

Predicting the crime rate based on the venues categories prevailing in London boroughs

PAVEL YAKOVLEV

MAY 8, 2020

1 INTRODUCTION

1.1 Background

The London boroughs are the 32 local authority districts. The Metropolitan Police website provides an interactive data dashboard showing the crime rates statistics presented on the map of London with data collected from April 2010. As we can see on the figure below, some areas remain safer than others persistently in years.

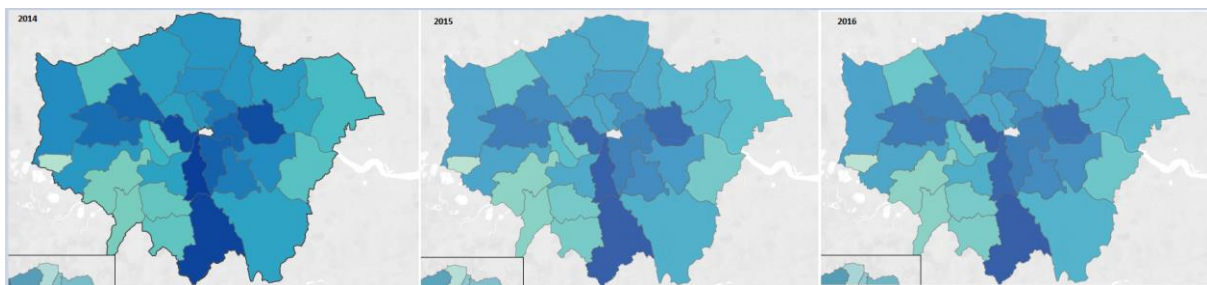


Figure 1. London boroughs crime rate for 2014 /2015/ 2016 years

There is no doubt that there are multiple factors making an impact on the crime rate including social, cultural, economic aspects within each borough. There is also an intuitive relationship between the most of these factors and categories of venues (cafes, restaurants, hotels, museums etc.) prevailing in a given area. In fact the demand on some certain places could generally represent the *character* of area and therefore may be used as an indicator to understand the trends in most important characteristics such as crime rate.

1.2 Problem

This project aims to model the crime rate behaviour by looking at the categories of venues prevailing in an area. The goal is to explore what is the common combination of categories in the boroughs with similar crime rate. It's also interesting to find out what categories have most significant impact on the crime rate.

1.3 Interest

If the relationship between crime rate and prevailing categories of venues could be modelled and used for prediction, it may become an important tool for local authorities to improve the criminal situation in their areas. As example, the local authorities can invest in providing more libraries or cultural venues to impact on the criminal situation in longer term.

This relationship can also be useful for property investors exploring areas with a good potential growth of house prices based on the future plans of local authorities or business projects in a given area.

2 DATA

The *crime rate* is computed as a ratio of *recorded crimes* to the *population* of a given area in 1000s. We are going to use the official data store of London government for both measures. It was decided to compute the crime rate on the **level of wards** (i.e. more granular than a borough level), it will allow to make a research on a more detailed statistical data.

$$\{crime\ rate\} = \frac{\{number\ of\ crimes\}}{\{population\ in\ thousands\}}$$

Additionally we will need *geographical coordinates* of London areas to get details of the *venues* located there.

2.1 Recorded Crime Feed

The main feed is 'Recorded Crime' produced by Metropolitan Police and published regularly on the official [London Government Data Store Portal](#). Each row represents a series of counts for recorded crimes of a given crime category on London ward's level. The time interval covers period from April 2018 to March 2020.



MajorText	MinorText	WardName	WardCode	LookUp_BoroughName	201804	201805	201806	201807
Arson and Criminal Damage	Arson	Abbey	E05000026	Barking and Dagenham	0	0	3	2
Arson and Criminal Damage	Criminal Damage	Abbey	E05000026	Barking and Dagenham	16	14	12	12
Burglary	Burglary - Business and Community	Abbey	E05000026	Barking and Dagenham	6	3	4	8
Burglary	Burglary - Residential	Abbey	E05000026	Barking and Dagenham	5	5	4	6
Drug Offences	Drug Trafficking	Abbey	E05000026	Barking and Dagenham	0	0	1	1
Drug Offences	Possession of Drugs	Abbey	E05000026	Barking and Dagenham	26	24	11	13
Miscellaneous Crimes Against Society	Disclosure, Obstruction, False or Misleading State	Abbey	E05000026	Barking and Dagenham	0	0	0	0
Miscellaneous Crimes Against Society	Fraud or Forgery Associated with Driver Records	Abbey	E05000026	Barking and Dagenham	0	0	0	0
Miscellaneous Crimes Against Society	Going Equipped for Stealing	Abbey	E05000026	Barking and Dagenham	0	0	0	0
Miscellaneous Crimes Against Society	Handling Stolen Goods	Abbey	E05000026	Barking and Dagenham	0	0	0	0
Miscellaneous Crimes Against Society	Making, Supplying or Possessing Articles for use i	Abbey	E05000026	Barking and Dagenham	1	2	0	0
Miscellaneous Crimes Against Society	Obscene Publications	Abbey	E05000026	Barking and Dagenham	0	1	0	0
Miscellaneous Crimes Against Society	Other Forgery	Abbey	E05000026	Barking and Dagenham	0	0	0	0
Miscellaneous Crimes Against Society	Other Notifiable Offences	Abbey	E05000026	Barking and Dagenham	0	0	0	0
Miscellaneous Crimes Against Society	Possession of False Documents	Abbey	E05000026	Barking and Dagenham	1	0	1	0
Miscellaneous Crimes Against Society	Profiting From or Concealing Proceeds of Crime	Abbey	E05000026	Barking and Dagenham	0	0	0	0
Miscellaneous Crimes Against Society	Threat or Possession With Intent to Commit Crimina	Abbey	E05000026	Barking and Dagenham	0	1	0	1
Possession of Weapons	Possession of Article with Blade or Point	Abbey	E05000026	Barking and Dagenham	2	3	0	0
Possession of Weapons	Possession of Firearm with Intent	Abbey	E05000026	Barking and Dagenham	0	0	0	0

Figure 2. Screenshot of Recorded Crime feed.

The following features are provided in the feed:

MajorText/MinorText - Major and Minor of crime category of a record.

WardName – Name of the ward.

WardCode – The ward code called GSS which is consistently used for other data feeds on wards level.

LookupBorough – The borough name.

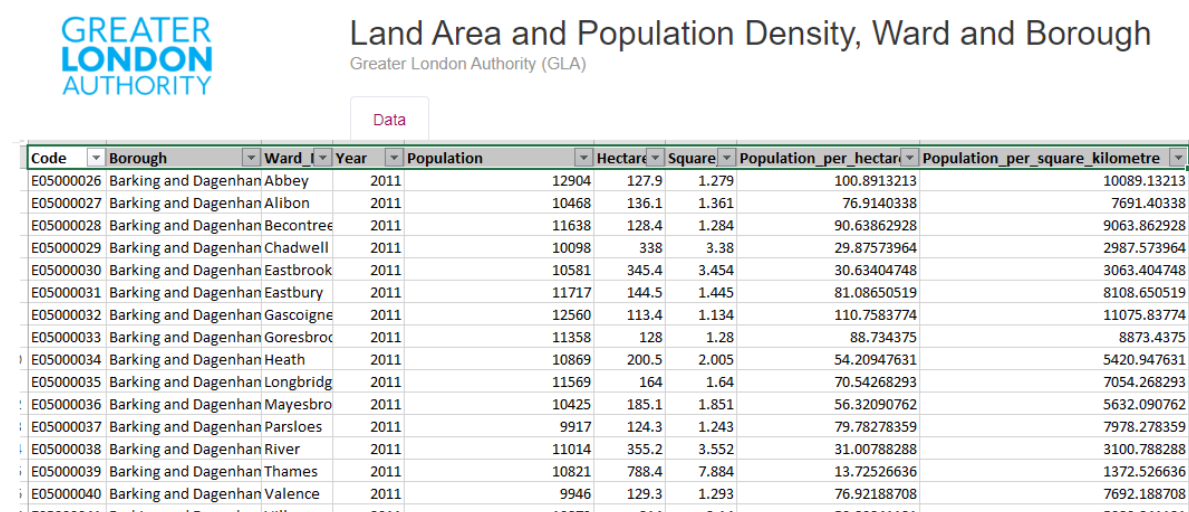
As we can see that the feed provides counts for a detailed list of different crime categories. As part of this project it was decided to focus on the statistics for one selected crime category, namely “***Violence against the person***”. We believe it is a most representative category among others.

Additionally we will be focusing only on the data for the full 2019 year, i.e. from January 2019 to December 2020.

It was also noted that there are 3 boroughs having a ward called ‘Village’. Although they have unique ward codes, we will need to make these wards distinct by using them in combination with borough names for the accurate presentation.

2.2 Population Density Feed

The Land Area and Population Density feed is gathered by Greater London Authority and is available on the official [London Government Data Store Portal](#). The population figures are provided starting from 2011 on the wards level.



Code	Borough	Ward	Year	Population	Hectare	Square	Population per hectare	Population per square kilometre
E05000026	Barking and Dagenham	Abbey	2011	12904	127.9	1.279	100.8913213	10089.13213
E05000027	Barking and Dagenham	Alibon	2011	10468	136.1	1.361	76.9140338	7691.40338
E05000028	Barking and Dagenham	Becontree	2011	11638	128.4	1.284	90.63862928	9063.862928
E05000029	Barking and Dagenham	Chadwell	2011	10098	338	3.38	29.87573964	2987.573964
E05000030	Barking and Dagenham	Eastbrook	2011	10581	345.4	3.454	30.63404748	3063.404748
E05000031	Barking and Dagenham	Eastbury	2011	11717	144.5	1.445	81.08650519	8108.650519
E05000032	Barking and Dagenham	Gascoigne	2011	12560	113.4	1.134	110.7583774	11075.83774
E05000033	Barking and Dagenham	Goresbrook	2011	11358	128	1.28	88.734375	8873.4375
E05000034	Barking and Dagenham	Heath	2011	10869	200.5	2.005	54.20947631	5420.947631
E05000035	Barking and Dagenham	Longbridge	2011	11569	164	1.64	70.54268293	7054.268293
E05000036	Barking and Dagenham	Mayesbrook	2011	10425	185.1	1.851	56.32090762	5632.090762
E05000037	Barking and Dagenham	Parsloes	2011	9917	124.3	1.243	79.78278359	7978.278359
E05000038	Barking and Dagenham	River	2011	11014	355.2	3.552	31.00788288	3100.788288
E05000039	Barking and Dagenham	Thames	2011	10821	788.4	7.884	13.72526636	1372.526636
E05000040	Barking and Dagenham	Valence	2011	9946	129.3	1.293	76.92188708	7692.188708

Figure 3. Screenshot of Population Density Feed

The following features are provided in the feed:

Code – GSS code of a ward. It matches to ‘WardCode’ field from recorded crimes feed (see 2.1 above).

Borough & Ward_Name – Borough and ward name for a record.

Year – Year of a record.

Population – The key feature of this field representing the number of habitants in a given ward for that year.

We are going to join both feeds by ward codes in order to compose a record suitable for computing **crime rate**.

2.3 Geographical coordinates of London wards

It turned out that there is no a clear direct feed available which could be used to get geographical location for each London ward. But there are a few useful websites which we can use to scrape required information. We will be using python package called [Beautiful soup](#) to deal with HTML pages. The process is described below.

As it was mentioned before that each line of data has 'Ward Code' a GSS code identifying this area, e.g.

CODE	BOROUGH	WARD_NAME
E05000026	Barking and Dagenham	Abbey

Firstly, we can send an HTTP request to [MapIt \(mySociety\)](#) server on the following URL (note that we use ward code in the text of link):

<https://mapit.mysociety.org/area/E05000026.html>

In its turn that page will contain a link to a service called 'Geometry (JSON)' providing the central location of each area (latitude and longitude), e.g.

<https://mapit.mysociety.org/area/8702/geometry>

```
{
  "max_e": 545296.3961,
  "srid_en": 27700,
  "area": 1282925.0015508249,
  "max_n": 184928.1042,
  "min_lat": 51.53359474114574,
  "max_lat": 51.54479573660534,
  "centre_n": 184358.43510997608,
  "max_lon": 0.09360140008380039,
  "centre_lon": 0.07793493806061133,
  "min_n": 183674.3031,
  "parts": 1,
  "centre_e": 544203.5366906121,
  "min_e": 543417.2972,
  "min_lon": 0.06664900830609423,
  "centre_lat": 51.53971138229575
}
```

If there is an error in this process, we will make a manual adjustment and populate required coordinates for areas where they are missing.

The **Figure 4** shows Greater London area covered by the coordinates we extracted by this way.

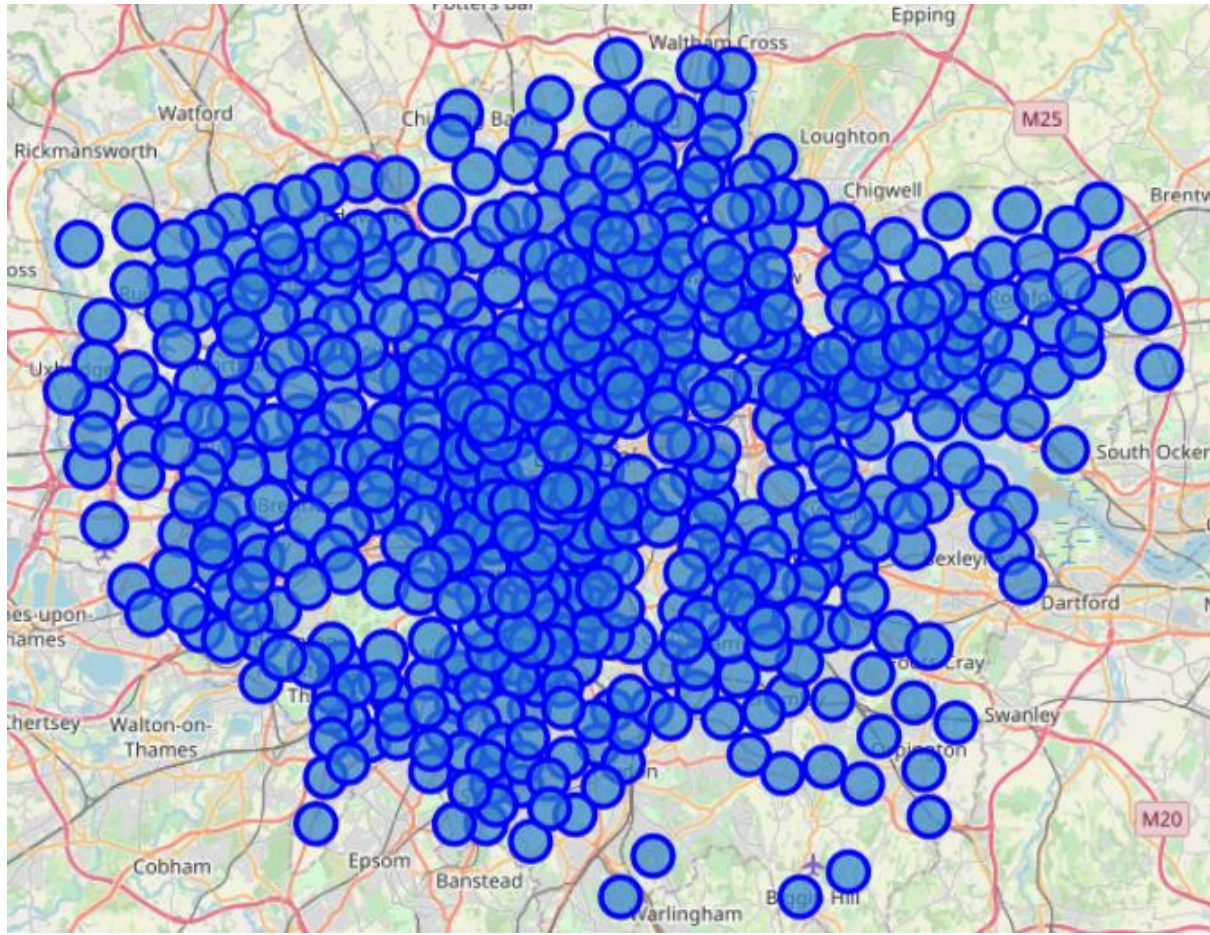


Figure 4. London wards coordinates on map

2.4 Venues information feed

As part of this project, we are going to use **Foursquare API** to get all the details for venues located in each London ward. The API end-point **explore** will provide 100 venues in the radius of 500 meters around the centre point of a ward.

The JSON result will include the list of venues with category names. The prevailing categories will be identified and used to resolve the problem of this project.

```

{
  'reasons': {'count': 0,
  'items': [{
    'summary': 'This spot is popular',
    'type': 'general',
    'reasonName': 'globalInteractionReason'
  }]}],
  'venue': {'id': '5628014b498e3a8dc4613c3a',
  'name': 'The Gym London Barking',
  'location': {'address': 'The Clock House, East Street',
  'lat': 51.53619293708999,
  'lng': 0.07860085726238432,
  'labeledLatLngs': [{
    'label': 'display',
    'lat': 51.53619293708999,
    'lng': 0.07860085726238432
  }],
  'distance': 394,
  'postalCode': 'IG11 8EQ',
  'cc': 'GB',
  'city': 'Barking',
  'state': 'Greater London',
  'country': 'United Kingdom',
  'formattedAddress': ['The Clock House, East Street',
  'Barking',
  'Greater London',
  'IG11 8EQ',
  'United Kingdom']},
  'categories': [{
    'id': '4bf58dd8d48988d176941735',
    'name': 'Gym',
    'pluralName': 'Gyms',
    'shortName': 'Gym',
    'icon': {
      'prefix': 'https://ss3.4sqi.net/img/categories_v2/building/gym_',
      'suffix': '.png'
    },
    'primary': True
  }],
  'photos': {'count': 0, 'groups': []},
  'referralId': 'e-0-5628014b498e3a8dc4613c3a-2'
},

```

Figure 3. Example of Foursquare API' JSON response

The category names provided by Foursquare API is too granular. For the purpose of this project, we would like to use higher level of categorisation. We will download the tree of available categories and perform the roll-up to the highest level.

3 METHODOLOGY AND ANALYSIS

In this project we will firstly perform clustering of London wards by crime rate. We will try to exclude the wards with outlined crime rates (too high). After this cleaning exercise, we will assign 3 grades of crime rates: *low, medium and high*.

Secondly, we will perform clustering of London wards by the combination of venue categories to see if similar wards have similar crime rate grade. We will exclude all the wards with too specific combination of venues to focus on standard cases.

Thirdly, we will try to observe a dependency between crime rate and the percentage of a given category on a series of scatter plots.

Finally, we will use **Decision Tree** and **Support Vector Machine** approaches to model crime rate grade value based on the venues profiles.

3.1 Clustering by crime rate

Initially, we will use **K-Means** clustering to split all London wards into 3 groups. The results are shown in the **Table 1**.

Category	Min rate	Max rate	Number of wards
Low	5.61	24.64	353
Medium	24.75	68.20	202
High	154.13	184.37	2

Table 1. Initial clustering by crime rate

We can see that High category contains only 2 wards, namely:

LookUp_BoroughName	WardName	WardCode	crime_rate
Westminster	St James's	E05000644	184.377478
Westminster	West End	E05000649	154.134234

It's a well-known fact that those wards are probably the most popular tourist attractions in the world and therefore they are quite exception and could be excluded from our data set to focus on more standard cases.

After this exclusion, the crime rate grades are finally allocated in a more balanced way as it's shown on Table 2.

Category	Min rate	Max rate	Number of wards
Low	5.61	21.26	272
Medium	21.36	35.35	242
High	35.77	68.20	145

Table 2. Final clustering by crime rate

3.2 Clustering by prevailing categories

It was decided to use **K-Means** clustering technique to split all wards into 10 groups by the similarity of prevailing venues. After this the crime rates for each corresponding cluster have been presented as a box plot (see **Figure 5**)

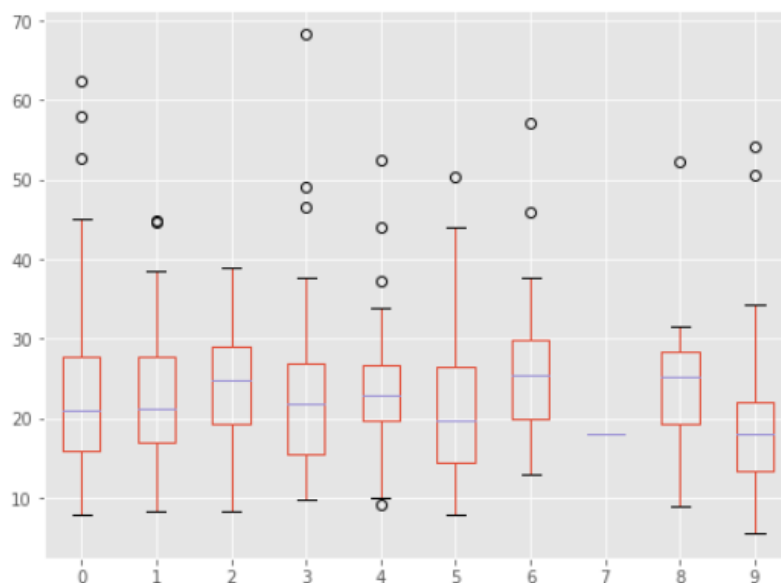


Figure 5. Box plot of crime rates per cluster.

One of the observation is that cluster 7 has a very limited number of wards in it, and indeed it's just one ward:

WardName	WardCode	crime_rate	crime_rate_category	Cluster Labels	Arts and Entertainment	College and University	Food	Nightlife Spot	Outdoors and Recreation	Professional and Other Places	Residence	Shop and Service	Travel and Transport
Chessington South	E05000405	18.152408	Low	7	0.782609	0.0	0.086957	0.043478	0.0	0.0	0.0	0.0	0.086957

As we can see it has 76% of venues in 'Arts and Entertainment' category, which is not grouping together with any other wards. It could be explained by the fact that this ward has **Chessington World of Adventure resort** and being quite unique. So we are going to exclude this cluster from our research.

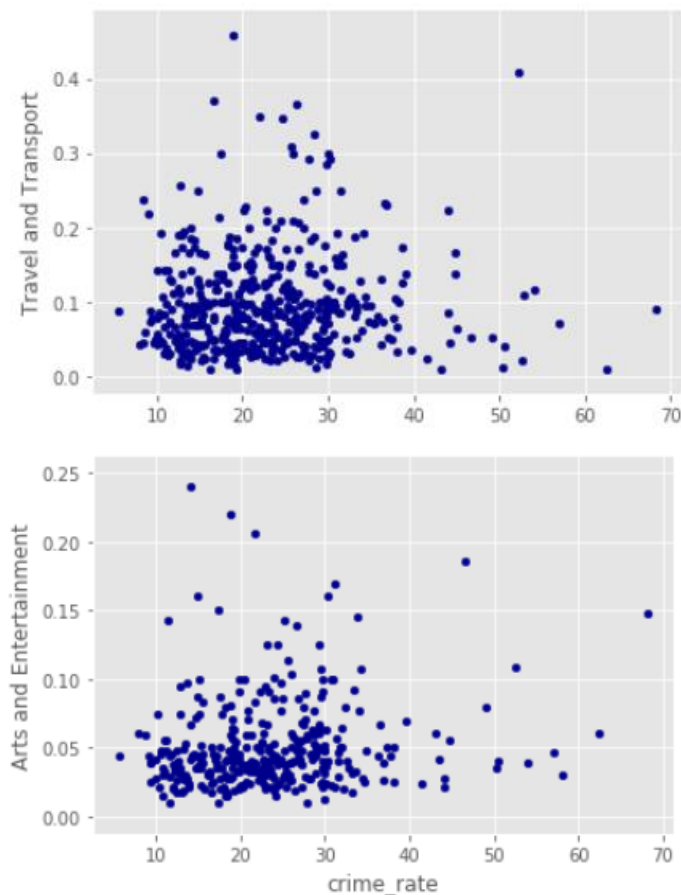


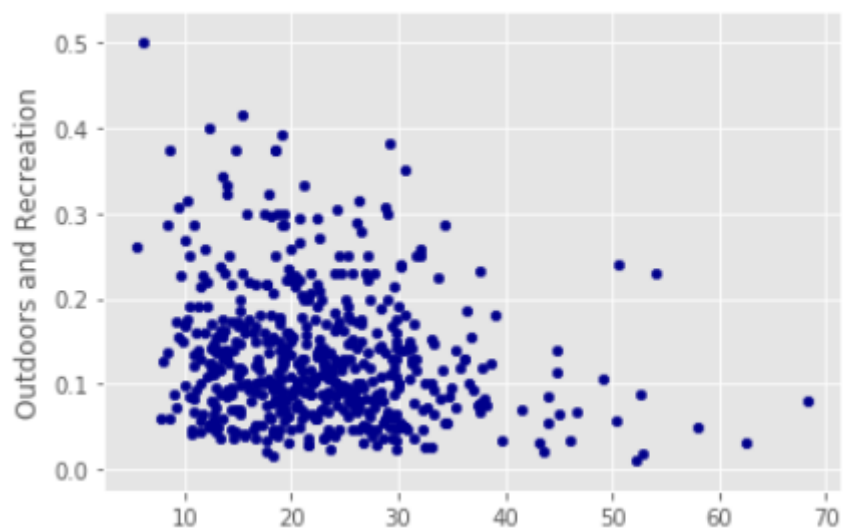
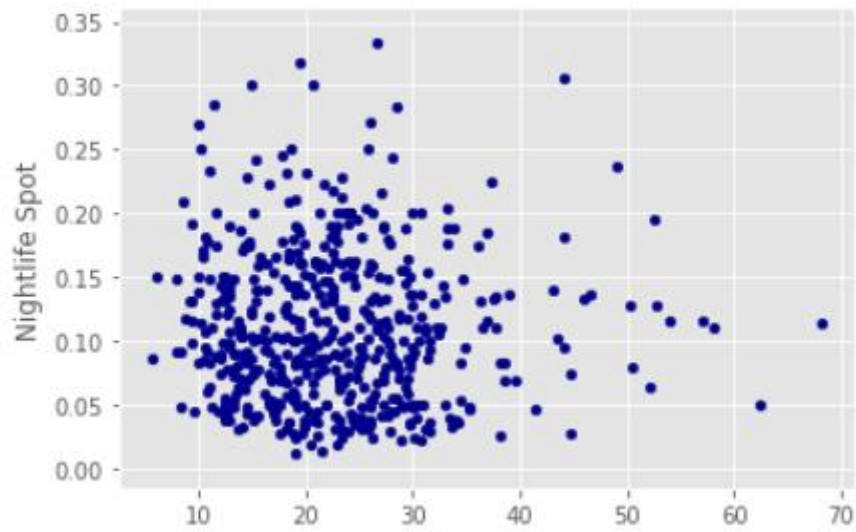
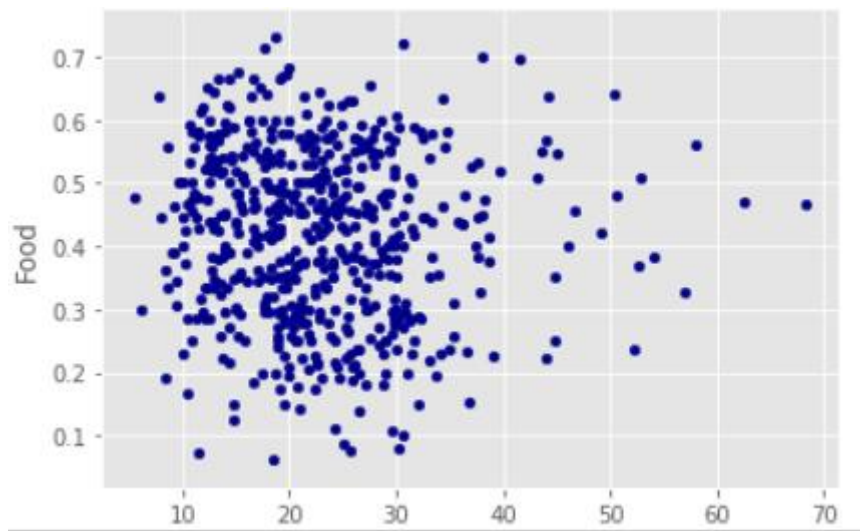
Figure 6. Chessington world of adventure logo.

The second observation is that crime rate value varies quite randomly within each of the cluster and we cannot see any definitive relationship between similar categories of venues and crime rate.

3.3 Exploring dependency between the category prevalence and crime rate

It is interesting to observe how a level of category presence in a given territory can impact on crime rate. Below you will find a series of scatter boxes showing this relationship.





As we can see, there is no clear relationship between category prevalence and crime rate in areas.

3.4 Decision Tree model

It was decided to use Decision Tree model for prediction purpose. The 30% of available were used as a testing data. The model was chose to build with maximum depth of 10.

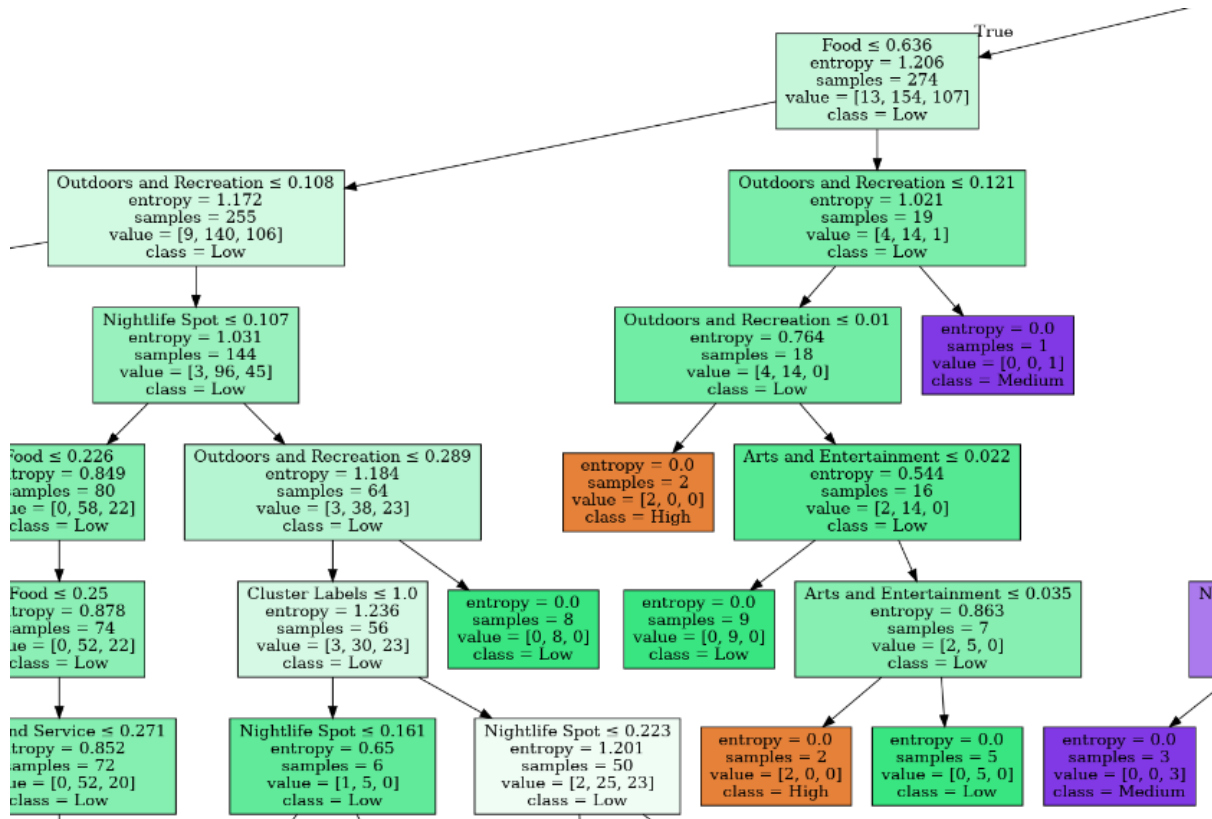


Figure 6. Snapshot of the decision tree visualisation.

The model accuracy was computed as 0.4910 value

3.5 Support Vector Machine

Additionally, Support vector machine technique was used for prediction. The F1 score of the model is 0.42117

4 RESULTS AND DISCUSSION

As part of process the data of more than 600 wards of London was collected – overall the information of more than 15000 venues of 429 categories were composed.

Initially the results were produced based on a very granular level of categories of venues provided by Foursquare API, it was found that it doesn't allow to cluster wards clearly (e.g. a Kebab Place comes separately from a Turkish Restaurant), so it was decided to roll up categories to the highest level to perform better clustering.

It was also discovered that some of the wards are very exceptional and they were excluded from our research.

The main result of this projects shows that there is no clear highly predictably relationship between prevailing categories and crime rate in London areas. In fact we can see wards with very similar profile of categories but being on opposite sides of crime rate range.

The resulted model shows average accuracy.

6 CONCLUSION

The possible explanation of the project results could be in the fact that crime rate is a very complex measure depending on plenty of factors of social, cultural, economic and historical aspects of an area.

As part of potential future work, we can perform a similar research for a different crime type (e.g. Theft) to see if there is a possible correlation could be found.

Additionally, it could be interesting to apply the findings of this project to another large municipality within the UK (Birmingham, Manchester) or even outside of the country.