Using Machine Learning Algorithms and Statistical Techniques to Understand Crime Trends in San Francisco

Name: Cem Yilmaz
Student Number: x18191681
Email: cemyilmaz185@gmail.com
National College of Ireland
School of Computing
Dublin, Ireland

I. Introduction

There exists currently a myriad of different publications and resources in the field which is a crossover of big data analysis and crime. Recognising and understanding crime patterns, understandably may be exploitable and utilised in crime prevention, detection and mitigation efforts.

Crime is an apparent problem across all cities. If it were possible to prevent the number of crimes occurring by using state of the art technology, one would assume the quality of city living experience would improve. Crime occurrences in a city impact the human experience in many ways. For example, according to a study analysing the correlation between the dollar evaluation of property and crime shows that they're negatively correlated. As crime increases, the monetary value of the home decreases (Ceccato and Wilhelmsson, 2019). Not only does crime have diminishing side effects on its inhabitants, but also on governmental monetary budget. It is estimated that crime costs the government of the United States annually anywhere between 690 billion dollars to 3.41 trillion dollars (WatchBlog: Official Blog of the U.S. Government Accountability Office, 2019).

By applying state of the art technology from the field of data science, many different research groups, and organisations in the field have put forward proposals to better understand crime patterns in a bid to mitigate the impact of crime in cities. For example, there exists a company namely Predpol which focuses solely on using data science tools to produce meaningful insights in the field of crime.

II. Objective of Project

The objective of this project is segmented into two different parts, part A and part B. The first and primary objective of this project is to apply machine learning algorithms to predict the type of crime that will take place based on various variables. The second part of this project is comprised of an exploratory analysis into the dataset of crime using statistical techniques to derive meaningful insights.

III. Skills and Knowledge Applied

Part A

During my course at the National College of Ireland, more specifically, in my data and web mining module, I have been exposed to the application of various supervised and unsupervised machine learning algorithms and have learned how to apply them. Due to the nature of this problem, which is a classification problem, I have only utilised classification based supervised machine learning algorithms I have learned. In this analysis, I have applied three different machine learning algorithms that are widely used in solving such problems. These three mentioned are Decision Trees, Naive Bayes and Support Vector Machines. Furthermore, the main technology used in this analysis was the statistical programming language R. Almost all of our modules required us to use R to carry out module related coursework. This has enabled me to gain a certain level of proficiency with the tool. Prior to this program, I did not have these skills mentioned.

Part B

Part B is mostly comprised of applying statistical analysis which I learned throughout the statistical modules covered in my degree. I have learned how to use R to gather and mine insights from big data from mostly the introductory module we covered called programming for big data. I have applied statistical tests to compare the means of many groups which I learned from the advanced business statistics module. I have used various tests for testing normality.

IV. Additional Skills and Experience Required

Additional skills and attributes required to carry out the project consisted of learning how to call google maps API into R, an improved understanding of the ggplot package to produce more detailed R plots and general new code in R that was required to carry out my analysis.

V. Literature Review Surrounding Crime and Data Science

There already exists a myriad of different papers publicly available on the use of machine learning for the use case of predicting crime type. Taking one example, one researcher has used machine learning algorithms Decision Tree, Gaussian Naive Bayes, k-NN, Logistic Regression, Adaboost, and Random Forest classification models. This study's aim is to predict the correct crime type such as theft vs assault based on various variables. This study reaches highest accuracy scores such 70% with Random forest classifiers and Adaboost (Shama, 2017). Similarly, another paper analysing crime trends in Chicago have arrived at similar conclusions as the previous paper. This particular paper found classification trees as well as K-nearest neighbours which wasn't used in the previous paper to be the most effective in predicting crime type (McClendon and Meghanathan, 2015). According to another paper analysed, artificial neural networks and KNN clustering are utilised to predict crime type (Barnadas, 2016). From analysing different research papers on the topic, it can be concluded that machine learning algorithms can be quite accurate in predicting the type of crime that will take place in the context of a classification problem.

VI. Literature Review on Selected Machine Learning Algorithms

Support Vector Machines or SVM

Support vector machines are widely used in classification problems. Support vector machines are linear machines which aims to classify predictor variables by its location in relation to a hyperplane (Torgo, 2011). However, it should be noted that support vector machines can also be used in regression problems (Statsoft.com, 2019). Advantages of using support vector machines is that they're quite memory efficient and run quite quickly. Support vector machines generally do not perform well when the data set is very large (K, 2019). I chose support vector machines as one of my mining techniques as they're quite robust meaning they can handle any data types and are widely used in classification problems.

Formula for Support Vector Machines (classifiers)

$$\frac{1}{2} w^T w + C \sum_{i=1}^{N} \xi_i$$

Decision Trees

Decision trees predicts the label that is associated with an instance by arriving at the leaf node from the root or the base node. The successor child is chosen based on a splitting of the input node (Shalev-Shwartz and Ben-David, 2017). There are many variations of a decision tree. A very common decision tree is the algorithm namely the C.50 algorithm which can be installed and unpacked by downloading its corresponding library in R. Decision trees can be both used in regression and classification problems. There are multiple advantages of using decision trees. Firstly, it often does not require data to be normalised before it is fed into it. Nor does it always require scaling or pre-processing. Decision trees may take quite some time to run if the data is big (K, 2019). I selected decision trees as they're very robust and can handle all data types.

Naive Bayes

Naive Bayes is a probabilistic classifier based on Bayes theorem that relies on creating predictions based on assumptions. The name, naive, is derived from the algorithms nature in producing predictions based on assumptions (Torgo, 2011). Nevertheless, Naive Bayes predictor models are widely used in the field of data science. The Naive Bayes algorithm calculates the probability of each class for a given case by using the following mathematical formula. Some advantages include low RAM and CPU usage when running this model. It may achieve better results with smaller datasets as it has a low propensity to overfit while other classifiers may over fit such as decision trees. On the other hand, disadvantages include not being able to analyse feature interactions. It is assumed that naive Bayes works better with normally distributed data which may be quite limiting (Catanzarite, 2018). I chose naive Bayes as its a machine learning algorithm that is widely used in classification problems and its ability in handling categorical variables.

Formula for Naive Bayes classifiers

$$P(A|B) = \frac{P(B|A)\ P(A)}{P(B)}$$

VII. Data Understanding

In this section, I will explore the nature of the dataset, the data cleansing process undertook and the features selected for the machine learning algorithms.

Nature of the Dataset

The dataset studied is taken from Kaggle. This dataset is published by the City and County of San Francisco. The dataset contains 878,000 records by 9 columns (Kaggle.com, 2019). The columns are as follows; CATEGORY, DESCRIPT, DAYOFWEEK, RESOLUTION, ARRESTED, X, Y, DATES, PDDISTRICT. The column names are quite explanatory. However, columns "X" and "Y" are not so clear. This is longitude and longitude, which has enabled me to create graphs and plots based on geographical location. Column "DESCRIPT" I have avoided using completely as it was an arbitrary text-based description of the crime. "RESOLUTION" and "ARRESTED" I have also decided not to use.

Data Cleansing

There were a number of different modifications I made to the dataset. First of all, most importantly, I created additional columns by making column "DATES" into 4 different columns namely, time rounded to the nearest hour, day in number format (e.g. 1, 28), month in number format (e.g. 1, 4,12) and year in number format (e.g. 2014). Secondly, I removed data for 2015 as it was clearly incomplete. I was receiving skewed data constantly, and it was due to the reduced number of observations recorded in comparison to its previous years. Thirdly, as there were 39 levels of crime types in total, I decided it would be best to reduce the number of crime types to a more manageable number both for part A of this analysis and part B. This I did by simply finding the top 10 most frequently occurring crime type and filtered out crime types such as "OTHER" subjectively and replaced it with the next most frequent crime in order. For the machine learning algorithms specifically, I have created two additional columns called "MONTH_TEXT" and "DAY_TEXT" which is the month and day in text format (e.g. January, February, Monday, Tuesday).

Feature Selection

I have used the same features for all three machine learning algorithms applied. The features used in the training models are "Category", "DayOfWeek", "Month_text" and "PdDistrict". After studying the data and testing it for normality, the category of crimes are generally quite normally distributed when summarised by month, day of the week and PD-district which would suggest these variables may be poor independent variables. I avoided using longitude and latitude as from testing it, it was taking quite a long time for the algorithms to run due the high number of levels.

VIII. Machine Learning

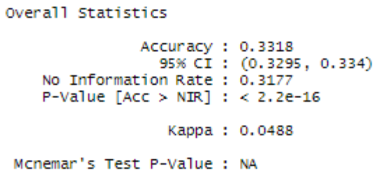
For the training models, I have applied the same process for splitting the dataset. I split the main data frame into both train and test in the ratio of 3:7. However, it must be noted that a much smaller sample was taken for the support vector machine learning algorithm as it was simply taking too long for it to learn. Confusion matrix was derived for each machine learning algorithm to compare the accuracy scores. The accuracy rate is calculated by adding up the correctly predicted values divided by the total number of predictions. These applied machine learning algorithms were selected based on the features and the nature of the problem. As it was a classification problem, the models applied are all suitable to make predictions for a classification problem. The kappa score of each result will also be examined to better understand the predictive capability of the model. A kappa score is used to test inter-rater reliability. A high kappa score will tell us that there exists a strong agreement between correctly predicted values and true values. In other words, it will tell us if the predictions were simply by chance or not. Kappa scores are well suited to evaluating multi-classification problems such as this one (McHugh, 2012).

Support Vector Machines - Model Evaluation

The support vector machines achieve an accuracy rate of 32.9%. The support vector machine achieved the lowest accuracy rate out of all three machine learning models.

Decision Tree - Model Evaluation

The decision tree model applied is the c50 model which is readily available on R by downloading and unpacking the required package. The decision tree achieved an accuracy rate of 33.18% which puts the decision tree c50 model best at predicting the crime type. The kappa score derived is quite low suggesting low levels of agreement.

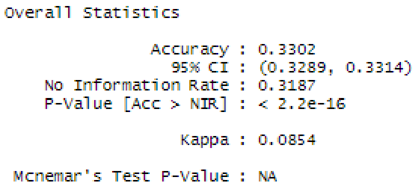


```
Overall Statistics

               Accuracy : 0.3318
                 95% CI : (0.3295, 0.334)
    No Information Rate : 0.3177
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.0488

 Mcnemar's Test P-Value : NA
```

Graph 8.1

Naive Bayes - Model Evaluation

The naive Bayes model comes second place in terms of accuracy rate achieved with an accuracy score of 33%.



```
Overall Statistics

               Accuracy : 0.3302
                 95% CI : (0.3289, 0.3314)
    No Information Rate : 0.3187
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.0854

 Mcnemar's Test P-Value : NA
```

Graph 8.2

Data Mining Results Discussion

It is apparent that none of the three applied machine learning algorithms have achieved high accuracy scores. This may deem the models as unusable in a real-life scenario as usually accuracy scores of much higher would be considered. In order of highest accuracy score achiever, the decision tree followed by naive Bayes and finally the SVM models were the algorithms applied.

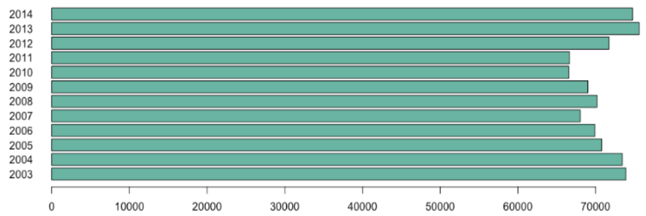
Further Testing and Reflection

For further analysis, I would like to experiment with pruning as a method of achieving better accuracy rates with the decision tree model. Furthermore, I would like to experiment with different independent variables and in different formats. For example, I would like to test using categorical data in numeric fashion to test whether it would change results. On that point, I would like to experiment using longitude and latitude points as independent variables. Thirdly, I would like to experiment with other machine learning algorithms such as KNN, neural networks and random forest classifiers to understand and learn how they would perform.

IX. Data Exploration

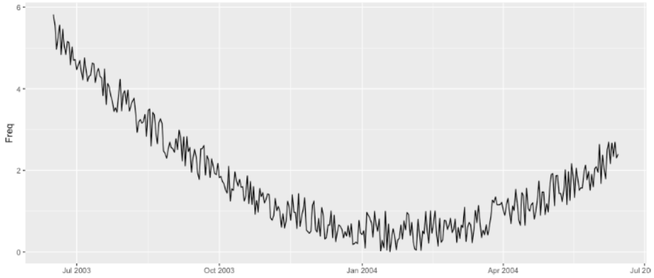
Year and Crime

The data has shown us that there is an upward trend in the number of crimes being committed in San Francisco. Even though the most recent year for crimes recorded is 2015, from plot (graph 9.1), one can see how the trend is positively rising upward. Other sources also agree on this point that the number of crimes committed have been on the rise and are expected to increase (Hoodline.com, 2019). From taking the last four years bar the year 2015 due to the lack of data collected in this particular year, it is observed that there was an increase in 0.12% increase in the number of crimes between the years 2010 and 2011, 7.7% between the years 2011 and 2012, 5.4% between the years 2012 and 2013, -1.11% between 2013 and 2014. We can see that the biggest spike occurred between 2010 and 2011. One could make many assumptions as to why this may have occurred. Studies examining the correlative relationship between crime and recession suggest that there exists a positive correlation meaning as recessions occur, the crime rates increase (Bell, 2015). These years were the height of the 2008 financial crisis. The standard deviation of the number of crimes occurring per year between 2010 and 2014 is 4,329, the mean number of crimes is 71,052 and after running a shapiro-test to test for normality, we can see that it is normally distributed as the p -value is greater than significant levels of alpha.



Graph 9.1

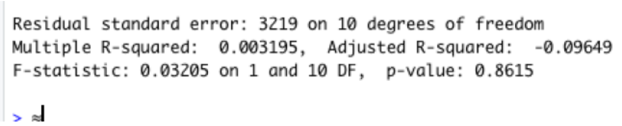
From the time series plot (graph 9.2), we can see that there was a decrease in the number of crimes between 2003 and 2004 and then an upward trend begins to start which is in line with the results of the bar plot.



Graph 9.2

Crime Rate Prediction using Linear Regression

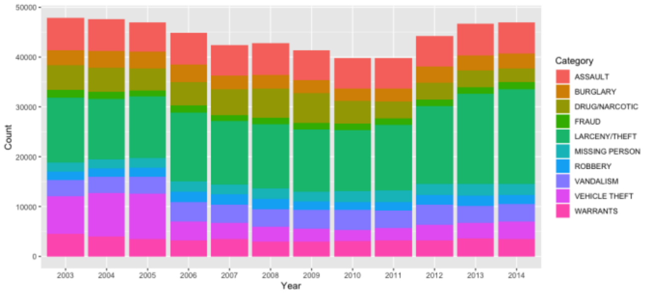
Linear regression works very well as a simplistic method of predicting an X value based on one singular independent variable. A linear regression achieves this by simply fitting a linear equation to the data which sits relative to both x and y values (Stat.yale.edu, n.d.). By applying a linear regression to the crime per year data frame which consists of year and the frequency of crimes occurring per year, we achieve the following linear equation. Freq = 70558.88 + 48.19*year. If we aim to predict the number of crimes that will occur in both 2015 and 2016 using the model, we arrive 71,185.35 and 71,233.54 crimes respectively. If we have a deeper look at the summary of our linear model, model1, we can see that the adjusted R squared is a negative. One could assume that this value for R adjusted squared was obtained by forcing the regression through specific points being numbers between 1 and 14 on the X axis or the year axis. The p-value being quite high suggests that changes in the predictor variable are not strongly associated with changes in the response variable (Blog.minitab.com, n.d.). By conducting a Cohen's correlation test, the test gives us a positive 0.05 which again assumes weak correlation. This tells us that the linear regression model is poorly fitted and it's likely that it will not produce accurate results.



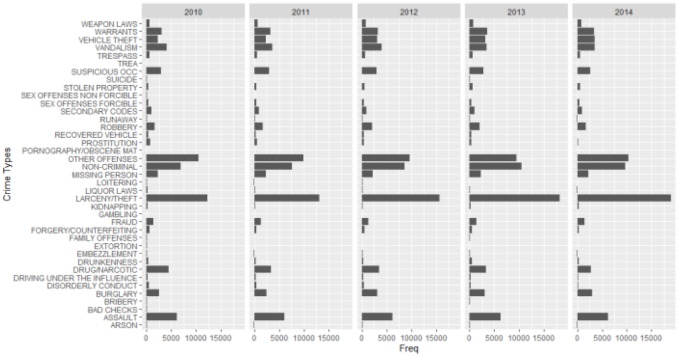
```
Residual standard error: 3219 on 10 degrees of freedom
Multiple R-squared:  0.003195,  Adjusted R-squared:  -0.09649
F-statistic: 0.03205 on 1 and 10 DF,  p-value: 0.8615

> 
```

Crime Category and Year

By creating a table for both categories of crime committed and year, we can observe the nature of crime category trends through the years. By color coding each category of crime, we can see that there exists a common trend amongst the crime. However, from by just observing plot (9.3), we can see that the number of vehicle thefts have greatly reduced. By having a deeper dive into the number of vehicle thefts that have occurred through the years, we can see that its distribution is not normal as it gave a P-value of 0.0018 from running a shapiro-test. In order to test whether the means of the observations are equal to each other, we need to find a suitable test. Because its distribution is not normal and there are more than 2 groups, we can use the Kruskal-Wallis rank sum test to test for levels of significance. The Kruskal-Wallis rank sum test tells us that we fail to reject the null hypothesis which is that all observations have equal means and that they're not significantly different. This would indicate that the number of vehicle thefts have indeed decreased over the years and there is little expectation that the trend will deviate much. It would be interesting to find out if the San Francisco police department deployed any special measures to reduce the number of vehicle thefts from year 2005.



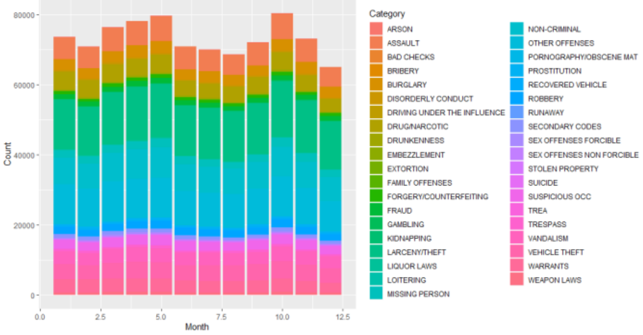
Graph 9.3



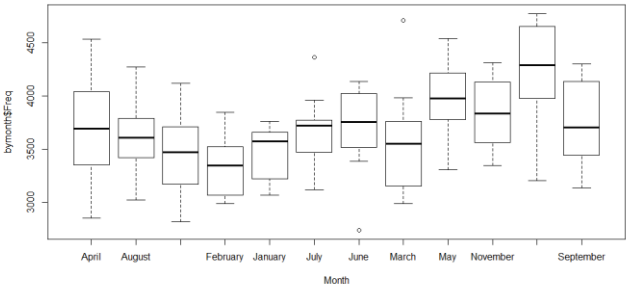
Graph 9.4

Month and Category of Crime

By creating a table consisting of categories of crime and month, we can gain insights into types of crime committed over the course of 12 months. By observing this plot (9.4), we can see that there doesn't seem to be great deviation in the number of crimes committed per category per month. Just by observing the colour coding, one can see that the colour bars are quite parallel to each other. Secondly, we can see that the 10th month or October witnesses the greatest number of crimes in a year and December the least number of crimes committed. What would be interesting is to make some assumptions as to why crimes formulate such trends when summarised by months of the year. Topics may include, weather, daylight savings, GDP levels per month to name a few. We can see from the box plot (9.5), that similar trends also exist.



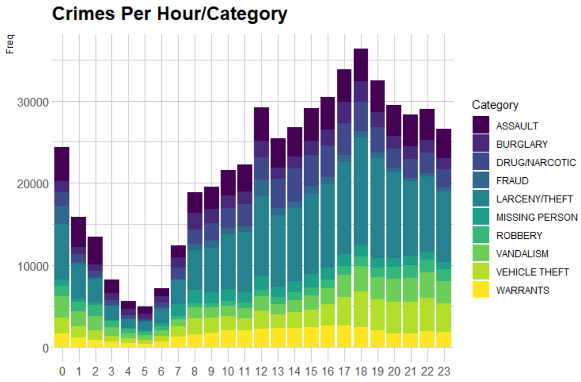
Graph 9.5



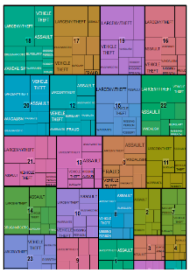
Graph 9.6

Hour and Crime Relationships

Hour and crime relationships can produce quite meaningful insights. From plot (9.7) not only can one recognise the frequency of crimes occurring per hour, but also the number of crimes committed per category of crime at different hours. It can be observed that a crime is least likely to take place in the early hours of 5:00 am and most likely to occur in the evening time 6:00 pm There is a 0.36 probability that a theft will occur at this hour and is 12.1 times more likely to occur at this hour when compared to the number of thefts occurring at 5:00 am. By observing the tree-map (9.8), one can see which crimes are most and least likely to occur at each hour. However, some drawbacks of this method of visualizing data includes restricting numeric data representations which makes it difficult to make objective arguments. By taking the whole distribution of crimes and testing it for normality, we receive a p-value of 2.2e-16 which is significantly low meaning non-normal. Plot (9.7) also shows us visually that it’s not normal. Clearly, the hours of day, have a significant impact on the number of crimes that will be committed. After applying a shapiro-test to the frequency of crimes occurring grouped by its respective category, we can see that three different crimes types are not normally distributed which would suggest large spikes if mapped out against the hour the crime takes place. These three crime types are assault, missing person and fraud. In plot (9.7), we can see that assaults more often take place in the evening and late at night and least likely to occur early in the morning like mentioned already. Taking another crime category example which have very contrasting number of times committed based on the hour is missing person. A missing person case is 12.6x times more likely to occur at 8:00 am in comparison to 4:00 am in the morning. This information may enable law enforcement officers to act accordingly to such cases.



Graph 9.7

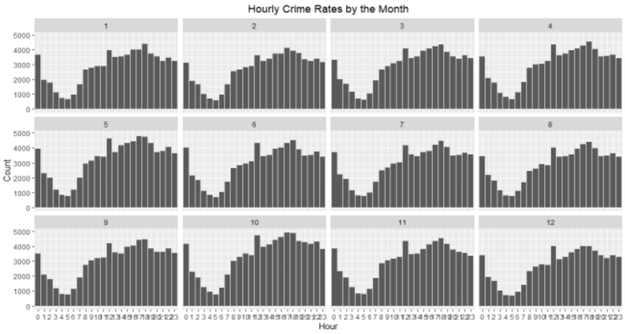


Graph 9.8 – Treemap of crime per hour

By applying an Anova test, we obtain a p-value that is significant. We have evidence to reject the null hypothesis that the means of the number of crimes committed per hour are equal. After performing a post hoc Tukey test, the greatest variation exists between the early hours of the morning between 2:00 am and 8:00 am afternoon to evening hours which is also evident from plot (9.7).

Hour and Month Categorisation

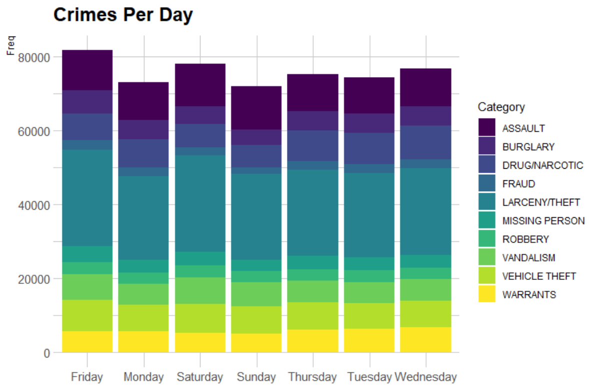
Just from observing plot (9.9), the number of crimes that occur grouped by month and hour, it seems that each month follows similar patterns. As there is no month that does not appear to follow the same trend, further statistical investigations have not been carried out.



Graph 9.9

Day of Week

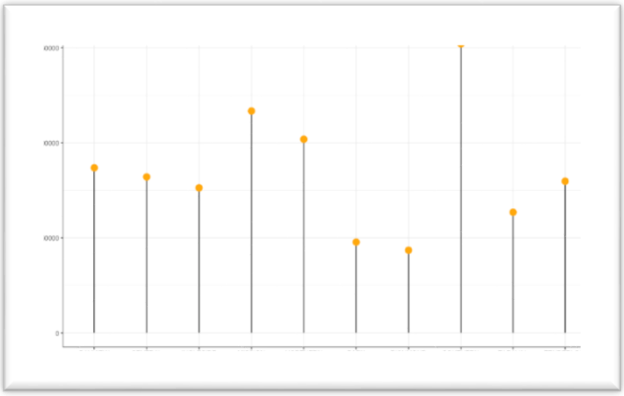
According to the data, 15% of all crimes occur on a Friday. This trend is graphically represented in plot (9.10). By applying a shapiro-test to test for normality we obtain a p-value 0.9889 which suggests high normality. Let’s now incorporate crime category into the equation to understand crimes committed per day and type. The category of committed crimes across the different days appear to be quite normally distributed. This would suggest why perhaps using day of the week as a feature in a machine learning model to be a poor choice. Whereas hour has clearly an impact on at least 3 different crime types significantly.



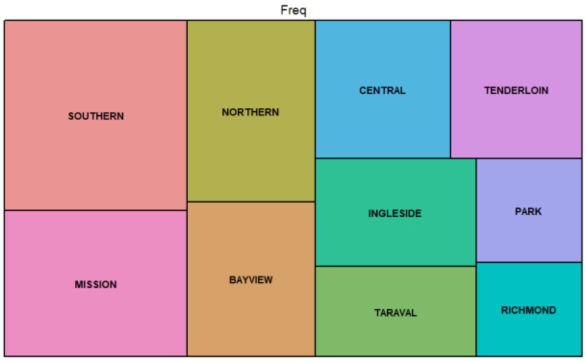
Graph 9.10

Crime and Location

From observing plot (graph 9.11), we can see clearly there are more crimes reported to police department district southern. Plot (graph 9.12), is an alternative way of presenting the same data. By analysing the type of crime committed per district, we have access to interesting data. From plot (see in images) we can see that the type of crime committed per district may differ. This could be interesting as it would suggest that different crimes happen in varying frequencies depending on the district. By applying shapiro-test to test for normality, it turns out, 60% of districts have non-normal distributions of crime category taking place. By applying an Anova test, we can see that the crime category has a significant impact on the number of crimes committed which allows us to reject the null hypothesis that the mean number of crimes committed per category and district have equal variance and means. This would suggest that PD-district could be a good variable to use in machine learning algorithms.



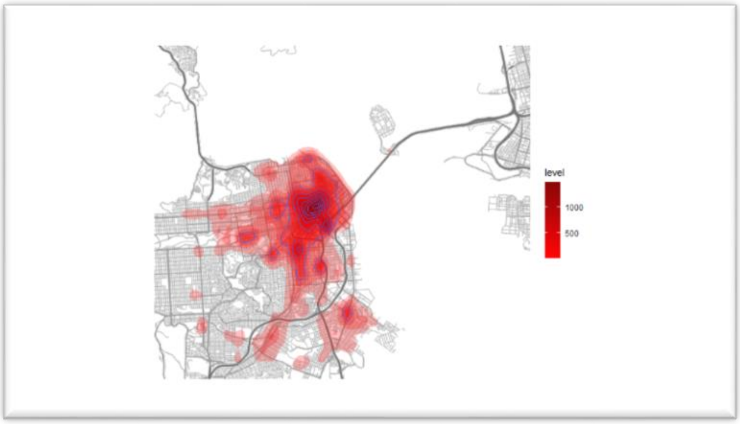
Graph 9.11



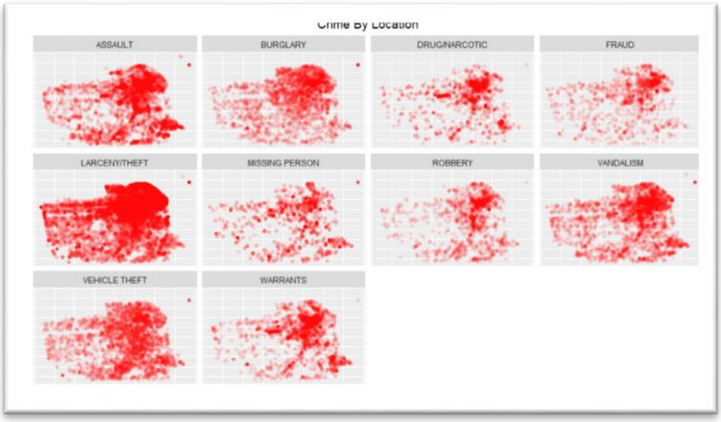
Graph 9.12

Maps and Crime

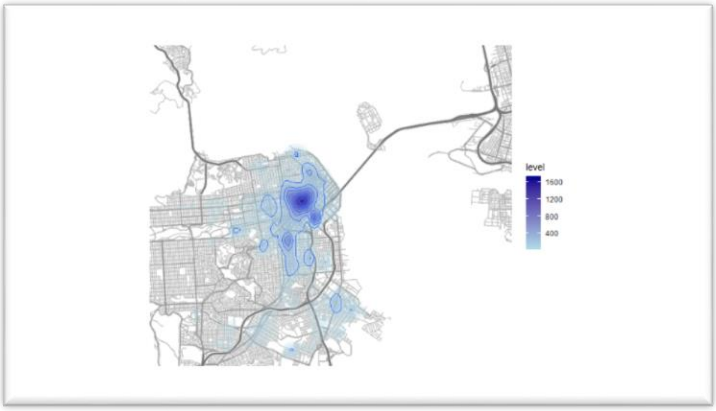
Interestingly, plots (*Graph 9.13*) and (*Graph 9.14*) both show that majority of crimes take place right in the city centre of San Francisco. On the other hand, plot (*Graph 9.14)* conveys the type of crime occurs more frequently due to location. Clearly, assaults take place more often than say warrants.



Graph 9.13



Graph 9.14

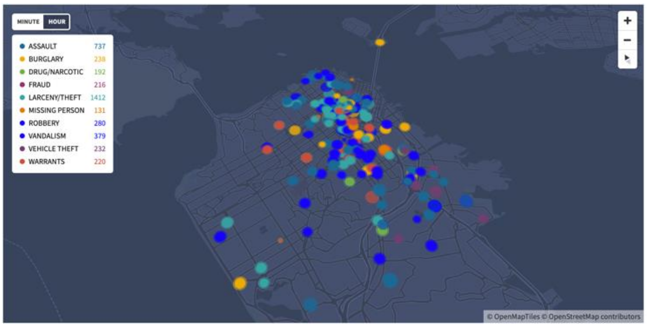


Graph 9.15

X. Conclusion

Part A - Machine Learning

Accuracy rates as low as what has been achieved would suggest a re-visit would be mandatory. Like mentioned already, I would like to explore methods of making models more precise. I would also experiment with other features. For example, from applying statistical methods to analyse data, we have witnessed that the category of crime across days is quite normally distributed. I have learned that many of the insights gathered from part B would have been beneficial in the feature designing and engineering process for the machine learning models.



Graph 9.16 – Animated Plot of Crimes in 2014
https://public.flourish.studio/visualisation/1020515/

Conclusion - Part B - Data Exploratory using Statistical Methods

It is evident that a lot of exploitable insights have been gathered from this dataset. The possibilities for finding insights are quite endless. What would help is having a strict goal from the beginning with certain questions to help the data analyst to remain on track.

XI. References

[1] WatchBlog: Official Blog of the U.S. Government Accountability Office. (2019). How Much Does Crime Cost?. [online] Available at: https://blog.gao.gov/2017/11/29/how-much-does-crime-cost/ [Accessed 2 Dec. 2019].

[2] Keenan, M. (2018). Home truths: On crime and punishment for property values - Independent.ie. [online] Independent.ie. Available at: https://www.independent.ie/life/home-garden/home-truths-on-crime-and-punishment-for-property-values-37267789.html [Accessed 2 Dec. 2019]

[3] Ceccato, V. and Wilhelmsson, M. (2019). Do crime hot spots affect housing prices?. Nordic Journal of Criminology, pp.1-19.

[4] Shama, N. (2017). A Machine Learning Approach to Predict Crime Using Time and Location Data. [online] Available at: https://pdfs.semanticscholar.org/3bb0/40430edf0ffdef0c93dadc04dcd7e6905637.pdf [Accessed 2 Dec. 2019].

[5] McClendon, L. and Meghanathan, N. (2015). Using Machine Learning Algorithms to Analyze Crime Data. Machine Learning and Applications: An International Journal, 2(1), pp.1-12.

[6] Barnadas, M. (2016). MACHINE LEARNING APPLIED TO CRIME PREDICTION. [online] Available at: https://core.ac.uk/download/pdf/81577388.pdf [Accessed 2 Dec. 2019].

[7] Torgo, L. (2011). Data Mining with R: Learning with Case Studies. 1st ed. Taylor and Francis Group, LLC.

[8] Statsoft.com. (2019). Support Vector Machines (SVM). [online] Available at: http://www.statsoft.com/textbook/support-vector-machines [Accessed 2 Dec. 2019]

[9] Shalev-Shwartz, S. and Ben-David, S. (2017). Understanding machine learning. Cambridge: Cambridge University Press.

[10] Torgo, L. (2011). Data Mining with R: Learning with Case Studies. 1st ed. Taylor and Francis Group, LLC.

[11] McHugh, M. (2012). Interrater reliability: the kappa statistic. [online] PubMed Central (PMC). Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900052/ [Accessed 2 Dec. 2019].

[12] K, D. (2019). Top 4 advantages and disadvantages of Support Vector Machine or SVM. [online] Medium. Available at: https://medium.com/@dhiraj8899/top-4-advantages-and-disadvantages-of-support-vector-machine-or-svm-a3c06a2b107 [Accessed 2 Dec. 2019].

[13] K, D. (2019). Top 5 advantages and disadvantages of Decision Tree Algorithm. [online] Medium. Available at: https://medium.com/@dhiraj8899/top-5-advantages-and-disadvantages-of-decision-tree-algorithm-428ebd199d9a [Accessed 2 Dec. 2019].

[14] Catanzarite, J. (2018). The Naive Bayes Classifier. [online] Medium. Available at: https://towardsdatascience.com/the-naive-bayes-classifier-e92ea9f47523 [Accessed 2 Dec. 2019].

[15] Hoodline.com. (2019). Crime is on the rise in San Francisco — here are the latest trends. [online] Available at: https://hoodline.com/2019/04/crime-is-on-the-rise-

in-san-francisco-here-are-the-latest-trends [Accessed 2 Dec. 2019].

[16] Bell, B. (2015). Do recessions increase crime?. [online] World Economic Forum. Available at: https://www.weforum.org/agenda/2015/03/do-recessions-increase-crime/ [Accessed 2 Dec. 2019].

[17] Stat.yale.edu. (n.d.). Linear Regression. [online] Available at: http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm [Accessed 2 Dec. 2019].

[18] Blog.minitab.com. (n.d.). How to Interpret Regression Analysis Results: P-values and Coefficients. [online] Available at: https://blog.minitab.com/blog/adventures-in-statistics-2/how-to-interpret-regression-analysis-results-p-values-and-coefficients [Accessed 2 Dec. 2019].

[19] Kaggle.com. (2019). San Francisco Crime Classification | Kaggle. [online] Available at: https://www.kaggle.com/c/sf-crime [Accessed 3 Dec. 2019].