DUBLIN INSTITUTE of TECHNOLOGY
Institiúid Teicneolaíochta Bhaile Átha Cliath

An Analysis of Crime Rates in Ireland and Prediction based on Social & Economic Factors

Project Report
DT265
Higher Diploma in Computing

Author: Gavin O'Neill
Supervisor: John McAuley

School of Computing
Dublin Institute of Technology

January 2017

School of Computing
Scoil na Ríomhaireachta

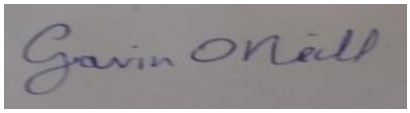
Abstract

This project was undertaken to provide an analysis of the crime rate in Ireland and explore its relationships with a variety of social and economic factors. The end goal is to see if these factors can be used to build a model, using the historic data available, to predict the crime rate going forward. Crime is a burden, financially and otherwise, to every society in the world and it's important that an effort is made to fight it in any way possible. An accurate predictive model would not only allow law enforcement organisations to properly prepare for the future, but would also give an indication of the most influential causes and present society with an opportunity to be proactive, instead of reactive, with this problem.

Declaration

I hereby declare that the work described in this dissertation is, except where otherwise stated, entirely my own work and has not been submitted as an exercise for a degree at this or any other university.

Signed:



_____

Gavin O'Neill

16th January 2017

Acknowledgements

I would like to thank my supervisor, John McAuley, for his help and guidance from the very beginning of this project. I would also like to express my gratitude to my lecturers in every subject, for passing on their knowledge without which I would not have been able to complete this work. Of course, my classmates have been great for helping and sharing what little we each know at this stage, and finally my family, friends and girlfriend for their support throughout.

Table of Contents

1. Introduction

1.1 Project Overview and Background

Crime has been a problem for every society for thousands of years. It affects victims directly, and others non-directly through the fear of becoming a victim, and trying to police crime bears a huge financial cost on everyone. One of the best ways to prevent crime is to figure out what factors cause it and then fight these factors. With that in mind this project looks at a range of social and economic factors to analyse which ones affect the crime rate and how this information can be used to predict future trends.

This project looks at the crime rate in Ireland since the turn of the millennium with the aim of building a predictive model based on social and economic factors. Prior to the commencement of this project I had originally thought about using cybercrime as the focus of this study, however, due to a lack of freely available data in that subject area I had to change course. As not stray too far from my original path, I decided upon crime in Ireland as my subject and the fact that it's where I'm currently living gives the study an additional interest to me.

For as long as there has been social order with laws and rules, there has been people willing break these rules and carry out criminal acts. This has in turn led to innumerous studies on crime and the many factors that contribute to it such as religion, politics, psychological, socioeconomic, geographical and others. Whilst this generally the case, when doing some background reading on this topic I couldn't find many studies on the links between crime rates and socioeconomic factors in Ireland specifically so it seemed like the perfect area to investigate.

The data used in this project is made up of crime statistics for Ireland from 2003 until 2016, which is broken down quarterly and by region thus allowing me to analyse trends over time and geographically. I hope to build an accurate model for predicting crime rates using multiple linear regression.

Initially data was gathered from online sources, pre-processed with Excel and an exploratory analysis was performed, using Tableau for visual representations of my findings. I then moved on to the next stage where I used the R language to do some further analysis and build the prediction model.

1.2 Project Objectives

The objective of the project is to study crime and socioeconomic data to find out if there are any correlations between the two. I want to see this on a national level primarily and break it down regionally if possible. One or more predictive models for crime rates will be built using multiple linear regression and to do this I will first need to learn more about the socioeconomic data such as any relationships the various factors have with one another. I hope to find out if an accurate predictive model can be built using socioeconomic data and see if this

information could be used viably in real decisions, like for the use of social programs to prevent crime or staffing number for law enforcement agencies.

1.3 Project Challenges

The first and biggest issue I faced during this project was the gathering of data. The crime data itself was not an issue as it is maintained by the Central Statistics Office (CSO) on behalf of An Garda Siochana and it was relatively easy to locate and download in csv file format from the CSO's website. The social and economic data, on the other hand, was a more difficult task that required basically every factor to be downloaded separately, pre-processed in some way and amalgamated into my main data file. There is no central source for a variety of data like this so a lot of time was spent exploring different sources to find what was needed.

Time and workload management was also difficult to balance because along with this project I had to complete weekly assignments for other modules, exams and a part-time job. Effective management in this area however, is a skill that will be important throughout my career and this has prepared me well for the future.

The final aspect that I found challenging in relation to this project was the fact that I had no knowledge of Data Analytics prior to beginning and it was difficult to make a proper plan regarding what I would need to do at each stage or even how to use the various languages or pieces of software needed to achieve my goals. In this way, the project became a tool to apply new knowledge as it was learned and I had the assistance of my teachers, supervisor and my own research to help along the way.

2. Background Research

2.1 A Look at Similar Studies

When I settled on crime as the topic for this report I decided that I needed to do some reading into the subject and the first thing that came to my mind was a study done in a book called Freakonomics (Levitt and Dubner, 2005). The authors carried out a study that showed an inverse correlation between the availability of abortion and crime rates. It was first pointed out by the authors that males aged 18 to 24 years old are the group most likely to commit crimes. They then went on to say that the legalisation of abortion in 1973 meant there was a reduction of unwanted babies being born. As a result, this led to reduced crime rates 18 years later, beginning in 1992 and decreasing sharply in the following years. Although abortion legalisation is not a factor I can consider for Ireland, it led me to include other unconventional variables in this study including the percentage of pupils that complete their leaving certificate, marriage rates, the number of people practising religion, etc.

Another interesting theory I read about was "Why food prices were crucial in the Arab Spring uprising". 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. These societies were not able to recover stability after the riots and 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., rent rates and foreign exchange rates.

Whilst looking for studies for predicting crime rates, I didn't come across any that were looking to forecast crime in the long-term but I did find quite a lot of information about how law enforcement agencies around the world are beginning to use the big data at their disposal for real-time crime prediction. An online article from Science magazine titled "Can 'predictive policing' prevent crime before it happens?" detailed the history of predictive policing from the early twentieth century until today. It states that scientists have been using statistics and geospatial analysis to determine crime risk levels since 1931. Then in the 1990s, various institutions began to embrace geographic information systems for mapping crime data and researchers used everything from basic regression analysis to complex mathematical models to try to forecast when and where the next crime wave would appear. It has only been in recent years however that computing power and storage has allowed them to build effective real-time predictive software. One such software package is PredPol, used by more than 60 police departments in the United States, which uses a proprietary algorithm that takes crime stats from the immediate and longer-term past to predict crime hotspots during a police officer's shift, and can be monitored on the laptop inside their car (Science | AAAS, 2016). This is just one of many different systems being developed and it's interesting to see what use can be made of the massive data that is held by law enforcement organisations worldwide.

2.2 An Overview of the Technologies Used

The technologies used in this project were Microsoft Excel, Tableau and the programming language R, via the RStudio application. Excel is used at the initial stage; all data was gathered in .csv file format and pre-processed here. I used Tableau at the exploratory stage of the analysis to produce visualisations of the data at my disposal. It allowed me to see trends in the data and relationships between the variables. It was an important step in the process as it gave me a better idea of which factors were going to be important to building the final prediction model and helped me when deciding whether or not it would okay to drop some variables if they were missing some amounts of data. Finally, R and RStudio were used for a final analysis of the data and building a predictive model. I used it to check the distributions of the variables, transform them if necessary, impute missing values if necessary, drop unwanted variables and

other data processing methods. It was also used to train and test the multiple linear regression models.

3. Data

3.1 Overview of the Data Sources and the Data Acquisition Approaches

The data used in this project was collected from a variety of online sources including the websites for the Central Statistics Office (CSO), An Garda Siochana, the Health Service Executive (HSE), Eurostat and the Department of Education. Their websites have open databases containing statistics that can be downloaded in many different formats. For the purposes of this project, all data was downloaded as comma separated value (csv) files that easily integrate with both Tableau and RStudio.

3.2 Data Preparation and Integration

The final dataset used consisted of 26 variables, 4 of which were dimensional values: Crime Type, Region, Year, Quarter; and 22 measurable values. The 4 dimensional values and the measure for Number of Offences came together in one crime statistics file, while the other 21 variables in the dataset were downloaded separately and added in. 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 isn't maintained in one common form and different pieces of information are often stored in incompatible date structures (some yearly, some quarterly or monthly, etc.) or only for certain regions. A lot of the social data has been collected from the national census which takes place every four or five years and there were missing records throughout the dataset for the years in between each census. Data that was available by region, but only had yearly totals, was divided into quarterly values that would correspond with the rest of the dataset.

To summarise data quality at this stage I have produced a tabular report (Table 1 below) for all 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:

Feature	Count	% Miss.	Min.	1st Qrt.	Median	Mean	3rd Qrt.	Max.	Std. Dev.
No. of Offences	3640	0	0	68	377	812.5	1011.2	10486	1.342481
Population	3640	0	439600	610700	820800	854745	1158800	1305300	2.869729
Live Register No.	3640	0	43554	70195	122808	134658	172530	331979	7.169566
Mean Disposable	3640	15.4	30186	37721	41512	45015	53519	61496	9.102584

Income									
Avg. Rental Rate	3640	30.8	343.1	740.8	791.7	821.7	958.1	1364.6	2.270835
CPI – All Items	3640	0	-4.5	0.2	2.2	1.569	3.5	4.9	2.472145
CPI – Housing, Gas, Water, Elec, etc.	3640	0	-22	0.6	1.5	4.008	9.7	20.4	9.879496
GBP per Euro	3640	0	0.6786	0.6843	0.7963	0.7712	0.8493	0.8909	8.144182
USD per Euro	3640	0	1.109	1.245	1.326	1.298	1.371	1.471	1.102810
Sunshine Hours	3640	0	141.7	242.6	344.6	373.7	509.2	686.6	1.475600
Rainfall (mm)	3640	0	97.7	197.8	247.2	258.1	306.3	626	9.047345
At Risk of Poverty Rate	3640	15.4	8.3	12.1	14.9	16.44	19.9	30.2	5.582120
Consistent Poverty Rate	3640	15.4	2.3	4.6	5.7	6.585	8.3	18.5	3.118047
% Daily Internet Users	3640	30.8	19	44	54	52.71	63	76	1.395128
% Pupils Finish Leaving Cert	3640	46.1	76	81.1	82.7	83.47	85.6	91.9	3.770490
Number Practising Religion	3640	76.9	246714	464296	784521	729906	1046088	1187176	3.112362
Gardaí Numbers	3640	0	13504	14969	16329	15936	16843	17799	1.379515
Marriages Number	3640	0	2571	4020	4940	5317	6058	8949	1.811262
Tobacco (gm) per Capita	3640	7.7	982	1210	1400	1442	1723	2076	3.205449
Alcohol (ltr) per Capita	3640	7.7	10.6	11.38	11.95	12.22	13.4	13.6	1.082693
% Population in Good Health	3640	15.4	82	82.7	82.8	83.02	83.20	84.30	6.221477

Table 1 – Data Quality Report

As you can see from the table above, there are 10 variables with more than 5 percent of their values missing and this is not a good position to be in because there cannot be any missing values when building a regression model. Going forward, a decision needs to be made about whether it is worthwhile imputing values or if it would be better drop some variables from the ABT. What I decided upon in the end was to cut some years from my dataset as it was visibly clear that certain years were responsible for much of the missing values. The resulting dataset covers the period 2007 to 2014, cut from 2003 to 2015, and now there are only two variables with missing values. The variables in question are '% Pupils Finish Leaving Cert' and 'Number Practising Religion' with 62.5 and 87.5 percent missing values respectively. These amounts of missing values are too large for accurate imputation and these variables are not expected to be influential in my predictive model so they will be dropped, using R, before modelling.

3.3 Final Dataset

The dataset to be used when beginning analysis consisted of the following variables:

- Region
- Crime_Type
- Year
- Qtr
- No_of_Offences
- Population_Num
- Live_Register_Num
- Mean_Disposable_Income
- Rent_Rates
- CPI_All_Items
- CPI_Water_Electricity_Gas
- GBP_per_Euro
- USD_per_Euro
- Sunshine_Hours
- Rainfall(mm)
- At_Risk_of_Poverty_Rate
- Consistent_Poverty_Rate
- Daily_Internet_Users
- Pupils_Finish_Leaving_Cert
- Num_Practising_Religion
- Gardai_Num
- Marriages_Num
- Government_Party
- Tobacco_gm_per_capita
- Alcohol_ltr_per_capita
- Pop_In_Good_Health

4. Analysis and Modelling

4.1 Exploratory Analysis

At this stage I conducted a comprehensive exploratory analysis of the data using Tableau. Here I had my first insight into the data, allowing me to understand what kinds of relationships I was dealing with and giving me an idea of the most likely important variables for my model.

Figure 1 below, displays the number of unemployed per year versus the number of offences. It was the first visualisation I created in Tableau as I suspected that unemployment would have a positive correlation with crime rates. Although the graph does show a correlation, it's not in fact the unemployment rate that acts as a precursor to future crime rates but appears to be the opposite way around. We can see that the crime rates begin to increase a couple of years before unemployment and decrease a few years beforehand too.

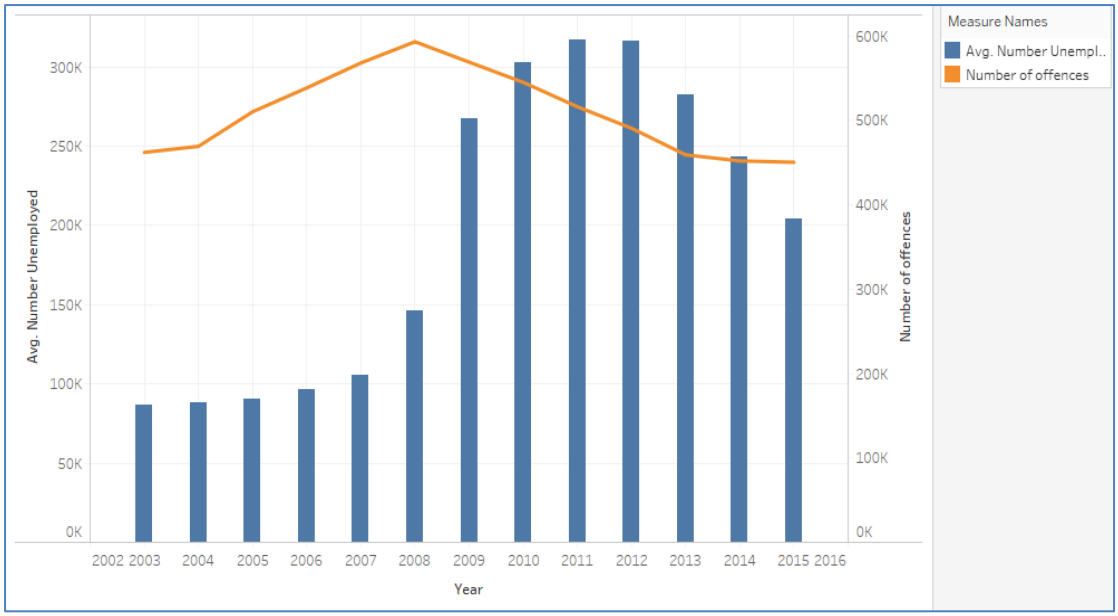


Figure 1 – Number_Unemployed vs No_of_Offences

In Figure 2, I've broken the same statistics down into regions. We can see that Dublin has a larger amount of both variables and how there are more pronounced variations in the data.

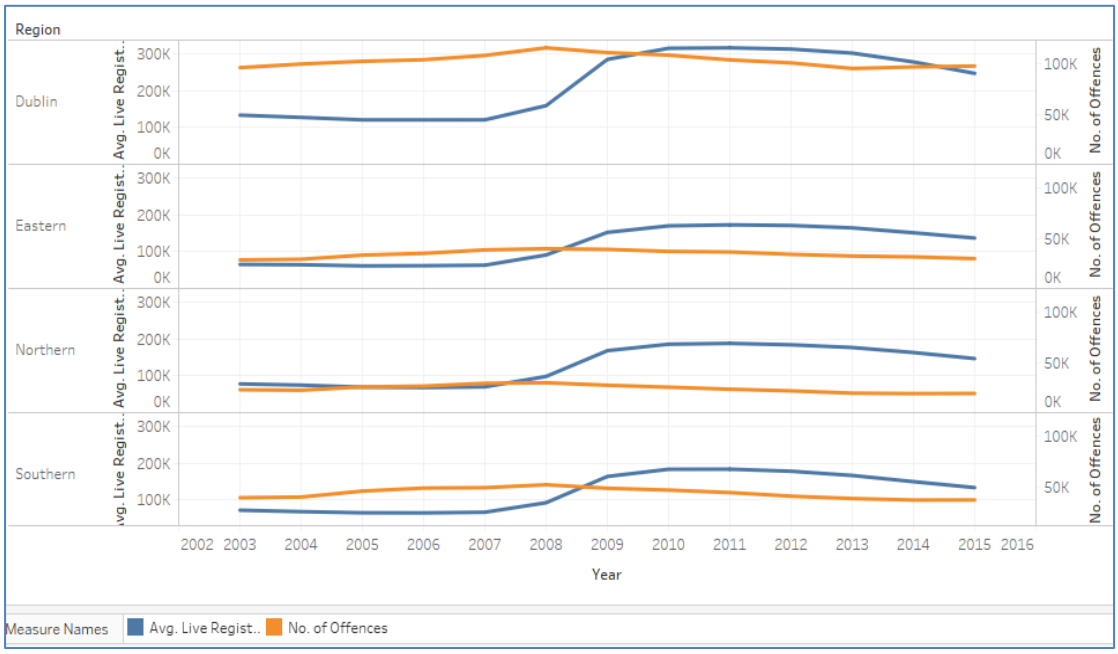


Figure 2 – Live_Register_Num vs No_of_Offences for each region

Figures 3 & 4 below, were generated to give an overview of how financial indicators would correlate with crime rates. Both Rent Rates (Figure 3) and the Mean Disposable Income (Figure 4) appear to be very closely correlated to changes in the crime rate and I believe they, and similar financial variables, could be important predictors in later models.

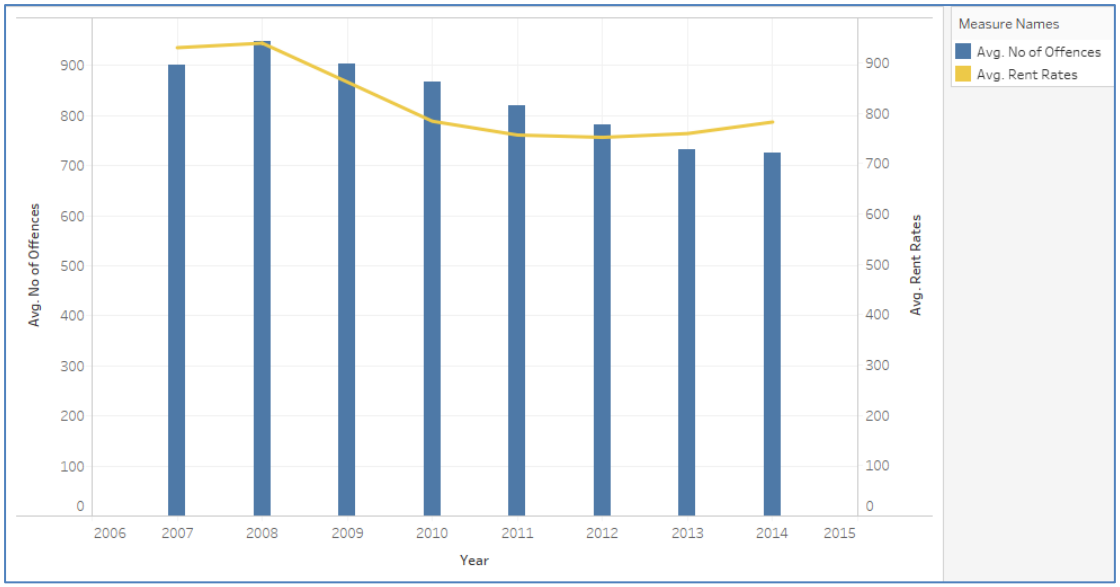


Figure 3 – Rent_Rates vs No_of_Offences

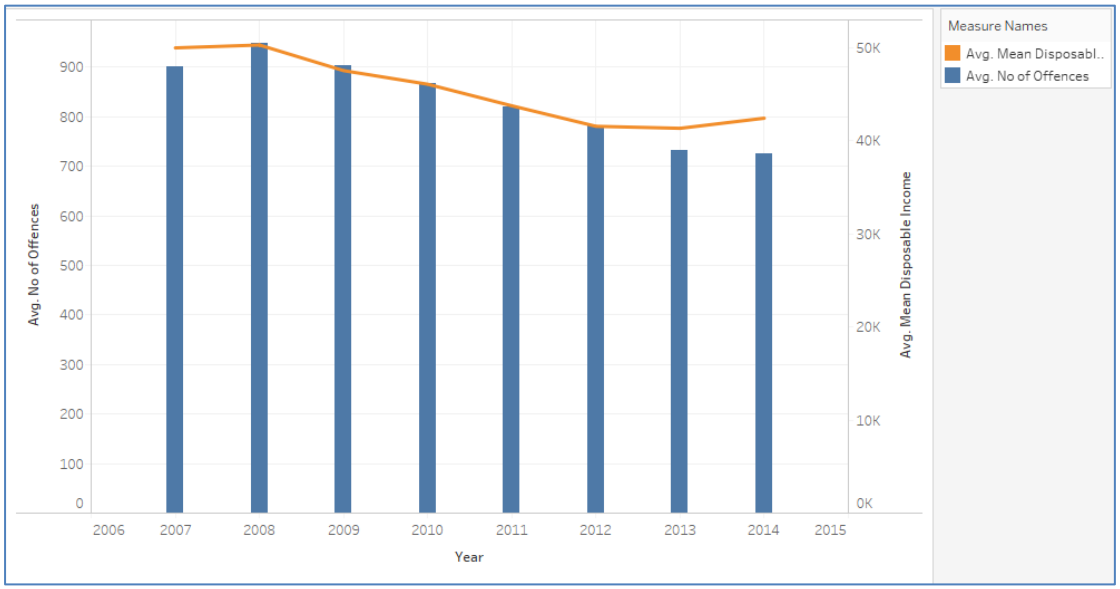


Figure 4 – Mean_Disposable_Income vs No_of_Offences

Next, Figure 5 visualises 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 had just begun and I would speculate this is due to people not going out to pubs getting drunk as much. There is also a significant drop in kidnapping and 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. Perhaps then, petty crime is not affected by the influencing factors in the same way as the other crime types.

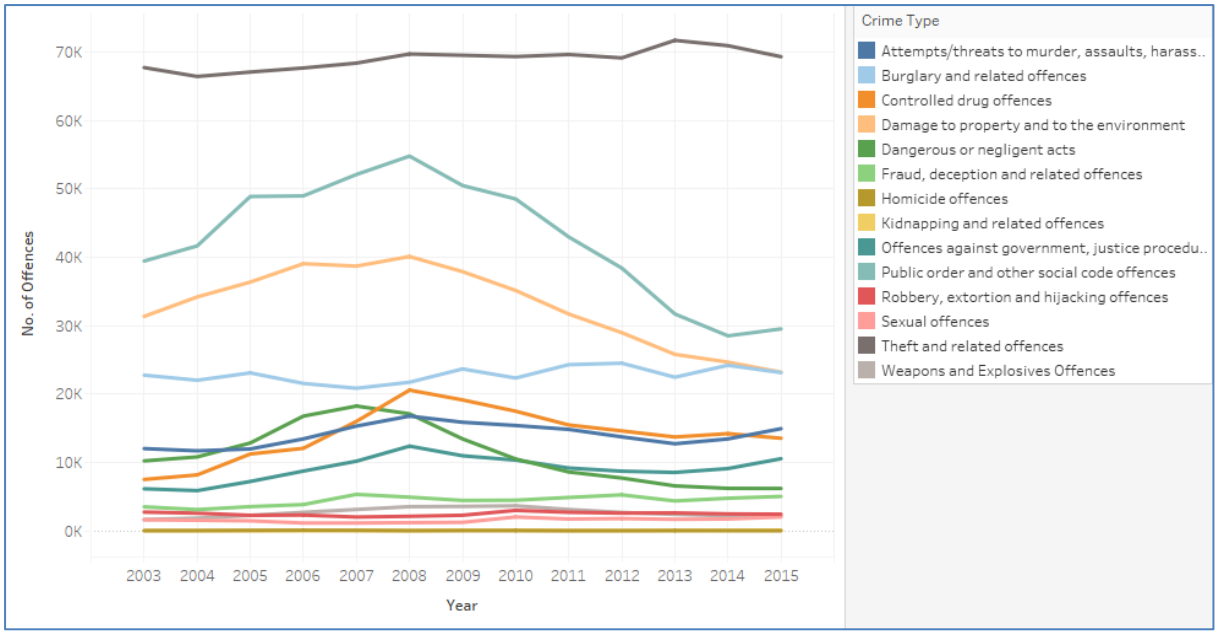


Figure 5 – number and type of offence over time

Figure 6 is a scatterplot that was used to view multiple dimensions and measures on a single visualisation. I wanted to see how the percentage of pupils finishing the leaving cert affected the total amount of each crime type committed and in each region. Crime types have been colour coded and regions marked with different shapes. You can see by the row of circles along the bottom of the cluster that Dublin has the lowest percentage of pupils finishing their leaving cert and in turn, also has the highest crime rates which is signified by the circles being further to the right side of the graph. In this visualisation, all the data points are close together and it can be difficult to differentiate one from another but in general I think this is an interesting way to explore multiple data features.

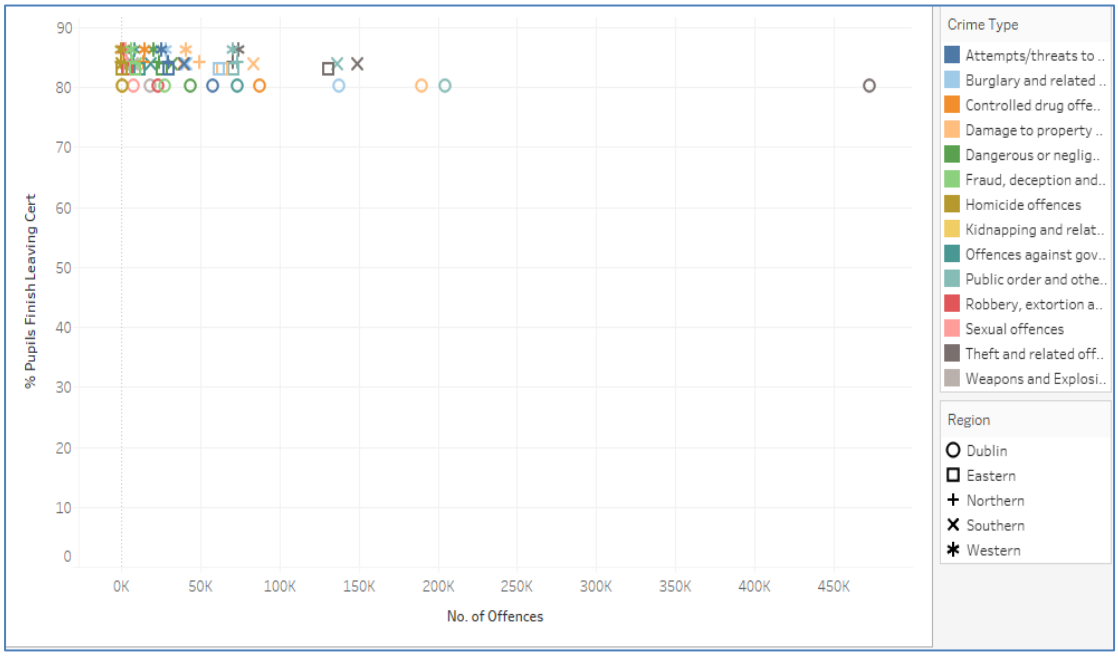


Figure 6 - exploratory scatterplot

Finally, a scatterplot matrix was used to see linear correlations between pairs of variables in the dataset. Figure 7 shows a segment of a scatterplot matrix created using Tableau, which was too large to fit on the page below. The matrix allows us to discover trends between two specific variables and will help when modelling the data later. A positive trend will have plots moving from the bottom left area of the grid to the top right. Trend lines have been included in this visualisation to assist with spotting the correlations. Two or more highly correlated variables have what is called multicollinearity. It is redundant to have more than one of these variables in our model and some will need to be removed. For example, we 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, we know that these relationships are between (1) CPI - All Items and Avg. Rental Rate; (2) CPI - All Items and CPI - Housing, Water, Gas, etc. One of these variables will be dropped from the dataset before modelling.

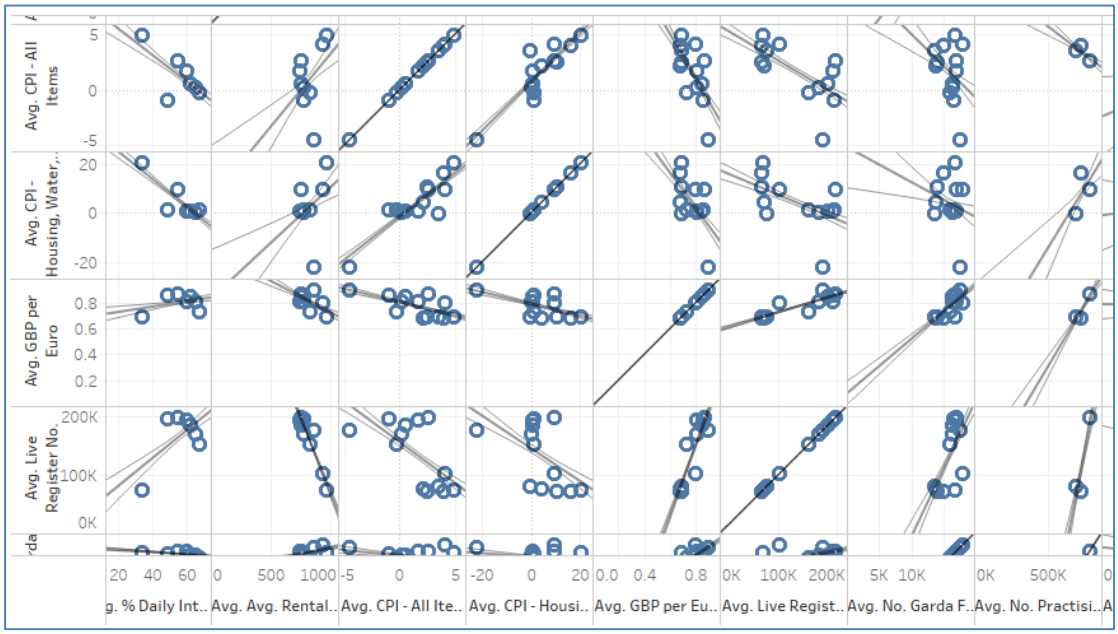
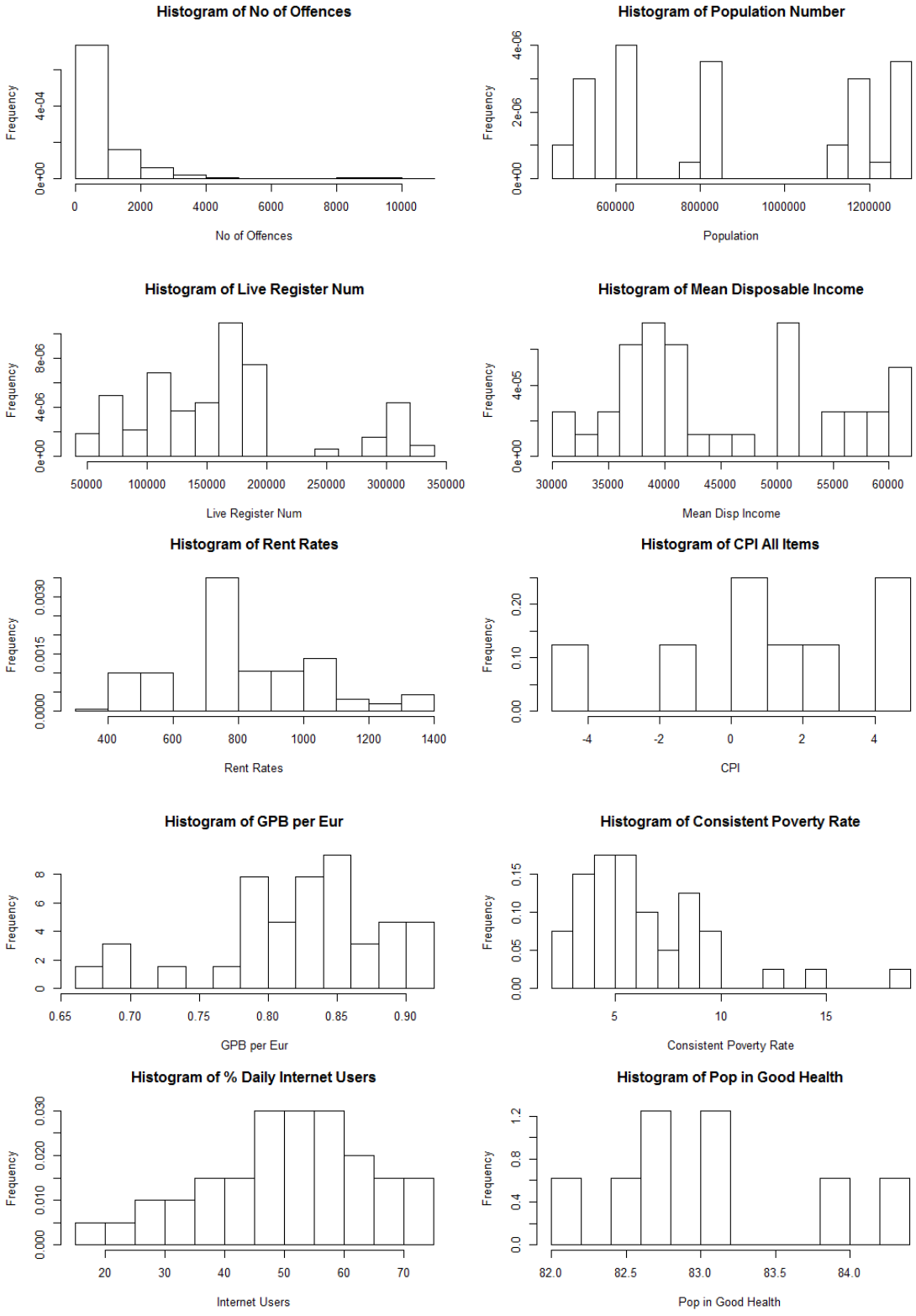
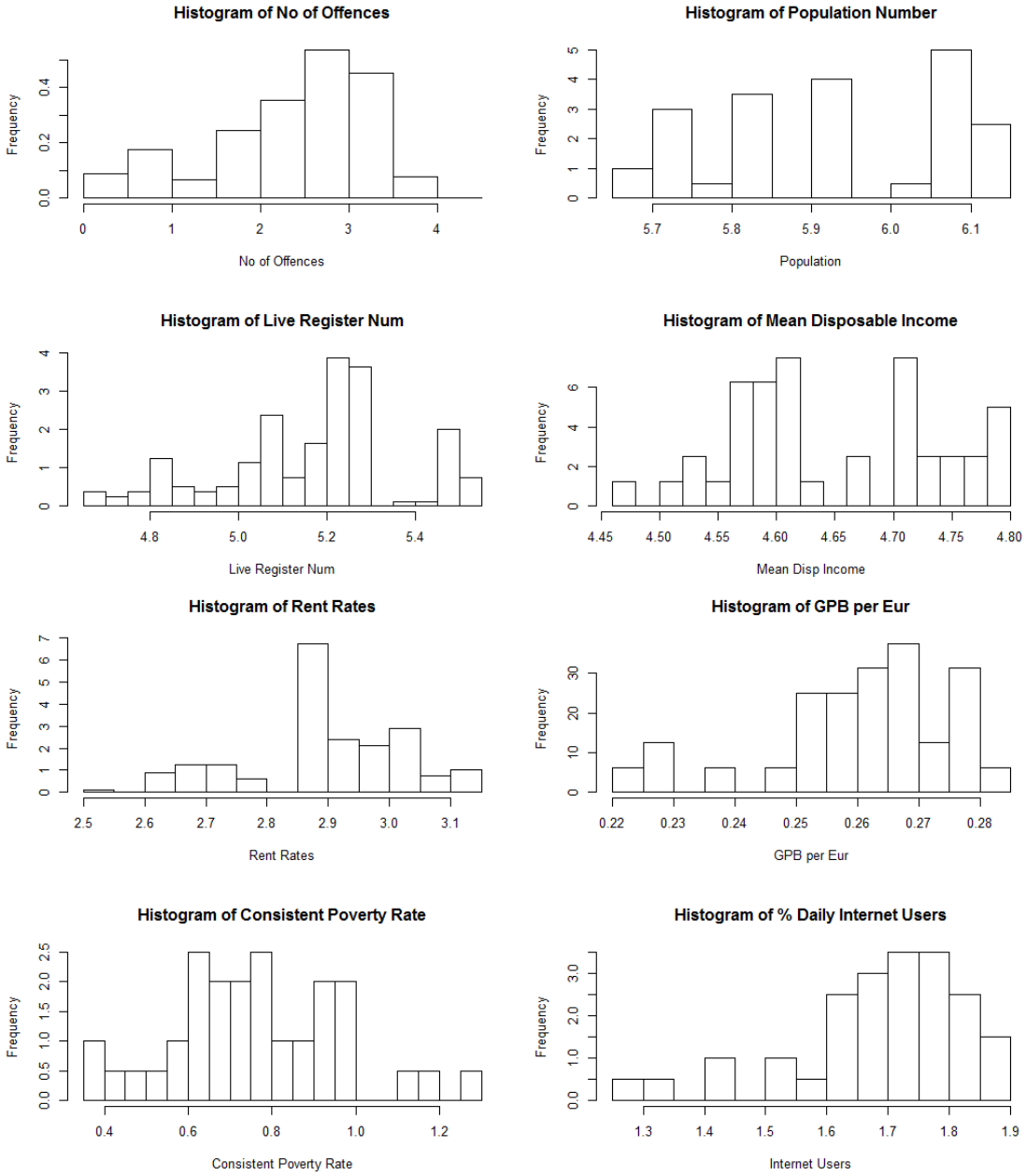


Figure 7 – portion of tableau scatterplot matrix

Following on from that, I generated histograms of the 10 main variables that I would be using for testing different models. A normal distribution is generally a bell shape and is seen to an extent below in the histogram of % Daily Internet Users. Some of the histograms have very wide distributions, peaks at their tails or have a skewed distribution.



At this point, a log transformation was carried out to normalise the data as otherwise the models I attempted to produce were producing very poor results. The histograms of the transformed variables are given below and it is pretty clear that the distributions have improved, although they are nowhere near perfect.



4.2 Modelling

4.2.1 Initial Steps

At this stage of the project, multiple linear regression (MLR) is used to build, train and test a predictive model for a target variable, No_of_Offences. The aim of linear regression is to model a continuous variable Y as a mathematical function of one or more X variables, so that we can use this model to predict Y when only X is known. The mathematical equation for this is:

$$Y = \beta_1 + \beta_2 X + \epsilon$$

where β_1 is the intercept, β_2 is the slope and ϵ is the error term, the part of Y the model is unable to explain.

To start, a model was built using all variables and the full dataset. Figures 8 and 9 below display the resulting plot and summary statistics for the model. In the summary statistics, we are given the values of:

- Multiple R-squared - is the proportion of variance in the dependent variable which can be explained by the independent variables
- Adjusted R-squared - is an adjustment of the R-squared that penalises for the addition of predictors to the model
- P-value – this tells us the statistical significance of the model. Ideally this will be below 0.05
- F-statistic – will tell us if a group of variables are jointly significant. The higher the better.

This is not a bad model but was only created to give an overview and contains some variables that will not be included in the final model such as Region, Crime_Type, Year and Qtr. Variables such as those will be omitted because we want the model to predict based only on socioeconomic factors.

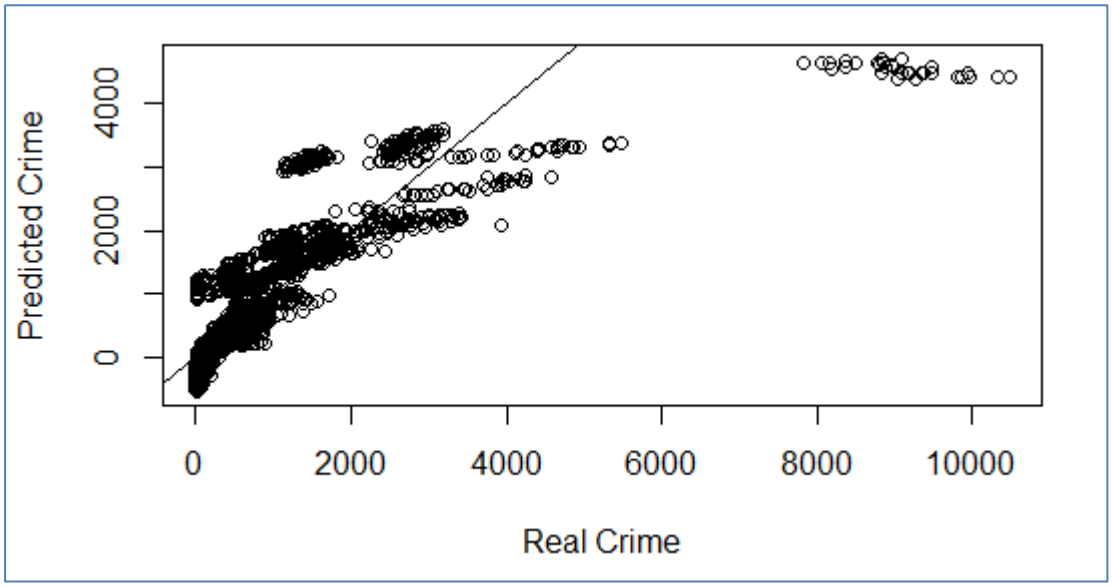


Figure 8 – plot of regression model for all variables

```
Residual standard error: 771.1 on 2202 degrees of freedom
Multiple R-squared:  0.7674,    Adjusted R-squared:  0.7634
F-statistic: 191.2 on 38 and 2202 DF,  p-value: < 2.2e-16
```

Figure 9 – summary statistics

My next step was to use the *corrplot* function that output the graph as seen below in Figure 10. The graph has every variable from the dataset and shows the correlations between them. On the upper right side the correlations are displayed using colour coded circles, blue for correlated relationships and brown

for inverse correlations. On the lower left these same relationships are signified with their numeric values. As mentioned before, highly correlated variables have multicollinearity and are redundant so, using the graph, I made the decision to drop CPI_Housing_Electricity_Gas, Tobacco_gm_per_capita and Alcohol_gm_per_capita.

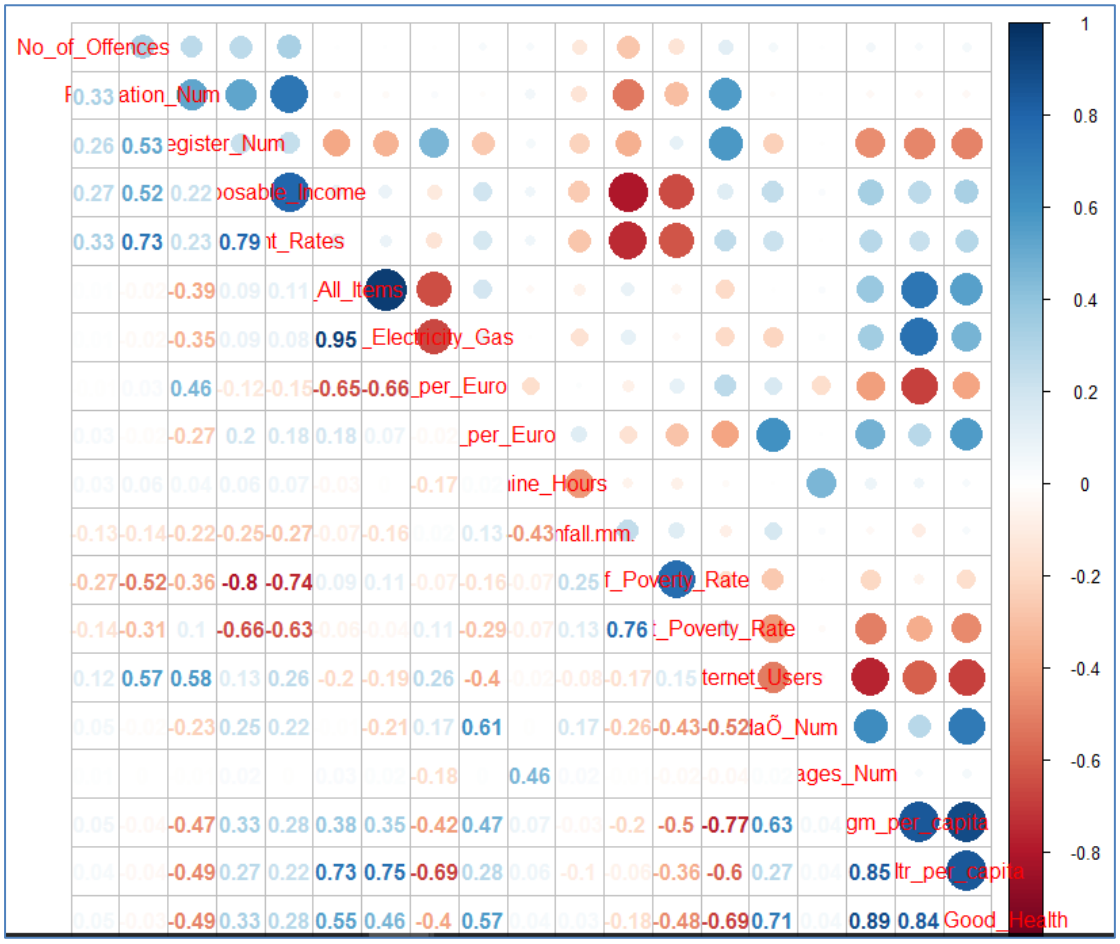


Figure 10 – correlation plot for crime dataset

All the variables went through an outlier check using a script that gave the number and percentage of outliers for that variable, it's mean value and a boxplot with any outliers present and also, without outliers. At the end of this process, it was determined that the mean values barely changed with or without the outliers and so any few outliers would be left as they are.

4.2.2 Choosing a Final Model

From this point a good deal of time was spent running models with different combinations of variables to find one with the best fit. The model I settled upon consisted of No_of_Offences as the dependent variable and the independent variables were Live_Register_Num, Rent_Rates, Daily_Internet_Users and Pop_In_Good_Health. My method for choosing the independent variables was to see, at each stage, which had the best p-values and dismiss the others. I also ran a stepwise regression at this time to see what final model R would decide upon

using an AIC (Akaike Information Criterion) measure and its final independent variables were Population_Num, Live_Register_Num, Rent_Rates, GBP_per_Euro and Daily_Internet_Users. Although Population_Num was a good predictor variable, I decided against keeping it because it isn't a social or economic factor and my own model had better results than the stepwise model when Population_Num wasn't included.

Figures 11 and 12 below show the summary statistics and plot for the model. As you can see there is an Adjusted R-squared value of 0.883 and a p-value of <2.2e-16. As the p-value is less than 0.05, we can reject the null hypothesis that $\beta = 0$ and assume that there is a significant relationship between the dependent and independent variables in the model.

```
Call:
lm(formula = No_of_Offences ~ . - 1, data = model3)

Residuals:
    Min      1Q  Median      3Q     Max
-2.3791 -0.4442  0.2040  0.6476  1.6432

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
Live_Register_Num     0.7500     0.1134   6.616 4.59e-11 ***
Rent_Rates            1.5135     0.1574   9.617  < 2e-16 ***
Daily_Internet_Users -0.6248     0.1796  -3.479 0.000513 ***
Pop_In_Good_Health   -2.5158     0.3281  -7.667 2.60e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8632 on 2236 degrees of freedom
Multiple R-squared:  0.8832,    Adjusted R-squared:  0.883
F-statistic:  4227 on 4 and 2236 DF,  p-value: < 2.2e-16
```

Figure 11 – summary statistics for final model

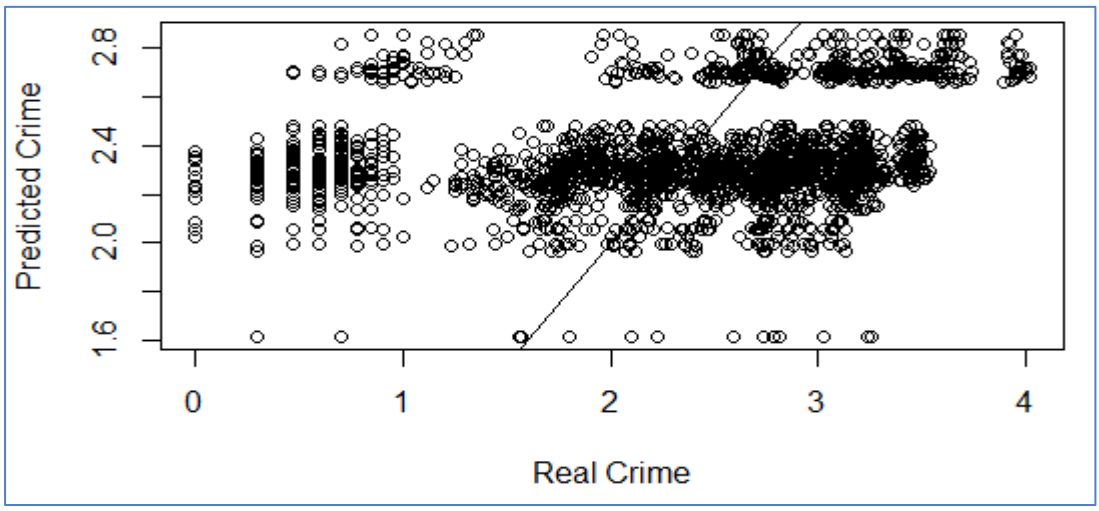


Figure 12

The diagnostic plots in Figure 13, show that the residuals seem to form a symmetrical band on both sides of the 0 line which suggests that the variances of

the error terms are equal. It also gives the indication that the relationship is linear. The normal Q-Q plot has some deviations from the straight line which could mean that we have some extreme values in the distribution.

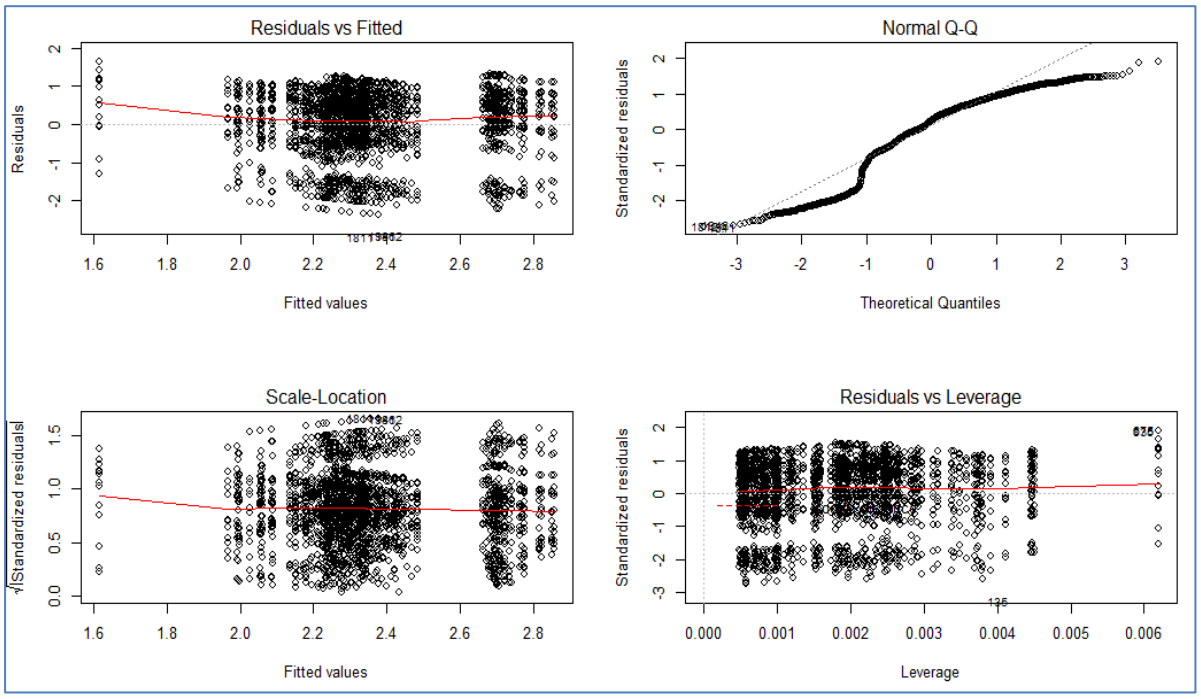


Figure 13 – diagnostic graphs for final model

4.3 Evaluating the Model

For the first part of test, the data was split 80:20 into a training and a test set. A MLR model was built on the training data and its predictive capability was examined using the test data and predict() function. You can view the summary output of the model built with the training data in Figure 14. The results are very similar to the model that was built using the full dataset previously. Beneath that in Figure 15, the output for the correlation accuracy between actual and predicted values is given, and it can be taken from this that the values the values are correlated and therefore the model is accurate. The final part of the model assessment here is the Min Max accuracy equation:

MinMaxAccuracy = mean(min(actuals,predicteds) / max(actuals,predicteds))

and gave a result of 74.78491%.

```
Call:
lm(formula = No_of_Offences ~ . - 1, data = trainingData)

Residuals:
    Min      1Q  Median      3Q     Max
-2.3732 -0.4491  0.2079  0.6556  1.6632

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
Live_Register_Num      0.7225     0.1275   5.665 1.71e-08 ***
Rent_Rates             1.6081     0.1768   9.096  < 2e-16 ***
Daily_Internet_Users  -0.5895     0.2034  -2.898   0.0038 **
Pop_In_Good_Health    -2.6183     0.3663  -7.148 1.28e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8667 on 1788 degrees of freedom
Multiple R-squared:  0.8821,    Adjusted R-squared:  0.8818
F-statistic:  3343 on 4 and 1788 DF,  p-value: < 2.2e-16
```

Figure 14 – summary statistics for training dataset

```
      actual predicted
4  0.6020600  2.126695
9  0.3010300  2.272654
10 0.6020600  2.291354
17 0.4771213  2.149484
20 0.4771213  2.117982
23 0.6020600  2.059390
>
```

Figure 15 – correlation between actual and predicted values

To ensure comprehensive testing on my model, I next carried out k-fold Cross Validation on it. This process entails splitting the data into *k* number of subsets, then performing the training on one group of subsets while testing on the other. This is repeated for each instance and an overall accuracy is provided. In Figure 16 we can see that all of the 10 fold lines are parallel and close to one another, therefore the model's prediction accuracy isn't varying too much for one particular sample and the lines of best fit don't vary too much with regards slope and level.

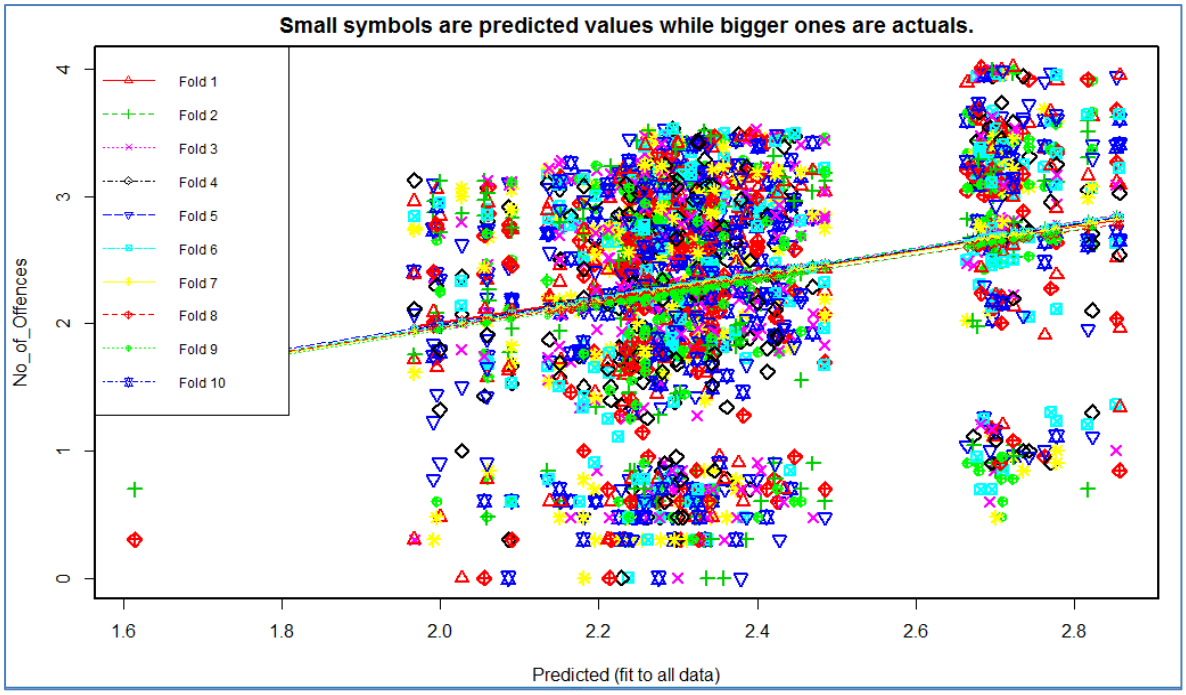


Figure 16 – k-fold cross validation output

5. Conclusion

5.1 Final Conclusion

The objective of this project was to ascertain whether any social and/or economic factors had an influence on the crime rate and find out if they could be used to predict the future movement of the rate. In this sense, I feel as though the project has been a success as many correlations have been found between the crime rate and socioeconomic variables, and an accurate prediction model was built and satisfactorily evaluated.

There were many different combinations of socioeconomic variables tested at the modelling stage and it became obvious which were important individually and worked well with others, and which were useless for prediction means. This was to be expected as the dataset was compiled by myself, using a mix of factors that I knew beforehand would be relevant and others that I included out of sheer interest but did not expect great results from. There are variations of the final model selected in this report that give very similar results to what I have recorded here and it would take a lot more time, and probably the use of other modelling techniques, to find the perfect model.

The four socioeconomic factors of Number of people on the Live Register, Rent Rates, % Population as Daily Internet Users and % Population in Good Health are a very interesting mix of predictor variables and not what I would have chosen as the final combination before modelling, but it goes to show it's not always the obvious answer is the correct one. As mentioned in the earlier chapter on

background research, I had read about studies where alternative, obscure reasons were found for events and a similar result has appeared here.

5.2 Future Work

From start to finish, this project has enabled me to apply all of the data visualisation and analysis techniques that I've learned throughout the past year. At this time, however, I'm still very much a novice in this area and with continued learning into the future I would hope to be able to produce an even better study on this topic. Obtaining a comprehensive collection of suitable data would be an obstacle to overcome though, as at the moment I don't feel enough is readily available from online sources. The trend in this area seems to be aimed at producing real time predictive models, based solely on historical crime data so is quite different to this project but I have enjoyed completing this project, the learning curve was steep and it has given me an appetite to progress even further.

Bibliography

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

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

3. Science | AAAS. (2016). *Can 'predictive policing' prevent crime before it happens?*. [online] Available at: http://www.sciencemag.org/news/2016/09/can-predictive-policing-prevent-crime-it-happens [Accessed 11 Jan. 2017].

4. Statistics How To. (n.d.). *F Statistic: Definition and How to find it*. [online] Available at: http://www.statisticshowto.com/f-statistic/ [Accessed 13 Jan. 2017].

5. R-statistics.co. (n.d.). *Linear Regression With R*. [online] Available at: http://r-statistics.co/Linear-Regression.html#k-%20Fold%20Cross%20validation [Accessed 13 Jan. 2017].

6. R-bloggers. (n.d.). *Visualising Residuals*. [online] Available at: https://www.r-bloggers.com/visualising-residuals/ [Accessed 14 Jan. 2017].

7. Dhana, K. (n.d.). *Identify, describe, plot, and remove the outliers from the dataset*. [online] R-bloggers. Available at: https://www.r-bloggers.com/identify-describe-plot-and-remove-the-outliers-from-the-dataset/ [Accessed 13 Jan. 2017].

Appendices

The figures below display analysis and modelling that did not make it into the main body of the text above.

Figure 17 displays the resulting output when AIC stepwise regression was used:

```
> step$anova # display results
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
No_of_Offences ~ Population_Num + Live_Register_Num + Mean_Disposable_Income +
    Rent_Rates + CPI_All_Items + GBP_per_Euro + Consistent_Poverty_Rate +
    Daily_Internet_Users + Pop_In_Good_Health

Final Model:
No_of_Offences ~ Population_Num + Live_Register_Num + Rent_Rates +
    GBP_per_Euro + Daily_Internet_Users


                      Step Df   Deviance Resid. Df Resid. Dev       AIC
1                                                2230   1647.524 -668.1329
2 - Consistent_Poverty_Rate  1 0.01221284      2231   1647.536 -670.1163
3      - Pop_In_Good_Health  1 0.11098831      2232   1647.647 -671.9654
4  - Mean_Disposable_Income  1 0.28946478      2233   1647.937 -673.5719
5           - CPI_All_Items  1 0.53377047      2234   1648.471 -674.8465
```

Figure 17

Figures 18 and 19 are the summary statistics and plot for the final model, without having carried out a data transformation. As you may notice the R-squared and F-statistic values are much lower and the plot distribution isn't as balanced:

```
Call:
lm(formula = No_of_Offences ~ . - 1, data = model3)

Residuals:
    Min      1Q  Median      3Q     Max
-2156.6  -598.9  -255.1   290.2  8719.7

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
Live_Register_Num     4.678e-03  4.529e-04  10.331  < 2e-16 ***
Rent_Rates            1.778e+00  1.228e-01  14.474  < 2e-16 ***
Daily_Internet_Users -1.037e+01  2.382e+00  -4.354 1.40e-05 ***
Pop_In_Good_Health   -1.028e+01  1.493e+00  -6.882 7.65e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1244 on 2236 degrees of freedom
Multiple R-squared:  0.3853,    Adjusted R-squared:  0.3842
F-statistic: 350.3 on 4 and 2236 DF,  p-value: < 2.2e-16
```

Figure 18

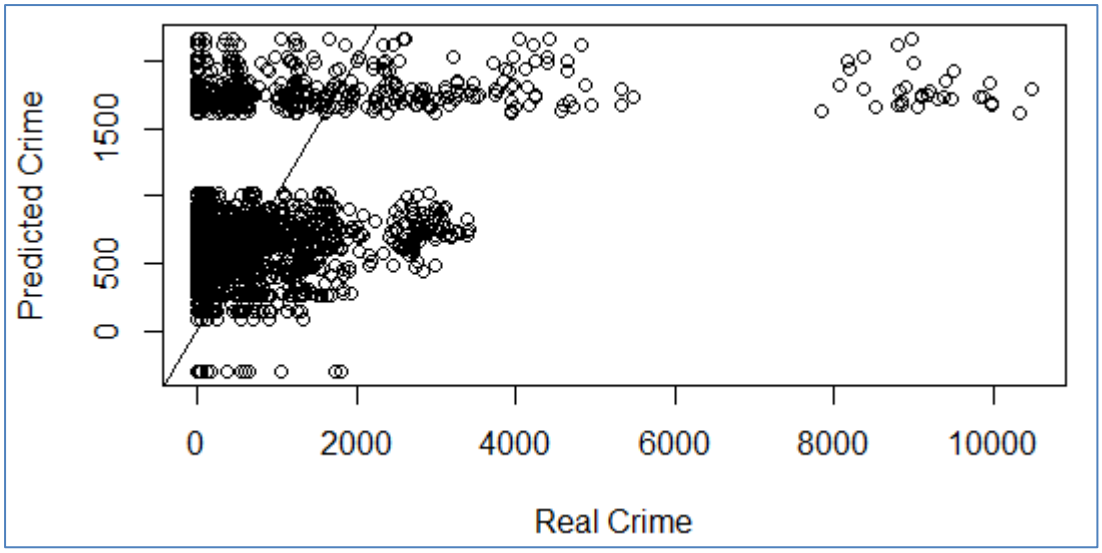


Figure 19

Training and testing was carried out on this model too. Figure 20 shows the actual and predicted values which as we can see are not correlated at all, and the model had a Min Max accuracy measure of only 36.208625%. Figure 21 is the output from a k-fold cross validation and the variation in the dotted lines tells us that the samples had varying degrees of accuracy.

	actual	predicted
4	3	243.0734
9	1	534.0889
10	3	593.3781
17	2	459.8296
20	2	408.9213
23	3	331.6958

Figure 20

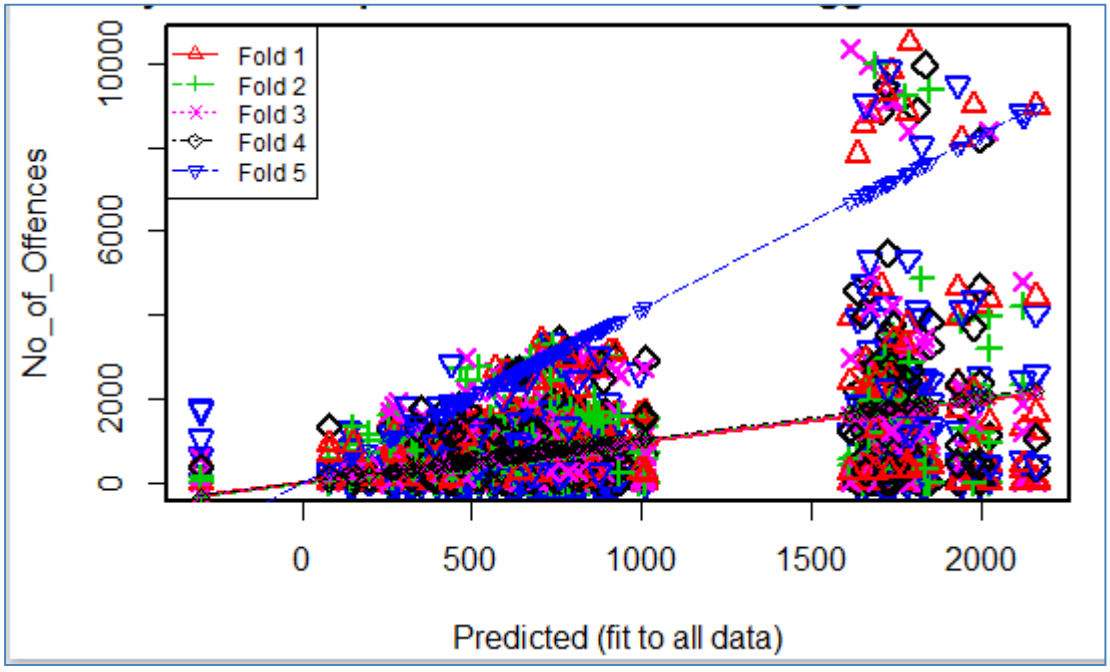


Figure 21

The summary statistics of one of the first models built using the transformed data can be seen in Figure 22 below. You may notice that the R-squared values are more or less the same as the final model chosen, however, there are quite a number of variables/coefficients with p-values much too high. You may also notice that the asterisks signal the best coefficients to help with variable selection.

```
Residuals:
    Min      1Q  Median      3Q     Max
-2.4198 -0.4324  0.2225  0.6518  1.3133

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
Population_Num            0.827959   0.244496   3.386  0.00072 ***
Live_Register_Num         0.782832   0.169264   4.625 3.96e-06 ***
Mean_Disposable_Income    0.251234   0.416165   0.604  0.54611
Rent_Rates                0.663964   0.335392   1.980  0.04786 *
CPI_All_Items             0.008980   0.008789   1.022  0.30705
GBP_per_Euro             -2.715773   1.835451  -1.480  0.13912
Consistent_Poverty_Rate   0.004767   0.163804   0.029  0.97679
Daily_Internet_Users     -0.847871   0.203216  -4.172 3.13e-05 ***
Pop_In_Good_Health       -3.921074   0.898601  -4.364 1.34e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8594 on 2231 degrees of freedom
Multiple R-squared:  0.8845,    Adjusted R-squared:  0.884
F-statistic:  1898 on 9 and 2231 DF,  p-value: < 2.2e-16
```

Figure 22