Analysis of Crime Rates in Ireland from 2003 - 2015

Student Name	Student ID	Supervisor Name
Gavin O'Neill	D15127205	John McAuley

Table of Contents

1. Project Statement

The objectives of this project are to analyse crime rates in Ireland since 2003 and look at various socio-economic factors that may contribute to these rates. Data will need to be collected from a variety of sources and combined into one dataset for analysis. I will analyse the data at both national and regional (Dublin, North, South, East and West) levels to get a better insight into how crime rates may vary depending on the area. I need to use visualisation techniques to find correlations between variables which will give me an idea of the variables that will be important in the predication stage. The prediction stage will involve building various models with the aim of being able to predict future crime rates based on the performance of a specific variable(s).

In the beginning I would have guessed that a good economy with plentiful jobs would lead to lower levels of crime and vice versa, however, since doing a bit of reading on this topic I have found that in fact the opposite may be true. It can be argued that a wealthier economy and population will entice more criminal activity from gangs that hope to take a share of this extra wealth available. This is just one question I hope to be able to answer over the course of this project along with creating a predictive model.

2. Research

When one thinks about factors that influence crime rates, you automatically think of the main issues such as income and education levels, employment status, location, etc. In this project, I hope to dive further into the data and provide the reader with alternative factors that are not so apparent at first. A couple of studies with outcomes similar to what I'm looking for are "Why food prices were crucial in the Arab spring uprising" and "How the legalisation of abortion led to decreased crime rates". These studies looked past the most obvious influencing factors and found some other underlying contributions.

As a quick backdrop to the Arab spring uprising, it's important to know that the Middle East and north Africa depend more on imported food than anywhere else in the world and are therefore affected badly when world food prices rise. In 2007/08 food prices spiked, with some staple prices doubling in price. This led to bread riots in 2008 affecting Bahrain, Yemen, Jordan, Egypt and Morocco. Three years later, all suffered political uprising. (The Economist, 2012). Owing to this, I have included in my dataset financial indicators such as the Consumer Price Index (CPI), the CPI for Housing, Water, Gas, etc., average rental rates and foreign exchange rates.

The second study mentioned above, showed an inverse correlation between the availability of abortion and subsequent crime. The authors of this study first pointed out that males aged 18 to 24 are the group most likely to commit crime. They went on to say that that the legalisation of abortion in 1973 meant there was a reduction in unwanted babies being born. This led to reduced crime rates 18 years later, starting in 1992 and decreasing sharply in the following years. These would have been the peak crime committing years of the unborn children. (Levitt and Dubner, 2005). As you may already be aware, abortion is not legal in Ireland at this time so this is not a factor I can take into account, however it has inspired me to include other unconventional variables including the percentage of pupils that finish their leaving certificate, marriage rates, number of people practising religion, etc.

My plan going forward is to first use Tableau to visualise the collected data, displaying the variables present, how they interact and any correlations between them. I will then use R to build predictive models based around the dominant predictor variables discovered in the previous stage of analysis. The models will be evaluated and a champion model decided upon as a conclusion.

3. Approach and Methodology

For this interim submission, I have gathered and processed almost all data needed, allowing me to conduct a comprehensive exploratory analysis of the data using Tableau. At this point in time there are some variables that I would like to add to my dataset providing I can locate them in a usable format, however in the meantime I will begin with what I have. In order to get as much knowledge as possible on the factors influencing crime, I have created two datasets to explore at this stage. The reason for this is that a lot of the data is on a national level whereas I would like my analysis to be broken down into regions so as to give more depth to the study. My main data set includes these regions along with quarterly and yearly observations but my secondary dataset only contains national and yearly statistics. I have used both datasets for the exploratory visualisations that will follow in this report, however I will only use the main dataset when reporting on data quality as only this will be used for model building.

Figure 1 below displays the number of unemployed per year versus the number of offences. It was one of the first visualisations I created in Tableau as I was of the opinion that when unemployment went up or down crime rates would follow. Although the graph shows a correlation, it is not in fact the unemployment rate that predicts future crime rates but appears to be the opposite way around. You can see that the crime rates begin to increase a couple of years before unemployment and decrease a few years beforehand too. As I mentioned above, Figure 1 was produced using the national level dataset and I have followed this with Figure 2 to give you an idea of the further breakdown I can get when using the main dataset with regional dimensions. In Figure 2, I'm able to see how Dublin has a larger proportion of each variable and how there are more pronounced variations in the data.

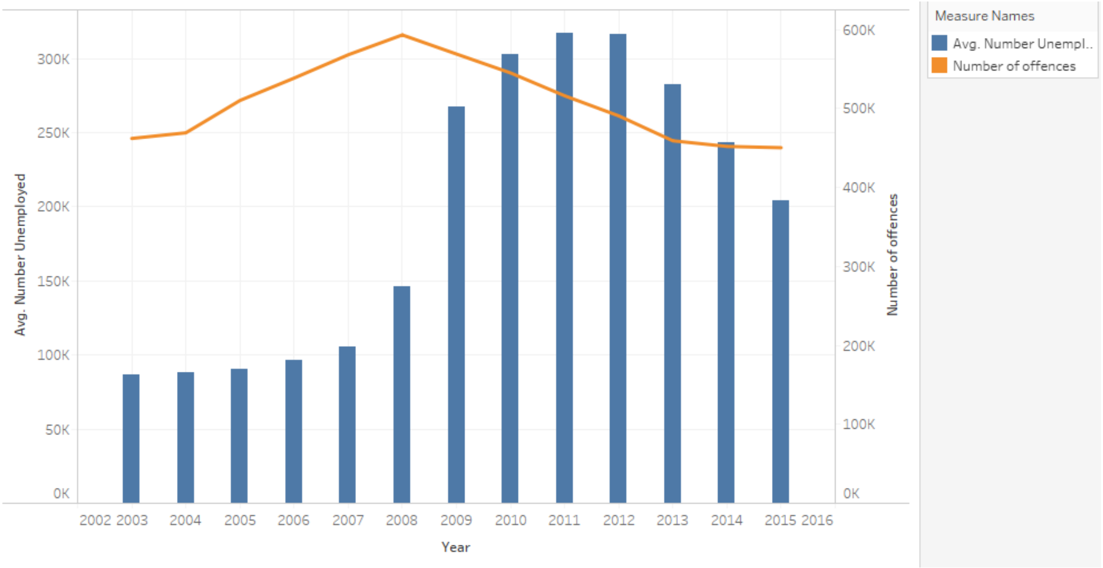


Figure 1

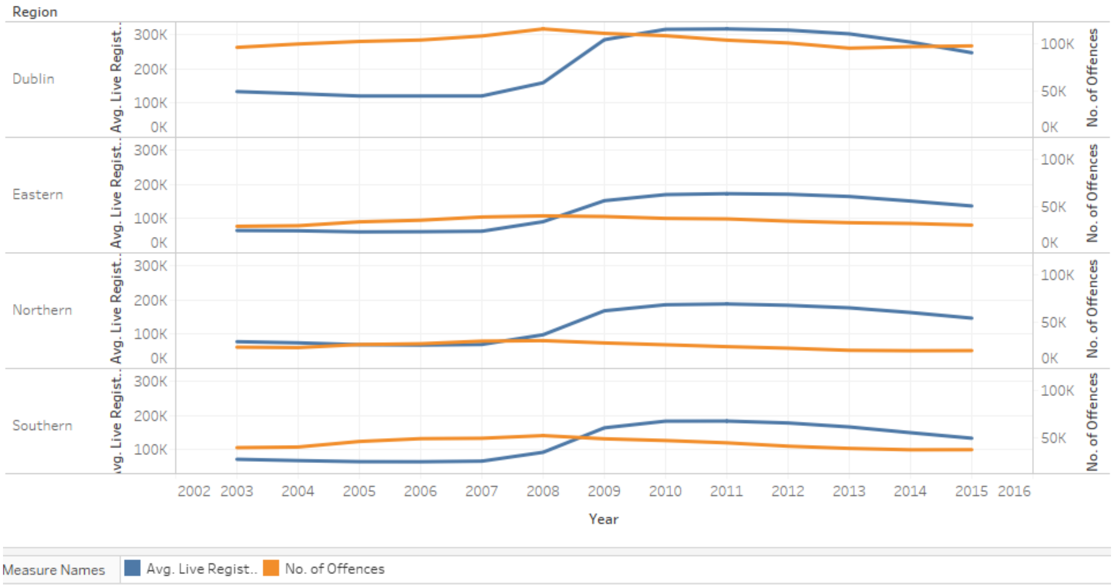


Figure 2

I produced some visualisations of the data including charts, graphs and scatter plots which has given me an overall view of the most obvious trends and influential variables. I have been able to see how crime rates vary depending on the region, time of year (broken down into quarters), etc. My findings in this stage have given me an indication of which variables are going to be important when I move on to creating a predictive model in the next stage. I'm targeting my analysis on socio-economic factors so I have collected data on items such as live register numbers, how many people practise religion, % of the population at risk of poverty, migration rates, education levels, rental rates, consumer price index figures, foreign exchange rates, number of police, etc.

Figures 3 & 4 below were generated as I wanted to get an overview of how financial indicators would correlate to crime rates. Both Avg. Rental Rates and the CPI appear to be very closely correlated to changes in the crime rate and I will definitely be looking into collecting more data in these areas as I believe they, or similar variables, could be important predictors in later models.

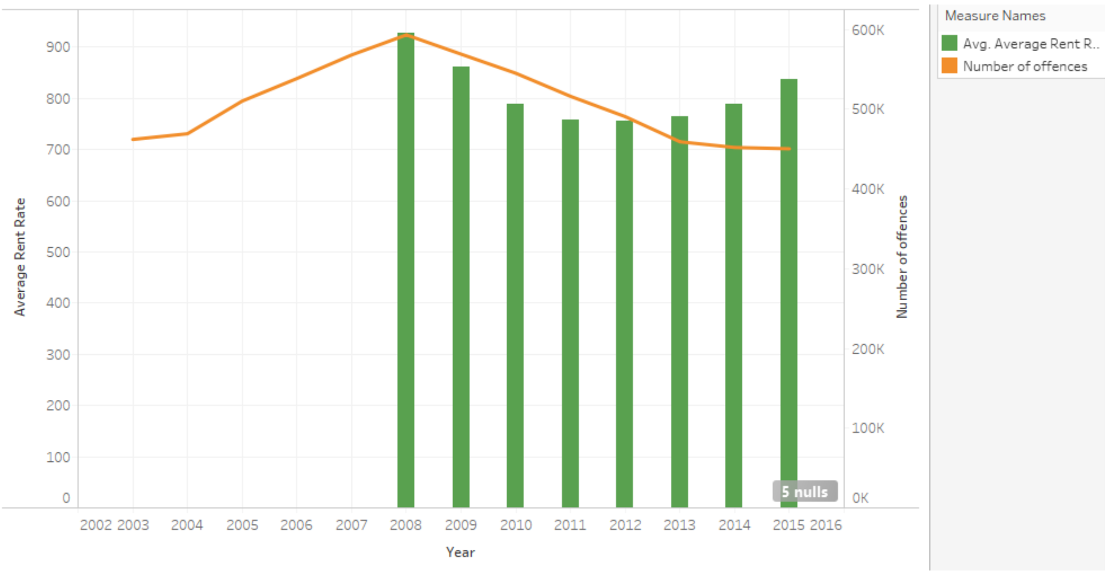


Figure 3

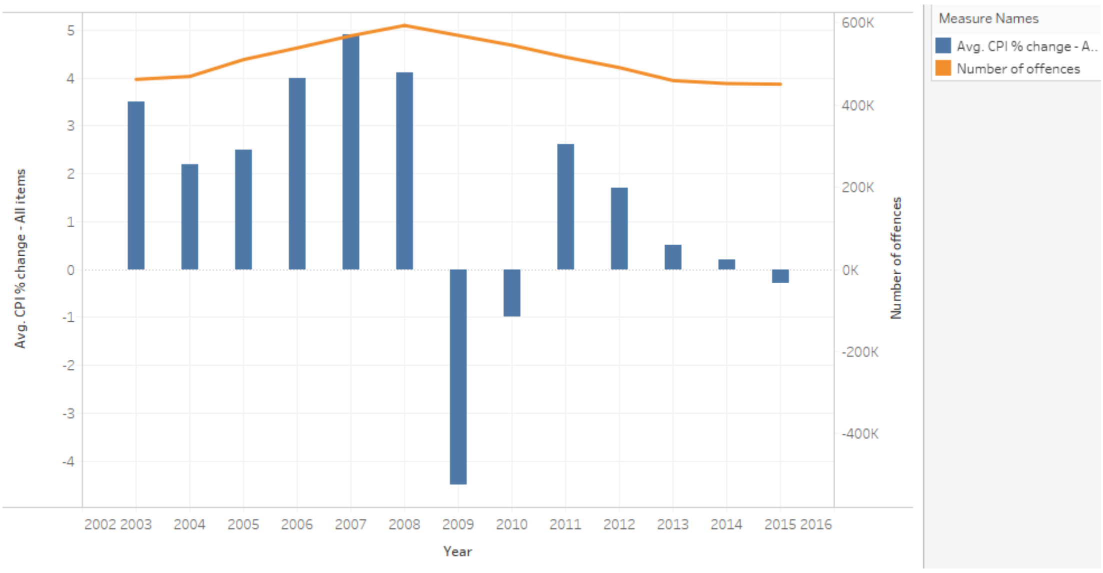


Figure 4

I generated Figure 5 below so that I could see a breakdown of the different crime types and how they have increased or decreased over time. Public order offences had a big decrease after 2008 when the recession just began which I would guess is due to people not going out to pubs as much and getting drunk less. There is also a significant drop in Kidnapping related offences which were a common activity for criminal gangs during the Celtic Tiger years. It is also interesting to note that while most crime types decreased after 2008, Theft and Burglary related incidents actually seem to increase slightly so perhaps petty crime is not affected by whatever influencing factors are involved.

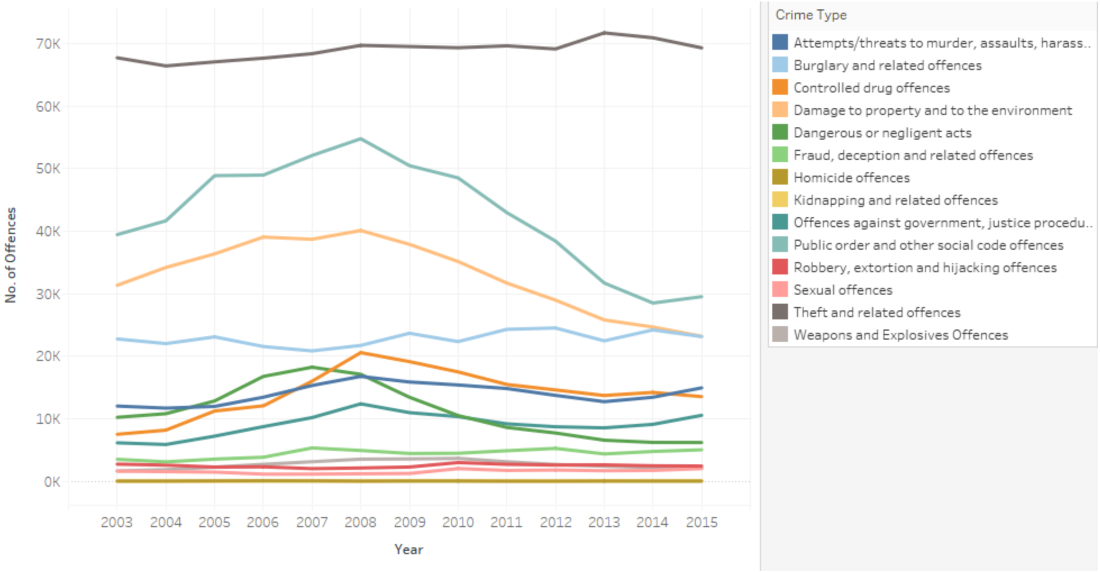


Figure 5

Figure 6 is a scatterplot that I need to use to be able to view multiple dimensions and measures on the same visualisation. I wanted to see how the percentage of pupils finishing the leaving cert affected the total number of each crime type committed and in each region. Crime types have been colour coded and regions marked with different shapes. You can see from the lowest horizontal line that Dublin has the lowest percentage of pupils finishing their leaving cert and in turn also have the highest crime rates which is signified by the points being more towards the right side of the graph. In this visualisation the points are pretty close together so it can be difficult to differentiate one from another but in general I think this is a very good way to explore multiple data features.

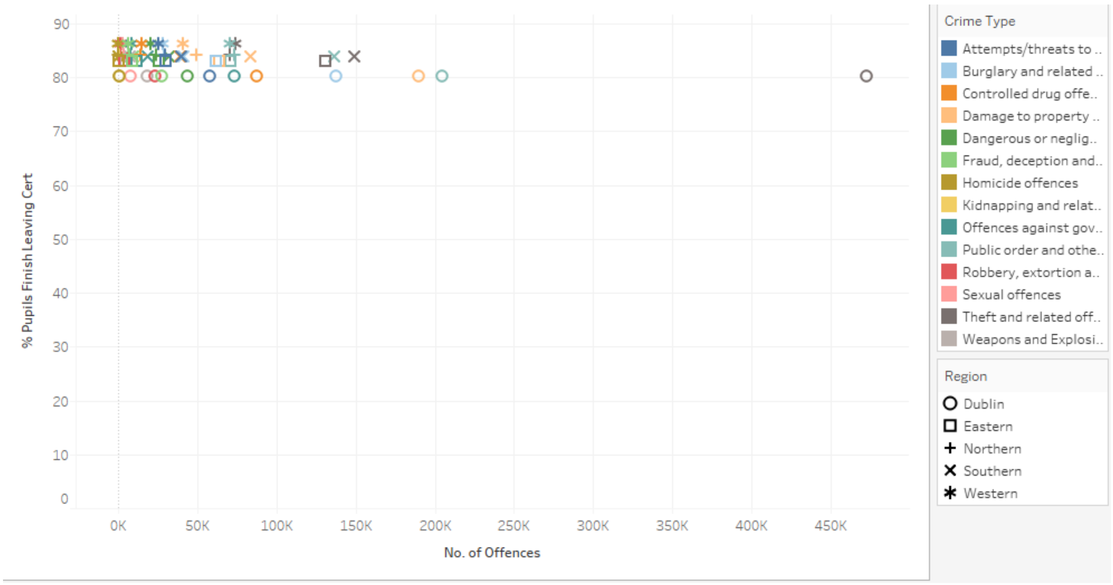


Figure 6

The final item I will discuss in this chapter relates to the use of a scatterplot matrix to see linear correlations between pairs of variables in the dataset. Figure 7 below is shows a segment of a scatterplot matrix I created using Tableau which was too large to fit completely here on the page. I can use this to discover trends between two specific variables which will help me when modelling the data later on. A positive trend will have plots moving from the bottom left area of the grid to the top right. I have included trend lines in this visualisation to help me to find any correlations. For example, I can see that along the top row, in the second and fourth grids there are positive relationships being displayed. Following the grids to the left and bottom, I then know that these relationships are between CPI –All Items & Avg. Rental Rate and CPI – All Items & CPI – Housing, Water, Gas, etc.

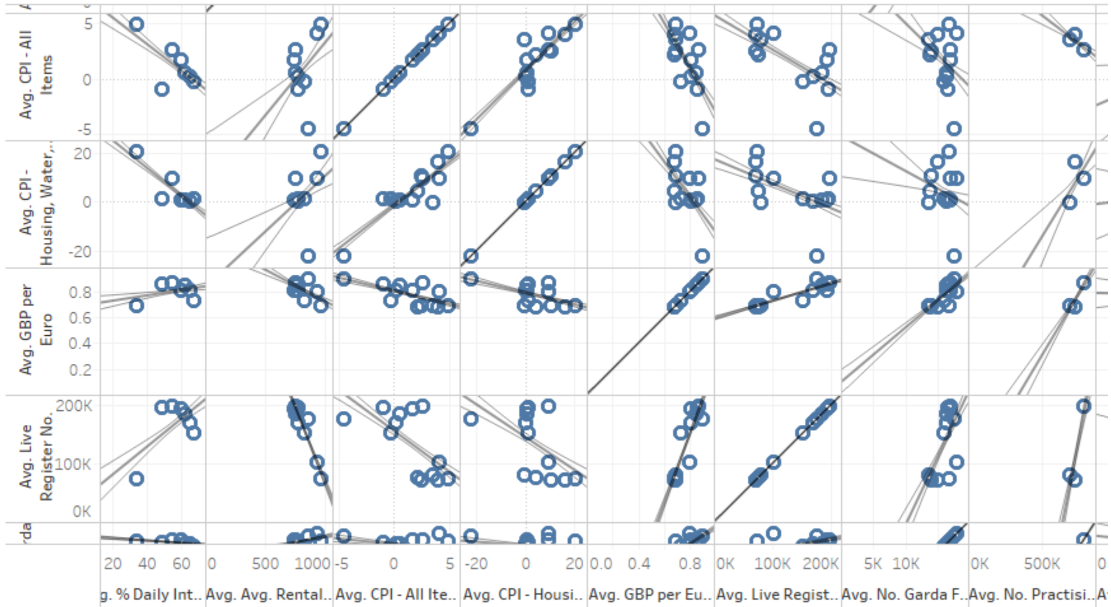


Figure 7

4. Data Profiling and Data Pipeline

The data used in this project has been collected so far from the websites for the Central Statistics Office (CSO) of Ireland, An Garda Siochana and the Department of Education. Their websites have open databases containing a variety of information that can be downloaded in various formats. For the purposes of this project, all data was downloaded as comma separated value (csv) files which can be formatted to function with both Tableau and R.

The data files didn't always lend themselves to merging without complications and a lot of time was spent on data cleaning and pre-processing. The data on the CSO website is not maintained in one common structure so different topics of information are often broken down into incompatible date structures (some yearly, some quarterly or monthly, etc.) or for only certain regions. A lot of the social related data has been collected from the national census which takes place every four or five years so there is missing data throughout the dataset for the years in between each census. After completing my exploratory analysis of the data I will decide which variables cannot be used and which will need to be imputed, using R, in order to allow a continuous analysis. Data that was available by region but only had yearly totals, was divided into quarterly values so that it would correspond with the other data in my Tableau stage of analysis.

To summarise data quality at this stage I have produced a tabular report (Table 1 below) for all of the continuous variables in the Analytics Base Table (ABT) and this is accompanied by visualisations that show the distribution of the values for each variable in the ABT. Each row, from left to right, has statistics for the number of instances (Count), the percentage missing, cardinality, minimum, 1st quartile, median, mean, 3rd quartile, maximum and standard deviation.

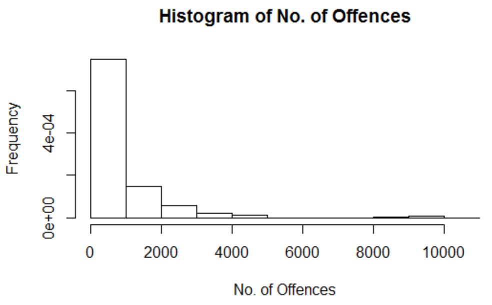
Continuous variables:

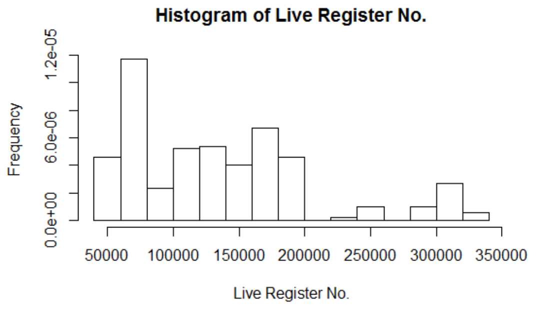
Table 1

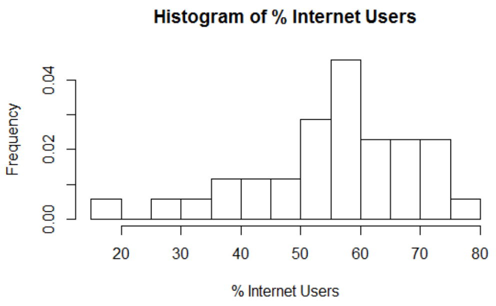
		%			1st			3rd		Std.
Feature	Count	Miss.	Card.	Min.	Qrt.	Median	Mean	Qrt.	Max.	Dev.
No. of Offences	3640	0		0	68	377	812.5	1011.2	10486	1342.481
Live Register No.	3640	0		43554	70195	122808	134658	172530	331979	71695.66
% Internet Users	3640	46.2		19	50	57	56.09	66	76	NA
No. Practising Religion	3640	76.9		246714	464296	784521	729906	1046088	1187176	NA
No. Garda Force	3640	0		13504	14969	16329	15936	16843	17799	1379.515
% Pupils Finish Leaving Cert	3640	46.2		76	81.1	82.7	83.47	85.6	91.9	NA
Sunshine Hours	3640	5.8		141.7	244	335.6	373.5	509	686.6	NA
Avg. Rental Rate	3640	36.5		436.6	740.9	788.1	813	937	1364.6	NA
CPI – All Items	3640	0		-4.5	0.2	2.2	1.569	3.5	4.9	2.4721
CPI – Housing, Gas, Water, Elec, etc.	3640	0		-22	0.6	1.5	4.008	9.7	20.4	9.8795
GBP per Euro	3640	0		0.6786	0.6843	0.7963	0.7712	0.8493	0.8909	0.07884
USD per Euro	3640			1.109	1.245	1.326	1.298	1.371	1.471	0.09891

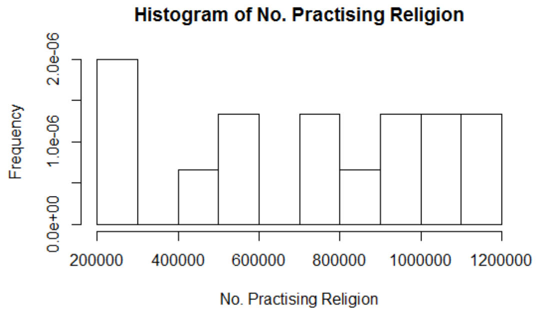
As you can see from the table above, there are some variables with a high percentage of missing values so going forward I need to decide whether it is worthwhile imputing values or if it would be better drop the variable from the ABT. Some of these values relate to features that we know don't generally vary much so imputation might not actually distort the data at all.

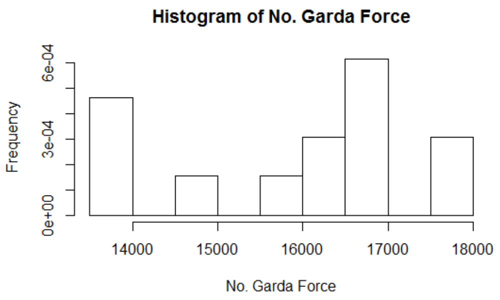
Following on from that I generated histograms of each variable to give a visualisation of the distribution that I'm going to be working with. A normal distribution is generally a bell shape and is seen to an extent below in the histograms of % pupils finishing leaving cert, CPI for Housing, Water, etc., % of Internet Users. Some of the histograms have peaks at their tails or appear as a plateau or have a skewed right distribution such as the histograms of No. of Offences and Live Register No. At this point I will need to consult with my supervisor for advice on what steps to take next such as determining best-fit distribution or if I must consider a normalising transformation.

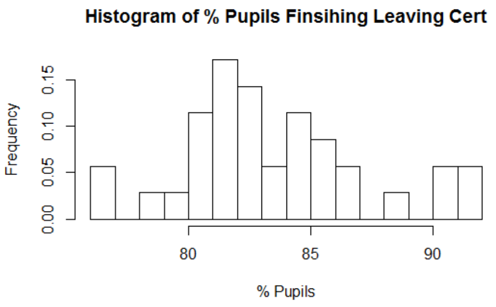

Histogram of No. of Offences
Frequency
0e+00
4e-04
0
2000
4000
6000
8000
10000
No. of Offences

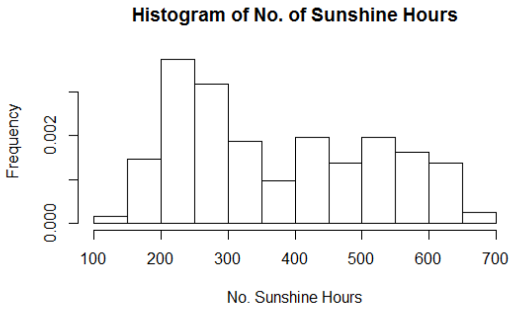

Histogram of Live Register No.
Frequency
0.0e+00
6.0e-06
1.2e-05
50000
100000
150000
200000
250000
300000
350000
Live Register No.

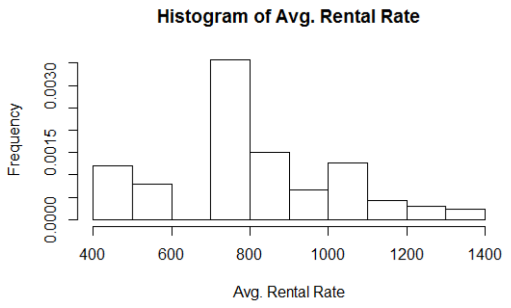

Histogram of % Internet Users
Frequency
0.00
0.02
0.04
20
30
40
50
60
70
80
% Internet Users

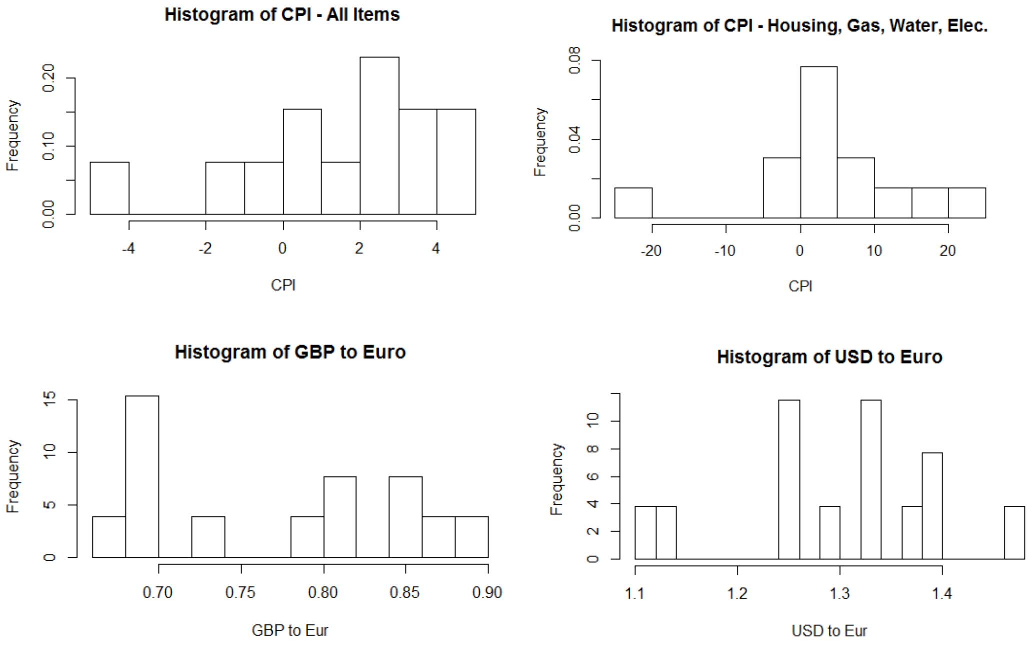

Histogram of No. Practising Religion
Frequency
0.0e+00
1.0e-06
2.0e-06
200000
400000
600000
800000
1000000
1200000
No. Practising Religion


Histogram of No. Garda Force
Frequency
0e+00
3e-04
6e-04
14000
15000
16000
17000
18000
No. Garda Force


Histogram of % Pupils Finsihing Leaving Cert
Frequency
0.00
0.05
0.10
0.15
80
85
90
% Pupils


Histogram of No. of Sunshine Hours
Frequency
0.000
0.002
100
200
300
400
500
600
700
No. Sunshine Hours


Histogram of Avg. Rental Rate
Frequency
0.0000
0.0015
0.0030
400
600
800
1000
1200
1400
Avg. Rental Rate



5. Evaluation

The end goal of this project is to build a predictive model, in this case using linear regression. Linear regression works quite simply by fitting a predictive model to a set of x and y values. After developing the model, if presented with another value for x without the accompanying y value, the fitted model should be able to make a prediction of the value of y. This is a simple explanation for the task ahead and realistically I will be using multi linear regression because I have multiple predictors to take into account. Going forward I will discuss with my project supervisor about what models I should develop that should be able to give me the best predictive results. After each model has been tested and scored, I will finally be able to select the champion model.

6. Issues and Risks

One of the main issues facing me within this project is having time to do enough research, data collecting and processing, analysis, etc., while also juggling other coursework and outside work. I have had this issue in the past, specifically in the first semester of this course and it is feasible to get everything done, however my time management will need to be very efficient. I normally make a plan each week of what I have coming up next and then allocate certain days and times for project work as I see fit. The second issue has been a lack of knowledge of what I would eventually learn with regard to Tableau and R so it was difficult to make a plan in the beginning of what to do and when.

It also had an effect on research as I didn't initially have any knowledge of predictive modelling or even general data analysis. This has meant that I have been getting to understand how to do the project as the weeks have passed and from this point onwards I feel more equipped to carry out my tasks.

The issue that has cropped up more than most is actually the data collection because almost all the files for each variable have to be downloaded separately, formatted for use in the main dataset (mostly calculations by hand) and copied in. I really underestimated the time that this would take in the beginning and it has taken hours upon hours to get a decent sized dataset put together. I don't think there is anything I could have done to avoid it but I will know to look for larger initial datasets in future projects.

7. Future Work

For the last two weeks in November I plan to start analysing the dataset using R with the aim of building a predictive model. While doing my initial visualisations of the data I have already realised that economic factors are going to be more influential than their social counterparts and because of this I would like to add some more financial related variables. I'm giving myself a cut-off time of the end of November for any alterations to my dataset so as to allow enough time for analysis during December.

Time has been very tight so far with every week having either an assignment or exam, while doing my work placement part-time in the evenings and weekends but I think that having an exam week in December with no exams will be beneficial especially as all other assignments will already be finished. Ultimately, I plan to complete my analysis and modelling by this time in mid December giving me ample time to finish off writing the report before the January 12th due date.

	September	October	November	December	January
Proposal	🟦				
Research	🟦	🟦			
Collect Initial Data		🟦	🟦		
Format		🟦	🟦		
Visualise & Investigate		🟦	🟦		
Submit Interim Report			🟦		
Finalise Data Collection			🟦		
Process for Analysis			🟦	🟦	
Build Predictive Model				🟦	
Complete Final Report				🟦	🟦

Bibliography

The Economist. (2012). Let them eat baklava. [online] Available at: http://www.economist.com/node/21550328. [Accessed 7 November 2016].

Levitt, S. and Dubner, S. (2005). Freakonomics.