UK Police Crime Data Analyses

Pre-processing and EDA

Initially following major preprocessing steps were taken test our claims:

1. **Merging:** Primary dataset (all_crimes_18_hdr) was merged with the dataset 4 (LSOA_pop_v2) based on LSOA feature. The resulting dataset now had all the necessary features like crime type, their locations (district wise) and the population in that area.
2. **Missing values:** Rows with missing LSOA id were simply dropped instead being filled using techniques like Backfill etc., because when the dataset is merged based on LSOA feature, any missing value in LSOA would imply missing values in population, location, and other features as well. A justification for simply dropping these rows is that they account for about 3% of the original data and would have a negligible impact on the bias and variance of the data.
3. **Location and Crime Type:** We did a simple chi-square test to determine whether location and crime-types were correlated. Null hypothesis was rejected implying that crime type depends on the location. In addition, we listed all the locations and the crime that occurred most frequently in that location. We found that either “Anti-social behavior” or “Violence and sexual offences” was the most frequent, given any location.

Claim 1: (True) The violent crime is increasing with time.

The claim seems to be true. Even though the data suggests that crimes follow a cyclical pattern in which crime tends to be high in the months of June, July and august, a simple glance at the peaks of these cycles suggests that every consecutive year the number of crimes tend to be higher. Additionally, the troughs of each cycle seem to be increasing with time proving that violent crime also increases with time. Irrespective of the monthly fluctuations, there seems to be a general year on year increase in the trend of crime rates. This can be seen in the figure 1.1, given below:

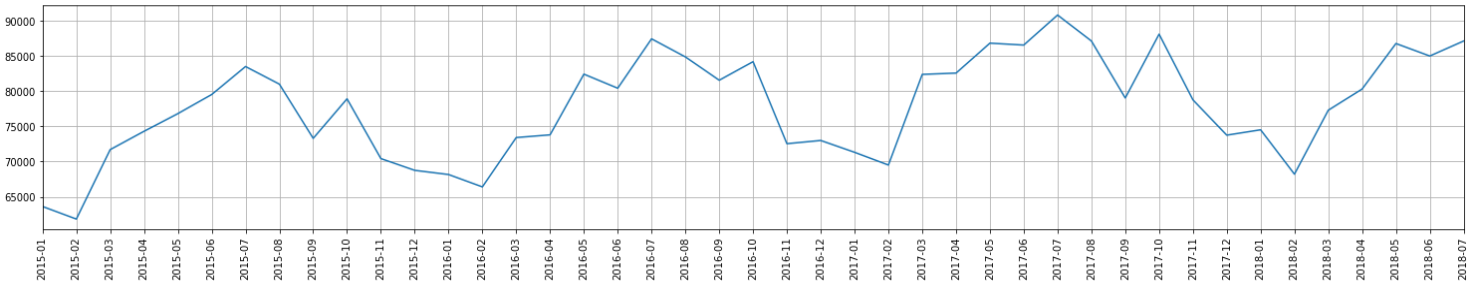


Figure 1.1: Line graph between crime rate and time.

Claim 2: (True) In Birmingham per head crime rate is higher than anywhere else in UK.

Eyeballing the histograms (figure 1.2) plotted below, it is easily understood that crime rate in Birmingham stands at around 60459, which is way higher than any of the other locations. Only Manchester and Leeds come close to competing with the number of crimes in Birmingham, but they still fall short.
The red line in the following graphs show the frequency of crimes in Birmingham:

Figure 1.2: Histogram of crime types and locations in UK

Prediction, Clustering and Outlier Analyses

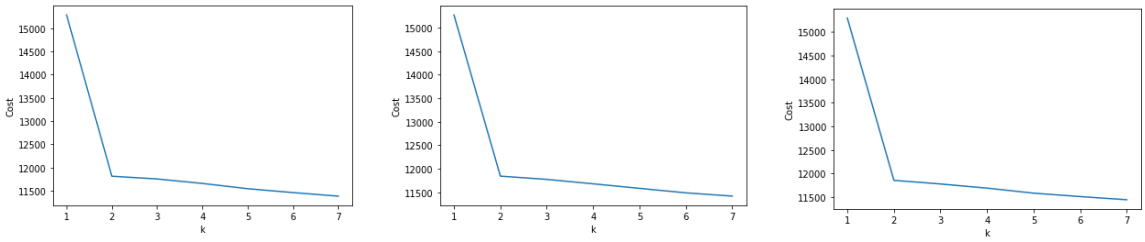
1. Perform cluster analyses on the data, primarily on location and crime types:

For cluster analyses, we used K-modes algorithm for clustering based on location and crime types. Both variables being categorical, made K-modes a very viable choice for clustering here. To determine the optimal value of K, we first randomly sampled 10,000 rows and ran the clustering for different values of K, and then plotted the graph. After which elbow method was used. We did this three times to ensure convergence to the real value of K. This gave out 2 as the optimal value of K.

Analyses of both clusters:

Examining both the clusters, cluster 1 has significantly high levels of crimes. For example, 'Violence and sexual offences" is about 60 times higher in cluster 1 than in cluster 2. Same can be seen for other crimes in cluster 1 when compared to cluster 2. We note that locations falling in cluster 2 has a very high rate of weapon possession. While in cluster 1, Anti-social behavior tops the list of crimes committed. Cluster 1 has also high levels crimes that fall in the category of 'Violence and sexual offences' and 'Criminal damage and arson'. It's worth noting that locations falling in cluster have no instance of "Anti-Social Behavior". Weapon possession in cluster 1 contributes the least to the sum of all crimes in locations belonging to cluster 1. An interesting theory originates from the cluster, that Higher Weapon possession made induce an effect which reduces the other types of crime in those locations. Cluster 1 can be regarded as "high threat zones" because of the intensity of crimes. While locations in cluster 2 can be considered relatively safer even though Weapon possession in cluster 2 is relatively high.

Elbow method:



['Anti-social behaviour', 'Violence and sexual offences', 'Criminal damage and arson', 'Drugs', 'Robbery', 'Possession of weapons']

```
[49] newdf[newdf['cluster_id2']==0]['crime_type'].value_counts()

0    1512153
1     483560
2     117363
4      52594
3      28393
5      22328
Name: crime_type, dtype: int64
```

```
newdf[newdf['cluster_id2']==1]['crime_type'].value_counts()

5    1118791
1       7974
4       2176
2       1465
3        679
Name: crime_type, dtype: int64
```

2. Given a crime predict its outcome:

For predicting the outcome of a crime. We used the variables listed below, which are a mix of numeric and categorical type.

1. crime_type (Categorical) Values: 0-5
2. Variable: Males; measures: Value (Numeric)
3. Variable: Females; measures: Value (Numeric)
4. Variable: Lives in a household; measures: Value (Numeric)
5. Variable: Lives in a communal establishment; measures: Value (Numeric)
6. Variable: Schoolchild or full-time student aged 4 and over at their non-term-time address; measures: Value (Numeric)
7. Variable: Density (number of persons per hectare); measures: Value (Numeric)
8. Region (Categorical)

After which we first converted the numerical variables into categorical using binning techniques. Due to computing power constraint, we selected 3 bins for each numerical variable. Then, using a train – test split of 20%, we trained a decision tree classifier based on entropy. Like the ID3 algorithm it chooses roots recursively based on the variable which has the least Entropy. This classifier would classify an outcome given the above variables. Outcome variable had missing values which were filled by "No information" as the missing values itself had meaning. To justify this, we can think of a crime type whose outcome usually is ambiguous and then it is given a missing value. The model had an accuracy of 73% and can output the outcome of a given crime.

3. Additional tasks that could have been done.

We used only those variables which gave us the most important information. But this doesn't mean that all the variables that were dropped were useless. Each variable had a purpose. For the variable time, we could examine the patterns like, in which segment of time would the most crimes occur. Similarly, we could also find the context of the most violent crimes depending on locations, by using context variable. Also using the locations variable, we could map each crime to a specific location and make a map which depicted hotspots of each type of crime. This map could also by dynamic in the dimension of time, changing throughout the day. This would utilize a lot of computing power and hence require a lot of time. This is why this technique of svisualizing the patterns was not done under the given time constraint.

4. Interesting findings:

a. **Cyclic nature of the crimes:** As shown in the figure above in claim 1, crimes follow a cyclic pattern in UK. According to the data, there is an increase in the number of crimes in the summers of 2015, 2016, 2017 and 2018. On the other hand, the number of crimes decrease during winters.

b. **Relationship between crime and population**: It can be seen from the four graphs in part two that the following areas have the highest crime rates in UK: Birmingham, Manchester, Leeds, Bradford and Liverpool. These areas with crime rate happen to be the most populated areas in the UK. This shows that population and crime rate are highly correlated in UK.