Classifying Crime Risk Using the K-Nearest Neighbor Classifier City of Leeds, England.

IBM Capstone Assignment

Justin Simpson

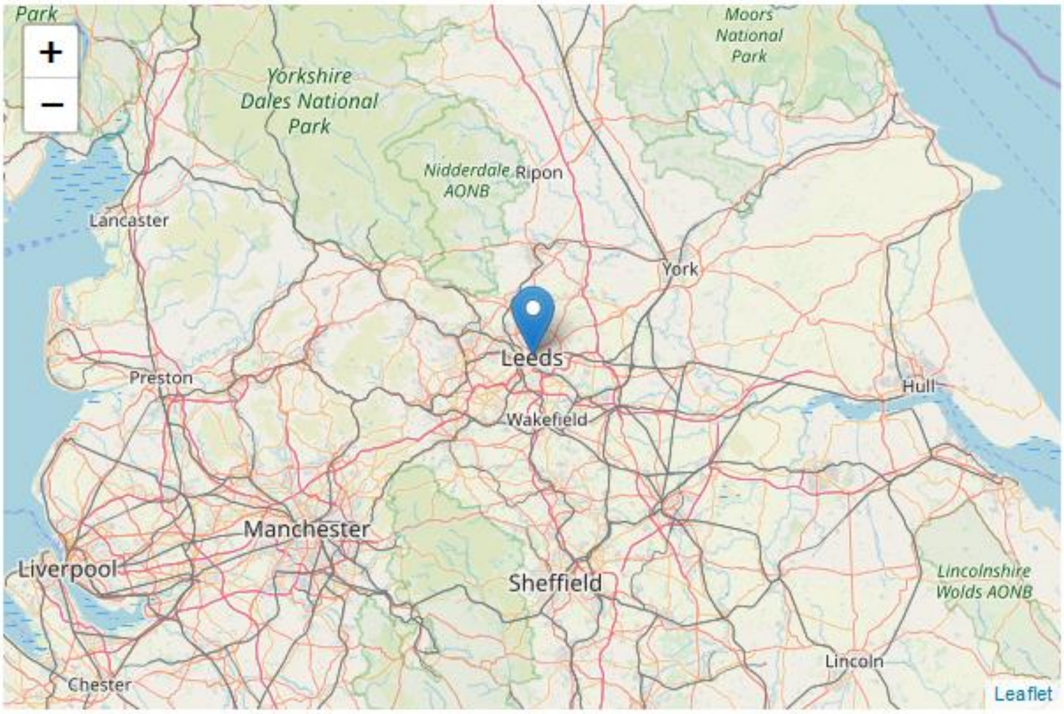
Table of Contents

1. Introduction

The city of Leeds and surrounding areas are located in the North of the United Kingdom within the county of West Yorkshire, approximately 170 miles north of London. Leeds has a diverse economic base and has a large multicultural population.

Neighbourhoods within the Leeds area vary with respect to their composition in relation to both the levels and type of crime reported and the numbers and type of facilities, businesses and land use types that exiat in any given location.

The City Of Leeds



2. Business Problem

An identifiable business problem relates to the ability to classify any specific geographic location within the Leeds area according to potential levels of crime based on the composition of the surrounding area. In other words:

"Can we determine if an area around a specific geographic point is likely to have an above or below average risk of crime, based on the numbers and types of locations surrounding it ?"

The ability to answer this question could benefit a range of potential stakeholders. People looking to move home could assess what the levels of crime would be at a potential address. Local Police could identify crime hotspots, while planning authorities could determine if any new developments or change of use applications would, lead to increased (or decreased) crime in the vicinity.

3. Data

In order to answer the question posed, multiple data sets were required. These data sets related to crime within Leeds, geographic data relating to Leeds and location type details. To answer the question we need to obtain the number and types of venues and crimes which fall within a set distance of a geographical point.

Once retrieved and processed, the data will be used to create a classification model using the K-Nearest Neighbour classifier. The objective of the model is to classify any geographical point as high or low risk in relation to exposure to crime.

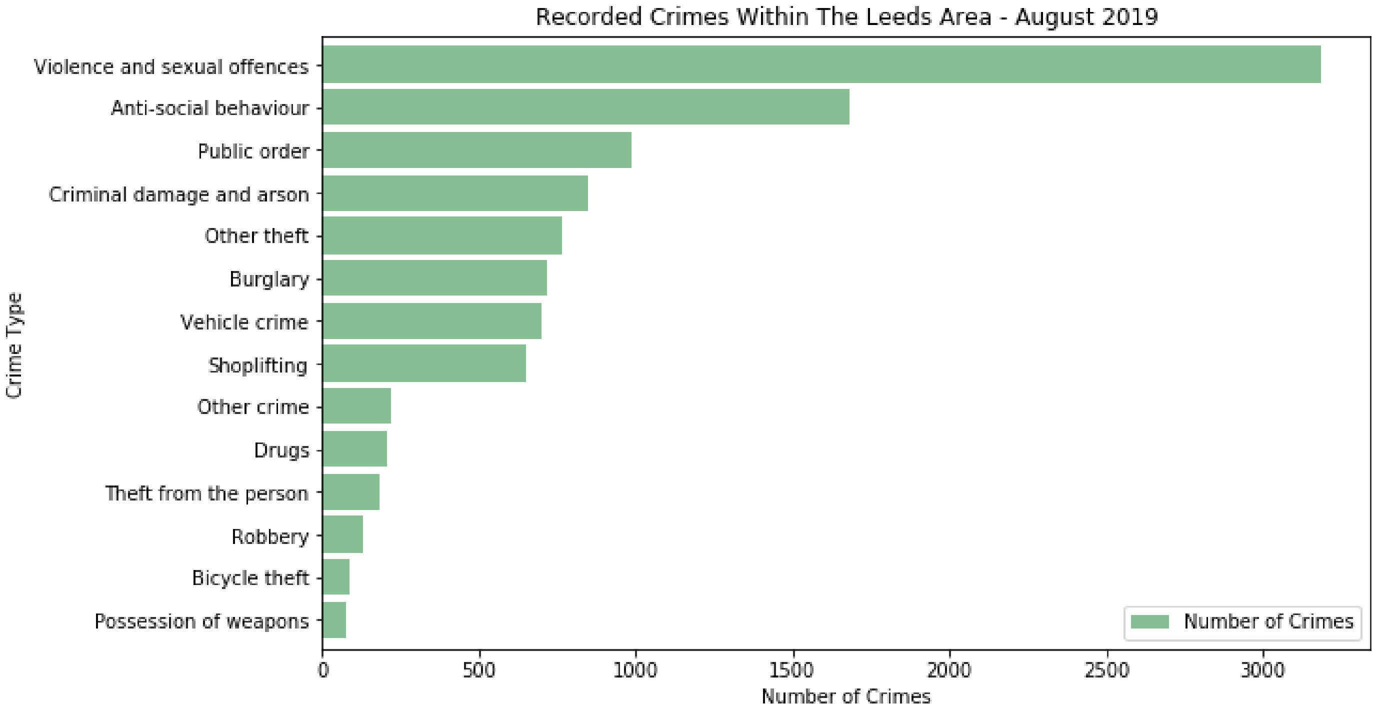
Crime Data:

Crime data relating to Leeds was obtained online from publicly available datasets published by the police.
(https://data.police.uk/data/)

This data was supplied by Police Force area and provided information with respect to the type of crime, the latitude and longitude of the crime and local area (LSOA) that the crime occurred within. The data was also supplied based on Month, in this case the most recent data-set was chosen, relating to August 2019.

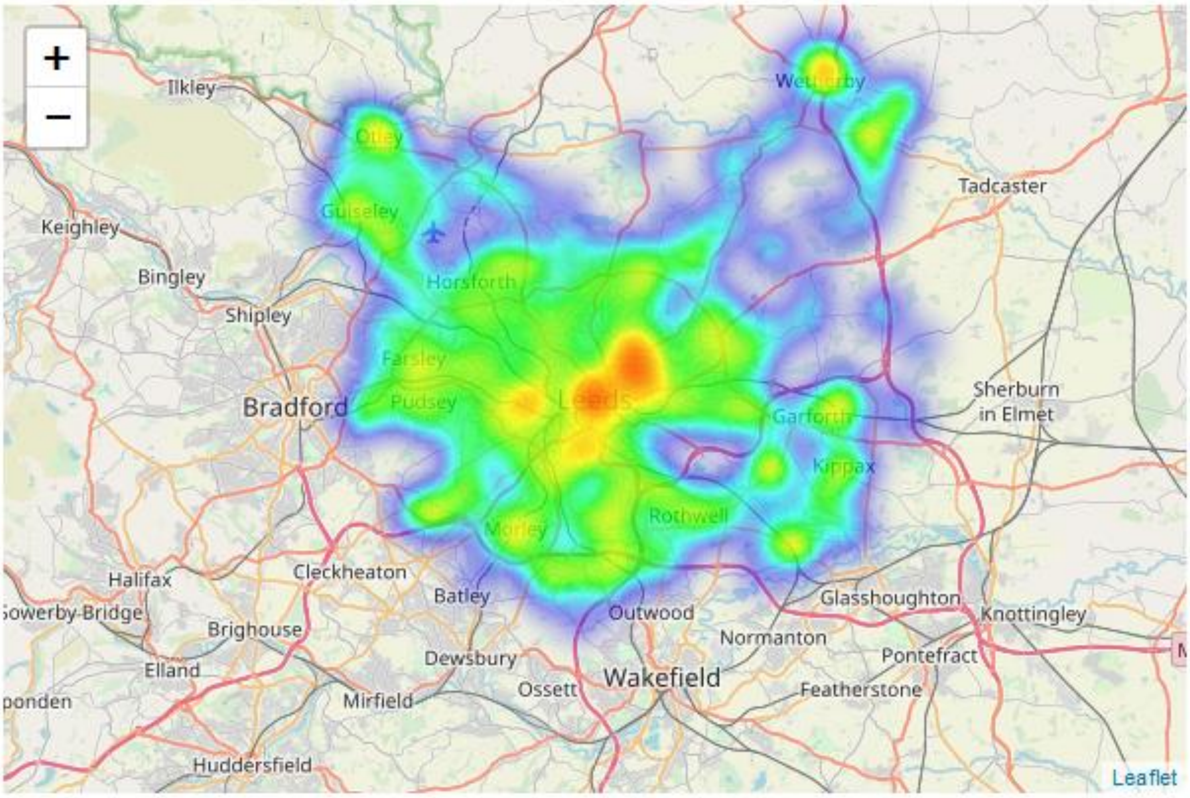
As data relating to the whole of West Yorkshire was included in the data-set, crimes relating to the Leeds areas had to be extracted, which was done based on the area codes that made up the Leeds area. The data was checked for missing values in any required columns (none were found)

The data-set contained details of 10,441 crimes, grouped by 14 distinct crime classifications. Details of the breakdown of these crimes is shown in the chart below, which shows that Violence and Sexual Offences were by far the most common crimes recorded..



The distribution of these crimes was also variable with urban centre areas showing higher concentrations of crime, especially in the Leeds city centre areas. The distribution of crime is shown on the heat-map below. The orange, yellow areas having the highest crime levels and the blue the lowest.

Distribution of Crime Within Leeds – August 2019



Leeds Geographic Data:

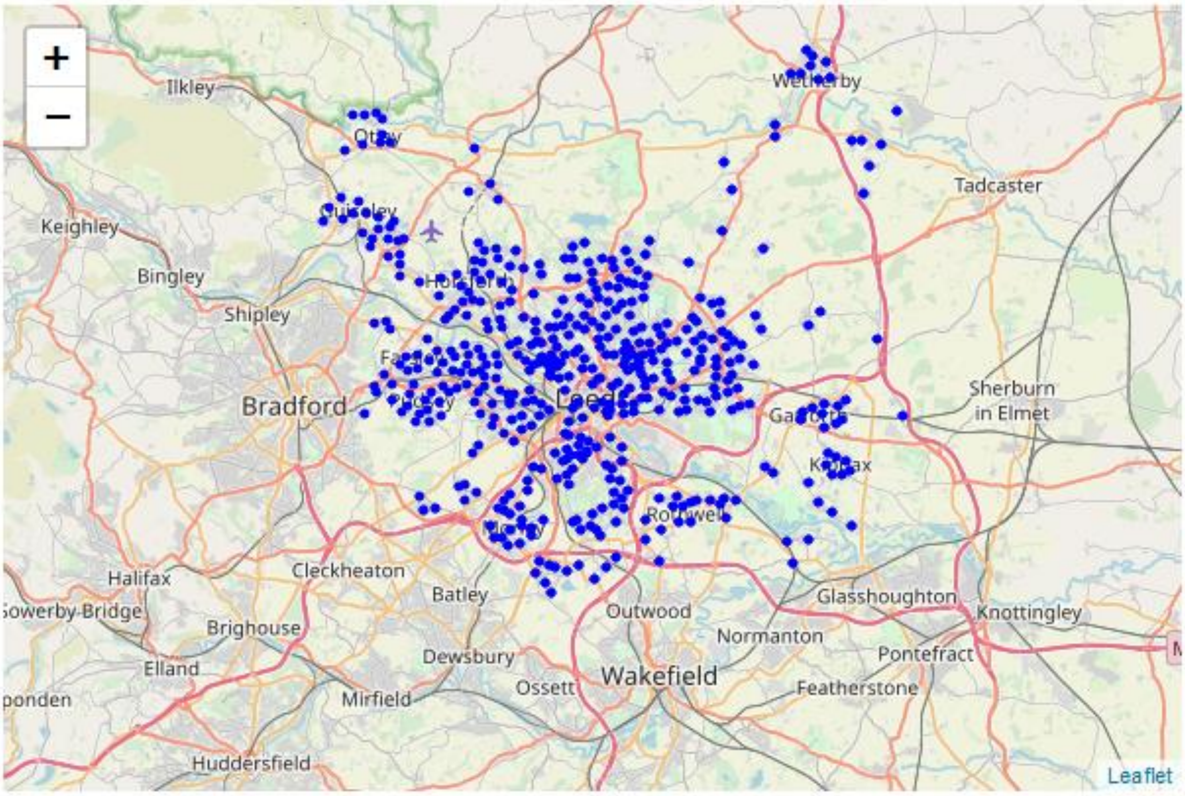
Publicly available Geographic Data relating to Leeds was obtained from The Office for National Statistics (https://geoportal.statistics.gov.uk/datasets/). This data comprised of geographic locations for the centroid of each LSOA (Lower Layer Super Output Area) within the Leeds area.

The data provided the locations of 482 geographic points within the Leeds area. Each LSOA code was unique and this was combined with a further set of data which assigned more user friendly Electoral Ward names to the area (Although each ward could consist of many LSOA codes, it provided a reference that would be potentially more meaningful).

These centroid locations were used as they fell within the geographical area that this project covered and related to already established points. The distribution of these points can be seen in the following map.

These points will be used as the basis for future analysis and comparison between areas.

Distribution of Geographic Points Within Leeds



Crimes within 500m Radius of Geographic Point

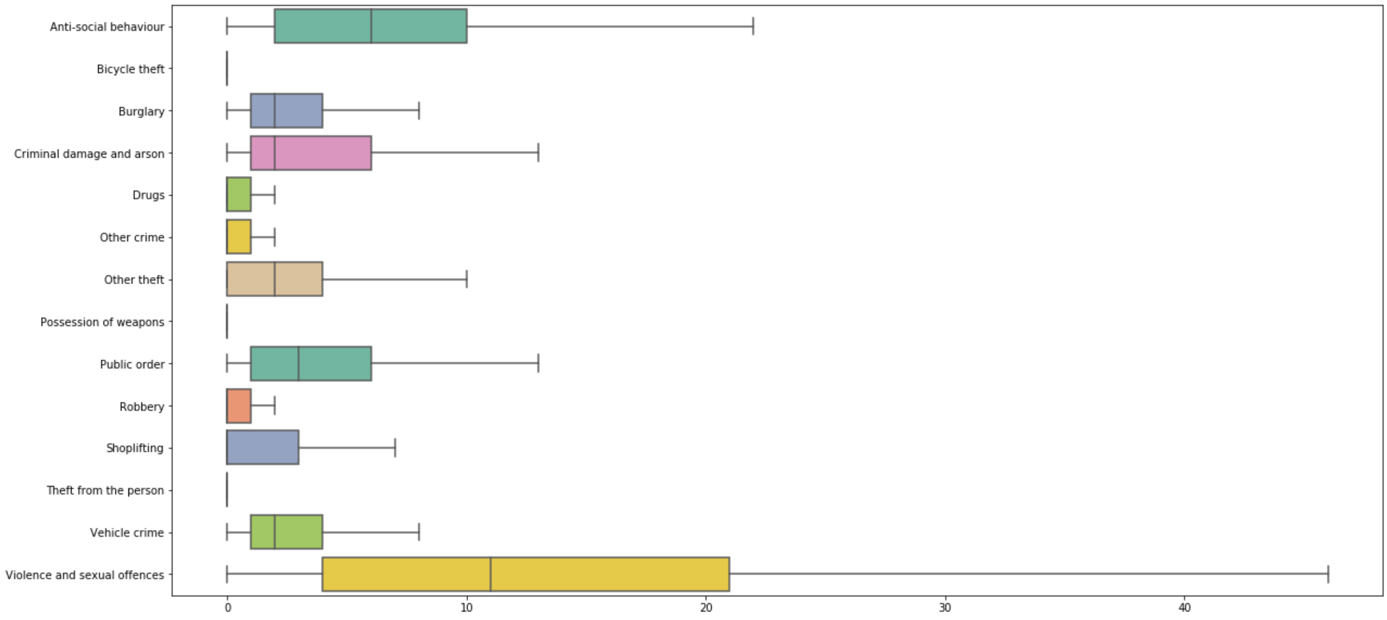
The next stage of data manipulation was to determine the number and types of crimes that fell within a radius of a given point. For each geographical location the crime data was iterated over, and if the crime fell within a radius of 500m from the geographical point, it was added to that points

(LSOA) data.

The data was then manipulated and aggregated in order for each LSOA point to be shown as a row and crime types as a columns. Showing the number of crimes by type within 500m of each geographic point..

The distribution of these crimes by area can be seen in the following chart. Outliers have been removed for illustrative purposes, but left within the data as they reflect valid data.

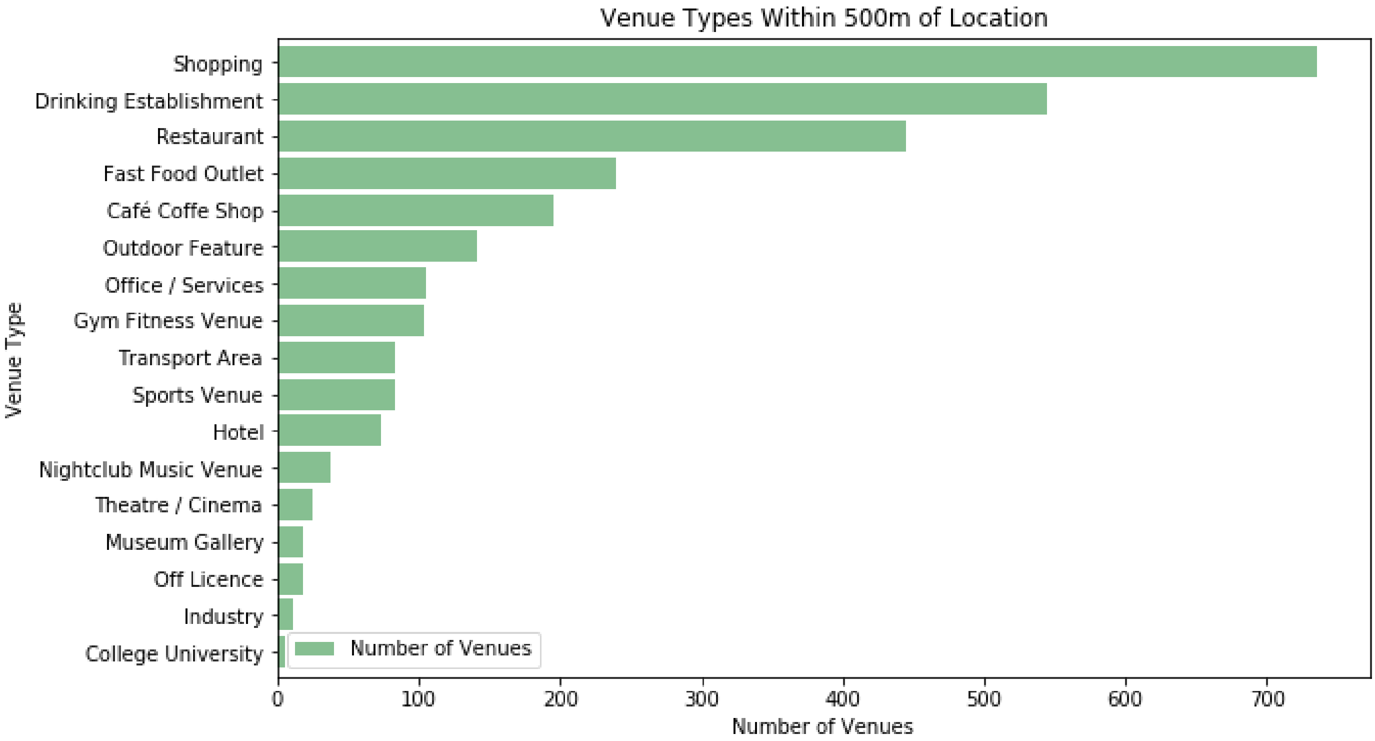
Box Plot Showing disribution of Crime Types by Area



Venues within 500m Radius of Geographic Point

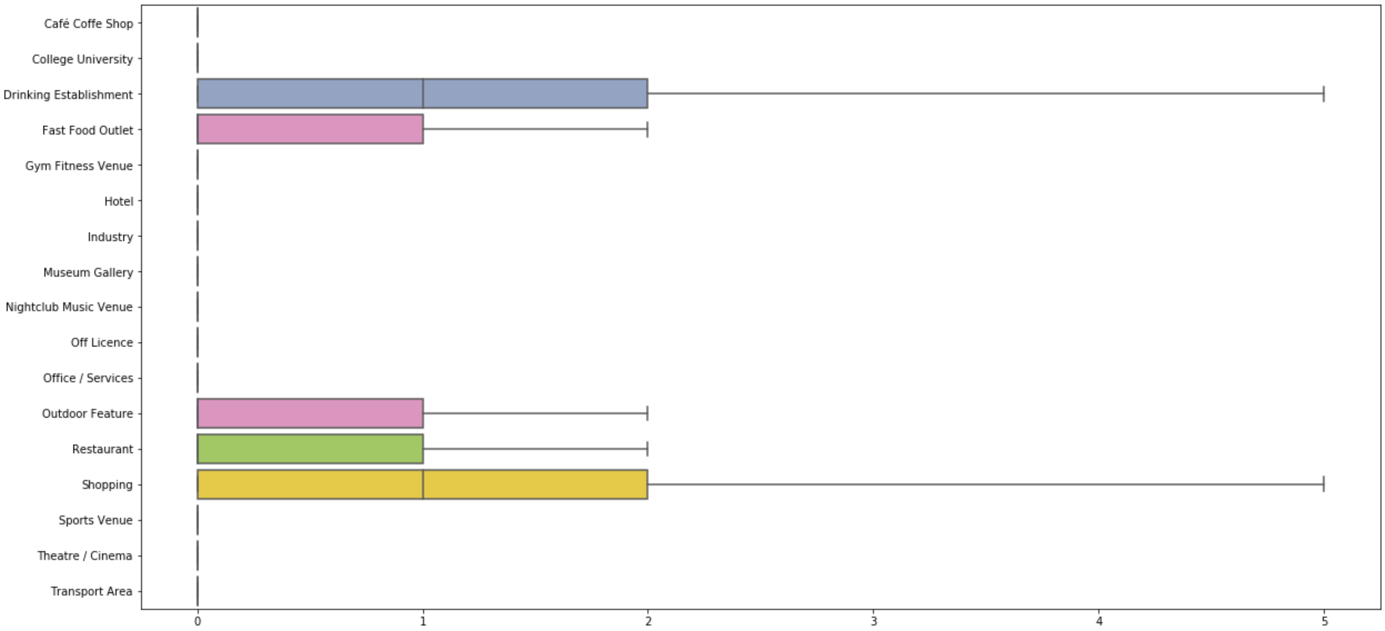
Collating venues types within a 500m radius of each point was the next step to be undertaken. Data available from Foursquare (https://foursquare.com/) was interrogated, with details of up to 100 locations within a radius of 500m being retrieved for each geographical point.

230 Unique venue types were retrieved through this process. In order to reduce this number to a more manageable and meaningful list of venue types, the 230 unique venues were mapped to a revised list of 17 high level categories. Numbers of venues within each category are shown below



As with the crime data, this venue data was aggregated and manipulated in order to show venue types by column and the number of each venue by area (row). The chart below shows the distribution of these venues by area (outliers have been removed, but remain in the data as they are valid)

Box Plot Showing the Distribution of Venue Type by Area



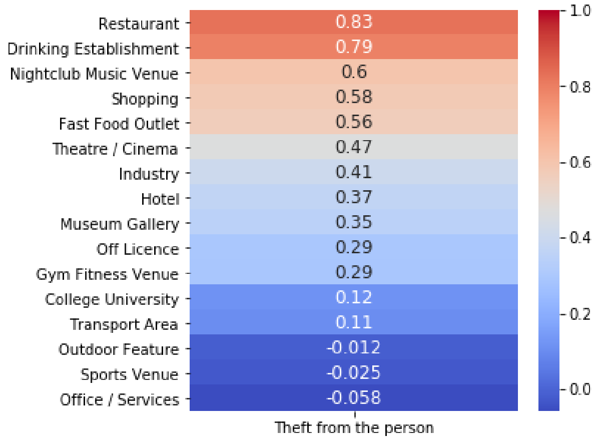
Amalgamantion of Data

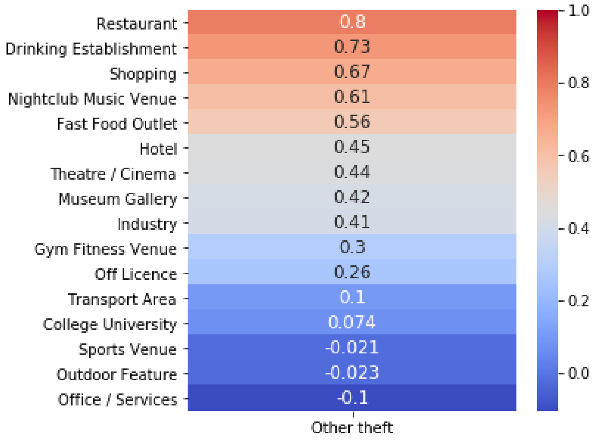
The final step in the preparation of data was to combine all the data into one table. The location data was merged with the crime data already process. A final data set containing all the data required to proceed with the analysis was created, detailing crime and location type numbers within 500m of each geographic location.

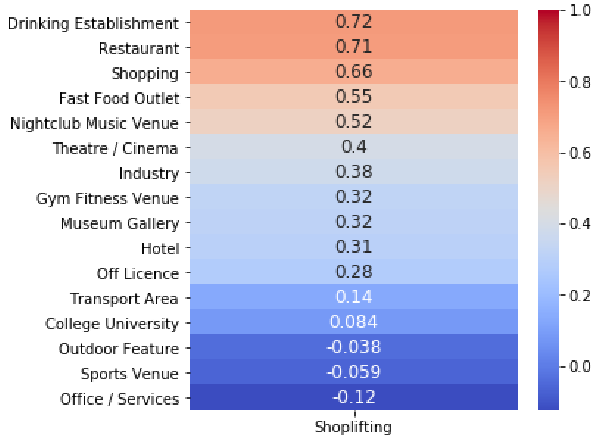
4. Methodology

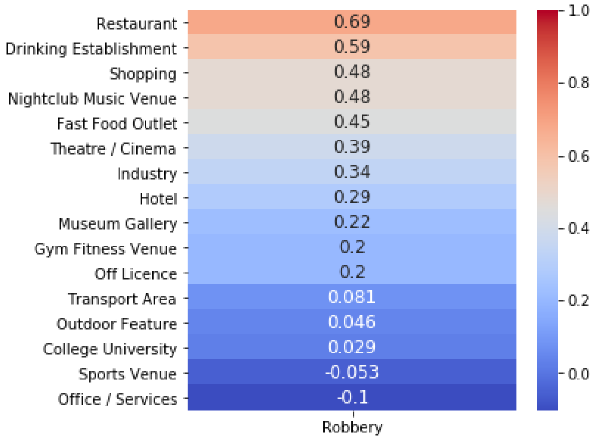
In order to obtain a reflective overview of any relationships that existed in the data, the data was standardised. Following on from this, relationships in the data between crime types and location types could be identified and from this, a feature set could be identified, using the most relevant locations and modelling performed on the most relevant crime types.

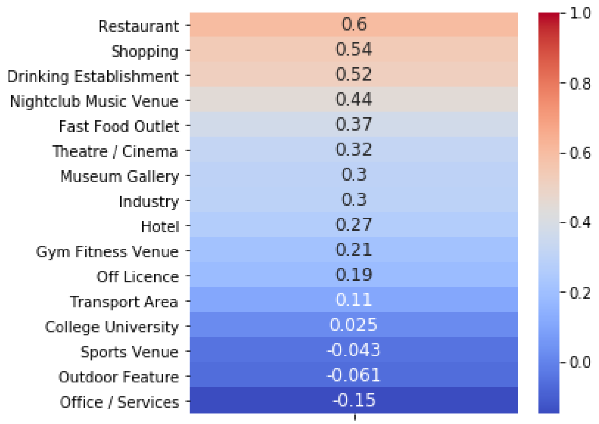
Investigation as to whether there were any relationships between crime type and locations. Correlation analysis was performed for each crime type against location type. All results can be seen in the charts below.



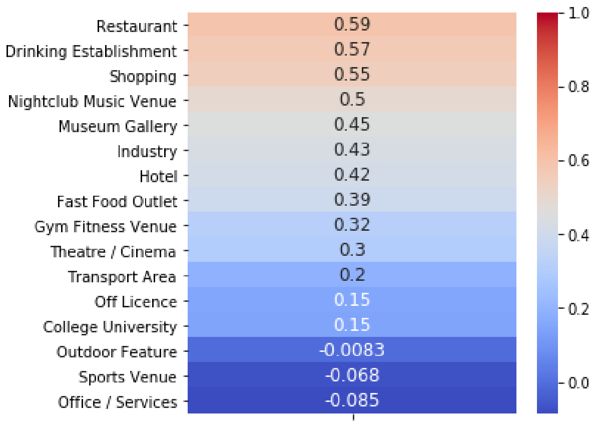




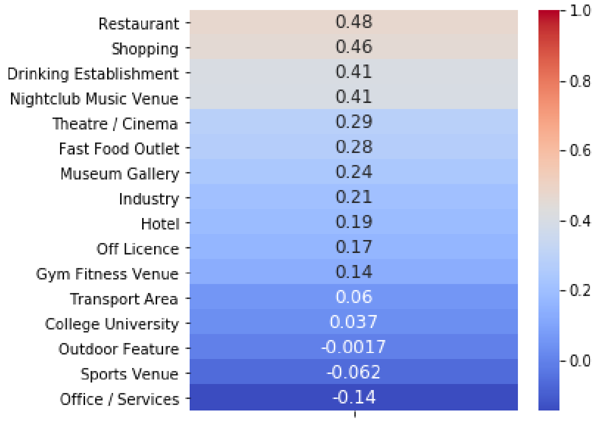




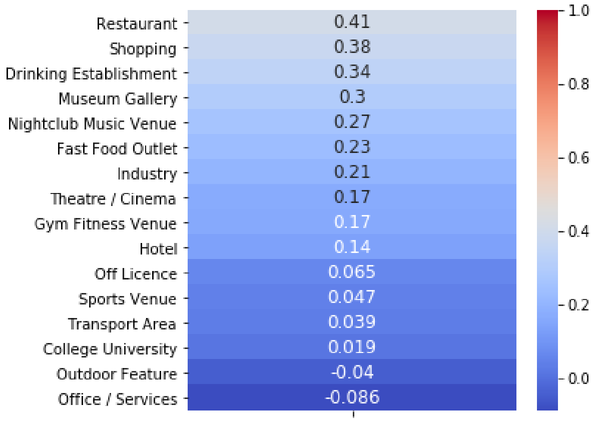
Public order



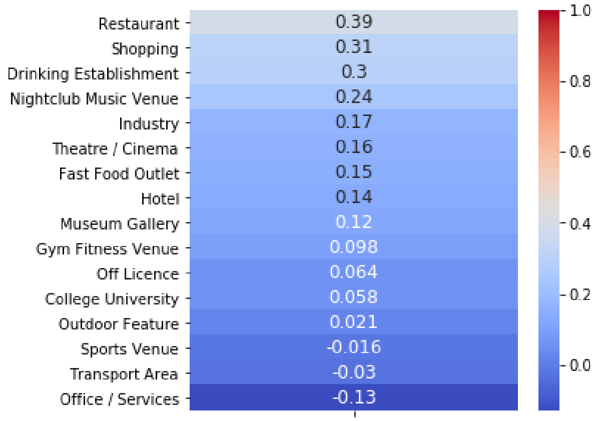
Bicycle theft



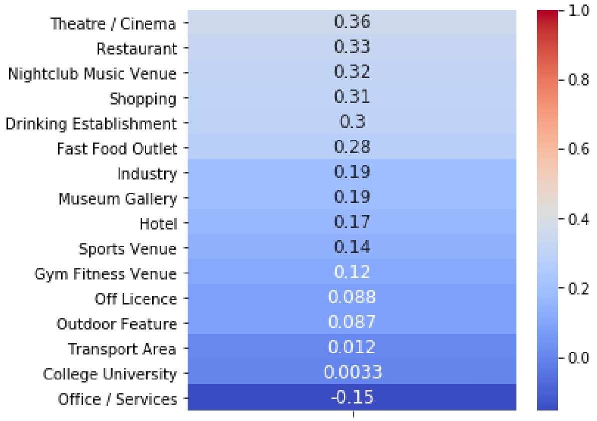
Violence and sexual offences



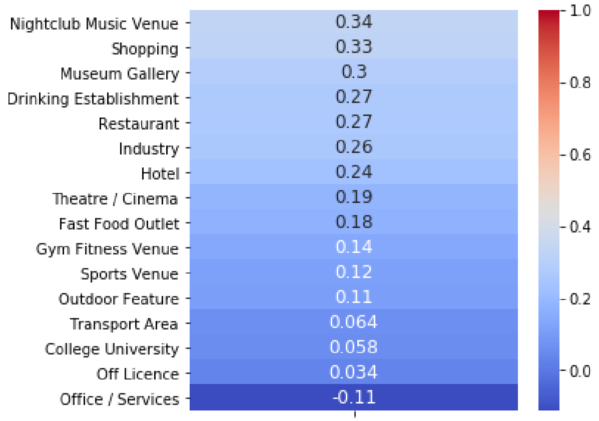
Drugs



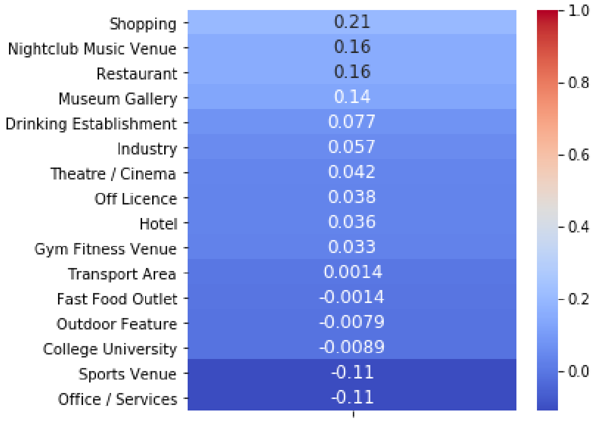
Possession of weapons



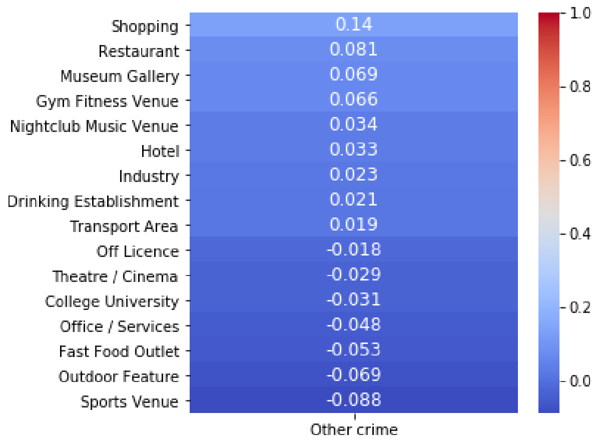
Burglary



Vehicle crime



Criminal damage and arson



The crime types identified as having a stronger positive relationship with specific outlet types are outlined in the table below. These outlet types were the features chosen to base the classification model on. Shoplifting was excluded from the model as this by default would be linked to the frequency of certain outlet types. Bicycle theft was also excluded due to the low frequency and not being relavenr to the study.

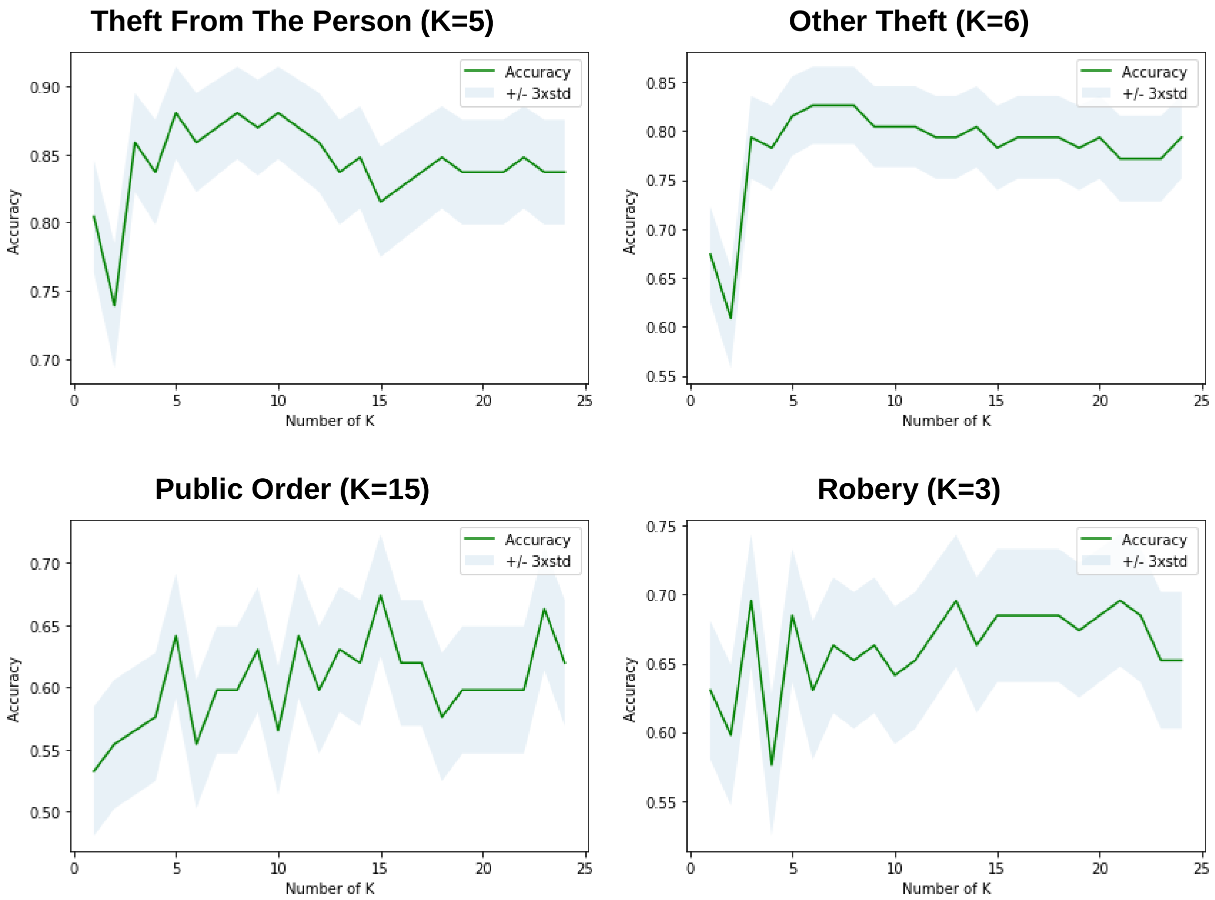
Crime Type	Outlet Type
Theft from the Person Other Theft Robbery Public Ordert	Restaurant Drinking Establishment Nightclub / Music Venue Shopping Fast Food Outlet

A classification of either a High or Low crime level was then assigned to each of the geographic points to allow for modelling. This was based on the mean value for each crime type to be analysed. A high rate being assigned to value greater or equal to the mean, and a low rate assigned to values less than the mean.

The data was then split into training and test sets and K-Nearest Neighbour Classification modelling was performed on the training set in order to create a model to be used against the test set.

In each case the best number of “K” had to be identified in order to provide the most accurate results and then use against the test set. The final values chosen to run the models on can be seen in the following charts.

Testing of the test data against these models was then performed.


Theft From The Person (K=5)

Other Theft (K=6)

Public Order (K=15)

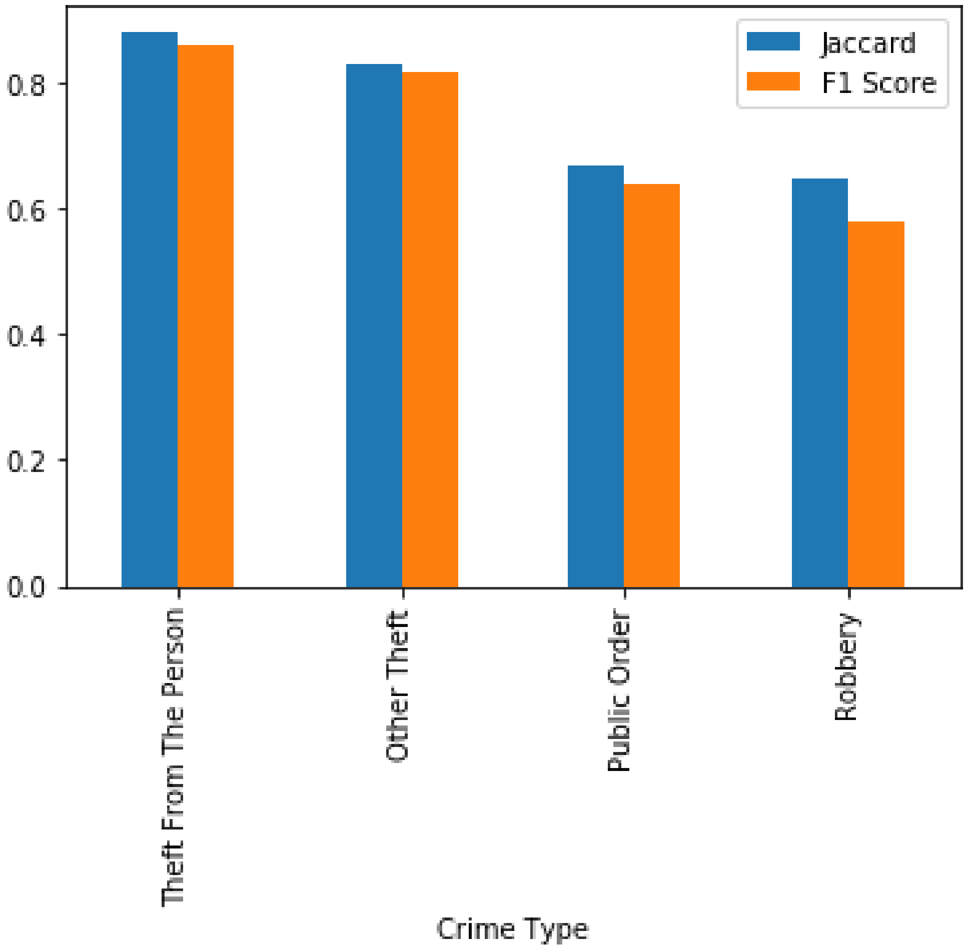
Robery (K=3)

5. Results

The results of the modelling can be seen in the table and chart below.

Crime Type	Jaccard Score	F-1 Score
Theft From The Person	0.88	0.86
Other Theft	0.83	0.82
Public Order	0.67	0.64
Robbery	0.65	0.58

Model Test Accuracy



For the four crime types which were used in the model the results for each of them showed a high level of accuracy, which theft from the person providing the highest accuracy score at 0.88 (Jaccard Similarity Score) and Robbery having the lowest at 0.65 (Jaccard Similarity Score).

6. Discussion

The implementation of the models generated good results with a high accuracy when classifying the four crime types chosed. With some confidence the model could classify a given geographic location as a high or low crime risk in relation the certain crime types.

The availability of the location data however would act as an influencer in these results. Using Foursquare as the location data provider resulted in relatively low numbers of location types being returned. Using an alternative source such as Google Places, may provide for a higher number of locations being included in the study. This would lead to a more representative reflection of what locations were in the vicinity and make for a more accurate data set and a more robust model.

This could also mean that further crime types could be included in the modelling as opposed to just four. If this was the case, investigation into the more frequently occurring crime types, not included in this study, could be made (eg Violence and Sexual Offences).

The crime data used was the available published data, which was already categorised If data relating to Violence and Sexual offences could be broken down further this may have an impact on the data relationships and the modelling outcomes. Ie do the crimes camcel each other out when looking at them in relation to nearby locations)

Additional classification models could also be tested in order to determine if a more accurate model could be utilised.

7. Conclusion

In conclusion the implementation of the K-Nearest Neighbour classification model provided a high level of accuracy when classifying a location as high or low risk in relation to a number of crime types.

The level of accuracy and validity of the models could potentially be improved if more complete information on nearby locations could be utilised and certain crime types being broken down into further categories.

Using additional classification models may also result in greater levels of accuracy being returned.