Capstone Project Report – The Battle of Neighbourhoods

Best areas to stay at while visiting Istanbul

Paweł Nowakowski

Introduction

When planning a trip to a new place we often have a hard time deciding which area would be the best to stay during the visit. Some of the main concerns are accommodation cost and safety of the area. Additionally, because we come as tourists, we want to be as close to tourist attractions and recommended eateries as possible. Due to the number of criteria, making an optimal choice might not be that straightforward.

Here, an analysis of different criteria to be considered while travelling to Istanbul is performed. Based on that analysis, a classification of districts of the area is made to indicate the most optimal areas to stay at.

Data

As mentioned in the introduction, several data sets are necessary to fully explore the given problem. For the case of travelling to Istanbul, the following datasets were used:

- List of districts of Istanbul with simple statistics such as population, area and population density obtained from https://en.wikipedia.org/wiki/List_of_districts_of_Istanbul
- Geospatial data of the borders of Istanbul districts for visualization, obtained from https://gadm.org/
- Average rent prices for districts of Istanbul obtained from https://www.realtygroup.com.tr/average-rent-price-in-istanbul-is-1486-tl/. Due to the lack of free API or data sets for prices of short time accommodation, such as hotels, hostels or Airbnb, an assumption was made that apartment rental prices in each district is positively correlated with short term accommodation prices.

- Number of crimes reported per district obtained from "*Ergun, N., & Yirmibeşoğlu, F. (2007). Distribution of Crime Rates in Different Districts in Istanbul. Turkish Studies, 8(3), 435–455. doi:10.1080/14683840701489324"*. Crime data was used to calculate a crime index for each neighbourhood, based on crime numbers, population and weighted by seriousness of the crime. Described more in methodology section.

- Top attractions in Istanbul together with their score and location coordinates were obtained from Triposo API. This data was used to calculate district-attraction score, based on the distance from district to attraction and weighted by the inverse of its score. Described more in methodology section.

- Top eateries in Istanbul together with their score and location coordinates were obtained from Triposo API. This data was used to calculate district-food score, based on the distance from district to eatery and weighted by the inverse of its score. Described more in methodology section.

*NOTE: For attractions and eateries, Triposo API was used instead of Foursquare API. Foursquare API does not have option to explore sightseeing places ('tourist sight', 'Popular with visitors' or similar searches return tourist information, hotels, restaurants, etc. all mixed) and getting ratings of restaurants required separate request for each place which is a premium call (limited use for free accounts). According to a response from staff in the discussion forum for week 4, the usage of other API instead of Foursquare is allowed.

Methodology

In order to determine the optimal areas to stay, the K-means clustering method was used. K-means method partitions unlabelled observables (districts) into provided number of clusters based on supplied features, such that each observable is assigned to cluster with the nearest centre. This method perfectly suits the given problem, where districts are to be grouped based on their similarities.

For this, gathered data needs to be transformed into useful insights.

- The first feature that was used is accommodation prices. Because the data that was found represents averaged rent prices for each district, no additional alteration of the existing data set was necessary.

- Another feature used for classification was crime index, which quantifies safety of the district. Its calculation was based on the Crime Severity Index described by Canada's national statistics office: https://www150.statcan.gc.ca/n1/pub/85-004-x/2009001/part-partie1-eng.htm. The crime index was calculated by taking into consideration the number of reported cases for a type of crime weighted by a "seriousness weight" of that crime, summed over all crime types and divided by population of the district:

$$crime\ index = \frac{\sum \#cases \cdot weight}{population}$$

The “seriousness weight” was derived from sentences given by court for the type of crime committed (incarceration rate and length of the prison sentence). The more serious the crime, the higher the “seriousness weight” it obtained. Here, weights given for offence types in Canada (https://www150.statcan.gc.ca/n1/pub/85-004-x/2009001/t001-eng.htm) were used. To make it easier, the crime index was normalized, where the most dangerous district obtained $crime\ index = 100$ and *crime index* = 0 would indicate no crimes.

- Next two features that were used are district-attractions and district-food score. They quantify how good the location of the district is, based on the distance to top attractions and eateries, respectively. To obtain these, the haversine formula was used to calculate the distance between the centre of the district to each location:

$$a = sin^2\left(\frac{\Delta\varphi}{2}\right) + cos\varphi_1 \cdot cos\varphi_2 \cdot sin^2\left(\frac{\Delta\lambda}{2}\right)$$
$$c = 2 \cdot atan2\left(\sqrt{a}, \sqrt{1-a}\right)$$
$$d = R \cdot c$$

where φ is latitude, λ is longitude, *R* is Earth’s radius. Haversine formula calculates the distance between two points assuming perfectly spherical shape of Earth. Additionally, calculated distance was weighted by the inverse of the score the location obtained on Triposo. This way it is reflected that even if a given place is located further, but if it has a better rating, it would still be chosen to visit over a closer place with lower rating. Weighted distances to all attractions and eateries were subsequently averaged for each district to obtain distance-attractions and distance-food score, respectively. Since described scores reflect distances and inverse of the ratings, districts with lower scores are considered to be better to stay for tourists.

While investigating gathered data, it was noticed that the used crime statistics dataset has missing values for a few districts. Therefore, before moving on to the fitting process, a data imputation had to be performed to be able to include all the districts in classification. To treat districts with missing data neutrally, any missing values were replaced with the average of crime indices from districts with crime statistics.

Before performing classification, scikit-learn’s StandardScaler was used to normalize all features. This is to ensure that all the features had equal impact on the fitting process.

The elbow method was used to determine the number of clusters that districts should be divided into. For this, K-means fitting was performed for a range of cluster numbers recording associated errors. Of course, the more k-clusters chosen, the lower the obtained fitting error. However, to avoid overfitting, a trade-off between the number of clusters and error has to be chosen. As shown in Figure 1, adding more clusters above 5-6 does not reduce the error significantly. Therefore, for this analysis 5 k-clusters were chosen.

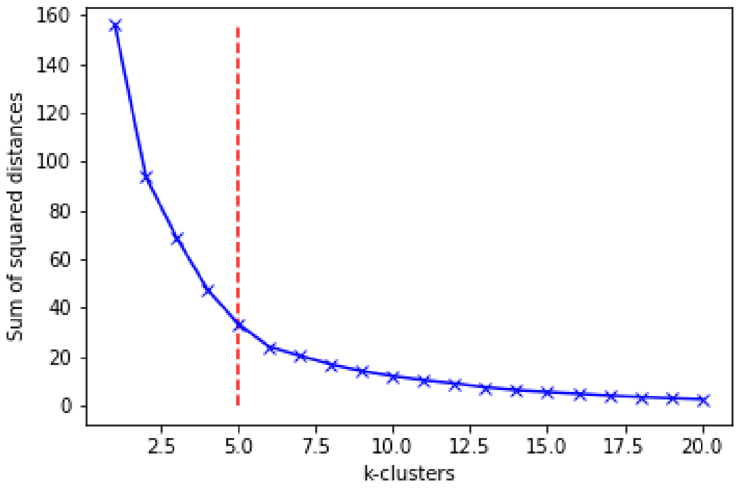
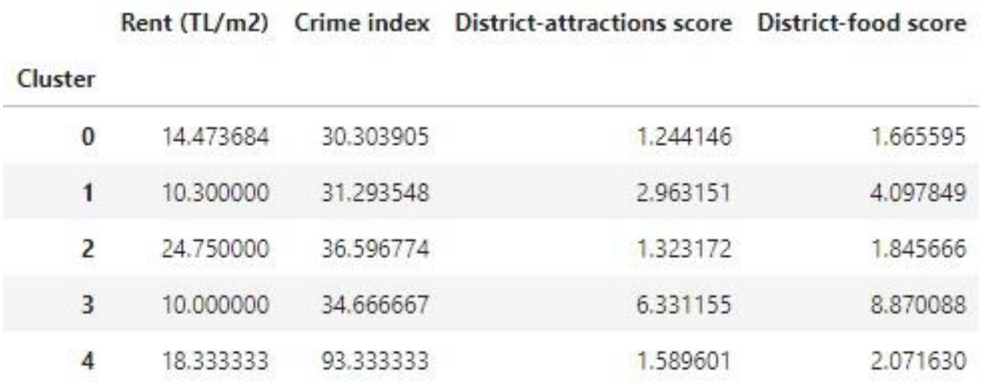


Figure 1. Elbow method to determine number of clusters to be used for K-means clustering. Red dashed line shows chosen number of clusters

Results and discussion

As mentioned in the previous section, 39 districts of Istanbul were classified into 5 clusters. Of all the districts, 19 were assigned to cluster 0, 10 to cluster 1, 4 to cluster 2, 3 to cluster 3 and 3 to cluster 4. To determine which cluster consists of the most optimal districts to stay at, clusters were rated according to average value of each feature for districts in a given cluster shown in Table 1. As it was described in methodology section, each feature was designed such that districts with lowest values of the features are better to stay at (ie., lower rent, lower crime, shorter distance to best rated places). Thus, cluster 0 in Table 1 is considered to contain districts best to stay at, because it is 3rd in rent prices, 1st in crime index, 1st in district-attractions score and 1st in district-food score, which adds up to 3+1+1+1=6 and is the lowest from all clusters.

Table 1. Juxtaposition of averaged features for each cluster, sorted, from cluster of districts best to stay at, to cluster of districts worst to stay at.



	Rent (TL/m2)	Crime index	District-attractions score	District-food score
Cluster				
0	14.473684	30.303905	1.244146	1.665595
1	10.300000	31.293548	2.963151	4.097849
2	24.750000	36.596774	1.323172	1.845666
3	10.000000	34.666667	6.331155	8.870088
4	18.333333	93.333333	1.589601	2.071630

Figure 2 shows boxplot charts visualizing basic statistics for each feature of the obtained clusters.

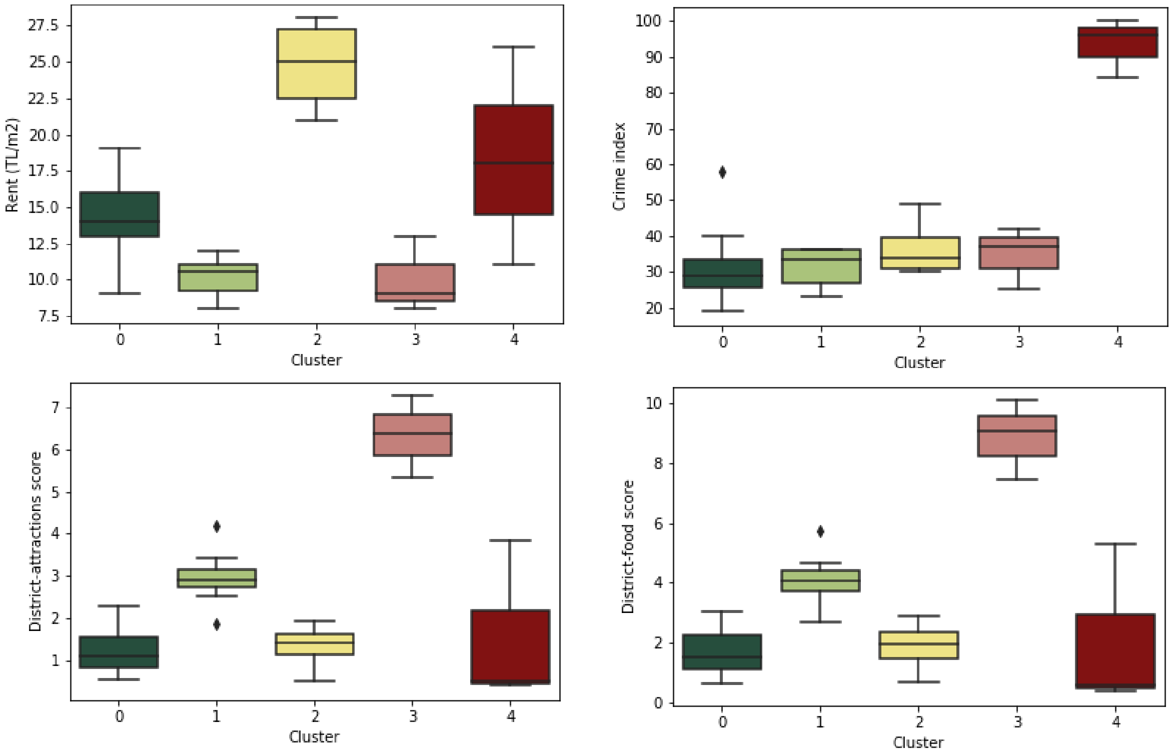


Figure 2. Boxplots showing statistics of clusters for each considered feature. Top-left: average rent prices; top-right: crime index; bottom-left: district-attraction score; bottom-right: district-food score.

In order to spatially visualize performed analysis and the statistics associated with the city of Istanbul, gathered information were overlaid on a few maps. Figure 3 shows the classification of districts into clusters, Figure 4 shows the location of the top 100 eateries and sightseeing places, Figure 5 shows average rent prices per m^2 and Figure 6 shows calculated crime indices. All the gathered statistics overlaid on an interactive map can be viewed at:

https://nbviewer.jupyter.org/github/nowacowski/Coursera_Capstone/blob/master/Istanbul/istanbul_map.html

Comparing boxplot charts with associated maps, a few conclusions can be drawn for the determined clusters.

- Best districts to stay at, classified as cluster 0, are located just outside the strict city centre.

- As it could be expected, the clusters with the lowest rent prices consist of districts located the furthest from the city centre. However, the cluster with the highest rent (significantly higher than other clusters, top left chart of Figure 2) consist of districts not in the strict city centre, but rather along the European coast between European and Asian parts of Istanbul.

- Looking at crime index boxplot (top right chart of Figure 2), a striking difference can be noticed between cluster 4 and other clusters. Districts in cluster 4 on average obtained around 3 times higher crime index than all other clusters. Two of the districts from cluster 4 are in the strict city centre, however, the third district is located far from the city centre. Better inspection of the crime statistic data show that districts from cluster 4 located in centre indeed have one of the highest number of committed crimes, however, the high crime score obtained by the other district might be caused by the low population of that district.

- District-attraction and district-food scores box plots for clusters seem to show almost identical statistics (bottom charts of Figure 2). Investigating the map on Figure 4, it can be concluded that this is because eateries and sightseeing places are located in the close proximity to each other. Additionally, because all top places are congested in the city centre, the calculated scores are highly dependent on the distance of the districts from the city centre.

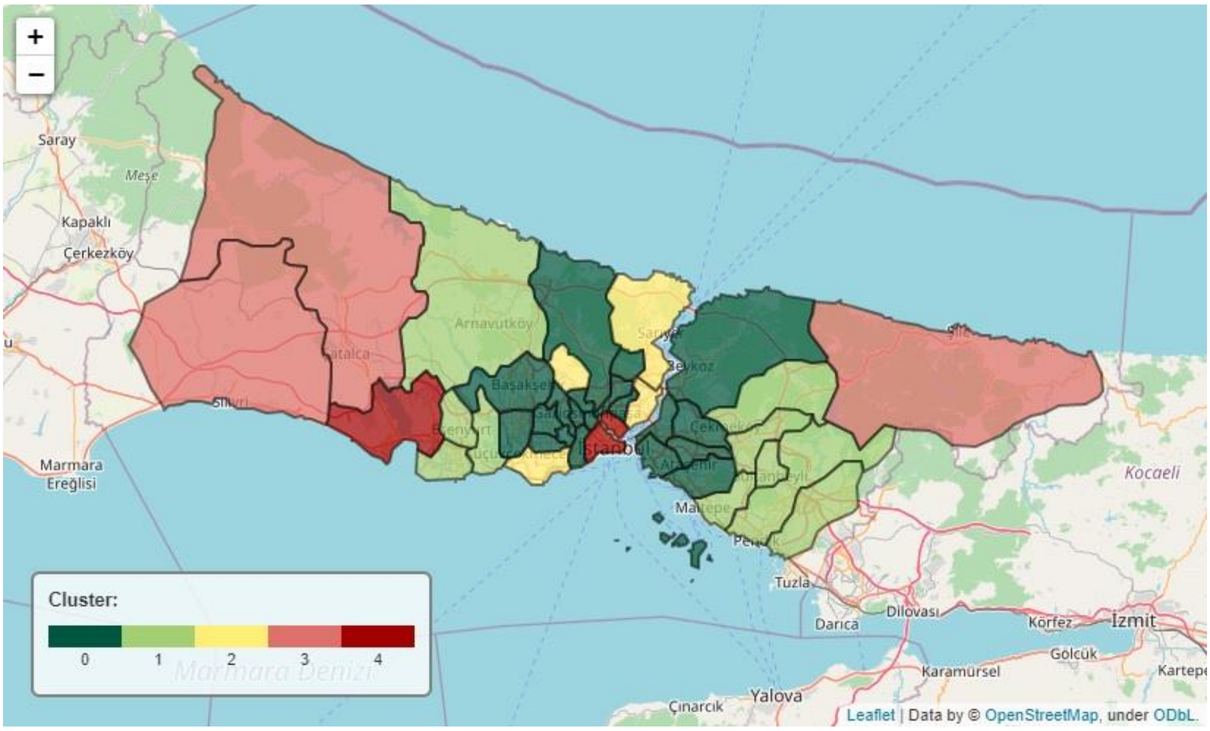


Figure 3. Map of Istanbul divided into districts coloured according to the assigned cluster.

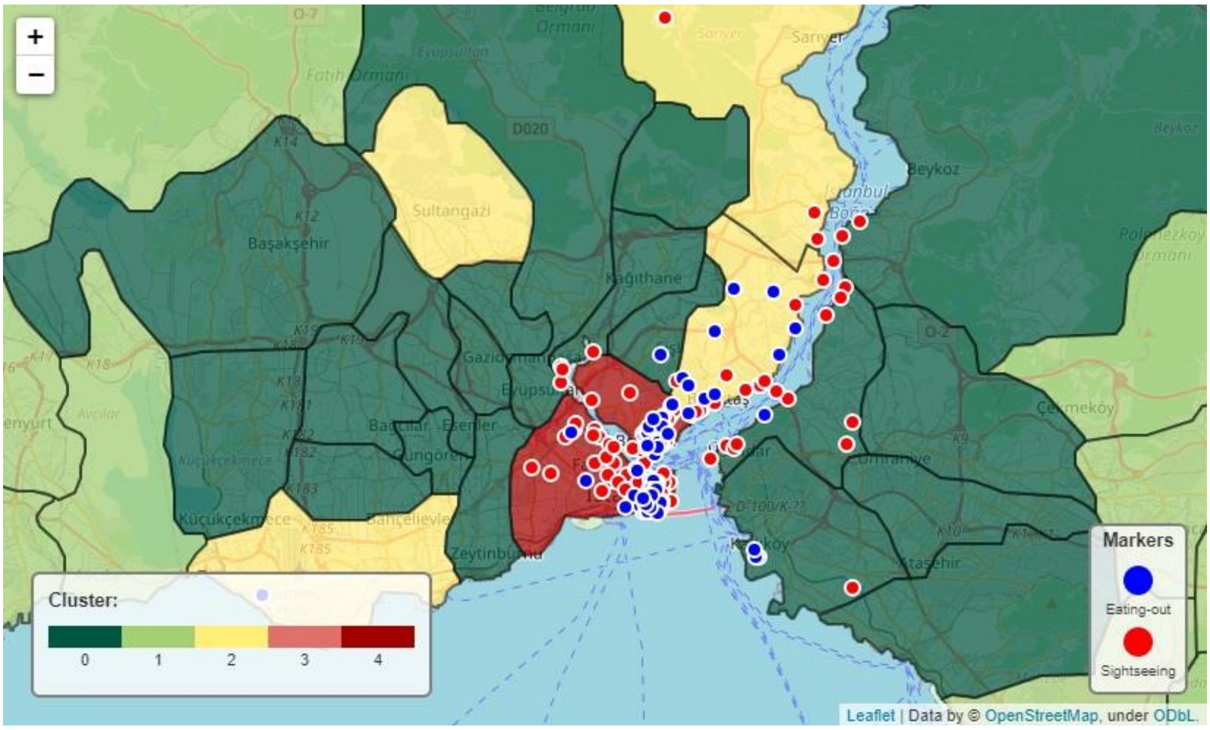


Figure 4. Close-up map of Istanbul divided into districts coloured according to the assigned cluster, with marked top 100 eating out (blue dot) and sightseeing (red dot) places.

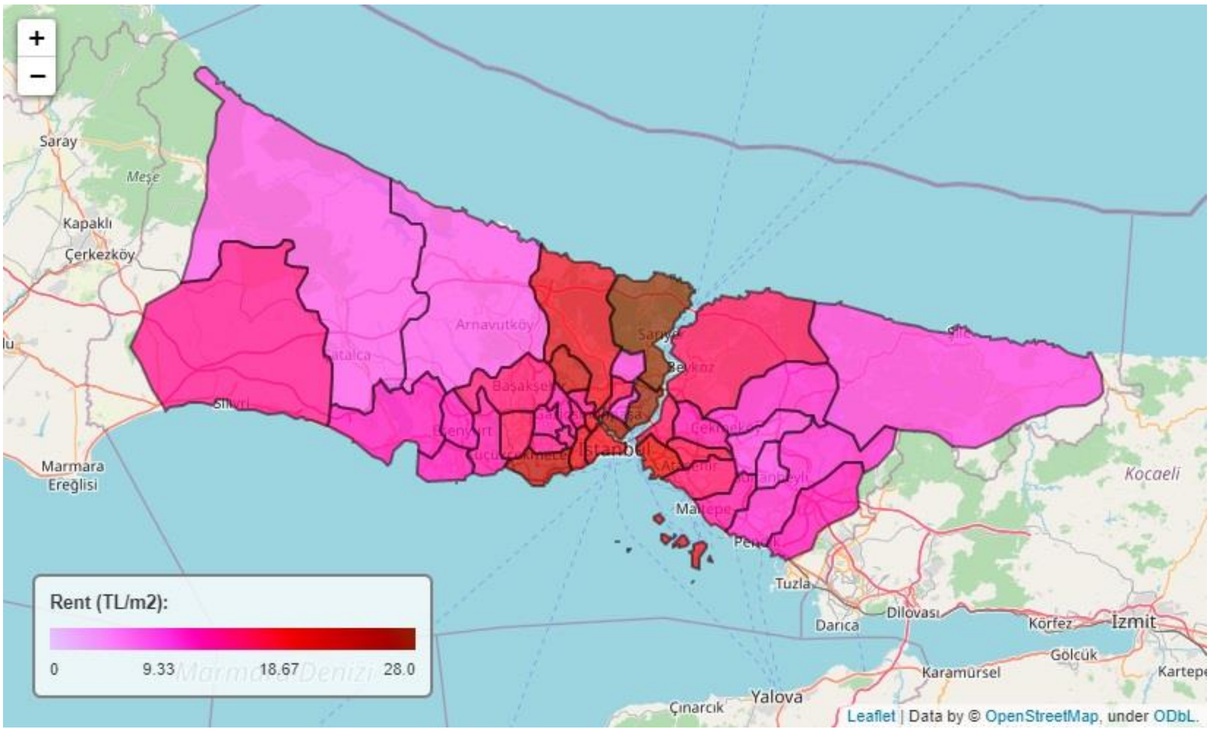


Figure 5. Map of Istanbul divided into districts coloured according to the average rent per m^2 as showed on the legend.

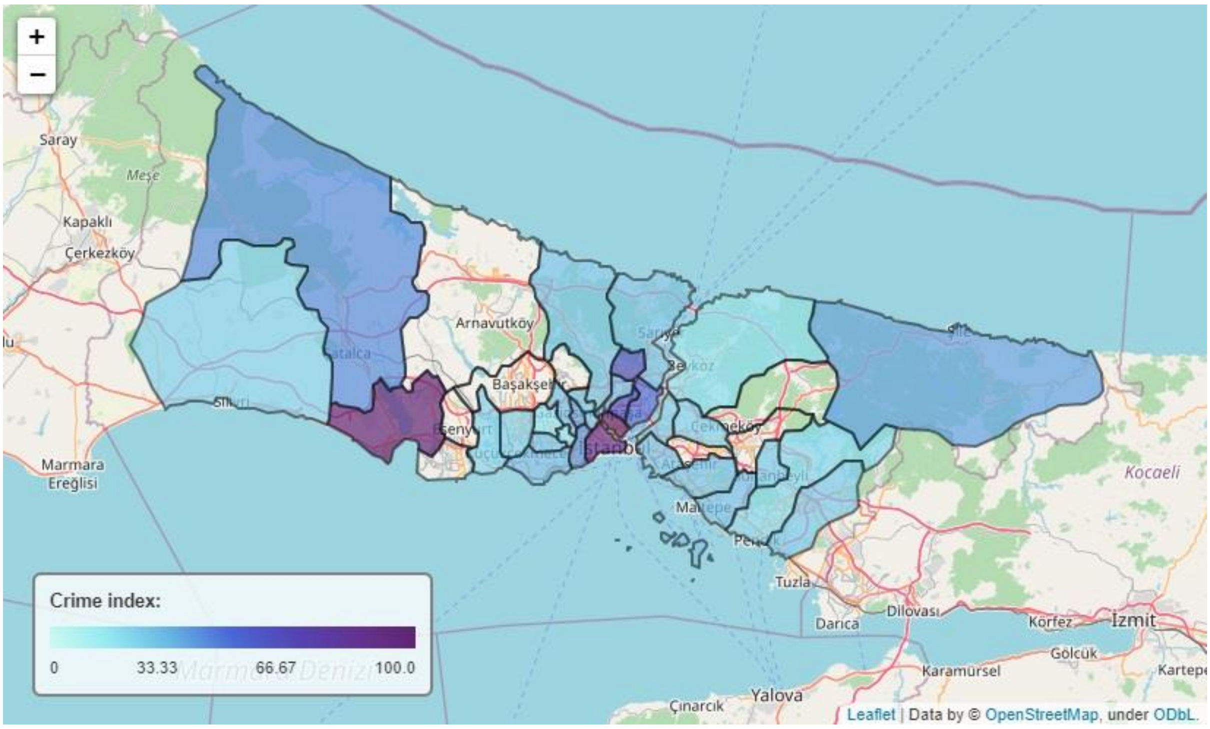


Figure 6. Map of Istanbul divided into districts coloured according to the calculated crime index as showed on legend. Uncoloured district correspond to districts for which no crime statistics were found.

Conclusions

By implementing K-means algorithm, 39 districts of Istanbul were classified into 5 groups, based on price, safety, and proximity to best eateries and sightseeing places. By sorting obtained results, the best and worst districts to stay at while visiting Istanbul were deducted.

The best to stay turned out to be districts located just outside of the city centre, while the worst, due to the high prices and crime, are districts in the strict city centre.

Although the final results seem to agree with the common view that it is the best to stay close to the centre but not in the centre itself, the performed analysis could still be improved. Used crime statistics data were from over 10 years ago. During that time a lot could change, therefore, using more recent data would be more beneficial. For the accommodation prices, long term rent prices were used instead of hotel prices per night. Although it is assumed that both would be correlated, without having both data one cannot be completely certain about it. As for the district-attraction and district-food scores, distance in straight line was used. Instead, commute time would be better, differentiating districts hard to get to, e.g. located on islands.

Jupyter notebook containing step by step process leading to creating of this report can be viewed at:

https://nbviewer.jupyter.org/github/nowacowski/Coursera_Capstone/blob/master/Istanbul/Istanbul_travel.ipynb