# Final project - The Battle of Neighborhoods

## Coursera | Applied Data Science Capstone

by Martin Kovarik

## Week 5



Table of Contents

[Final project - The Battle of Neighborhoods 1](#_Toc61376656)

[Coursera | Applied Data Science Capstone 1](#_Toc61376657)

[Week 5 1](#_Toc61376658)

[Assignment 2](#_Toc61376659)

[1a) Introduction where you discuss the business problem and who would be interested in this project. 3](#_Toc61376660)

[1b) Data where you describe the data that will be used to solve the problem and the source of the data. 4](#_Toc61376661)

[1c) Methodology section 5](#_Toc61376662)

[1d) Results section 8](#_Toc61376663)

[1e) Discussion section 8](#_Toc61376664)

[1f) Conclusion 8](#_Toc61376665)

### Assignment

1. A full report consisting of all of the following components (15 marks):
   1. Introduction where you discuss the business problem and who would be interested in this project.
   2. Data where you describe the data that will be used to solve the problem and the source of the data.
   3. Methodology section which represents the main component of the report where you discuss and describe any exploratory data analysis that you did, any inferential statistical testing that you performed, if any, and what machine learnings were used and why.
   4. Results section where you discuss the results.
   5. Discussion section where you discuss any observations you noted and any recommendations you can make based on the results.
   6. Conclusion section where you conclude the report.

### 1a) Introduction where you discuss the business problem and who would be interested in this project.

As a use case for this project, I chose the young family looking for a 2 bedroom apartment to rent in London. As London offer immense pool of apartments to rent, it is a good idea to:

**[i]** either ev,

**[ii]** or have a tool at a disposal which would help the family quickly evaluate the rental opportunity.

In the first case, the criterion to limit the search can be simply the monthly rent rate. Especially a young family, which may need to change the apartment with the increasing family size, will hesitate to pay high rent for an apartment which is just temporary for them. We can help them by categorizing the London boroughs by a median of local rents and visualize those on the map of London. This way they may limit their search more effectively to boroughs which better correspond to their rent rates expectations.

In second case, we can help them evaluate the rental opportunity by describing the given borough typical venues. Or even provide them set of clusters of boroughs which share similar characteristics. This way they can immediately get a basic idea what to expect in the given borough where the apartment is situated.



### 1b) Data where you describe the data that will be used to solve the problem and the source of the data.

To achieve the goal described in the *description* section, we will use 3 sources of data.

1. First we will gather list of **London's boroughs and their coordinates** from Wikipedia web page. (<https://en.wikipedia.org/wiki/List_of_London_boroughs>).

This will allow us to visualize the boroughs on the map of London.

*Example of final output from this section:*

*Table

Description automatically generated*

2. Secondly, we will use Foursquare API to extract information about **venues** available in given London's boroughs. Applying the one hot encoding along with data transformation into easily readable dataframe we create a simple overview of the typical venues in each borough. Then using clustering method K-means, we will group the boroughs into clusters with similar features. All of those actions will help us get an additional insight about nature of each borough.

*Example of final output from this section:*

Table

Description automatically generated

3. Thirdly, we will use the use the public data about **rental rates** in London. (<https://www.ons.gov.uk/peoplepopulationandcommunity/housing/adhocs/12435privaterentalmarketinlondonoctober2019toseptember2020>)  
We will use this to create an easily understandable, visual overview of the rental rates in London, to allow our user group refine their decisions about where to focus their apartment search.

*Example of final output from this section:*

Table

Description automatically generated

### 1c) Methodology section

Methodology section which represents the main component of the report where you discuss and describe any exploratory data analysis that you did, any inferential statistical testing that you performed, if any, and what machine learnings were used and why.

#### [i] London's rental rates visualization

As a first step we will combine two data sets described in *Data* section, **Rental rates** and **London's boroughs with their coordinates.**

Table

Description automatically generated

To visualize that on the map of London, we will also need London’s "Geo" data used to plot a *choropleth* map. We can easily find then on the internet. To streamline the process, the "Geo" file is stored in the project directory ('london\_boroughs.json').

To get nicely focused map, we use Nominatim module from *geopy* library to get geo coordinates of London.

Now we have all the elements to plot a heat map showing us the London's boroughs colored based on the Median of the monthly rental rates (in £).

Map

Description automatically generated

#### [ii] Describing London's boroughs with boroughs typical venues

In second phase we are going to describe the London's boroughs via typical venues existing in the area. For that we will use data from Foursquare database.

Table

Description automatically generated

Applying *one hot encoding* and pandas *groupby* method, we create overview of typical venues for each London's borough. For a convenience we transform the dataframe into easily understandable table.

Graphical user interface, application

Description automatically generated

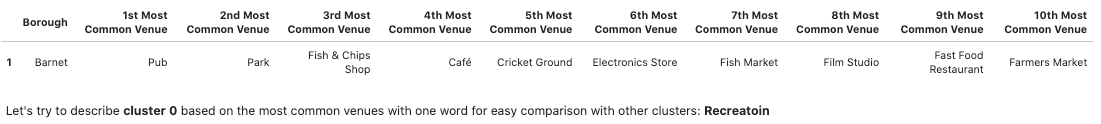
#### [iii] Clustering the boroughs based on similarity of their venues

Last step is to cluster the boroughs based on their similarity, using the Foursquare data. Using the K-means method we cluster the boroughs into 4 clusters and visualize the results on the map of London.

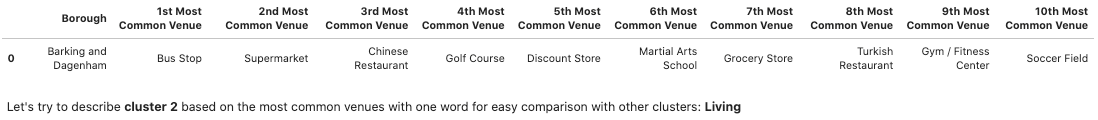
Diagram

Description automatically generated with medium confidence

Looking at the dataframes of each cluster, we can make a qualified assumption about the nature of each cluster.



Table

Description automatically generated

Graphical user interface, table

Description automatically generated with medium confidence

### 1d) Results section

Our family looking for a two-bedroom apartment can now refine their search using the insights we provided for them. Heat map with rental rates, provides a predictable picture, which shows lower rents in the boroughs on the periphery of the London. Still, the scale of differences is quite illustrative and can help them decide about limiting their efforts just for apartments in boroughs with lower rental rates.

As an additional tool for their decision about their new apartment location are the clusters of boroughs, where we showed that certain boroughs share similar characteristics. They can use that insight to either refine their search parameters or decide about which apartment they are going to move to, in case they will need to decide between similar rental opportunities.

The foursquare data, where we calculated the 10th most common venues in each borough, can also serve them as additional tool to evaluate the given apartment based on character of the borough described by those venues.

### 1e) Discussion section

Doing the research for this project, I realized that using machine learning method and power of visualization, we can provide our clients an immense amount of information, which would be either not accessible for them or they would have to put a significant effort into gathering those data by themselves.

### 1f) Conclusion

The biggest limit of this project is simply the limitation in the scope of the information I decided to use and the depth of analysis of my demo users.

In real world scenario, applying a design thinking method and employing additional data resources, would provide even stronger apparatus for client's decisions.