



National College of Ireland

Software Project

Data Analytics

2020/2021

Tiblet Woldegiorgis

17394526

X17394526@student.ncirl.ie

Analysis of the Crime Rate in London Using
Machine Learning and Data Mining Models

Technical Report

Contents

Executive Summary.....	2
1.1. Background	3
1.2. Aims.....	4
1.3. Technology.....	5
1.4. Structure	6
2.0 Data.....	7
3.0 Methodology.....	8
3.1 Selection.....	8
3.2. Pre-processing.....	9
3.2.1 Handling of Missing values	9
3.2.2 Merging of datasets.....	9
3.3 Transformation.....	10
3.4. Data Mining	10
3.5. Interpretation/Evaluation	11
4.0 Analysis	12
5.0 Results.....	13
6.0 Conclusions	21
7.0 Further Development or Research	22
8.0 References	23
9.0 Appendices.....	25
9.1. Project Plan	25
9.2. Ethics Approval Application (only if required)	26
9.3. Reflective Journals	30
9.4. Other materials used	35

Executive Summary

Crime is one of the biggest problems faced by countries that negatively impacts society and the economy. Numerous studies have been carried out to identify the causes of crime rate in London. Compared to the years used for this analysis, currently crime rate in the London has been increasing year on year, with the steepest increase occurring after the year 2016. This report investigates factors that contribute the crime rate in London. There could be many factors that contribute to crime rate but seen in this study factors are unemployment rate, homelessness rate and employees earning below the living wage in London area.

The datasets were collected from public source such as (kaggle.com) and (data.gov). These datasets were pre-processed manually using excel, and furthered pre-processed in Jupyter using Python. Machine learning and data mining models like Pearson's correlation coefficient, multiple linear regression analysis, and k-means clustering used in different analytics tools such as Python and R and Tableau for visualisation.

Pearson's correlation coefficient is a statistical test that measures the statistical relationship, between two constant variables. It is suitable for this analysis since the goal is finding a correlation between each of the three factors namely Unemployment, homelessness, Employees below the living minimum wage as related to the crime rate in London.

Multiple linear regression is a statistical model that assesses the connection between one dependent variable and one or more independent variable utilizing a line. Multiple linear regression is also suitable for this analysis since it helps us to identifying a relationship between each of the variables in dataset. K-means Clustering is used when there are unlabelled data point like such as data without defined categories or groups so the algorithm works iteratively to assign each data point to one of K groups based on the features that are provided. For this analysis K-means was used to see the dataset be groups in to 3 clusters.

The analysis indicated that Crime Rate and Low Income have a strong negative correlation while homelessness rate had a strongly and positively relation in correction analysis. Unemployment is also related positively to crime but weaker than homeless. Multiple linear regression analysis also show that these factors have a strong relationship as indicated by the F-statistic and the P-values.

K-means clustering grouped the datapoints in three cluster which we can label as Low Crime Area, Medium Crime, and High Crime Area.

The result of all the analysis were consistent in that Crime Rate in London between 2012 and 2016 was highly influenced by unemployment rate, low wage rate, and homelessness , but at different and varying levels.

1.1. Background

London is the fourth largest city in Europe, it has a population of 8.982 million. It's known for its museums and royal palace. As per this report currently it suffers from high crime rate especially a high intensity of knife crime and drug activities. Various examinations are being completed consistently to comprehend the reason for crime and viciousness in London with the aim of improving public security. Different categorizing of public along classes and ethnicity lines were tried to comprehend their associations the rising crime in the city. The perplexing idea of crime makes a wide extension to contemplate different elements which might be a reason for crime.

One main factor to take from theoretical evidences is that places that attract in huge quantities of individuals for non-crime purposes such as tourism and recreations can be utilised for crime activity. Areas that are generally private but just display areas of interest under the encompassing populace may be places with a higher extent of crime attractors to animate crime, yet less generators to draw in volumes of individuals. (Malleeson et al.; 2016; Andresen 2016). The driving factor to crime activities could be anything but unemployment and low income are the known drivers. According to a study carried out show utilized area level information to research the connection between crime and unemployment and it demonstrate that there is a precise positive connection between most crime and unemployment. (Carmichael et al.; 2001 Ward 2001)

Comparing these studies with the crime statistics in London for the period 2012-2016, both criminal damage and robbery crime have continued to increase, while burglary and drugs crimes keep fluctuating as shown in the Figure 1.1.1.

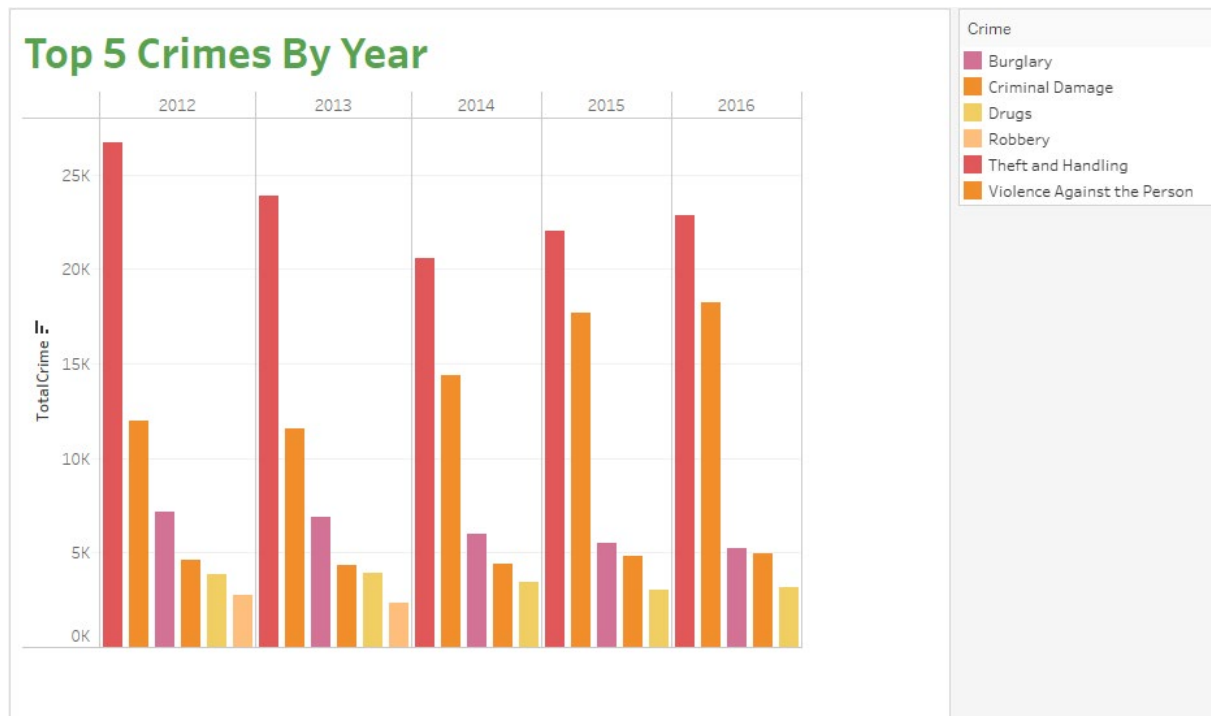


Figure 1.1.1: Distribution of Crimes in London (2012-2016) Tableau

1.2. Aims

The aim of this Project is to investigate the factors contributing to the crime rate in London. This is achieved by using machine learning and data mining models. The investigation aims to identify whether crime rate in London is as a result of unemployment, homelessness and below the living wage. As explained above, to achieve these aims, several analyses were carried out using correlation, Regression, Clustering and Classification.

The research question for this study is “can machine learning and data mining approaches help explain crime rate in London using factors such as unemployment, homelessness and low income?”

The motive of this investigation is to answer the question by merging various factors such as unemployment, homelessness and employees earning below the living wage in London. This data investigation of the crime rate in London might help in obtaining helpful knowledge which will be beneficial to the London metropolitan police department and other government officials in mitigating crime in the city. The following objectives are defined in this study to meet the above research question:

- Data pre-processing and merging of multiple datasets on London crime data.

- Selection of the best features contributing to crime prediction model using a combination of multiple linear regression and clustering.
- Transformation and scaling of features using effective techniques such as one hot encoding

1.3. Technology

The analysis included using various machine learning and data mining models, such as Pearson Correlation which is used to assess the relationship strength between variables like unemployment and crime rate. A high correlation implies that at least two factors have a solid relationship with one another, while a low correlation implies that the variables are not really related. Regression which is a supervised learning model was also used to identify which factor affect a subject of interest and allow to determine which factors matter most. It is used to identify the variables that can be disregarded, and how so to see how these elements impact the others. Clustering which is an unsupervised learning technique was used in distinguishing and gathering comparative data in datasets without worry for the particular result. Finally, Classification, another supervised learning model was used to recognize and relegate categories to an assortment of data to consider more precise analysis.

The methodology process selected for this study was KDD, which is alluded to as Knowledge Discovery in Database and is characterized as a technique for finding, changing, and refining significant data and patterns from a raw data set to be used in various spaces or applications.

Different data analytic tools were used to carry out these machine learning algorithms, data mining techniques and statistical operation, technology. Excel which is an easy, famous and broadly utilized analytical tool practically in all ventures. I used Excel to download the data from the website then pre-processed the data by resolving missing values by replacing with mean average, and removing fields that weren't necessary for the analysis. Pivotal tabling and many Excel formulas were used to reorganise and share the datasets adding columns needed for the analysis as well as create tables and for calculation.

Another tool intensively used was Python. Python is an Object-Oriented scripting language to peruse, compose, keep up. It is a free open-source tool. Jupyter notebook another free, open-source which was used to write Python code was used. I used Python for manipulating, processing, cleaning, and crunching the data. Python was used to analyse the dataset for correlation and regression.

Another tool I used was R, it is the main examination tool in the business and generally utilized for measurements and information displaying. I used R to analyse the data and to facilitate various statistical data mining techniques through its packages and achieved algorithms such as classification and clustering.

Lastly I used Tableau public for better visualization. The public version is and makes data visualization, maps, dashboards and look better. I used this tool for a better visualization of the clustering and correlation of the data.

1.4. Structure

The remainder of the report structure contained data. This part of the document contains detail description of the dataset used, how was the data sourced and compiled. It outlined all the exploratory data analysis conducted as well as the statistical measures that was used. Finally it describes the data visualisation tools used.

The methodology part contains description of how it was reached from the data to the analysis. Using the KDD methodology it explains the methods step by step so that the research could be easily reproduced on this or another data set. It further explains any decisions made in relation to important features of the modelling approach.

The analysis part contains description of the approaches that was employed in the analysis of the data, explains why these approaches were chosen when other options were available and finally clarify the decisions that were made in relation to important features of the modelling approach such as why was a particular attribute chosen as a predictor in the model.

The result part presents the results of the analysis using tables and figures to support and clarify were appropriate evaluation metrics, description of figures and tables indicated. Following that the conclusion part includes the problem that was addressed and the solutions to the problem. It outlined key findings of the project and what the end product is, It also bears the impacts they have on society's wellbeing, and how they are going to benefit the end-users.

Finally the last part of the documents contains the further development or research part where it is explained how, with additional time and resources, one can carry out further development and research.

2.0 Data

There was four datasets for this analysis, and all four the dataset used are secondary datasets. These datasets are collected by someone else but are made available on the web. One of them was collected from Kaggle (www.kaggle.com), a website that allows users to publish and download datasets and more. The rest of the three datasets were collected from London datastore (www.data.london.gov.uk), a website that allows access the data that the GLA and other public sector organisations hold.

The first dataset is the London Crime data, 2008-2016 and was obtained from (Kaggle.com) this dataset is a row data. This data covers the number of criminal reports by month, LSOA borough, and major/minor category from Jan 2008-Dec 2016. It contains 13M rows and 7 columns, which are soa_code, borough, major_category, minor_category, value, year and month. For the analysis the year 2012 to 2016 was selected, downloaded as a CSV and pre-processed manually in excel and furthered pre-processed in Jupyter and other tools as needed for analysis.

The second dataset is the Unemployment Rate by Ethnic Group & Nationality, Borough, and was obtained from (data.london.govuk). The dataset contains unemployment rates broken down by ethnic group from the year 2005 to 2019. The data is taken from the Annual Population Survey (APS), produced by the Office for National Statistics. For this analysis to be effective and match the other datasets the year 2012 to 2016 was selected, downloaded as a CSV and pre-processed manually in excel and furthered pre-processed in Jupyter.

The third dataset is the homelessness provided by borough obtained from (www.data.london.gov.uk). The dataset contains homelessness rates broken down by ethnic group from the year 2004 to 2017 and Source from DCLG P1E Homelessness returns (quarterly). It's essential for the analysis that the dataset is the same which is why the year 2012 to 2016 was selected, downloaded as a CSV and pre-processed manually in excel and furthered pre-processed in Jupyter.

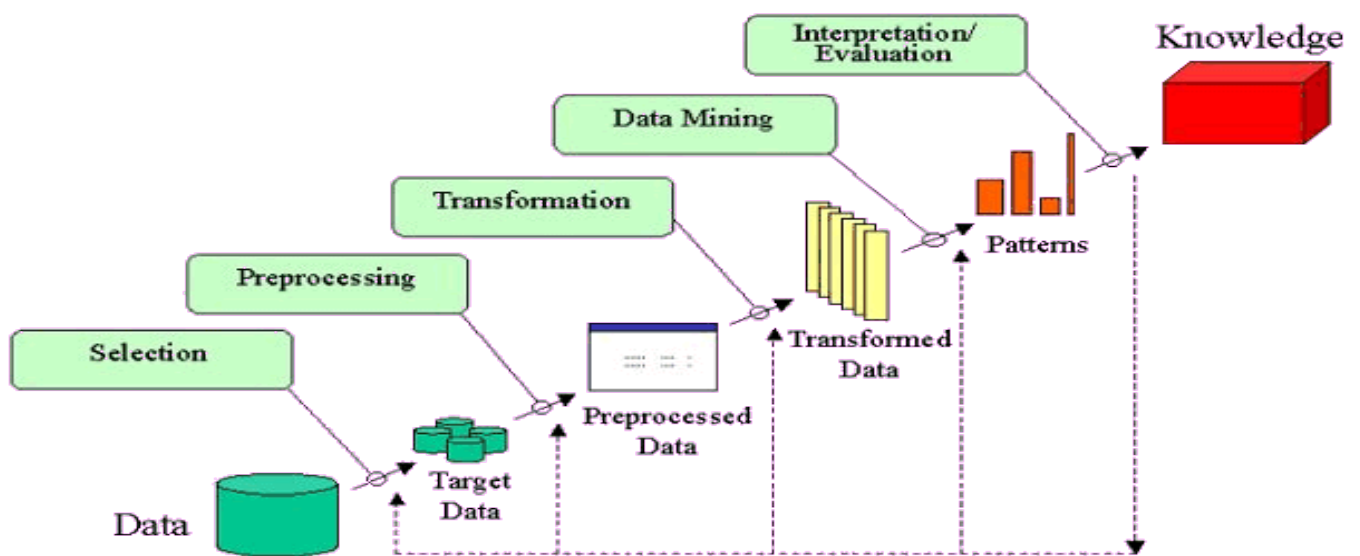
The fourth and final dataset is Employees earning below the London Living Wage obtained from (www.data.london.gov.uk). Percentage of respondents in work earning less than the London Living Wage (LLW) taken from the ONS Annual Survey of Hours and Earning (ASHE). Data is also provided by borough from 2005 to 2018. Also includes employees earning below the UK Living Wage by region in London. Like the other dataset the year 2012

to 2016 was selected, downloaded as a CSV and pre-processed manually in excel and furthered pre-processed in Jupyter.

3.0 Methodology

The various stages of this analysis implementation follows a design methodology resembling the stages of knowledge discovery in (KDD). The KDD process, as introduced in (Fayyad et al, 1996) is the way toward utilizing DM techniques to extricate what is considered information as indicated by the particular of measures and edges, utilizing a data set alongside any necessary pre-handling, sub sampling, and change of the data set.

Figure 3 represents the crime rate analysis design methodology covering the 5 stages namely: Data Selection, Pre-Processing, Transformation, Data mining and Interpretation/Evaluation. The results generated could be then used for making effective decisions.



3.1 Selection

For this research the dataset has been selected from various platform like Kaggle and londo.gov.uk website. This data represents crucial part in accomplishing the goals illustrated in this report. Four datasets are accumulated from numerous sources as demonstrated in the Table 3.1.1, for this analysis. All the datasets except, London crime rate, are downloaded as flat files in the form of CSV (comma separated values) from (Kaggle.com) and (londo.gov.uk) websites, where the information is consistently refreshed and made accessible openly for scientists and scholars to carryout studies. London crime rate dataset is extracted through

Application Programming Interface (API) from Kaggle API and saved as a CSV document. Out of crime data from 2008 to 2016 only records from 2012 to 2016 are considered for this study, the unemployment rate data, homelessness rate data and employees earning below the living wage have been pre-processed first to be merged with crime data.

3.2. Pre-processing

Data pre-processing was one of the essential periods of this examination, where data manipulation exercises such as detecting, correcting or removing and replacing of missing, and inaccurate data records, and merging of datasets. This stage is vital in anticipation of calculations and models to run easily and in order to obtain consistent data. The data was first pre-processed manually in excel once the data was ready it was furthered pre-processed in Jupyter using python.

3.2.1 Handling of Missing values

- The crime dataset consisted crime data from 2008-2016 with 13 million records, for this analysis the data was selected from 2012 to 2016. Within that data there were many missing data values, and repeated 0 or null values. The month column was dropped since it wasn't needed for the analysis. To make effective these data values were removed all the other dataset contained rates. Since this research focuses on merging all the datasets based on rates and the crime data didn't have rate values so the population was divide by the crime number to get the crime rate. Data point for City of London was removed from all datasets as it was missing in all years in some datasets
- The Unemployment rate dataset consisted data from 2005 to 2019. The dataset also contains unemployment rate in terms of race in London. To match with the other dataset, I selected from 2012 to 2016. The dataset contained other data such as employment rate and economic activity rate which were removed since the unemployment rate data was the only data needed. There were no missing values
- The homelessness rate dataset consisted data from 2004 to 2018, the dataset contained homelessness rate in terms of race in London, and data from 2012 to 2016 was selected to match the other dataset. No missing values were found.
- Finally, the employee earning below the London living wage dataset consisted data from 2005 to 2008 with no missing values.

3.2.2 Merging of datasets

The four datasets in consideration had some relevant attributes based on which the merging has been carried out. These were primarily geographical attributes (Area name and area code) or

date-time attributes (date expressed as year). The merging phase were carried out in Jupyter using python

3.3 Transformation

For improving the crime prediction model accuracy and remove multicollinearity, it is essential to understand and obtain the best features from unemployment, homelessness and employees below living wage by locations and race attributes, that describes the target variable crime count.

- In multiple crime studies, Pearson Correlation Coefficients are compared to understand the linear relationship of several factors with crime occurrences, in the data considered in this work does possess linear relationship.
- Excel’s functionalities such as pivot table was used. IF functions and lookups were also to clean, organise and ready the dataset for the analysis. Descriptive analysis and data discovery methods were rigorously used to see if the datasets have missing values
- Top 10 features namely: Crime rate, NewCode, Below Wage Pay Rate, Homeless-White, Homeless-NonWhite RaceUnknown, Total#Homeless, RateOfHmless, Area, Year Population, Non White Rate, White Rate, AvUnEmpRate are selected.

3.4. Data Mining

For this stage of the methodology is consist of three stage design flow has been followed, comprising of data, modelling and visualization stages as shown in Figure 3.4.1

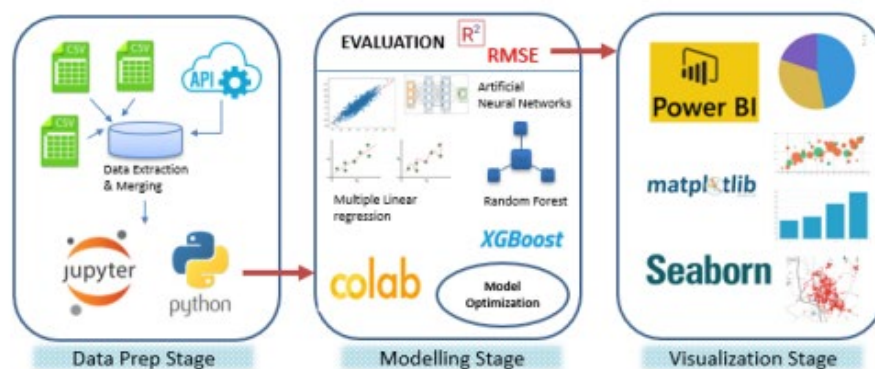


Figure 3.4.1: Crime analysis – Design flow

- The data planning phase comprises of the each and every phase achieved explicit to data gathering, merging, and exploratory data analysis, include designing and highlight choice stages. Based on the data source, every data was downloaded as a CSV records

or extricated utilizing API associations utilizing Python programming on Jupyter Notebooks.

- Modelling stage deals with implementation of multiple machine learning algorithms and data mining techniques such as Correlation, Linear Regression, Classification, and Clustering were undertaken. Modelling and optimization have been done on Jupyter and Tableau.
- Lastly, the results obtained were presented in the form of plots, and graphs as desired for visualization purpose.

3.5. Interpretation/Evaluation

For this stage using the results of the previous stages carry out interpretation and evaluation of the mined patterns.

Implementation of Pearson’s Correlation

Person correlation was experimented on the dataset with default and tuned parameters and evaluated train and test accuracy. A look at the result of correlation analysis shows some important insights. It provides us a very interesting facts about the data. By looking at the below Pearson Correlation model we can say that low wage is negatively correlated to crime while unemployment and homeless are both have a positive relationship, though homeless showing more stronger relations (0.74). From the data we can easily conclude the as wage decreases crime increases, while as homeless and unemployment increase crime also increases.

```
#Correction between Quantity, sales profit and discount
pearsoncorr = PartData.corr(method='pearson')
```

```
pearsoncorr
```

	Below Wage Pay Rate	RateOfHmless	AvUnEmpRate	Crime Rate
Below Wage Pay Rate	1.000000	0.104840	-0.342447	-0.279014
RateOfHmless	0.104840	1.000000	0.332387	0.739915
AvUnEmpRate	-0.342447	0.332387	1.000000	0.511067
Crime Rate	-0.279014	0.739915	0.511067	1.000000

Implementation of Linear Regression in SKLearn

SKLearn is practically the best standard with regards to AI in Python. It has many learning calculations, for relapse, arrangement, grouping and dimensionality decrease.

```
from sklearn import linear_model
lm = linear_model.LinearRegression()
model = lm.fit(X,y)
```

The lm.fit() function fits a linear model. This model is used to make predictions

```
lm.score(X,y) #this the percentage of explained variance of the predictions
0.6983934314728626
```

This is the R^2 score of the model. this the percentage of explained variance of the predictions. If you're interested,

```
lm.intercept_#We can create our model using this intercept and Coefficients given in
#the model for the variable (-0.0014,1.3375 and 0.0022)
0.06384270028823437
```

```
lm.coef_
array([ 0.          , -0.00141065,  1.33751145,  0.00222536])
```

These are all (estimated/predicted) parts of the multiple regression equation I've mentioned earlier.

Implementation of K-mean Clustering Models

The clustering model was implemented in R using RStudio Notebook. There are methods such as Elbow Method, to decide on the number of clusters. In this project I used three clusters chosen. The crime dataset has 160 samples that were picked, given that the analysis is based on wage, homeless, unemployment and crime. We can label the cluster Low Crime Area, Medium Crime Area and High Crime as showing in the results section below.

4.0 Analysis

Overall for this investigation there are three approaches that were applied. These approaches are Pearson Correlation, Multiple Linear Regression and K-means Clustering.

The Pearson Correlation Coefficient is the test statistics that actions the statistical relationship, or association, between two persistent variables. It is known as the best strategy for estimating the relationship between factors of revenue since it depends on the technique for covariance. It gives data about the extent of the affiliation, or correlation. The Pearson Correlation Coefficient method of analysis was implemented in this analysis to determine and judge the strength of potential relationships between unemployment versus crime rate in London, homelessness versus crime rate in London and employees living below the minimum wage versus crime rate in London.

Multiple Linear regression extends basic linear regression to incorporate multiple illustrative variables. Multiple linear regression method of analysis was implemented to estimate the relationship between two or more independent variables such as unemployment, homelessness

and employee living below the minimum living wage, and one dependent variable which is the crime rate data.

K-means clustering method of analysis was implemented in this analysis k-Means takes data points as input like the crime rate data and groups them into k clusters this is putting the crime data in a group of area.

Other option like decision tree, random forest and kNN are other option to employ in the analysis. These options are all types of classification algorithm, Random forest. In order to carry out these models the needed data types and the analysis would have to be different in order for the model to give the right prediction

The difference edge is a basic baseline approach to feature selection. It eliminates all features which variance doesn't meet some limit. It eliminates each of the zero-variance features, i.e., features that have similar value in all samples. The data that was selected contained missing values as well as unwanted columns therefore this was the best feature selection method for this analysis. For this analysis the top 10 features namely: Crime rate, NewCode, Below Wage Pay Rate, Homeless-White, Homeless-NonWhite RaceUnknown, Total#Homeless, RateOfHmless, Areaio87, Year, Population, Non White Rate, White Rate, AvUnEmpRate are selected.

5.0 Results

As stated in the previous sections this project was to analysis, investigate and research and report on the impact of homelessness, unemployment, and low wage on the crime rate that is seen in London. All dataset were downloaded from kaggle and London datastore. In order to analysis and report the dataset has to normalised and standardised to fit to the analysis. For this analysis four related datasets are needed. Though all the four datasets were available, they were in some way different from each other. Some of the datasets had a rate value while others have raw numbers. In such cases the rate values were calculated from the existing numerical data values to have a similar value for analysis. The analysis was done on a five years dataset to see the changes of crime rate over five years as related to other factors in the same time frame.

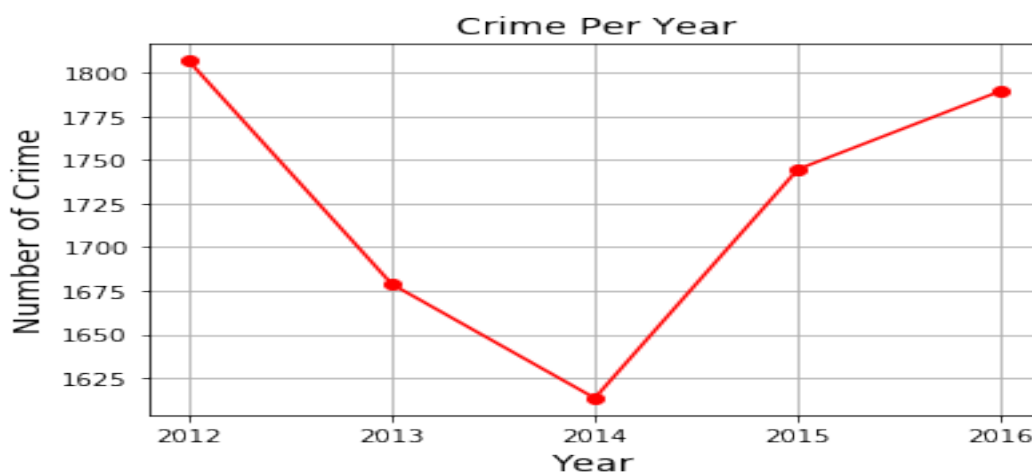
There was huge data cleansing and manipulation needed in all the dataset. For this purpose, Excel's functionalities such as pivot table was used, if functions and lookups to clean, organise and ready the dataset for the analysis. Descriptive analysis and data discovery methods were rigorously used to see if the datasets have missing values, or if any of the dataset violet important assumptions for the type of analysis envisaged to be used. After investigating the datasets data

point for City of London was removed from all datasets as it was missing in all years in some datasets. There few data points which have missing values for one or two years. In such cases the mean average was calculated and replaced the null values.

Finally, all the dataset were clean and ready for further analysis. At the end of the clean up the dataset every dataset has rate values for respective dataset group. This means homeless dataset has homeless rate, unemployment dataset has unemployment rate and so on.

The aim of the report, as mentioned earlier is to find out the relationship between homelessness, unemployment, low pay and crime rate in London for the years 2012 to 2016. Different dataset analysis techniques were used to find out the impact on has on the other. Correction analysis and multiple simple regression analysis were made in Python and Excel. In this section I report the analysis that us been done in Python.

The datasets have undergone lots of manipulations in Python. Datasets were merged, and then sliced to create a needed portion of dataset for the analysis. Lots of field headings were renamed for clarity. A new data fields were created to facilitate merging. For example, as the codes were repetitive across city areas every year a combo field from year and code was created to have a unique row identifier across the dataset. Z-test was used to see if there are any outliers exist that can affect the data analysis. All Z-test scores that fall ≥ -3 and ≤ 3 were removed.



Different analysis was made on the different dataset individually and the following figures were evident from the data. According to the graph above the number of crimes in London was fallen in year 2014 while risen sharply afterwards.

The following graph shows ever increasing number of homelessness in London. As we can see it was steadily increasing from 2012 to 2015, but showing a decline in 2016. While it is clear

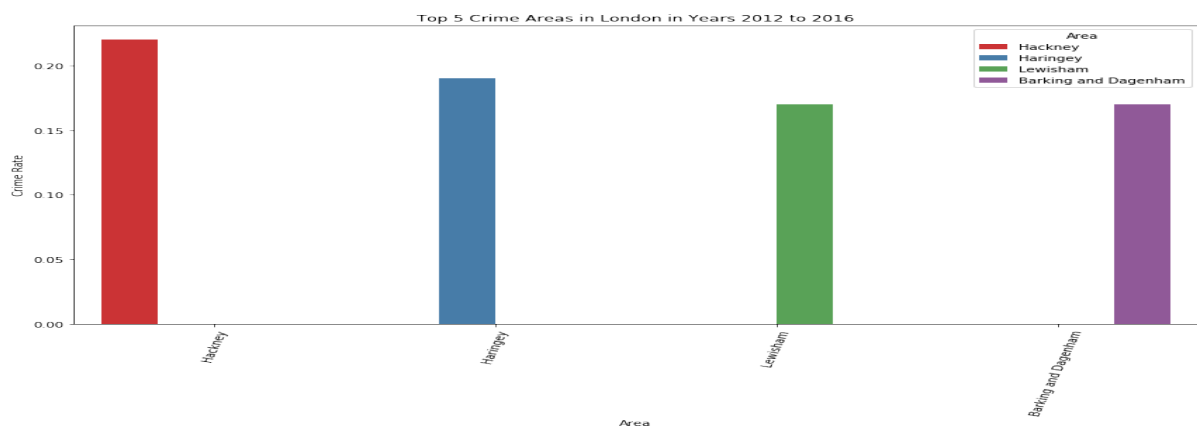
visually clear from the graph below that homelessness and crime seems not related, but we will see next the influence homelessness has on crime rate in London.



The following tables shows the average homeless per race. Here we are not going to base our analysis on race but it gives us some perspective on how homelessness is feed and other important factors behind that. While the average number of homeless non-white is near double that is homeless white, the rate of unemployment for both is nearly the same. The table shows on average there is 1727 crimes every year in London, with a increasing rate of 0.087 every year.

Homeless-White	180
Homeless-NonWhite	321
RateOfUnEmplyedNonWhite	8.30
RateOfUnEmplyedWhite	8.92
# of Crime	1727
Crime Rate	0.087

Hackney, Haringey, Lewisham and Barking and Dagenham are the most crime spot part of the London while Harrow, Sutton, Bexley and Kingston Upon Thomas were the lowest crime rated in the city. Hackney is the top crime hot spot while Kingston Upon Thomas is the safest crime free zone.



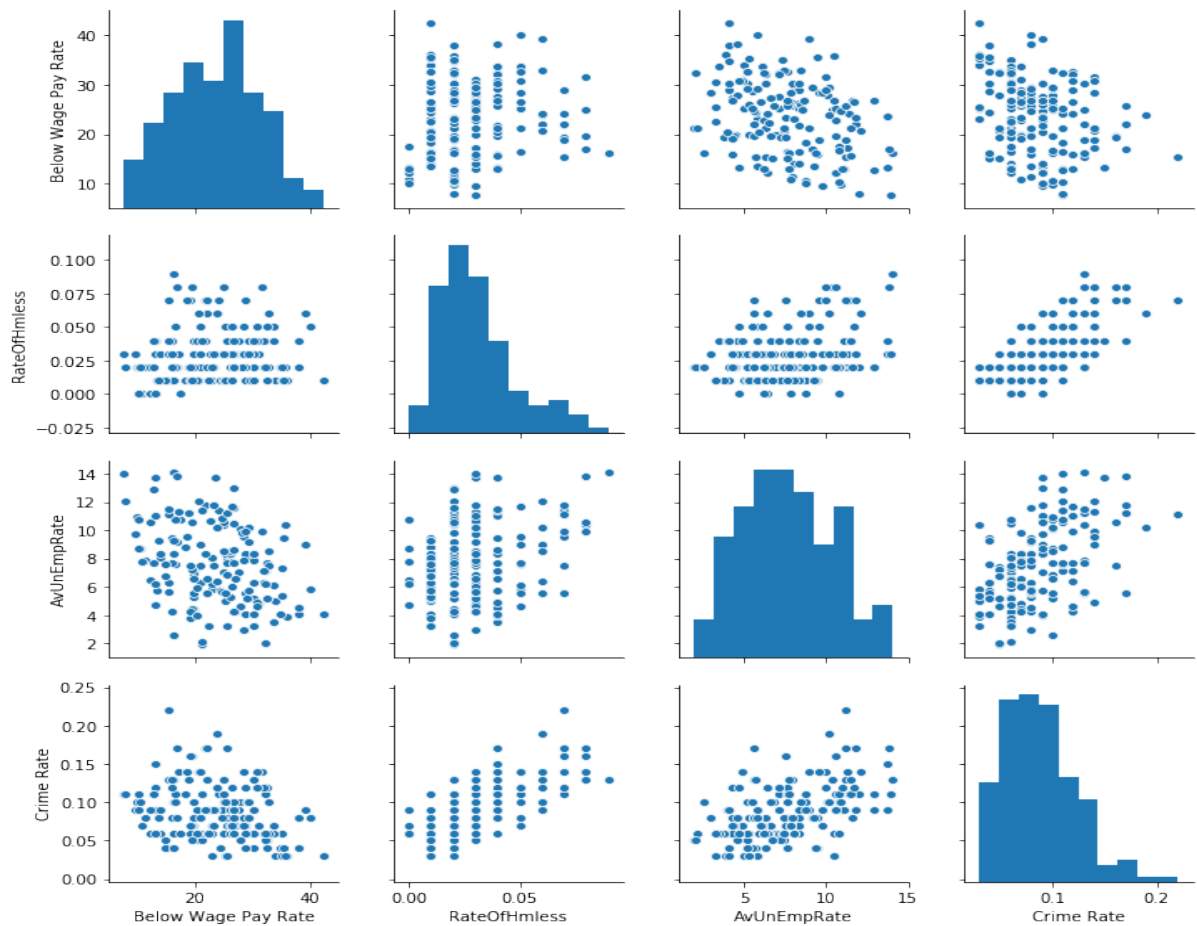
A look at the result of correlation analysis shows some important insights. Even though correlation analysis does not indicate any clue other than the existence of relationship, it provides us a very interesting facts about the data. The following graph show Pearson's Correlation analysis. By looking at the below Pearson Correlation model we can say that low wage is negatively correlated to crime while unemployment and homeless are both have a positive relationship, though homeless showing more stronger relations (0.74). From the data we can easily conclude the as wage decreases crime increases, while as homeless and unemployment increase crime also increases.

```
#Correction between Quantity, sales profit and discount
pearsoncorr = PartData.corr(method='pearson')
```

pearsoncorr

	Below Wage Pay Rate	RateOfHmless	AvUnEmpRate	Crime Rate
Below Wage Pay Rate	1.000000	0.104840	-0.342447	-0.279014
RateOfHmless	0.104840	1.000000	0.332387	0.739915
AvUnEmpRate	-0.342447	0.332387	1.000000	0.511067
Crime Rate	-0.279014	0.739915	0.511067	1.000000

The same data above was visualised below to show the strength or weakness of the relationship among the variable.



The plot above clearly indicates that crime rate and low wage are negatively but weakly related, white homeless is strongly and positively related. Unemployment is also related positively to crime but weaker than homeless.

As stated, the above correction does not show us whether these factors contributed or how much each contribute to crime rate. In other word correlation does not show causality. Because of this we will try to use other techniques to find out the share each factor's contributed to the increasing crime rate in London. Analysis of multiple Regression below will give us a clue.

Out[140]:

OLS Regression Results

Dep. Variable:	Crime Rate	R-squared:	0.698			
Model:	OLS	Adj. R-squared:	0.693			
Method:	Least Squares	F-statistic:	120.4			
Date:	Fri, 07 May 2021	Prob (F-statistic):	2.09e-40			
Time:	21:17:33	Log-Likelihood:	402.07			
No. Observations:	160	AIC:	-796.1			
Df Residuals:	156	BIC:	-783.8			
Df Model:	3					
Covariance Type:	nonrobust					
	coef	std err	t	P> t 	[0.025	0.975]
	const	0.0638	0.008	7.715	0.000	0.047 0.080
	Below Wage Pay Rate	-0.0014	0.000	-6.098	0.000	-0.002 -0.001
	RateOfHmless	1.3375	0.090	14.826	0.000	1.159 1.516
	AvUnEmpRate	0.0022	0.001	3.400	0.001	0.001 0.004
Omnibus:	17.503	Durbin-Watson:	1.651			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	24.312			
Skew:	0.637	Prob(JB):	5.26e-06			
Kurtosis:	4.423	Cond. No.	1.48e+03			

Looking into the R-Squared and the Adj. R-square it seems that all the variables included into the analysis equally contribute to the model. Thus, we are not removing any of the variables in the model. These values also show a positive and relatively strong Correlations. The F-statistic shows there is a statistically significant learner relationship between the predictors and the outcome. This is also supported by the P-values of the variables helping us to reject the Null hypothesis and can claim that there is a strong evidence between Crime, unemployment, low pay and homelessness. According to this model nearly 0.70(70%) of the data can be explained or predicted. We can model our data and prediction using this intercept (0.064) and the coefficients given (-0.0014, 1.3375 and 0.0022) in the model for the variable.

K-mean Cluster Analysis

K-means clustering is one of the simplest and popular machine learning algorithms. It is the unsupervised learning methods as opposed to supervised learning algorithms. The two are different in that we do not have a prior knowledge of the data and only the algorithms groups or clusters along the lines it sees fit to certain category based on the data. To do this the algorithm calculates the centroid. The process runs for multiple rounds to fit each dataset into a given

cluster which is decided by the user. To do this it uses the mean average and thus K-Mean for decision. In this example the decision was to use three clusters. There are methods such as Elbow Method, to decide on the number of clusters. In this project case it was three clusters that I picked

```
kc <- kmeans(alldata, 3)
```

Then the dataset should run for some rounds of calculation to determine in which cluster to assigned a datapoint. In this analysis the dataset was run for 50 times as indicated in the below code.

```
kc <- kmeans(alldata, centers = 2, nstart = 50)
```

The following plot and output are the output from the cluster analysis from the data.

K-means clustering with 3 clusters of sizes 6, 13, 13

Cluster means:

	Below.Wage.PayRate	Rate.Of.Hmless	Av.UnEmp.Rate	Crime.Rate
1	12.78000	0.01900000	8.840333	0.09033333
2	21.84923	0.03384615	7.677846	0.09276923
3	30.18000	0.02923077	7.049846	0.07892308

Clustering vector:

```
[1] 2 2 3 3 3 1 2 3 3 2 2 2 3 3 3 2 2 1 2 2 1 2 3 3 3 2 1 3 1 3 2 1
```

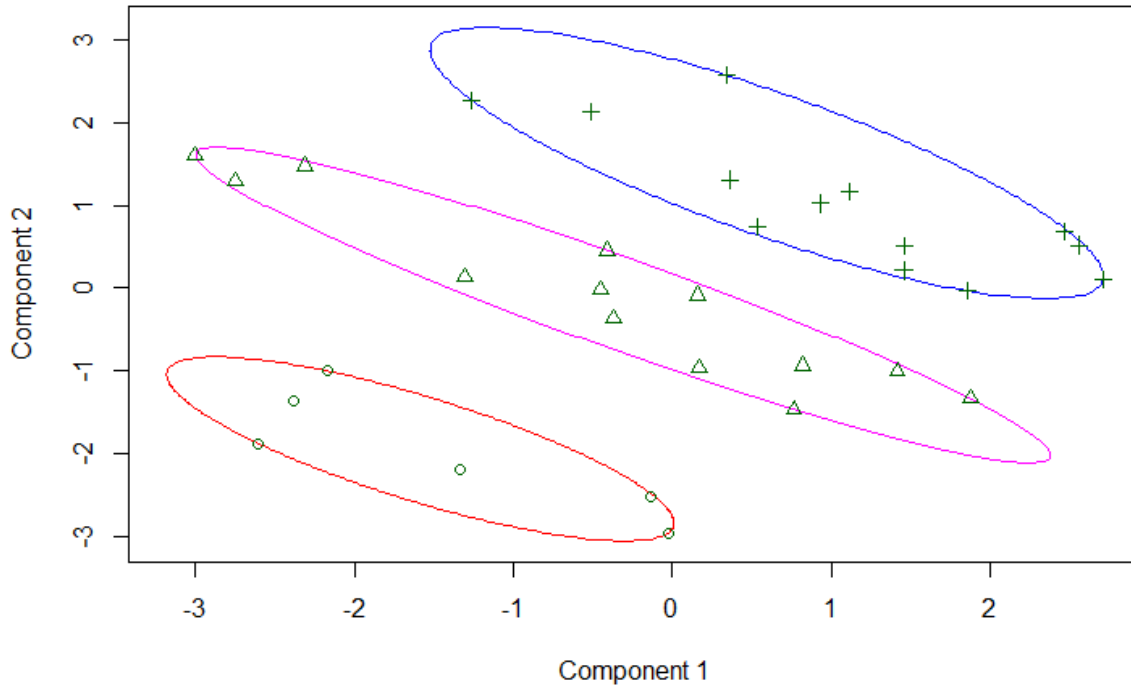
Within cluster sum of squares by cluster:

```
[1] 35.0211 104.7769 141.5472  
(between_SS / total_SS = 82.4 %)
```

Available components:

```
[1] "cluster" "centers" "totss" "withinss"  
[5] "tot.withinss" "betweenss" "size" "iter"  
[9] "ifault"
```

k-Means Cluster Analysis



This is the output from using $K = 3$ getting 89.98% of well-grouped data. $K = 2$ and $K = 4$ gave us less than this result which was 87.38% and 75.88%. Therefore $K = 3$ is the best cluster number to use as it was able to explain about 90% of the variability in the data.

Based on the output of the analysis which is indicated below we can label the three groups with a meaningful naming.

	Below.Wage	PayRate	Rate.Of.Homeless	Av.UnEmp.Rate	Crime.Rate
1	12.78000	0.01900000	8.840333	0.09033333	
2	21.84923	0.03384615	7.677846	0.09276923	
3	30.18000	0.02923077	7.049846	0.07892308	

Given that the analysis is based on wage, homeless, unemployment and crime we can label the cluster as follows

Cluster 1 = Lowe Crime Area as the average from the dataset shows lower values than other clusters.

Cluster 2 = Medium Crime Area as the average from the dataset shows medium values than other clusters.

Cluster 3 = High Crime Area as the average from the dataset shows higher values than other clusters.

The same dataset was used for clustering analysis in Tableau and the above plot is from cluster



analysis from Tableau. Both plots are very similar and can easily fit to the same description.

6.0 Conclusions

In this research the aim was to investigate if the crime rate in London is linked to unemployment, homelessness and employees earning below the living wage. This analysis was carried out using machine learning and data mining approaches such as Correlation, Multiple Linear Regression and Clustering. Correlation shown homelessness as having strongly and positive relationship with crime while unemployment has positive but weaker than homeless.

Multiple Linear Regression analysis shows there is a statistically significant learner relationship between the predicators and the outcome. This is also supported by the P-values of the variables helping us to reject the Null hypothesis and can claim that there is a strong evidence between Crime, unemployment, low pay and homelessness. K-means clustering analysis is based on wage, homeless, unemployment and crime we can cluster the whole London city in to tree clusters as Lowe Crime Area, Medium Crime Area, and High Crime Area .

This investigation on the crime rate in London might help in obtaining helpful information which may benefit to London metropolitan police department and other government officials.

7.0 Further Development or Research

The research presented in this report is just small-scale step towards the prediction of crime rate. Even though I've achieved the aim of this investigation, but with additional time and resources there are other things I would add to further strength the analysis. Selection of data for a period of five years probably limits the capability of the models and adding more data could result in better insights. Bring more factors such as peer, politics, religion, family conditions, the society at large could add more insight to explain crime rate in London. As these are outside the scope and reach of this project and due to time and resource constraints, I invite any interested researchers to take this project further.















8.0 References

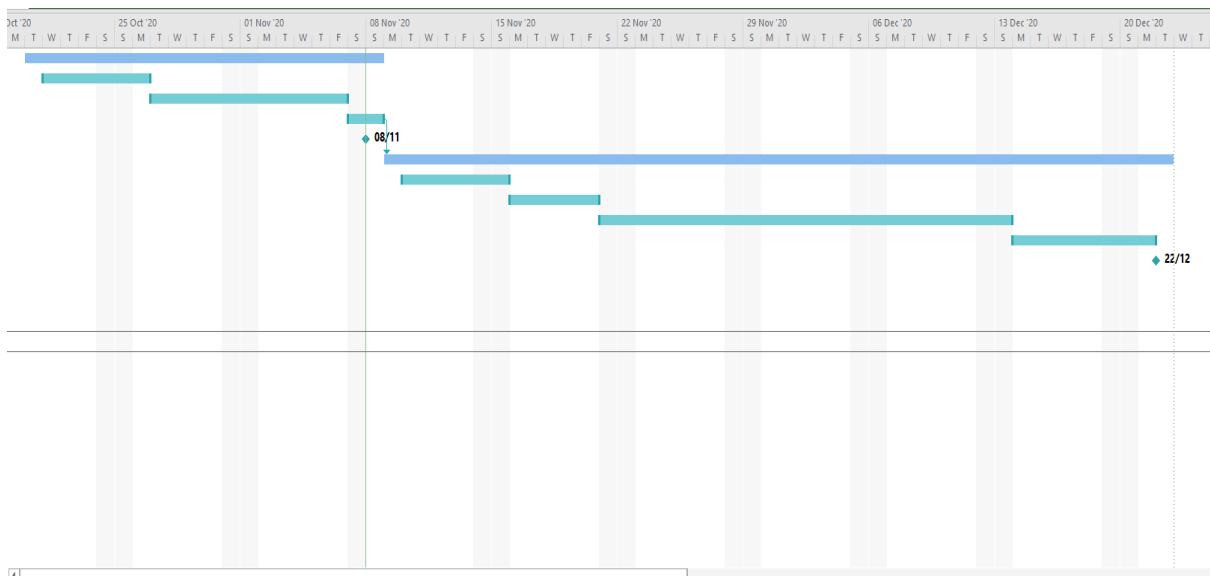
- Wieshmann, H., Davies, M., Sugg, O., Davis, S. and Ruda, S., 2021. [online] Available at: <https://www.london.gov.uk/sites/default/files/bit_london_violence_reduction_final_28_january_2020.pdf> [Accessed 16 May 2021].
- Antonio, J., 2021. *London Crime Data Analysis (Part I)*. [online] Medium. Available at: <<https://becominghuman.ai/london-crime-data-analysis-part-i-9b7062081f9a>> [Accessed 16 May 2021].
- Platt, D. and Lucinda, H., 2021. [online] Eprints.lse.ac.uk. Available at: <http://eprints.lse.ac.uk/68133/1/Newburn_Social%20Disadvantage%20and%20Crime.pdf> [Accessed 16 May 2021].
- Aghababaei, S. and Makrehchi, M. (2018) ‘Mining Twitter data for crime trend prediction’, *Intelligent Data Analysis*, 22(1), pp. 117–141. doi: 10.3233/IDA-163183.
- Jabeen, Nahid & Agarwal, Parul. (2021). Data Mining in Crime Analysis. 10.1007/978-981-15-6707-0_10.
- Palant, O. Y., Ortina, G. V. and Kucher, M. M. (2018) ‘Statistical Assessment of Socio-Economic Determination of Crime in Ukraine’, *Scientific Bulletin of Polissia*, 16(4), pp. 14–20. doi: 10.25140/2410-9576-2018-4(16)-14-20.
- Baciú, O. A. and Parpucea, I. (2011) ‘Socio-Economic Factors Impact on Crime Rate’, *Review of Economic Studies & Research Virgil Madgearu*, 4(2), pp. 5–20. Available at: <http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,cookie,shib&db=bsu&AN=69856548&site=ehost-live> (Accessed: 2 April 2021).
- Baumer, E. P., Wolff, K. T. and Arnio, A. N. (2012) ‘A Multicity Neighborhood Analysis of Foreclosure and Crime A Multicity Neighborhood Analysis of Foreclosure and Crime’, *Social Science Quarterly (Wiley-Blackwell)*, 93(3), pp. 577–601. doi: 10.1111/j.1540-6237.2012.00888.x.
- Fabio, A. et al. (2011) ‘Neighborhood Socioeconomic Disadvantage and the Shape of the Age-Crime Curve’, *American Journal of Public Health*, 101(S1), pp. S325–S332. doi: 10.2105/AJPH.2010.300034.
- Berrittella, M. (2018) ‘Organized crime and public spending: a panel data analysis’, *Economics of Governance*, 19(2), pp. 119–140. doi: 10.1007/s10101-018-0206-3.
- Misztal, M. (2020) ‘Application of the Partial Triadic Analysis Method to Analyze the Crime Rate in Poland in the Years 2000–2017’, *Folia Oeconomica Stetinensia*, 20(2), pp. 249–278. doi: 10.2478/fofi-2020-0047.
- dos Santos, M. and Kassouf, A. (2013) ‘A cointegration analysis of crime, economic activity, and police performance in São Paulo city’, *Journal of Applied Statistics*, 40(10), pp. 2087–2109. doi: 10.1080/02664763.2013.804905.
- AYHAN, F. and BURSA, N. (2019) ‘Unemployment and Crime Nexus in European Union Countries: A Panel Data Analysis’, *Journal of Administrative Sciences / Yonetim Bilimleri Dergisi*, 17(34), pp. 465–484. doi: 10.35408/comuybd.574808.
- Gakrelidz, N., 2021. *Predicting London Crime Rates Using Machine Learning*. [online] Blog.dataiku.com. Available at: <<https://blog.dataiku.com/predicting-london-crime-rates-using-machine-learning>> [Accessed 16 May 2021].
- Trevethan, Shelley. (2019). THE INTERSECTION OF SOCIAL AND ECONOMIC SYSTEMS WITH THE CRIMINAL JUSTICE SYSTEM. 10.13140/RG.2.2.17123.78881.
- Boysen, J., 2020. *London Crime Data, 2008-2016*. [online] Kaggle.com. Available at: <<https://www.kaggle.com/jboysen/london-crime>> [Accessed 22 December 2020].
- Office for National Statistics, 2021. *Economic Activity Rate, Employment Rate and Unemployment Rate by Ethnic Group & Nationality, Borough – London Datastore*. [online] Data.london.gov.uk. Available at: <<https://data.london.gov.uk/dataset/economic-activity-rate-employment-rate-and-unemployment-rate-ethnic-group-national>> [Accessed 16 May 2021].

- Ministry of Housing, Communities & Local Government, 2021. *Homelessness Provision, Borough – London Datastore*. [online] Data.london.gov.uk. Available at: <<https://data.london.gov.uk/dataset/homelessness>> [Accessed 16 May 2021].
- Office for National Statistics, 2021. *Employees earning below the London Living Wage (LLW) – London Datastore*. [online] Data.london.gov.uk. Available at: <<https://data.london.gov.uk/dataset/earning-below-llw>> [Accessed 16 May 2021].
- Malleson, N. and Andresen, M.A., 2016. Exploring the impact of ambient population measures on London crime hotspots. *Journal of Criminal Justice*, 46, pp.52-63.

9.0 Appendices

9.1. Project Plan

		Task Mode ▾	Task Name ▾	Duration ▾	Start ▾	Finish ▾	Predecessors
1			Second Month	15 days?	Tue 20/10/20	Sun 08/11/20	
2			Feedback	4 days	Wed 21/10/20	Mon 26/10/20	
3			Project Proposal	9 days	Tue 27/10/20	Fri 06/11/20	
4			Ethic Form	2 days	Sat 07/11/20	Sun 08/11/20	
5			Milestone	0 days	Sun 08/11/20	Sun 08/11/20	
6			Third Month	32 days?	Mon 09/11/20	Tue 22/12/20	4
7			Matrix on excel	5 days	Tue 10/11/20	Sun 15/11/20	
8			Meeting Supervisor	5 days	Mon 16/11/20	Fri 20/11/20	
9			Start Project	17 days	Sat 21/11/20	Sun 13/12/20	
10			Mid presentation	6 days	Mon 14/12/20	Mon 21/12/20	
11			Milestone	0 days	Tue 22/12/20	Tue 22/12/20	



DECLARATION OF ETHICS CONSIDERATION**School of Computing****Student Name:**Tiblet Woldegiorgis.....**Student ID:**17394526.....**Programme** Bachelor of Science(Honours) in Computing.....**Year:**2020/2021.....**Module:**Software Project.....**Project Title:** Analysis of the crime rate in London using machine learning and data mining models.....**Please circle (or highlight) as appropriate**

This project involves human participants	Yes / No
------------------------------------------	-----------------

Introduction

Secondary data refers to data that is collected by someone other than the current researcher. Common sources of secondary data for social science include censuses, information collected by government departments, organizational records and data originally collected for other research purposes. Primary data, by contrast, is collected by the investigator conducting the research.

A project that does not involve human participants requires ONLY completion of Declaration of Ethics Consideration Form and submission of the form on module's Moodle page

A project that involves human participants requires ethical clearance and an Ethics Application Form must be submitted through the module's Moodle page. Please refer to and ensure compliance with the ethical principles stated in NCI Ethics Form available on the Moodle page.

The following decision table will assist you in deciding if you have to complete the Declaration of Ethics Consideration Form or/and the Ethics Application Form.

Public Data	Y	Y	Y	Y	N	N	N	N
Private Data	Y	Y	N	N	Y	Y	N	N
Human Participants	Y	N	Y	N	Y	N	Y	N
Declaration of Ethics Consideration Form	x	X	x	X	X	X	x	
Ethics Application Form	X		X		X		X	

Please circle (or highlight) as appropriate

The project makes use of secondary dataset(s) created by the researcher	Yes / No
The project makes use of public secondary dataset(s)	Yes / No
The project makes use of non-public secondary dataset(s)	Yes / No
Approval letter from non-public secondary dataset(s) owner received	Yes / No

Sources of Data:

It is students' responsibility to ensure that they have the correct permissions/authorizations to use any data in a study. Projects that make use of data that does not have authorization to be used, will not be graded for that portion of the study that makes use of such data.

Public Data

A project that makes use of public secondary dataset(s) **does not need ethics permission**, but **needs a letter/email from the copyright holder** regarding potential use.

Some websites and data sources allow their data sets to be used under certain conditions. In these cases, a letter/email from the copyright holder is NOT necessary, but the researcher should cite the source of this permission and indicate under what conditions the data are allowed to be used. See Appendix I for examples of permissions granted by Fingal Open Data, and Eurostat website.

Where websites or data sources indicate that they do not grant permission for data to be used, you will still need a letter/email from the copyright holder. For example, see Appendix II for an example from the Journal of Statistics Education.

Private Data

A project that makes use of non-public (private) secondary dataset(s) must receive data usage permission from School of Computing.

An approval letter/email from the owner (e.g. institution, company, etc.) **of the non-public secondary dataset must be attached to the Declaration of Ethics Consideration**. The letter/email must confirm that the dataset is anonymised and permission for data processing, analysis and public dissemination is granted.

Evidence for use of secondary dataset(s)

Include dataset(s) owner letter/email or cite the source for usage permission

I am using the Kaggle data and this is a link to an evidence for use of secondary dataset

<https://www.kaggle.com/terms>

I am also using London datastores data and this is a link to an evidence for use of secondary dataset

<https://data.london.gov.uk/about/terms-and-conditions/>

CHECKLIST

Non-public/private secondary dataset(s) -Owner letter/email is attached to this form OR Citation and link to the web site where permission is granted – provided in this form	Yes / No Yes / No
--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	----------------------------------------

ETHICS CLEARANCE GUIDELINES WHEN HUMAN PARTICIPANTS ARE INVOLVED

The Ethics Application Form must be submitted on Moodle for approval prior to conducting the work.

Considerations in data collection

- Participants will not be identified, directly or through identifiers linked to the subjects in any reports produced by the study
- Responses will not place the participants at risk of professional liability or be damaging to the participants' financial standing, employability or reputation
- No confidential data will be used for personal advantage or that of a third party

Informed consent

- Consent to participate in the study has been given freely by the participants
- participants have the capacity to understand the project goals.
- Participants have been given information sheets that are understandable
- Likely benefits of the project itself have been explained to potential participants
- Risks and benefits of the project have been explained to potential participants
- Participants have been assured they will not suffer physical stress or discomfort or psychological or mental stress
- The participant has been assured s/he may withdraw at any time from the study without loss of benefit or penalty
- Special care has been taken where participants are unable to consent for themselves (e.g children under the age of 18, elders with age 85+, people with intellectual or learning disability, individuals or groups receiving help through the voluntary sector, those in a subordinate position to the researcher, groups who do not understand the consent and research process)
- Participants have been informed of potential conflict of interest issues
- The onus is on the researcher to inform participants if deception methods have to be used in a line of research

I have read, understood, and will adhere to the ethical principles described above in the conduct of the project work.

Signature:Tiblet Woldegiorgis.....

Date:07/11/2020.....

Appendix I

1) Fingal Open Data: <http://data.fingal.ie/About>

Licence

Citizens are free to access and use this data as they wish, free of charge, in accordance with the Creative Commons Attribution 4.0 International License (CC-BY).

Note: From November 2010 to July 2015, data on Fingal Open Data was published in accordance with the PSI general licence.

Use of any published data is subject to Data Protection legislation.

Licence Statement

Under the CC-BY Licence, users must acknowledge the source of the Information in their product or application by including or linking to this attribution statement: "Contains Fingal County Council Data licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) licence".

Multiple Attributions

If using data from several Information Providers and listing multiple attributions is not practical in a product or application, users may include a URI or hyperlink to a resource that contains the required attribution statements.

2) Eurostat: <https://ec.europa.eu/eurostat/about/policies/copyright>

COPYRIGHT NOTICE AND FREE RE-USE OF DATA

Eurostat has a policy of encouraging free re-use of its data, both for non-commercial and commercial purposes. All statistical data, metadata, content of web pages or other dissemination tools, official publications and other documents published on its website, with the exceptions listed below, can be reused without any payment or written licence provided that:

- the source is indicated as Eurostat
- when re-use involves modifications to the data or text, this must be stated clearly to the end user of the information

Appendix II

Journal of Statistics Education: http://jse.amstat.org/jse_users.htm

JSE Copyright and Usage Policy

Unlike other American Statistical Association journals, the Journal of Statistics Education (JSE) does not require authors to transfer copyright for the published material to JSE.

Authors maintain copyright of published material. Because copyright is not transferred from the author, permission to use materials published by JSE remains with the author. Therefore, to use published material from a JSE article the requesting person must get approval from the author.

9.3. Reflective Journals

Reflective Journals 01oct – November1

Student Name: __Tiblet Woldegiorgis____ **Student number:** __17394526__

WHAT?	My course is data analytics so it's important I picked a project that will give me data I can analyses, Considering the current climate and seeing how businesses are failing and employment are going down I thought it would be a good idea to analyses how covid-19 had an impact on retail business and compare it to previous years and predict how the future would be
SO WHAT?	Right now, we are going through a pandemic and I concluded that covid-19 either has a positive or a negative impact on retail business and I will be analyzing that using data I can find online like from the statistic office of Ireland as well as short questioner for store manager. This analysis hopefully will benefit retail sectors. I expect that getting data would be difficult considering the restriction in the country
NOW WHAT?	I am going to get feedback from my project supervisor and then do my project proposal afterwards I will start collecting data from online and preparing store managers.

Reflective Journals 01Nov – December1

The reflective Journal

My initial idea was to do “The impact of Covid-19 on retail Business” but upon doing more research on this topic I couldn’t find a large datasets that will help me for my analysis. Online resources like the central statistics office don’t have much datasets that I will need for my analysis due to the fact that Covid has only been here for a short period of time. I also talked to store managers to get data I will need but that won’t be possible so I decided that it was better to change my project.

The project I decided to is “Analysing the London crime rate and the factors of the crime”. London is the capital city of England there are 8.9 million people living in London. The city is filled with people from all over the world with different culture, background, religion, race and ethnicity. Over the years London has become one of the biggest city with a very high crime rate.

WHAT?	I have done research on the London crime rate read statistics and articles and I found that this would be a good project for me to execute. Right now I have found a dataset on the crime rate in London. I chose to do this analysis based on the year 2012 to 2016. I also have a datasets on the employment numbers during those years. I chose to bring employment into this because I believe that employment is one factor that causes people to do crime. Using this dataset I will analyse how employment in those year has gone down or up and see the crime increases or decreases. I will also bring in homelessness data as a factor for the analysis.
SO WHAT?	Right now I cleaned the datasets and made them accessible for my analysis. I will use that data as well as datasets that I will retrieve through an API and start doing my analysis. I also plan to bring more datasets that are factors and can be good for my analysis.

NOW WHAT?	For my midpoint presentation I plan to have all the datasets I have ready and start doing my analysis using machine learning.
-----------	-------------------------------------------------------------------------------------------------------------------------------

Reflective Journals 1st of January – 1st of February

Student Name: __Tiblet Woldegiorgis__ **Student number:** __17394526__

WHAT?	This month I got the feedback from the midpoint presentation and I went through it with my supervisor. From the feedback I received I noticed that I need to improve on my proposal of the project, need to build on my high level analysis , communication and develop the preliminary data analysis and technology I will be using. Over all it was good but I need to do much better if I want to get better marks for the final submission
SO WHAT?	Using the feedback I outlined and listed how I am going to apply the feedback I got and one plan I have is that I will have to use the different types of analysis I've done in my modules and apply it to my project, I also used different technologies in my other modules and decide to apply it to my project.
NOW WHAT?	For the next few weeks I have to do a lot of learning since I need more experience using language like R, python and tools like RStudio, tableu. I will have regular meeting with my supervisor to assist me with the project and get more feedback.

Reflective Journals 1st of February – 1st of March

Student Name: __Tiblet Woldegiorgis__ **Student number:** __17394526__

WHAT?	This month I was able to do more research for my project and found more datasets I can implement for my analysis. I was able to clean the dataset and get rid of any unwanted rows and columns. I continued with my education on R and python.
SO WHAT?	I am applying what I learned from my modules like business analysis and data mining for this project and I am finding doing calculation using R, to help me with that I am taking extra classes for further understanding. I meet with other student to exchange ideas and understand how to use tools.
NOW WHAT?	For the next few weeks I'll continue with my project, I have a meeting with a lecturer for my showcase feedback and using that feedback start making my showcase for the final submission of my project

Reflective Journals 1st of March – 1st of April

Student Name: __Tiblet Woldegiorgis__ **Student number:** __17394526__

	This month I continued with my final year project, I was able to complete one of my analysis using the data I got, I also spent a lot of my time on advancing my skill in r so I watched a lot of videos
--	----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

WHAT?	on r and Tableau which is a tool used for visualization since we never got any lessons from college it was my responsibility to find information and learn those skills.
SO WHAT?	I used the skills i learned on youtube and other website and applied it to my project, I am implementing another analysis on the data on jupyter using python, I ran into some problems which I am using resources to fix the issue
NOW WHAT?	For the next few weeks I have to complete a good amount of analysis and implement some work on my project since we're coming on the end of the year there will be a few assignments I have to upload for other assignments.

Reflective Journals 1st of April – 8st of May

Student Name: __Tiblet Woldegiorgis__ **Student number:** __17394526__

WHAT?	This month I was only able to implement a few things for my project because I had other group work and assignments I needed to give for other lecturers so I put the final project on hold. I also met up with my supervisors weekly to get advice on the project.
-------	--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

SO WHAT?	I currently finished the other assignments and was able to implement analysis and write up the report and send it to my supervisor to give me a feedback before I upload it. I ran into some problems on the analysis I was doing in r and now trying to fix that problem and waiting for a feedback from supervisor.
NOW WHAT?	For the next week I will finish with all the analysis and writing up of the report I need to do and when i get a feedback from my supervisor I will use that feedback to advance my report.

9.4. Other materials used

Any other reference material used in the project for example evaluation surveys etc.

