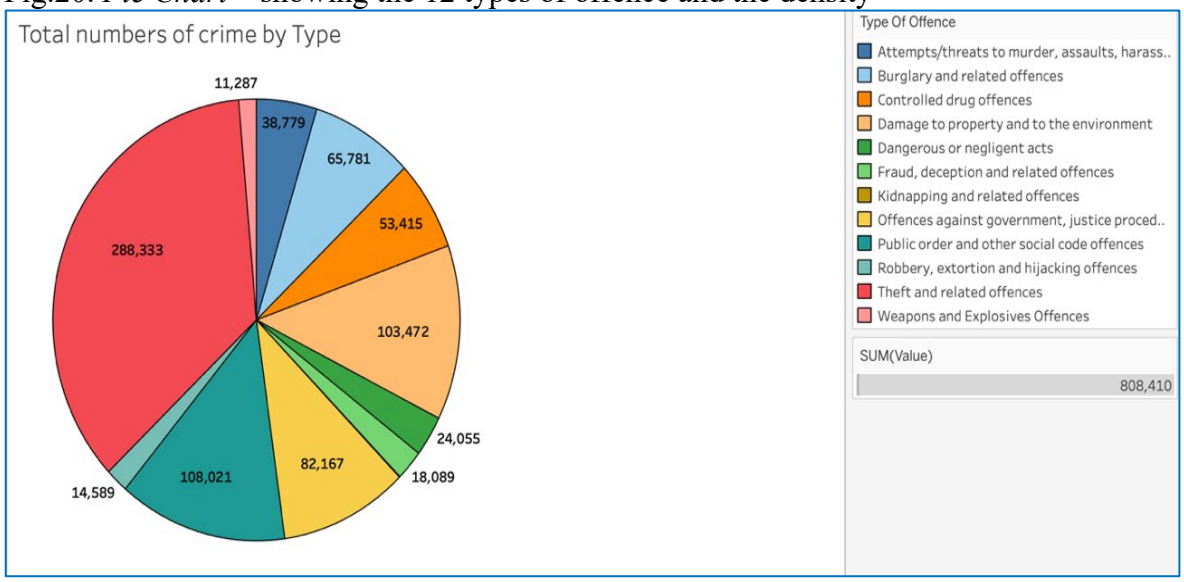


The plot above is similar to the one explained before except it contains an additional variable. The average crime values was chosen to be included in this for a quick analysis on whether or not yearly average crime values are affected by income. The dark blue symbolises the lowest number of average crimes, the pink/purple colour symbolises the middle number of average crimes and the yellow symbolises the highest number of average crimes.

An observation of this graph shows that income and crime don't seem to have any direct relationship with each other. In 2008 the crime value is at a peak, the income is the 5th highest on the chart from a total of 16 observations. In 2012 income is at the lowest and the average crime value seems to be between 200-220. The income then starts to pick up and peaks in 2019 at close to €47500. At this time, the value of average crimes is close to that of 2012. Between these years at 16 average crimes is at the very lowest at around 180 crimes.

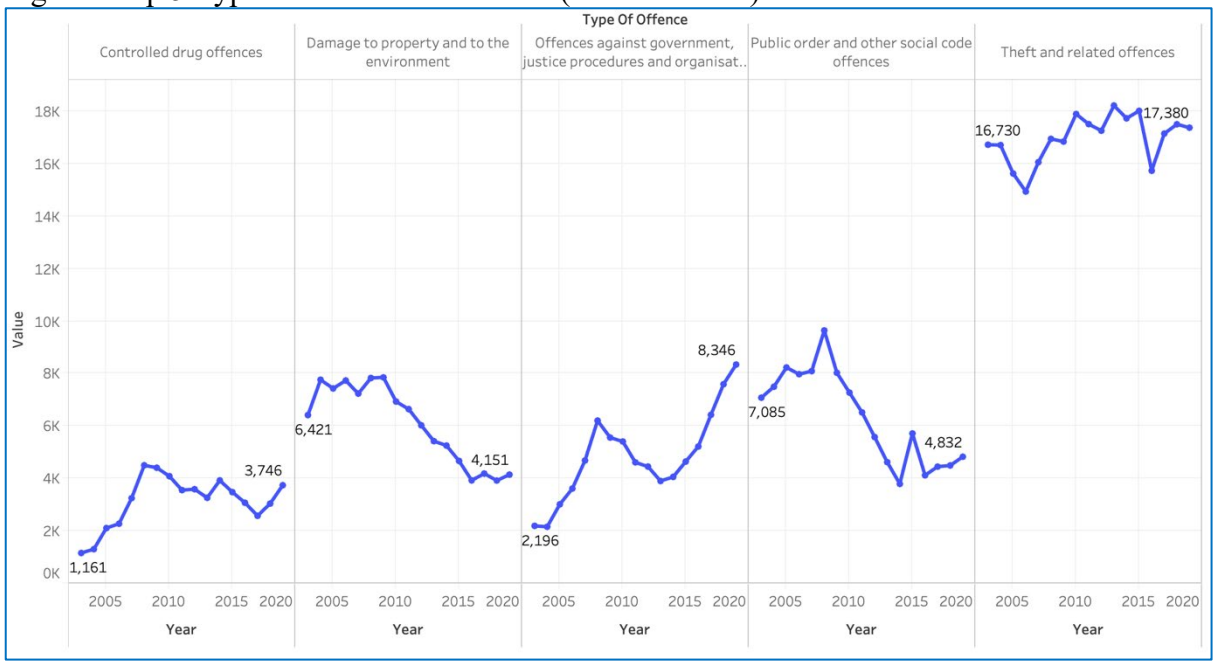
It can be concluded from this graph that income has no effect on the value of average crimes, but this does not mean it has no relationship with particular types of crime. These will be further explored in this analysis.

Fig.20. *Pie Chart* – showing the 12 types of offence and the density



The pie chart above represents the density for each type of crime based on the reported incidents of the crime. The labels are the total numbers of crime for that type and the colour matches the key displayed on the right hand side. A sum was also calculated which shows there was a total of 808,410 crimes reported within the 17 years(2003-2019). This is an average of around 47,000 crimes per year. This is a very high value for just 40 Garda Stations in Dublin and proves the point that crime is a problem.

Fig. 21. Top 5 Types of crime – 2 variables(Year & Value)



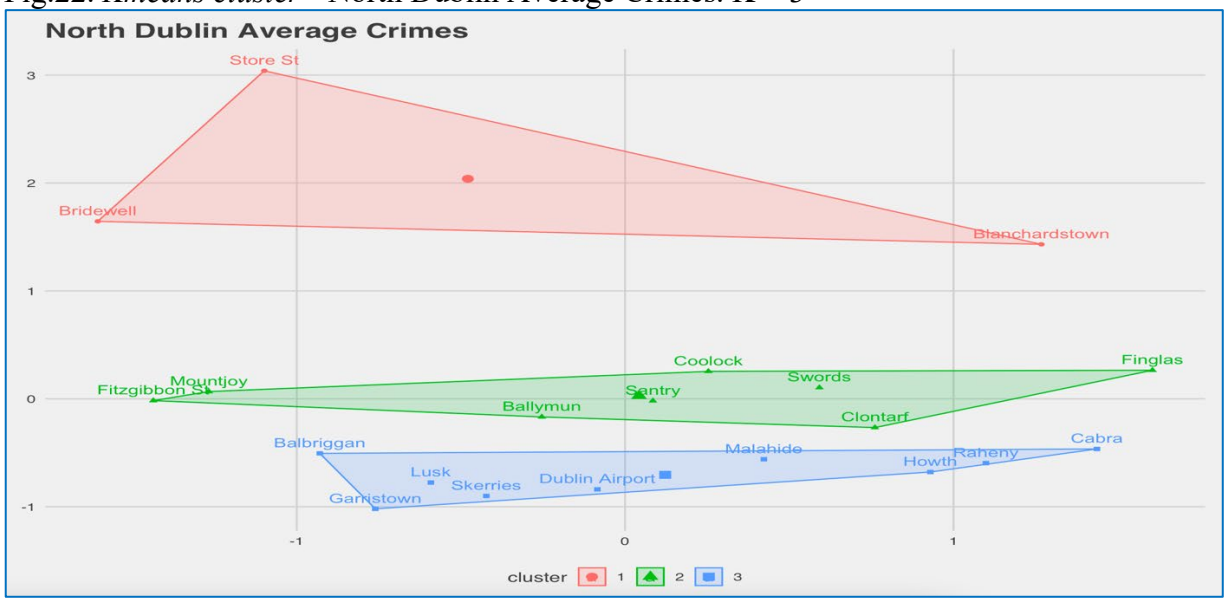
Comparing the *Theft and Related Offences* peak year 2013, to income in 2013. It can be concluded that there is no direct association with the activity of crime in regards to income when compared using the graphs.

Damage to Property and to the Environment has decreased by a couple of thousand from start to finish and when compared alongside the education dataset there is no major change in the educational attainment to suggest that education has any link to this decrease.

5.2 Data Mining

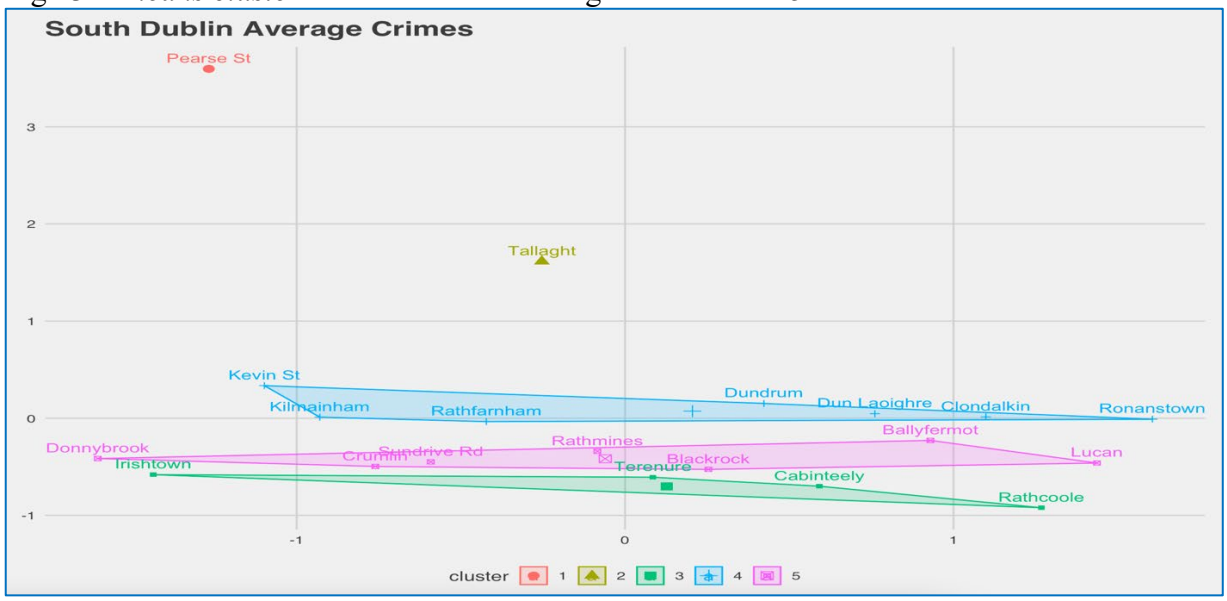
K-Means clustering was the first data mining technique to be implemented in this analysis. A K-Means cluster was run on each data set to show the groups within the data. The K value signifies the number of clusters to be outputted for the visualisation and this was manually selected for each one. The K-Means clusters group the data based on the nearest neighbour mean value. Variables with similar means will be grouped closely.

Fig.22. *Kmeans cluster* - North Dublin Average Crimes. K = 3



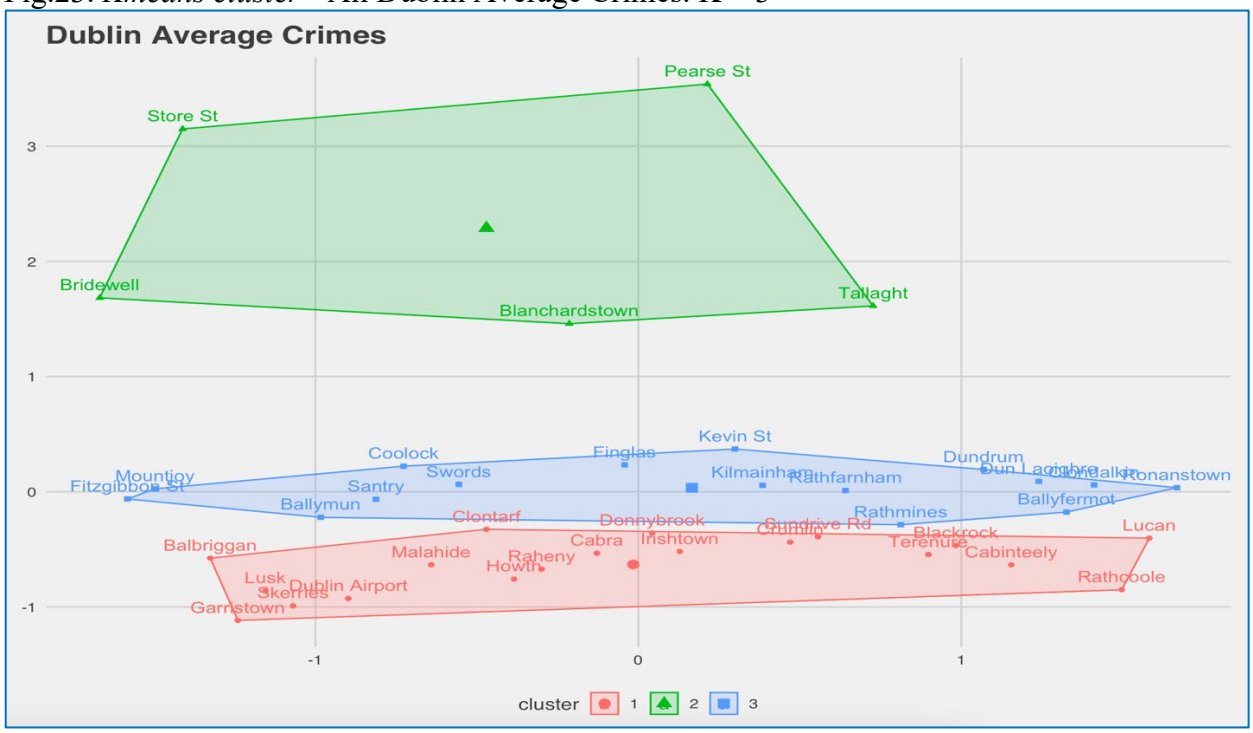
The cluster above shows 3 clusters for North Dublin. Cluster 1 contains 3 variables, cluster 2 contains 8 variables and cluster 3 contains 10 variables. The higher the cluster on the graph, the higher the value.

Fig.23. *Kmeans cluster* - South Dublin Average Crimes. K = 5



The south cluster contains 5 groups as there are two major outliers to the rest of the data. Pearse Street and Tallaght have values so much higher than the rest meaning they needed their own cluster. The means of the values in the 4th blue cluster were too different to Tallaght and Pearse which is why it was decided to set the K value to 5 for this visualisation.

Fig.23. *Kmeans cluster* - All Dublin Average Crimes. K = 3



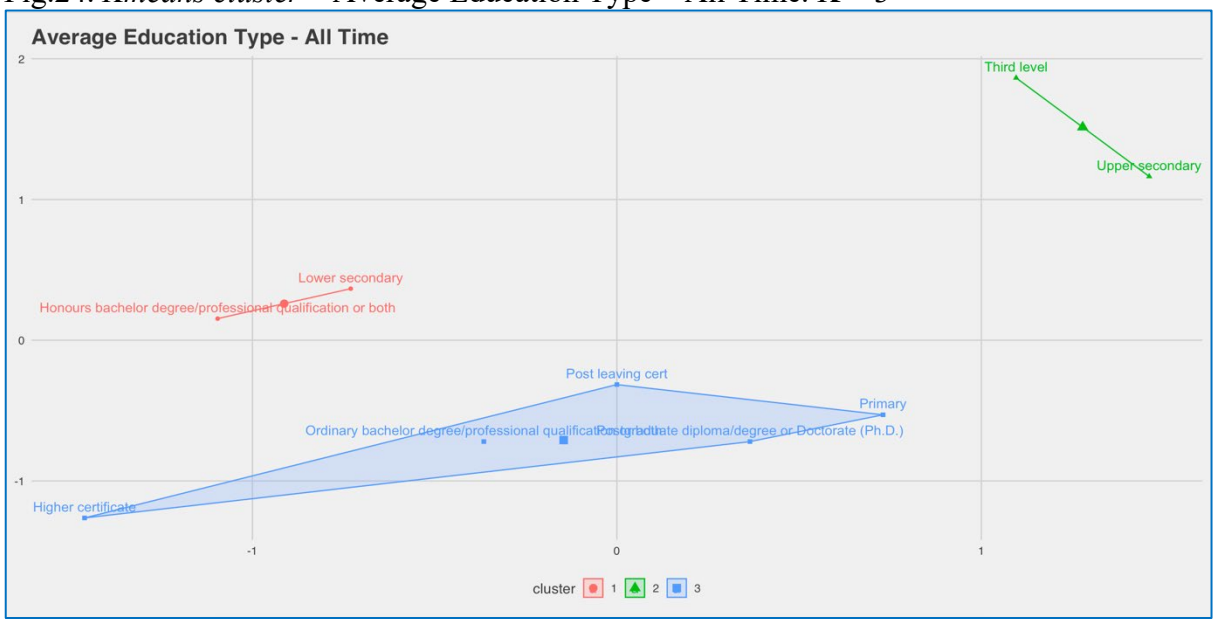
A cluster visualisation was made to show 3 clusters of data for the 40 Garda Stations in Dublin. The group with the highest number of values here was the 3rd blue cluster. From this cluster it can be observed that 5 Garda Stations have a high average value of crime reported in the analysis.

This can be justified as follows:

Store Street, Pearse Street and Bridewell are all located in the City Centre which is considered a shopping district. The highest number of crimes committed each year was Theft related crimes indicating that that played a major role in the value of these locations. Also, the population density in the inner city is also much higher than suburb areas which would also influence more crime. Having these three Garda Stations in a reasonably close proximity to one another means that the likelihood of seeing people committing a crime is higher than suburb areas that are limited to one station for many kilometres. This is due to the fact that the number of Guards and patrolling Garda is possibly higher meaning they have a higher chance of finding any sort of criminal offences being committed.

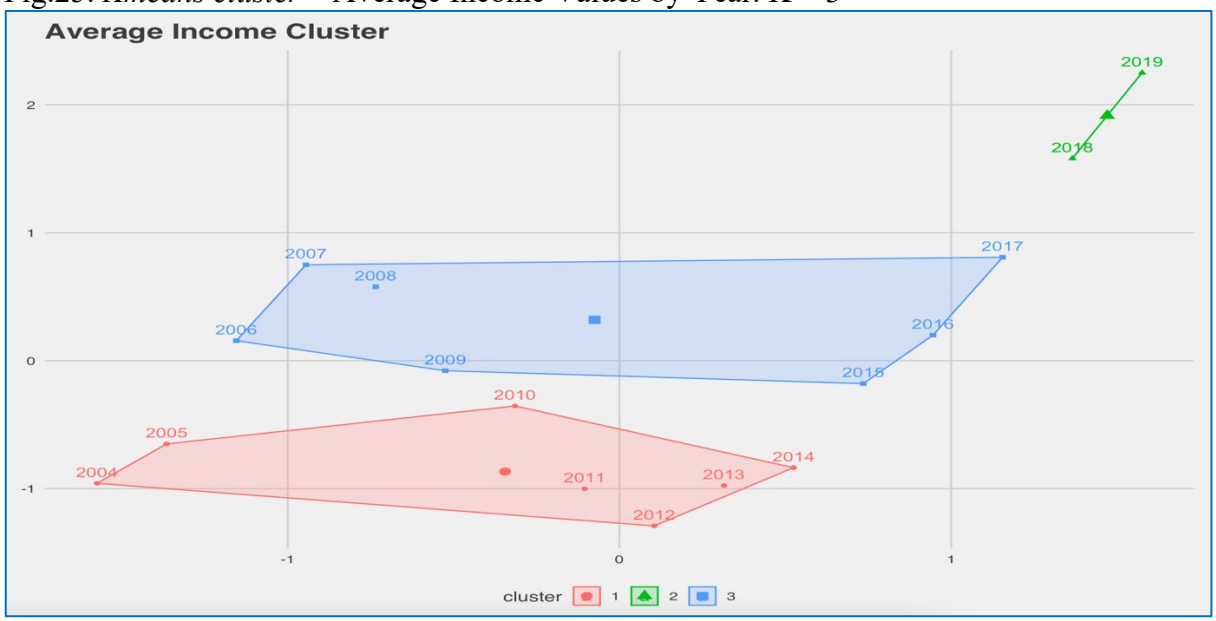
Blanchardstown and Tallaght are two areas which are isolated from other Garda stations for quite a distance. They have a larger district region which they need to respond to which would drive the number of crimes up for these stations. In areas like Coolock, Raheny and Clontarf that are more close to each other than other stations, they would each operate in their own radius which means crime from locations out of their jurisdiction won't be reported to the wrong police station.

Fig.24. *Kmeans cluster* – Average Education Type – All Time. K = 3



The cluster above is constructed based on the all-time values of each education type. There is a total of 3 clusters. Third level & Upper secondary are the two highest averages and have similar means. Lower secondary and Honours bachelor degree/processional qualification or both is contained in the second group and the third group shows the rest. The cluster groups in this have very different means from one another and it's visible from how the clusters are presented.

Fig.25. *Kmeans cluster* – Average Income Values by Year. K = 3



These clusters are representative of the income value for each year. 2018 & 19 make up the cluster of the highest income. The other two clusters have 7 values in each and they are dispersed similarly.

Linear modelling was used to find relationships between variables in the data, or check if they even exist. The North Dublin crime dataset was exported and reformatted in excel. This needed to be done as there was only one variable containing all types of crime, for 12 different types of crime. The dataset was reformatted and each crime type was given a column to be its own variable. The key for these columns is displayed below;

Fig.26. Key of Crime columns - with explanation as to what crime it the number symbolises.

The data was reformatted so that each variable could be tested individually instead of as one large column of crimes with thousands of rows. These are the crime types that were included in the model creation. Two models were constructed, a full model and a half model. The full model included all of the crime types and the half model only contained crime types that showed a level of significance within the model through observation of the asterisk at the end of each output.

The *Flextable* library was used in R to style the tables for the models. This was done to make them more aesthetically pleasing and separate them from other visualisations created.

Fig.27. *Flextable* showing model output - Full Model

	Estimate	Standard Error	t value	Pr(>|t|)	
(Intercept)	40,566.936	258.361	157.016	0.0000	***
Crime.1	28.620	2.917	9.810	0.0000	***
Crime.2	-6.287	1.981	-3.174	0.0017	**
Crime.3	-2.096	1.292	-1.623	0.1056	
Crime.4	-5.390	1.529	-3.526	0.0005	***
Crime.5	14.236	3.206	4.440	0.0000	***
Crime.6	-4.395	4.639	-0.947	0.3442	
Crime.7	-33.681	72.072	-0.467	0.6406	
Crime.8	0.158	0.164	0.960	0.3378	
Crime.9	-1.081	0.744	-1.454	0.1470	
Crime.10	-19.292	8.406	-2.295	0.0224	*
Crime.11	-0.394	0.446	-0.883	0.3777	
Crime.12	14.732	9.949	1.481	0.1397	
*Signif. codes: 0 <= '***' < 0.001 < '**' < 0.01 < '*' < 0.05 < '.' < 0.1 < '' < 1*					
Residual standard error: 2322 on 307 degrees of freedom					
Multiple R-squared: 0.3621, Adjusted R-squared: 0.3372					
F-statistic: 14.52 on 307 and 12 DF, p-value: 0.0000					

The linear model above is a full model which means it contains all of the crime variables. The dependant variable for the models is Income. The independent variables are the crime types.

From the full model output above, it is evident that 6 out of the 12 variables have significance towards the income value, based on the observation of the asterisks. These asterisks suggest a level of statistical significance with the regression coefficient. These were the values which were later chosen for the half model.

The Multiple R-Squared value for this model shows the level of explanation in which variance in the model can be described. In this model the value is 0.3621. This means that a total of 36% of the variance in this model can be described. In this case, a low Multiple R-Squared value does not affect the conclusion of the fit of the model. Due to the fact that the Adjusted R-Squared isn't much higher than the Multiple R-Squared, this shows that overfitting was not an issue for this model. *(Frost, 2020)*

The F-Statistic in this case is not statistically significant. This means that there is not much of a relationship between the dependent variable and the independent variables. From these results and with the inclusion of the P value < 0.05, it can be concluded that this model is statistically significant

Due to the P-Value being less than alpha at 0.05, this suggests that this is a good fit model. The result of this model would be to reject the null hypothesis.

Fig.28. *Flextable* showing model output - Half Model

	Estimate	Standard Error	t value	Pr(>|t|)	
(Intercept)	40,593.309	254.918	159.241	0.0000	***
Crime.1	22.014	2.194	10.035	0.0000	***
Crime.2	-5.276	1.795	-2.939	0.0035	**
Crime.4	-4.290	1.381	-3.108	0.0021	**
Crime.5	11.469	3.176	3.611	0.0004	***
Crime.10	-27.416	7.420	-3.695	0.0003	***
*Signif. codes: 0 <= '***' < 0.001 < '**' < 0.01 < '*' < 0.05 < '.' < 0.1 < ' ' < 1*					
Residual standard error: 2392 on 314 degrees of freedom					
Multiple R-squared: 0.3078, Adjusted R-squared: 0.2968					
F-statistic: 27.92 on 314 and 5 DF, p-value: 0.0000					

The crime variables chosen in this all have some level of statistical significance with the regression coefficient as implied by the asterisks.

The Multiple R-Squared in this model is 0.3 indicating 30% of the variance can be described. This is less than that of the full model. The Adjusted R-Squared is again close to the Multiple R-Squared which indicates that overfitting didn't occur.

The F-Statistic is also not statistically significant in this model.

The P-Value < 0.05 indicating that there is significance in this model and that the model is a good fit.

The full model would be considered the best model in this case as the Adjusted R-Squared value is the highest. This means that the most information captured between these two models exists in the full model. The Akaike Information Criterion value was calculated for both models. *(Bevans, 2020)*

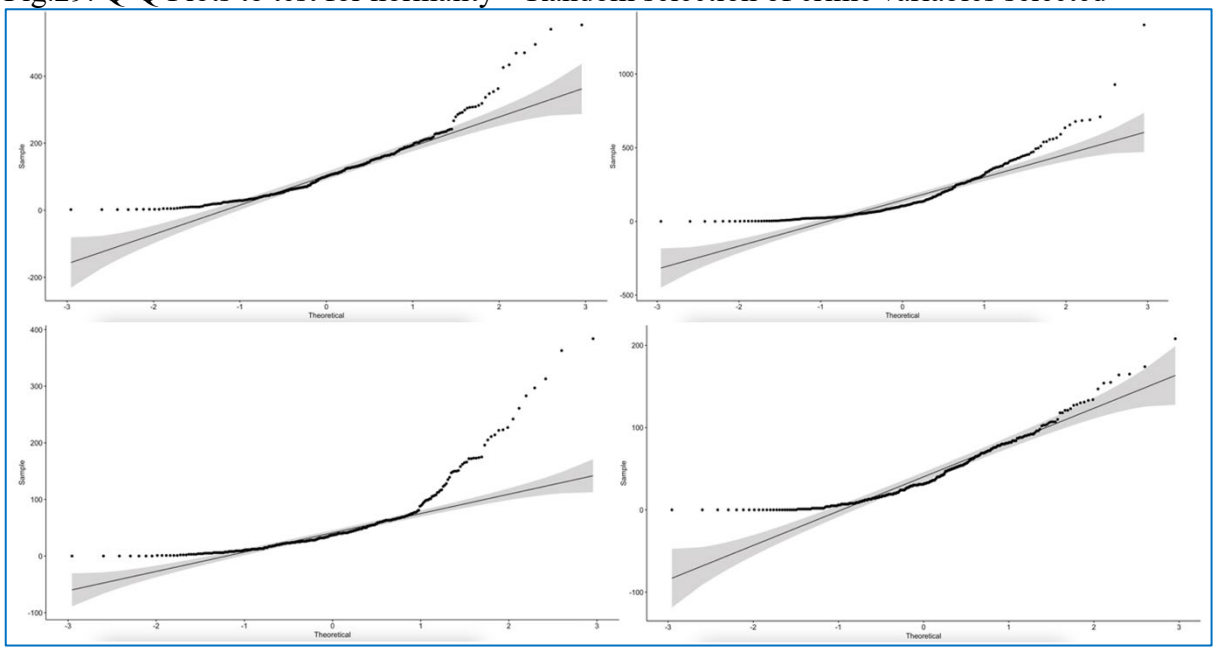
AIC for the full model = 5883.013

AIC for the half model = 5895.166

Although the value is only slightly smaller for the full model, this still suggests that it is a better fit due to less prediction error.

5.3 Statistical Tests

Fig.29. Q-Q Plots to test for normality – Random selection of crime variables selected



The Q-Q plots above show that the data distribution is not-normal. There are many outliers on each of the plots. There are too many outliers to even remove to check if outliers are effecting the normality of the data.

To be absolutely sure the data distribution was not normal, a Shapiro-Wilk normality test was also done. The P-Value of the Shapiro-Wilk test is used to indicate whether data is normally distributed or not.

Fig.30. Shapiro-Wilk Normality Test

Shapiro-Wilk normality test data: crimeOffences_northDublin_reformatted$Crime.1 W = 0.8739, p-value = 1.59e-15	Shapiro-Wilk normality test data: crimeOffences_northDublin_reformatted$Crime.3 W = 0.80256, p-value < 2.2e-16
Shapiro-Wilk normality test data: crimeOffences_northDublin_reformatted$Crime.6 W = 0.7362, p-value < 2.2e-16	Shapiro-Wilk normality test data: crimeOffences_northDublin_reformatted$Crime.10 W = 0.89844, p-value = 8.281e-14

The P-Value for all tests is < 0.05 which concludes that the data in each of the crime variables is not-normal. This was used as an indicator for what tests should be run on the data. The tests which are conducted for data that is not normally distributed are non-parametric tests.

Fig.31. Spearman's Correlation between variables that may have a link to income

Correlations

			Income	Crime4	Crime5	Crime6
Spearman's rho	Income	Correlation Coefficient	1.000	-.118*	-.004	.109*
		Sig. (2-tailed)	.	.030	.940	.045
		N	340	340	340	340
	Crime4	Correlation Coefficient	-.118*	1.000	.845**	.574**
		Sig. (2-tailed)	.030	.	.000	.000
		N	340	340	340	340
	Crime5	Correlation Coefficient	-.004	.845**	1.000	.573**
		Sig. (2-tailed)	.940	.000	.	.000
		N	340	340	340	340
	Crime6	Correlation Coefficient	.109*	.574**	.573**	1.000
		Sig. (2-tailed)	.045	.000	.000	.
		N	340	340	340	340

Spearman's correlation was done over Pearson's-R correlation due to the data being not normal. Spearman's correlation is a non-parametric test. The crimes chosen for this data set were:

Crime 4 - Robbery, extortion and hijacking offences
Crime 5 – Burglary and related offences
Crime 6 - Theft and related offences

These were chosen as they may have some relationship with crime based on logic that money can attract robbery, burglary, and theft related offences. The correlation was run to look for a relationship between the crimes and income.

Crime 4 (P=.030 < 0.05. Correlation Coefficient= -.118 < 0 indicating no correlation but statistical significance between the variables.

Crime 5 (P=0.9 > 0.05. Correlation Coefficient= -0.004 < 0 indicating that there is no statistical significance between the variables and no correlation.

Crime 6 (P=0.45 < 0.05. Correlation Coefficient= .109 <0 indicating that there is no correlation again, and statistical significance between the variables.

No correlation was present between these selected variables and income. It has previously been suggested through the visualisations comparing crime and income that there is no relationship evident between the two for the selected datasets, this can now be concluded.

6.0 Conclusions

An advantage of this project is the information retained in the visualisations. The visualisations created can give an insight to people who want to know information about the topic of crime in Dublin. This analysis provides the service of information retrieval at a level that can be understood by the general public. This is something that will be appreciated by viewers with no data science background. This is a key advantage that this project provides. The complex tasks are done in the back end so that the viewer can view aesthetically pleasing graphs to gather their information, rather than using maths and coding.

Limitation was a problem in this analysis and was the biggest disadvantage. The two datasets that were chosen contained not only a limited amount of information, but also a limited amount of observations. The income data for example only contained 10698 observations for a total of 4183 households. It is not clear where these households were based but there is a small chance that the locations are primarily Dublin, as they are labelled Ireland. This hinders the accuracy of the dataset completely. The crime dataset contained locations but the income one did not. This was the best available income dataset that could be used, which further explains the point made previously that American data may have enabled more testing and gave better results with correlation between datasets. The education dataset also didn't specify a location which indicates that accuracy with the results from using education as a cause of crime may not be accurate. It can be concluded that from this analysis, neither income or education have any effect on the rate of crime. This may be down to the fact that the data used for each of the datasets didn't contain a location, or it could be just a general fact. However it is definitely clear that the accuracy of the results are hindered by the accuracy of the combined data.

Although there were limitations as to what could be proved from the supplementary datasets in relation to crime, a lot of information can still be interpreted from this study. Crime in North Dublin is rampant. It is an every-day issue that occurs in the public eye and has become somewhat of a norm in reality. Theft related offences are the most common with the highest volume, and these are only the offences that are recorded.

Committing crime is a job for some. People rely on criminal activity to generate income, this is particularly evident in disadvantaged areas such as Ballymun, Coolock and Finglas, where drug dealing is a big issue. The problem with these offences is that they are harder to catch. Theft in Dublin City is controlled by the use of cameras in stores, this allows security guards to monitor the store and analyse areas of crime in the store and common times that they are occurring. Controlled drug sales are harder to catch as there may be a shortage of cameras where these trades are occurring. This is a known issue although the total offences recorded in this data does not match the activity of this kind of crime. Drug crime is definitely one of the most popular forms of crime in Dublin City and access to such products is as almost easy as the access to legal goods sold in stores.

This analysis was chosen to focus on crime as it is a serious topic. Theft in a store may not lead to death, but the misuse of drugs or weapons and explosives can. An important observation to take from this analysis is that all crimes contained in the crime dataset were only reported offences. The values do not represent that unreported offences that occur. If every single offence was recorded the value could exceed the values by tenfold. Although only reported values were recorded, the numbers were still high showing the problem with crime.

All objectives that were set out were successfully met. This statement defines the overall completeness of this project. A limited amount of data tests could only be conducted, this was due to a limitation with the data that was used. If it was known at the start of this analysis that the data would propose problems down the line, other datasets may have been introduced for flexibility.

7.0 Further Development or Research

An extension on time would grant the opportunity of introducing additional datasets that could supplement the crime data. As there was a limited window in which data could be obtained for the proposal of this project, not every possible source was explored. More datasets that have potential relationships with crime could have been introduced and statistically tested to look for correlation between them.

Also with more time, a request of information could have been made to the Central Statistics Office for guidance on what exact locations were used to populate the income and education datasets. This would have gave more insight into the practical use of the selected datasets. The final conclusion of this study is that the supplementary datasets income and education, have no relationship with crime in Dublin. However, this may be incorrect based on the location of the households used in each dataset. Using areas that the supplementary datasets belong to could deliver different results, however this was just an observation made from this study.

The focus of crime in this analysis could be scaled focusing on areas outside of Dublin to show the distribution of crime across the country on an interactive map similar to what was used in this. This would give an interesting observation on whether crime in cities differs from crime in suburb areas. Due to the population density of Dublin compared to surrounding areas, the values of crime would still be indefinitely larger. However the type of crime may differ when dealing with new locations. Theft may not be the highest totalling type of crime in areas with a small shopping district. R scripts for a task like this would be much larger and would require far more time to complete.

Another important study which could be implemented is the comparison of Dublin to a location of similar population in England such as Manchester. If the same kind of data could be obtained on England's crime this would be a great element to add to this study.

Machine learning is an area that was initially planned to be included in this analysis for crime prediction but couldn't be due to the lack of time series to be analysed. The crime data only contained years. Prediction would need variables such as dates to be successful and accurate. This was dismissed due to limitations mentioned previously with the data but is something that could definitely be looked at if a dataset containing dates was to be located.

Further research would include a search for events that took place that can be linked to crimes in particular Garda Stations such as armed robberies etc. to give a better insight as to the type of incidents each Garda Station is dealing with. The crime dataset is limited in what it can actually show whereas the research and study of news articles combined with the crime dataset could lead to a deeper understanding of why crimes are higher in certain years than others, as it was proven the supplementary datasets can't show any indication for this matter.

The course content for this analysis was taught at the same time the project was being developed. This meant that no prior background knowledge to the subject of data analysis was obtained before starting this project. This caused a lot of difficulty as it became apparent that particular statistical tests required certain data, some visualisations needed certain data and the overall information that could be retrieved from the dataset was limited due to this. If the topic of data analysis was explored in previous years of study, this project would have included far less limitation as datasets would have been carefully selected and challenges wouldn't have been a first time occurrence. The programming language R was being learned in other course subjects simultaneously to applying it to this project which caused issues with a lack of knowledge. After completing this project and upskilling in R Studio, dataset limitations are known now that weren't at the start. This means that a more information dense analysis with accurate conclusions could be produced in the future leading to a better study of the subject.

8.0 References

Bevans, R., 2020. *An introduction to the Akaike information criterion*. [online] Scribbr. Available at: <https://www.scribbr.com/statistics/akaike-information-criterion/> [Accessed 12 May 2021].

Brew, T., 2020. *Social Conditions: Income And Poverty Rates*. [online] Central Statistics Office. Available at: <https://data.cso.ie/table/SIA13> [Accessed 22 December 2020].

Crilly, S., 2020. *Educational Attainment: Persons Aged 15-64*. [online] Central Statistics Office. Available at: <https://data.cso.ie/table/EDQ01> [Accessed 22 December 2020].

Farmer, C. and Wasser, L., 2020. *Creating Interactive Spatial Maps in R Using Leaflet*. [online] Earth Data Science - Earth Lab. Available at: <https://www.earthdatascience.org/courses/earth-analytics/get-data-using-apis/leaflet-r/> [Accessed 10 May 2021].

Frost, J., 2020. *How To Interpret R-squared in Regression Analysis*. [online] StatisticsByJim. Available at: <https://statisticsbyjim.com/regression/interpret-r-squared-regression/> [Accessed 12 May 2021].

Kang, H. and Kang, H., 2017. Prediction of crime occurrence from multi-modal data using deep learning. *PLOS ONE*, [online] 12(4), p.e0176244. Available at: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0176244> [Accessed 11 May 2021].

S. Kim, P. Joshi, P. S. Kalsi and P. Taheri., 2018. *Crime Analysis Through Machine Learning.* [online] IEEE. Available at: <https://ieeexplore.ieee.org/document/8614828> [Accessed 22 December 2020].

Scriven, S., 2020. *Crime & Justice: Recorded Crime Offences Under Reservation*. [online] Central Statistics Office. Available at: <https://data.cso.ie/table/CJA07> [Accessed 22 December 2020].

Technik, D., 2019. *Shapiro-Wilk Test for Normality in R | R-bloggers*. [online] R-bloggers. Available at: <https://www.r-bloggers.com/2019/08/shapiro-wilk-test-for-normality-in-r/> [Accessed 11 May 2021].

Varan Nath, S., 2020. *Crime Pattern Detection Using Data Mining*. [online] Cs.brown.edu. Available at: <http://cs.brown.edu/courses/csci2950-t/crime.pdf> [Accessed 22 December 2020].

9.0 Appendices

9.1 Project Proposal

National College of Ireland

Project Proposal
Analysis of Crime In North Dublin
08/11/2020

Technology Management
Data Analysis
2020/21
Ryan Johnston
X17437624
X17437624@student.ncirl.ie

Contents

1.0 Objectives

The dataset I will be using is the reported incidents of crime by Garda Station in Ireland. The dataset is made up of the total number of incidents by type, for each garda station ranging from 2003-2019. The focus of this study is on Dublin, mainly North Dublin but also using South Dublin as a comparison to crime activity and influences. This dataset has its own limitations as to what can be proven, but backed with other datasets a strong analysis can be made.

Aims;

- Assess background motives of crime that may have relevance to the study. For example, what factors may cause a higher volume of crime e.g. poverty, educational attainment etc. Trying to find a link between those to crime using data sets.
- To calculate the top 5 areas of crime in North and South Dublin
- Calculate the mean of north vs south to determine which area crime is occurring more and interpret the results.
- Calculate the top 5 types of crime of all time based on the overall value of each.
- Create visualisations that simplify the values for the viewer allowing for more efficient understanding and interpretation.
- Create a map visualisation showing locations of each garda station with interactivity.
- Construct a Shiny Interactive dashboard containing a map and data table that updates values based on the user selection of a slider bar.
- Carrying out statistical data analysis tests that apply to the data and topic.

2.0 Background

The reason I chose this dataset is to analyse an existing problem which is evident on a daily basis in all areas of the world; crime. Growing up in an area considered disadvantaged, crime can be seen first person on a regular basis. Theft, vandalism, drug dealing, fraud, assault and harassment are a handful of areas, within a bigger picture of what goes on all around us. It is at the hands of the Gardaí to ensure the safety and to enforce law abidance on the public; however sometimes people take matters into their own hands.

I will discuss the topic of Crime focusing on the motives behind peoples actions. What are the reasons why individuals go down the wrong route? This will be discovered within the finished analysis.

Possible motives:

Rewarding: *"Throughout history people have tried to explain why people would commit crimes. Some consider a life of crime better than a regular job." (JRank, N/A)* The life of crime brings fast, tax free, undeclared revenue into the hands of the participants. The minimum wage in Ireland as of October 2020 is €10.10 per hour, where it will soon increase by 10cent in January 2021. *(RTE, 2020)* The unlawful sale of drugs can exceed this number exponentially. This is how people fall into the trap. Young people get themselves involved in this before they can legally get their first job, resulting in the thought behind getting a proper job unappealing. This is just one area of crime that has been known to drive high rewards, and the risk of getting caught is hardly ever considered. This is the start of the lavish lifestyle loop for some, the creation of expensive habits.

Lack of Education: The 2016 census conducted by the CSO revealed that 57% of Irish prisoners have qualifications only up as far as a Junior Cert, many of whom haven't even done the Junior Cert. *(Jack Power, 2020)* For some it's not the money that attracts them to a life of crime, it's just as basic as *income* as a whole. A valid leaving cert is a valid requirement for a lot of jobs. It is the most basic form of qualification in the educational sector. The discipline of going to school is something that lacked in their lives and whether that was their parental situation or social influence of friends is a possible link to their downfall.

Poverty: *"Criminal action is the way for poorer people to acquire economic goods, which could not be attained legally." (Sileika & Bekeryte, 2012)* Crimes such as theft, burglary and fraud are examples of how someone experiencing poverty can obtain items or funds that they can't afford legally. Poverty can be linked with debt, that may occur from the loss of a job or addiction to a) gambling, b) drugs or c) alcohol. These are just some of the factors that can influence the risk of poverty upon an individual.

Peer Influence: The transition to adolescence is generally when people start to develop outlooks and perspectives put together by observing their social environment, this is typically when peer influence is evident at its peak. Peers such as; childhood friends, school classmates & workmates sometimes have the ability to influence a certain outlook on an individual. Things that are brand new to the individual like particular; sports, music & fashion may be easier to influence due to the individual having no logical background of the areas. *(Sullivan. J, Christopher & Jolliffe, Darrick, 2012)*

Suppose a group of young boys in school all support Manchester United, if a new person develops a friendship but has never watched football before, he may be more inclined to 'follow the crowd' & develop a liking for the same football club himself. This is a minor form of where individuals can be influenced *(Welsh & Farrington, 2012)*

Easy Access: Easier access to certain things can be an influence on crime. In Ireland, the access to a firearm is not as common as America, it can be difficult to get a license to hold one as you need sufficient reasoning and your area of residence can come into question. *(JRANK, N/A)*

Unemployed individuals are not the only ones who commit crime, those who work directly with products and money have easy access to such things and can easily perform dishonest actions. The use of cameras in a workplace can help to prevent this however they cannot prevent every instance.

These are some of the many factors that can motivate people to commit crimes consciously or subconsciously. These are some areas that will be researched during the course of this project to hopefully reveal answers as to whether or not these factors have an effect on certain areas in North Dublin.

3.0 Technical Approach

Research will be done internally within datasets and externally using alternative datasets and sources such as academic papers, web articles and books.

Requirements Capture; the list of objectives outlined is the bare minimum requirements as to what needs to be answered. This will allow for a deeper analysis which will answer potential questions that may arise. The dataset of the crimes will be compared with datasets in relation to the income, education attainment & other factors that influence crime.

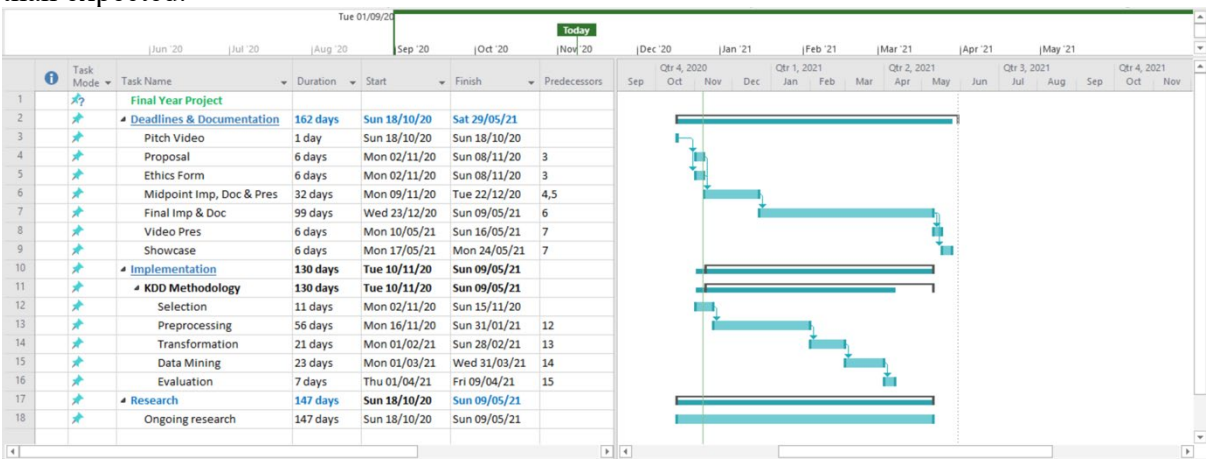
Implementation: I will use R and R studio to manipulate and analyse the chosen datasets. The KDD methodology will be followed to ensure the data is completely ready for the visualisation. The creation of a Gantt chart in Microsoft Project will allow me to stay on top of schedule allocating certain timeslots to focus on tasks and due dates.

4.0 Special Resources Required

No special resources are required for this project.

5.0 Project Plan

The Gantt chart has been created using Microsoft Project to outline the expected processes and plan when they have to be done. As of now this is only an estimate and some processes may be subject to change due to completing some earlier than others and some taking longer than expected.



	Task Mode	Task Name	Duration	Start	Finish	Predecessors
1		Final Year Project				
2		Deadlines & Documentation	162 days	Sun 18/10/20	Sat 29/05/21	
3		Pitch Video	1 day	Sun 18/10/20	Sun 18/10/20	
4		Proposal	6 days	Mon 02/11/20	Sun 08/11/20	3
5		Ethics Form	6 days	Mon 02/11/20	Sun 08/11/20	3
6		Midpoint Imp, Doc & Pres	32 days	Mon 09/11/20	Tue 22/12/20	4,5
7		Final Imp & Doc	99 days	Wed 23/12/20	Sun 09/05/21	6
8		Video Pres	6 days	Mon 10/05/21	Sun 16/05/21	7
9		Showcase	6 days	Mon 17/05/21	Mon 24/05/21	7
10		Implementation	130 days	Tue 10/11/20	Sun 09/05/21	
11		KDD Methodology	130 days	Tue 10/11/20	Sun 09/05/21	
12		Selection	11 days	Mon 02/11/20	Sun 15/11/20	
13		Preprocessing	56 days	Mon 16/11/20	Sun 31/01/21	12
14		Transformation	21 days	Mon 01/02/21	Sun 28/02/21	13
15		Data Mining	23 days	Mon 01/03/21	Wed 31/03/21	14
16		Evaluation	7 days	Thu 01/04/21	Fri 09/04/21	15
17		Research	147 days	Sun 18/10/20	Sun 09/05/21	
18		Ongoing research	147 days	Sun 18/10/20	Sun 09/05/21	

6.0 Technical Details

The language of implementation will be R as mentioned previously. The two possible languages at the beginning were R and Python but I have decided to go with R using R studio due to the broader range of statistical libraries. R has proven to be more efficient at data visualisation than python and the aim is to make the data interactive and to use libraries such as GGPlot2, Shiny and Leaflet. *(DataDrivenScience, 2018)*

7.0 Evaluation

The selected data will be processed & manipulated using R studio following the KDD method. The data when ready, will be visualised to aid in the insight of the research analysis. Tableau will also be used to display visualisations.

8.0 References

JRank, Law. (N/A) accessed at: 2020, 11, 07. *"Causes of Crime"* Retrieved from *JRank*: https://law.jrank.org/pages/12004/Causes-Crime.html

Miley, Ingrid. (2020, 06, 10) *"Government Approves 10c an hour Minimum Wage Increase"* retrieved from *RTE*: https://www.rte.ie/news/business/2020/1006/1169795-minimum-wage/

Power, Jack. (2020, 10, 05) "*Highest Education for Half of Prisoners is Junior Cert or Less – CSO"* retrieved from *IrishTimes*: https://www.irishtimes.com/news/crime-and-law/highest-education-level-for-half-of-prisoners-is-junior-cert-or-less-cso-1.4372566

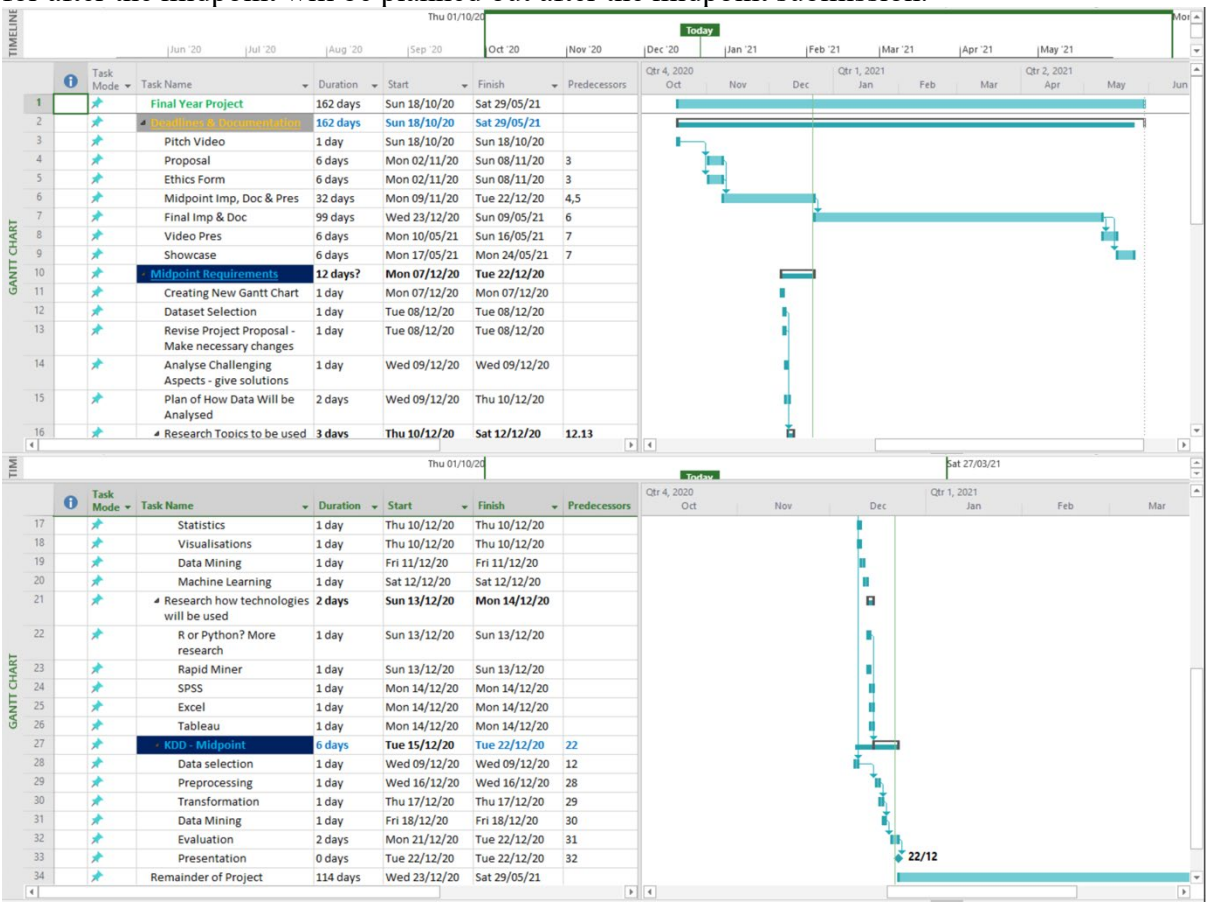
Science, Data-Driven. (2018, 01, 31) *"Python vs R for Data Science: And the winner is.."* retrieved from *Medium*: https://medium.com/@datadrivenscience/python-vs-r-for-data-science-and-the-winner-is-3ebb1a968197

Sileika, Algis. & Bekeryte, Jurgita. (2012, 09, 15) *"Theoretical Issues of Relationship Between Unemployment, Poverty and Crime in Sustainable Development"* retrieved from *Journal of Security and Sustainability Issues*: http://jssidoi.org/jssi/uploads/papers/7/Sileika_The_theoretical_issues_of_unemployment_poverty_and_crime_coherence_in_the_terms_of_sustainable_development.pdf

Welsh, C. Brandon & Farrington, P. David. (2012, 03) *"The Oxford Handbook of Crime Prevention"* retrieved from *GoogleBooks:* https://books.google.ie/books?hl=en&lr=&id=ZRo1DQAAQBAJ&oi=fnd&pg=PA207&dq=crime+peer+influence&ots=ssPS38cAsQ&sig=rN1eTCbhl0rIzbdfawpts9i-MgA&redir_esc=y#v=onepage&q=crime%20peer%20influence&f=false

9.2 Project Plan

A Gantt chart has been created to help me to allocate the correct time for each task. The tasks for after the midpoint will be planned out after the midpoint submission.



	Task Mode	Task Name	Duration	Start	Finish	Predecessors
1		Final Year Project	162 days	Sun 18/10/20	Sat 29/05/21	
2		Deadlines & Documentation	162 days	Sun 18/10/20	Sat 29/05/21	
3		Pitch Video	1 day	Sun 18/10/20	Sun 18/10/20	
4		Proposal	6 days	Mon 02/11/20	Sun 08/11/20	3
5		Ethics Form	6 days	Mon 02/11/20	Sun 08/11/20	3
6		Midpoint Imp, Doc & Pres	32 days	Mon 09/11/20	Tue 22/12/20	4,5
7		Final Imp & Doc	99 days	Wed 23/12/20	Sun 09/05/21	6
8		Video Pres	6 days	Mon 10/05/21	Sun 16/05/21	7
9		Showcase	6 days	Mon 17/05/21	Mon 24/05/21	7
10		Midpoint Requirements	12 days?	Mon 07/12/20	Tue 22/12/20	
11		Creating New Gantt Chart	1 day	Mon 07/12/20	Mon 07/12/20	
12		Dataset Selection	1 day	Tue 08/12/20	Tue 08/12/20	
13		Revise Project Proposal - Make necessary changes	1 day	Tue 08/12/20	Tue 08/12/20	
14		Analyse Challenging Aspects - give solutions	1 day	Wed 09/12/20	Wed 09/12/20	
15		Plan of How Data Will be Analysed	2 days	Wed 09/12/20	Thu 10/12/20	
16		Research Topics to be used	3 days	Thu 10/12/20	Sat 12/12/20	12,13
17		Statistics	1 day	Thu 10/12/20	Thu 10/12/20	
18		Visualisations	1 day	Thu 10/12/20	Thu 10/12/20	
19		Data Mining	1 day	Fri 11/12/20	Fri 11/12/20	
20		Machine Learning	1 day	Sat 12/12/20	Sat 12/12/20	
21		Research how technologies will be used	2 days	Sun 13/12/20	Mon 14/12/20	
22		R or Python? More research	1 day	Sun 13/12/20	Sun 13/12/20	
23		Rapid Miner	1 day	Sun 13/12/20	Sun 13/12/20	
24		SPSS	1 day	Mon 14/12/20	Mon 14/12/20	
25		Excel	1 day	Mon 14/12/20	Mon 14/12/20	
26		Tableau	1 day	Mon 14/12/20	Mon 14/12/20	
27		KDD - Midpoint	6 days	Tue 15/12/20	Tue 22/12/20	22
28		Data selection	1 day	Wed 09/12/20	Wed 09/12/20	12
29		Preprocessing	1 day	Wed 16/12/20	Wed 16/12/20	28
30		Transformation	1 day	Thu 17/12/20	Thu 17/12/20	29
31		Data Mining	1 day	Fri 18/12/20	Fri 18/12/20	30
32		Evaluation	2 days	Mon 21/12/20	Tue 22/12/20	31
33		Presentation	0 days	Tue 22/12/20	Tue 22/12/20	32
34		Remainder of Project	114 days	Wed 23/12/20	Sat 29/05/21	

The project plan was constructed in Microsoft Project to set an estimate of the time allocations to stick to, to achieve success of the analysis. The project plan was created at the start of the analysis and doesn't reflect the exact dates that everything was completed. Sticking to schedule was a big challenge faced in this as the R Studio programming environment was all a learning process. Some challenges were faced that caused a lot of back tracking in which time was underestimated. There was no updated plan that could be constructed when facing an issue as the completion time was of complete uncertainty.

When this plan was created there was no prior knowledge to what exactly would be included in this analysis as it was a completely new topic. With no background knowledge of a data analytics project, it was impossible to plan each and every step required.

Ongoing learning was a key area of this analysis and could fall under every single heading on the project plan Gantt chart. It was impossible to construct an accurate plan as to exactly what steps were to be performed, even after research. There were too many unknowns at the start which were evident during the study and tackled at the time issues arose. This was the best way to stay efficient as following the plan caused a lot of time spent off schedule completing the tasks.

Rapid Miner was a technology for data mining that was initially going to be used but the plan to use this was discarded as it wasn't needed for the study.

9.3 Monthly Reports

September
This month was spent brainstorming possible topics and ideas for the project. Areas of interest were explored here and potential datasets were kept noted. It was apparent that most things I am interested in such as music, fitness and gaming had a lack of good datasets. Data was the main exploration in September and just finding something that I would like answers to myself, this is when the idea of focusing on crime in the County that I'm from arose.

October
After realising that crime would be a potential area of focus for this project I spent time searching for crime datasets. I spent this month looking for a crime dataset that contained information that would give a good insight into the topic alone. I first explored the Irish data.gov website and found no potential datasets. After researching further I noticed a crime dataset on the Central Statistics Office. This was selected for the project proposal. The next step was to look into different factors that may influence crime. This is when two other datasets were discovered that would be included in the study. A project pitch video was recorded and submitted this month including the topic chosen.

November
The start of this month was spent preparing for the project proposal which was due at the start. The proposal was to determine if the project subject was a viable option, and outlined what objectives may be met from completing the project. After working on the proposal and awaiting confirmation that the project was valid, more research went underway. This research was to confirm that two datasets that were questionable for the project, were valid. After confirming these datasets would be used the technologies were the next step. Technologies were researched to find what would work best to meet the objectives of the project. The selected technologies at this point were R Studio, R, GitHub, SPSS, Excel and Rapid Miner.

December
This month was spent preparing for the midpoint submission which was due. Another large project was also due this month and was to do with R studio and the KDD which helped with the learning of R. The other project development had been started prior to this project so I gained a valuable background of knowledge as to what could and needed to be done in R Studio for the midpoint. A leaflet map visualisation was constructed for the midpoint alongside other visualisations.

January
This month was spent further researching other potential datasets that could supplement the main crime one. Time was spent searching for an education and income dataset that contained locations to improve the accuracy of the analysis however none were found. It was decided that the income and education dataset selected would be used for the rest of the analysis.

February
The KDD methodology was applied to the two supplementary datasets. Visualisations were completed and new data frames constructed in R which helped with the understanding of what could be shown with the use of the new datasets. It was decided to create visualisations that could be compared or combined with visualisations from the crime dataset such as point plots of the average mean household disposable income or percentage of educational attainment by type, which had year ranges and could be compared to crime to determine if there was any correlation between the results. It was evident at this point that accuracy of the datasets may not have the best accuracy, however they were the best datasets for the topics chosen that could be found.

March
This month was spent implementing a Shiny interactive dashboard which includes North and South Areas on a map, a data table below the map and a slider bar in which a year could be selected to produce data corresponding to the selected year. The Shiny dashboard creation came with many errors which slowed the analysis down. The reactive aspect of the dashboard caused a setback many times. By the end of this month the dashboard was half working and was not performing as expected. A long time was spent trying to resolve the error of this and the month finished with it still unresolved. A project profile was also created this month which included pictures and information about the project to be included in the project showcase in May.

April
The start of this month was spent trying to fix the error with the half working dashboard. After spending a lot of time the dashboard was fixed successfully and fully functioning. The rest of the month was spent creating new visualisations for all datasets to help produce a final conclusion. Final implementation and documentation was also started here. A project poster was made to give insight as to what the project is about and contained an abstract, project explanation, visualisations from the project and logos of the technologies used.

May
This month was spent preparing for the final submission. The project code was cleaned up removing stuff that wasn't needed for the final implementation and cleaning it up keeping relevant chunks together. This was done for 3 different R Scripts. The final documentation was also run over and the final version was completed in time for submission. Interpretation of statistical tests, visualisations and data mining were also done this month, to be added into the documentation. A project presentation was also done this month giving insight to what the project was about, why it was carried out and what it concluded.

9.4 Other materials used
No other materials were used for the implementation of this analysis.