Battle of Neighborhoods, Amsterdam

Pranjal Biyani

October 21, 2019

1.Introduction

1.1 Background

Amsterdam is the capital city of the Netherlands, with a population of around 2.4 million in the metropolitan area. The city also is the commercial and cultural capital of the Netherlands and is considered an alpha- world city. Amsterdam is also one of the most popular tourist destinations in Europe, receiving more than 4.63 million international visitors annually. Along with Frankfurt, the city is one of the gateways to Europe for travelers from the East.

With the influx of travelers to the city increasing by every year, there has also been a surge in the number of places of accommodation to cater to the increasing demand. Along with the traditional options like hotels and lodgings, cheaper and flexible alternatives like Airbnb stays and hostels have become more popular among tourists. This leaves the travelers with myriad of options to choose from but at the same time present a dilemma in deciding which location serves them best considering all factors like affordability, proximity to places of interest etc.

1.2 Business Problem

The project attempts to provides the travelers to Amsterdam with a comprehensive comparison of different neighborhoods of the city to help them when deciding the location of their stay. We also classify the neighborhoods into clusters of similar characteristic and thus providing our target audience a better idea upfront about the type of neighborhoods they are looking at.

For this project, we use the Airbnb listing data for Amsterdam along with the location data from Foursquare API listing the happening venues around the neighborhood of stay. The project targets mainly solo travelers and back-packers who, it is presumed, would benefit the most from the exercise.

2. Data acquisition and Preparation

2.1 Data Sources

As the main dataset, we use the Airbnb listing for Amsterdam available here and list of neighborhoods of Amsterdam available here on Kaggle website. This dataset consists of information like neighborhood, number of listings, listing price, and number of reviews, geospatial data of the listing etc.

Further to add more useful features to the data set before we compare the neighborhoods, we leverage the location data from Foursquare API to find out happening venues based on general categories of interest to travelers. First, the coordinates of the neighborhoods are obtained by using the Python's

Geopy library and the coordinates are later used to send Foursquare API requests to bring in nearby venues information for a selected list of venue categories.

As the target audience are the tourists to the city, for this analysis we have identified the below main venue categories of interest to tourists. There are many subcategories under each of the categories, the main category “Food” for example, has subcategories like Chinese restaurant, Turkish restaurant etc.

a. Arts & Entertainment

b. Food

c. Nightlife Spot

d. Outdoors & Recreation

e. Shop & Service

Using the Foursquare API, we bring in the number of venues of each of these categories to help us classify the neighborhoods.

2.2 Data cleaning

As we are mainly targeting solo travelers and backpackers, who are more likely to go for single or shared rooms rather than full apartments, we work with a subset of the full data set and all full apartment listings will be removed. For the same reason, we will remove any listing where the minimum night stay required is more than a week as our target will be mostly looking for short term stay.

We also remove any listing that was last reviewed before year 2018 treating this as an indicator of the listing being inactive. Listings with zero reviews till now also are removed citing irrelevance. We have one record with zero price value and is removed as bad data as we have only a single case like this.

Finally, a quick stats-describe check on the price column along with its histogram plot revealed one extreme outlier of value 5000 and is promptly removed.

2.3 Feature Selection

After cleaning, we derive 3 relevant features from the main Airbnb dataset aggregated for each neighborhood and create a mini dataset with 3 features. These are

1. Number of Listings

2. Average Price

3. Average number of reviews.

We now combine this dataset with the venue count for each main category of venues that we have already identified and prepare the full data set with all relevant features.

	Neighborhood	Latitude	Longitude	number_of_listings	avg_price	avg_review_count	Arts & Entertainment	Food	Nightlife Spot	Outdoors & Recreation	Shop & Service
0	Bijlmer-Centrum	52.317257	4.950483	66	103.409091	41.090909	22.0	56.0	13.0	20.0	74.0
1	Bijlmer-Oost	52.321193	4.975637	59	88.457627	34.000000	9.0	9.0	4.0	5.0	15.0
2	Bos en Lommer	52.378521	4.848738	708	117.156780	25.929379	9.0	4	7.0	25.0	46.0
3	Buitenveldert - Zuidas	52.336020	4.865890	128	137.617188	20.507812	9.0	48.0	9.0	22.0	35.0
4	Centrum-Oost	52.358983	4.924509	1108	185.080325	41.322202	12.0	69.0	26.0	26.0	73.0
5	Centrum-West	52.373730	4.895691	1582	191.352086	45.010746	59.0	100.0	100.0	57.0	100.0
6	De Aker - Nieuw Sloten	52.346370	4.796833	87	113.114943	48.471264	3.0	17.0	4.0	4.0	13.0
7	De Baarsjes - Oud-West	52.366158	4.862715	2172	145.706262	27.480203	23.0	100.0	42.0	42.0	95.0
8	De Pijp - Rivierenbuurt	52.346579	4.917350	1462	154.714090	28.514364	4.0	39.0	6.0	19.0	25.0
9	Gaasperdam - Driemond	52.312666	4.989188	84	90.428571	30.333333	3.0	4.0	2.0	5.0	4.0

Figure 1. A snippet of the Airbnb Listings dataset from Kaggle

3. Methodology

3.1 Exploratory Data Analysis

To start out, we visualize the neighborhoods on a map of Amsterdam. There are 21 neighborhoods in the city which we will display on a map to get an idea about which are closer to the center and which are located towards the perimeters of the city limit, and how far they are located from each other.

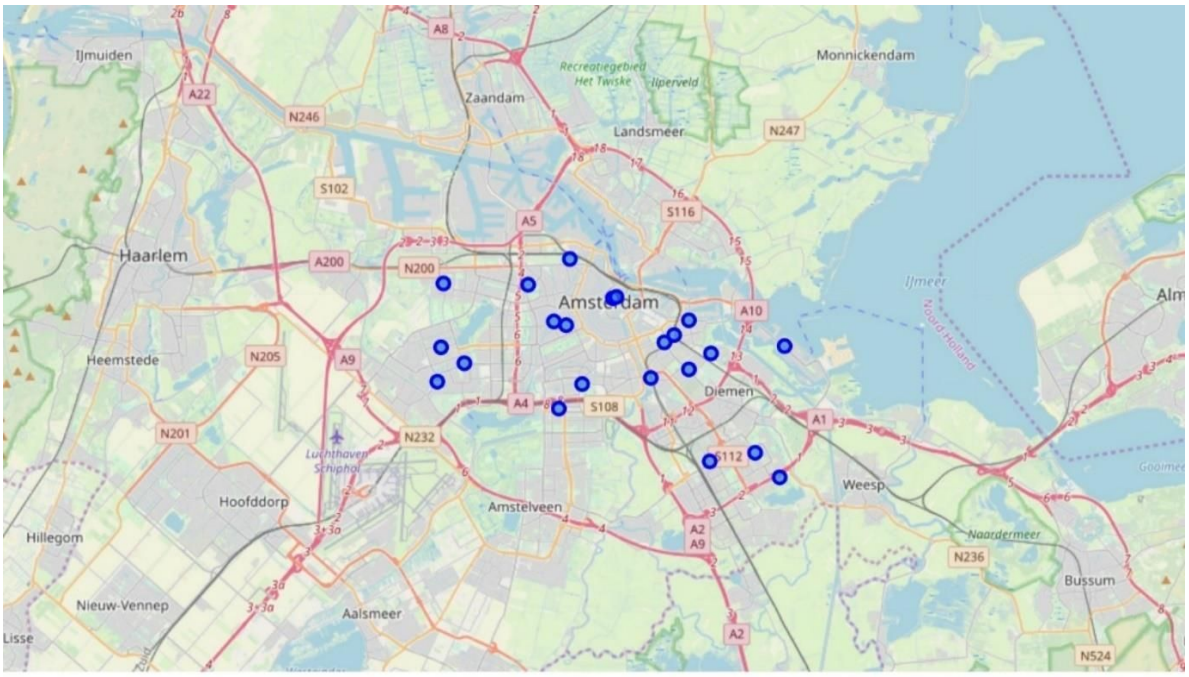

Figure 2. Amsterdam neighborhoods on a map of Amsterdam

As seen from the map, most of the neighborhoods are located ‘centrally’ and there are two distinct outer spatial clusters having three neighborhoods each. There also are two neighborhoods, “IJburg – Zeeburgereiland” and “Geuzenveld – Slotermeer”, which are further far from the center and also from other neighborhoods.

Now, we analyze the features that we derived from the Airbnb dataset by plotting these for the neighborhoods (Figure 3). It is seen that central neighborhoods like Centrum-West, Centrum-Oost, De Baarsjes - Oud-West, Noord-West, Oud-Noord score high for each category and these should form a cluster. The maximum number of listings is in the neighborhood Centrum-West and so it appears to be the most popular neighborhood for stay. Average prices are lower and are similar among the

neighborhoods which are not central. Maximum number of listings are with price range between 70-100 (Figure 4) and are in neighborhoods located non-centrally.

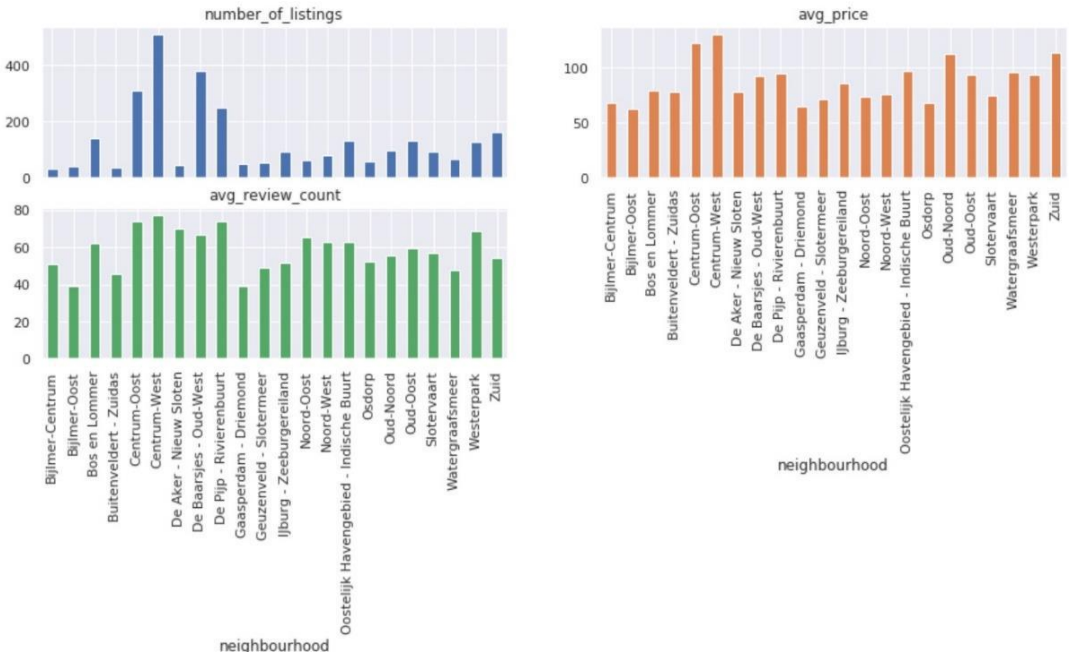

Figure 3. Airbnb data features plotted for neighborhoods.

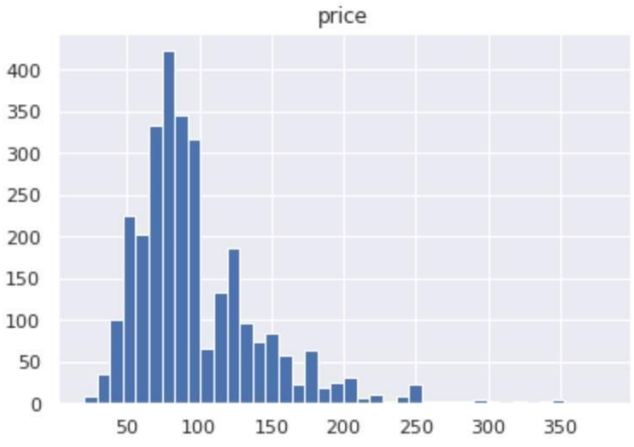

Figure 4. Histogram of average price.

3.2 K-Means Clustering:

Equipped with the insights gathered from the exploratory data analysis, we classify the neighborhoods into three distinct clusters of neighborhoods using the unsupervised K-mean clustering algorithm. The value of K is conveniently chosen to be 3 as we have only 22 neighborhoods to cluster and in order not to enforce the algorithm to create a cluster of one or two neighborhoods alone if we select a higher value of K. Before clustering, the feature sets are normalized using a Min-Max normalization algorithm as the features are of different scale and hence this may bias the clustering algorithm.

4. Results

The algorithm classified the neighborhoods with similar characteristic into 3 clusters of different sizes.

Cluster	**Labels Neighborhoods**
Bijlmer-Oost,	Buitenveldert - Zuidas, De Aker - Nieuw Sloten, Gaasperdam - Driemond, Geuzenveld -
	Slotermeer, IJburg - Zeeburgereiland, Noord-Oost, Osdorp, Slotervaart,
	1 Watergraafsmeer
2	Centrum-West, De Baarsjes - Oud-West, Noord-West, Oud-Noord
	Bijlmer-Centrum, Bos en Lommer, Centrum-Oost, De Pijp - Rivierenbuurt, Oostelijk
3	Havengebied - Indische Buurt, Oud-Oost, Westerpark, Zuid

Figure 5. Final clusters and the neighborhoods.

Now to characterize each cluster, we analyze them based on the features. As we have different number of neighborhoods in different clusters, we take the average values of each features in each cluster display them. We also show the total number of neighborhoods in each cluster.

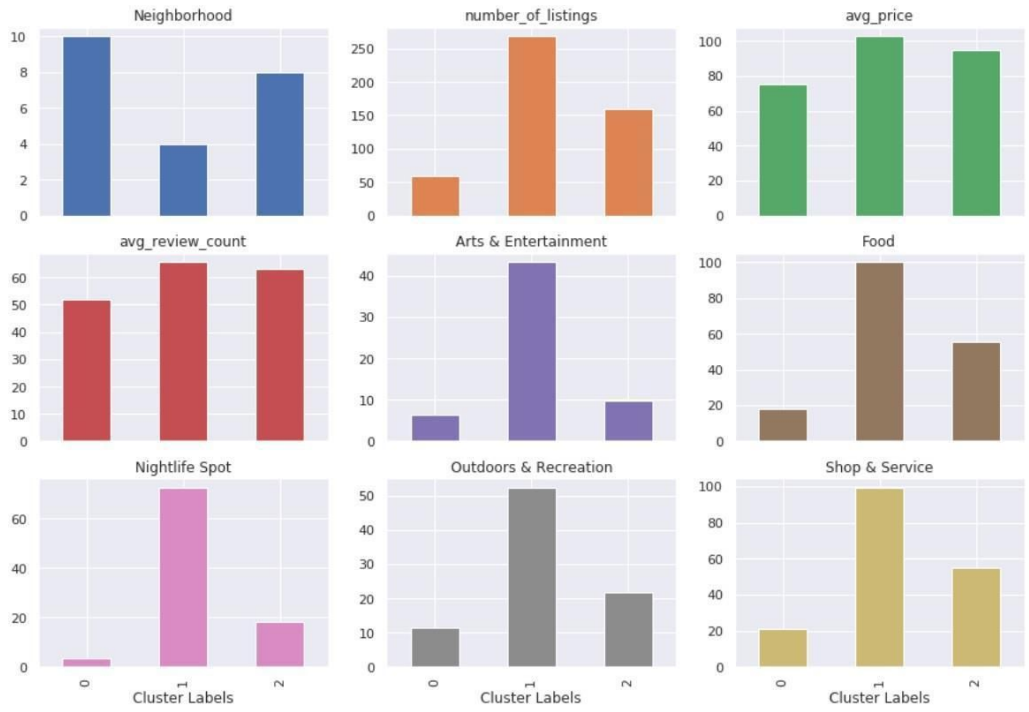

Figure 6. Bar graph of average feature values for each cluster.

The clsuter which stands out is clsuter 1 which on inspection turns out be having the four neighboorhoods located centrally as expected. So, clsuter 1 is the main central area and has the highest concentration of all types of venues we which considered but is also costlier on average.

However, cluster 2 has more(8), neighborhoods and is cheaper than cluster 1 on average and also matches or exceeds cluster 1 in total number of listings and venues of all categories

It is also revealed that even with in a cluster most of the feature values varies significantly (Figure 7).

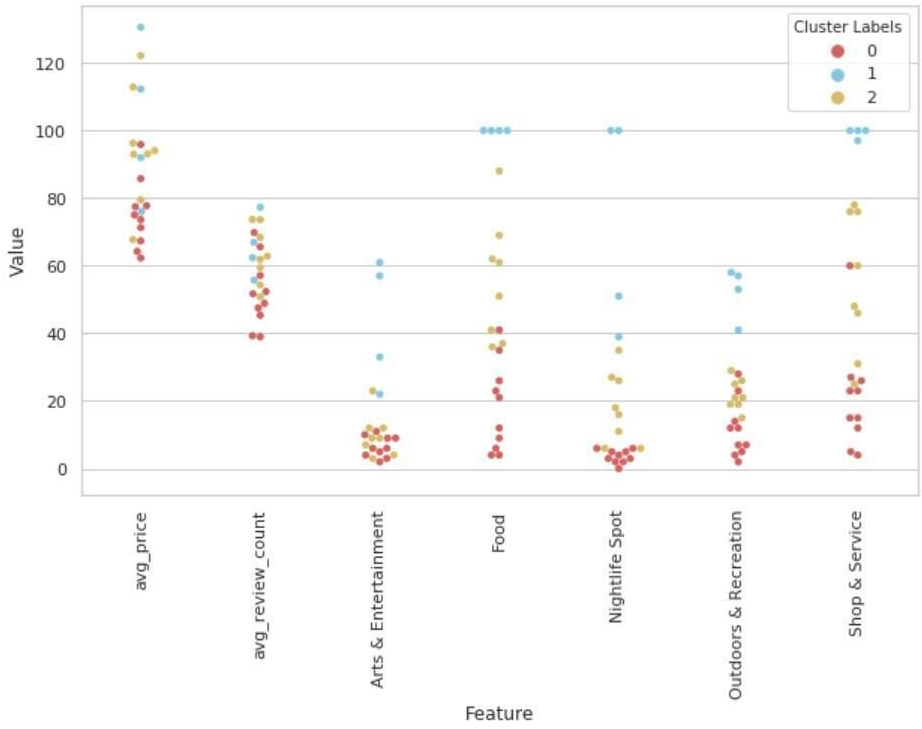

Figure 7. Swarm plot of neighborhood features in each cluster.

Finally, we color code the three clusters and display the clusters of neighborhoods on a map of Amsterdam along with the values of the features we based the clustering on (Figure 8).

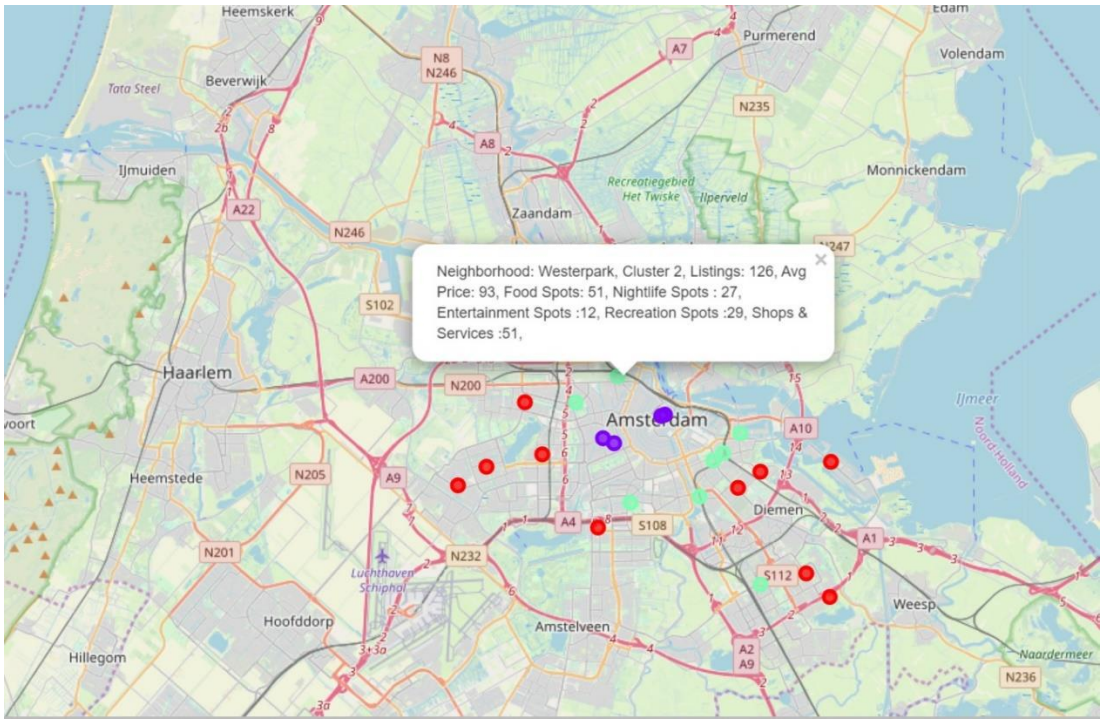

Figure 8. Final three neighborhood clusters on a map of Amsterdam.

We end this exercise by visualizing the venue concentration and other three feature values in each cluster's neighborhoods using a heatmap.

Figure 9. Feature heatmap Cluster 0

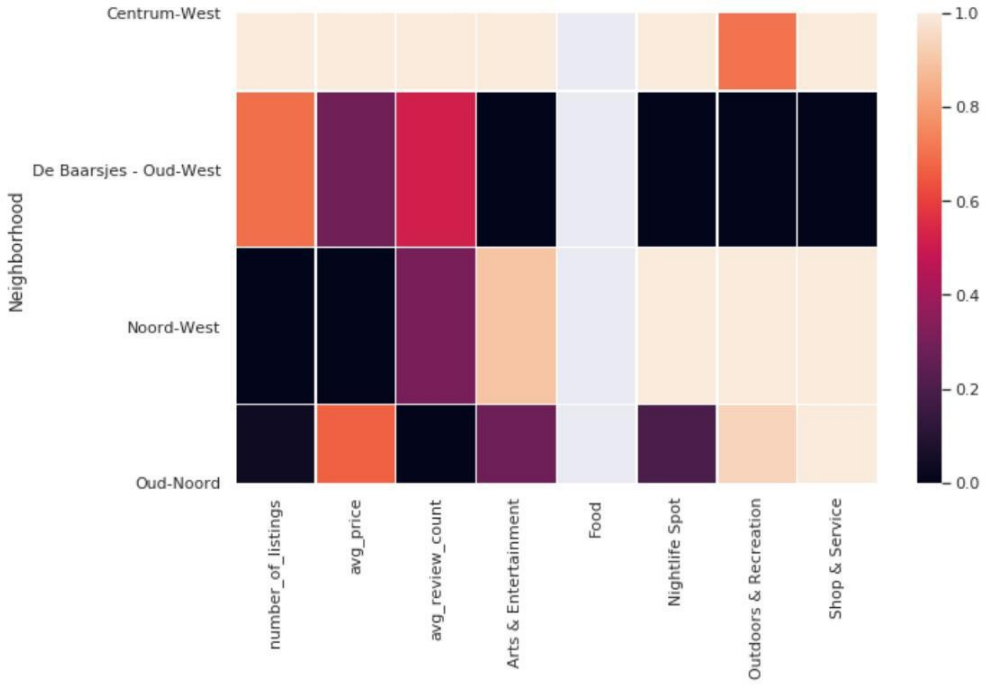

Figure 10. Feature heatmap Cluster 1

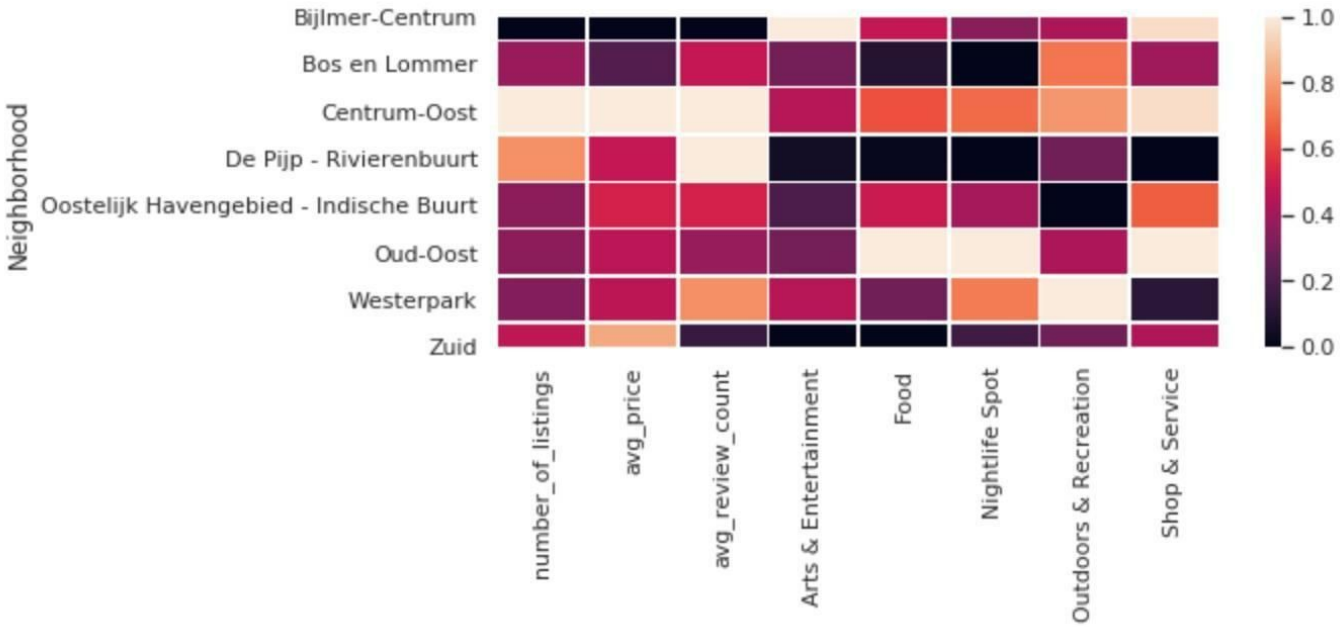

Figure 11. Feature heatmap Cluster 2

5. Discussion

The objective of this exercise was to provide a comprehensive comparative analysis of the different neighborhoods to our target audience by grouping the neighborhoods into clusters of similar characteristics. The intent was not to pick an outright winning cluster as the centrally located neighborhood cluster was expected to score higher in most of the categories. However, this analysis revealed that “cluster 1” neighborhoods combined had more venues of interest and was also on par with the central cluster in terms of popularity as measured from the number of listings and was also cheaper on average. We also saw that even with in a cluster most of the values varies significantly

Future recommendations

We had excluded listings of full homes/apartments from the analysis as we were targeting solo short term stay travelers. Including these can reveal different picture in terms of the number of listings shift between clusters as such facilities normally tend to be located non-centrally, reinforcing the case for such neighborhoods even more. Also, the data set on Kaggle website that we used was a little old, so we could try to obtain dataset with more recent listings and this may present a different case for each neighborhood.

6. Conclusion

The objective of this project to provide comprehensive comparison of different neighborhoods of Amsterdam, and to classify its neighborhoods into clusters of similar characteristic to give a better idea upfront to tourist about the type of neighborhoods they are looking to book their stay in. For this we used Airbnb listing data along with Foursquare location data to run a K-Means clustering algorithm on selected features.

We formed 3 different clusters of neighborhoods and characterized the clusters and analyzed the clusters individually. We also created visualized these clusters on a map of Amsterdam and displayed a snippet of venue and price info for each neighborhood on the map. We also made a few recommendations to be incorporated in future to the model to further help our analysis.