IBM Applied Data Science Capstone Project

The Battle of the Neighborhoods

Tomas Dmitrijevas

2020-08-12

1. Introduction

Vilnius – capital city and the largest city of Lithuania. In recent years Vilnius has achieved a favourable landscape for businesses to invest and grow in the city. With yearly growth of establishments of business Service Centers from all across Europe, Vilnius needs to adapt to a changing demand in services it can provide for business companies. As the number of highly diverse specialists relocating to Vilnius is growing, the need for new places to refresh and eat is growing as well.

Young and highly motivated business entrepreneurs are willing to help their city and open a new Pizza restaurant. There are many details and risks to be considered when opening a restaurant. The problem to decide on a most suitable city location for a new venture is in question.

For decision on a suitable place to be made, some of the factors need to be analysed:

- population of the neighborhood
- area of the neighborhood
- number of competitors in the neighborhood.

In this project, the most suitable location for a new Pizza restaurant will be analysed based on above mentioned criterias and insights for business owners will be provided based on conducted modeling of Vilnius neighborhood locations.

2. Data description

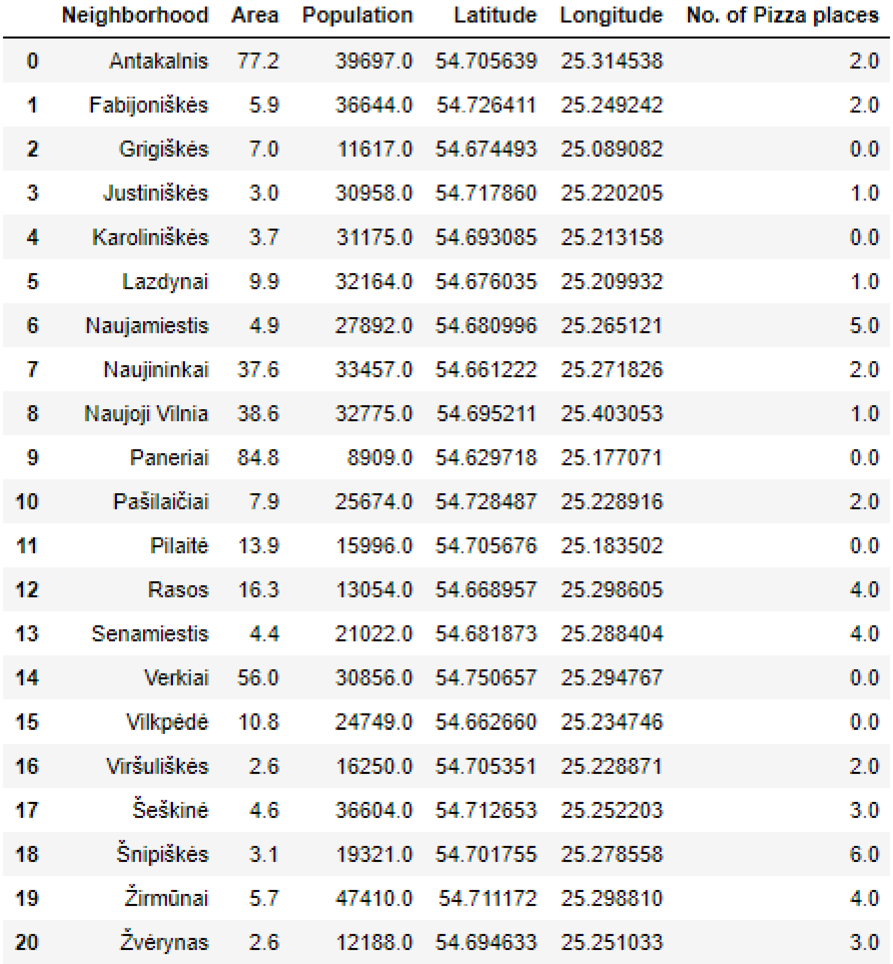
Data for the project was retrieved from from several public data sources. Data sources that are used for further data analysis of the project:

1. "Neighborhoods of Vilnius". Source: https://en.wikipedia.org/wiki/Neighborhoods_of_Vilnius
    - List of neighborhoods of Vilnius, Lithuania
    - Area, km2 of Vilnius neighborhoods
    - Population of Vilnius neighborhoods
2. Foursquare API:
    - Data of the venues that are located in Vilnius neighborhoods in defined radius
3. Geolocator API
    - Latitude and Longitude coordinates of analysed locations

All retrieved data was investigated, cleaned and adjusted, if needed, using these data manipulation methods:

1. Unnecessary symbols cleaning
2. Unnecessary data columns dropping
3. Conversion of features to a correct data type
4. Value sorting
5. Value aggregation

After data cleaning steps were taken, dataframe for further analysis was generated.



	Neighborhood	Area	Population	Latitude	Longitude	No. of Pizza places
0	Antakalnis	77.2	39697.0	54.705639	25.314538	2.0
1	Fabijoniškės	5.9	36644.0	54.726411	25.249242	2.0
2	Grigiškės	7.0	11617.0	54.674493	25.089082	0.0
3	Justiniškės	3.0	30958.0	54.717860	25.220205	1.0
4	Karoliniškės	3.7	31175.0	54.693085	25.213158	0.0
5	Lazdynai	9.9	32164.0	54.676035	25.209932	1.0
6	Naujamiestis	4.9	27892.0	54.680996	25.265121	5.0
7	Naujininkai	37.6	33457.0	54.661222	25.271826	2.0
8	Naujoji Vilnia	38.6	32775.0	54.695211	25.403053	1.0
9	Paneriai	84.8	8909.0	54.629718	25.177071	0.0
10	Pašilaičiai	7.9	25674.0	54.728487	25.228916	2.0
11	Pilaitė	13.9	15996.0	54.705676	25.183502	0.0
12	Rasos	16.3	13054.0	54.668957	25.298605	4.0
13	Senamiestis	4.4	21022.0	54.681873	25.288404	4.0
14	Verkiai	56.0	30856.0	54.750657	25.294767	0.0
15	Vilkpėdė	10.8	24749.0	54.662660	25.234746	0.0
16	Viršuliškės	2.6	16250.0	54.705351	25.228871	2.0
17	Šeškinė	4.6	36604.0	54.712653	25.252203	3.0
18	Šnipiškės	3.1	19321.0	54.701755	25.278558	6.0
19	Žirmūnai	5.7	47410.0	54.711172	25.298810	4.0
20	Žvėrynas	2.6	12188.0	54.694633	25.251033	3.0

Figure 1. Dataframe used in project analysis

3. Methodology and Analysis

To analyse defined problem of the project a suitable analysis method need to be identified. The goal of analysis is to identify neighborhoods which are suitable for a new Pizza place in the city. All regression models are out of focus as the final goal is not of linear nature. Classification models are selected to be used as the problem requires neighborhoods to be labeled/clustered. Based on unlabeled data that is used in the problem solving of the project, it was decided to use k-Mean clustering model to segment all neighborhoods of Vilnius in accordance of their similarity.

3.1. Visual and Data Analysis

To analyse data of neighborhoods, first visualization was created. The map with centers of all listed and analysed neighborhoods was plotted on top of the map of Vilnius city. After visual investigation of generated map, it was noticed that some neighborhoods centers are significantly misaligned with the neighborhood labels on the Vilnius map. Neighborhoods location coordinates were not completely correct in the data set that was scraped from the web. Google Geolocator API was used for identifying correct longitude and latitude coordinates of neighborhood centers.

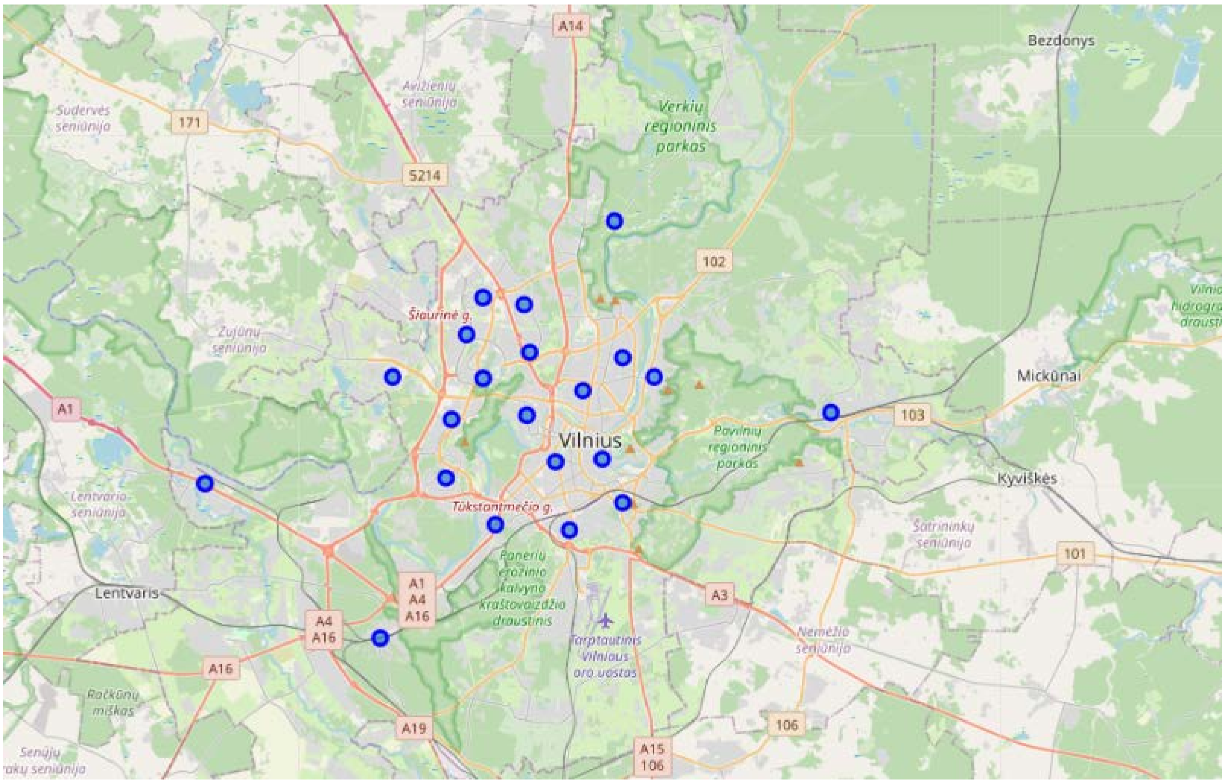


Figure 2 Vilnius map with centers of the neighborhoods

One more iterative cycle of plotting Vilnius city neighborhood centers was executed and plot was again inspected visually. No unexpected discrepancies in plotted centerpoints were noticed and analysis was continued with retrieving of all venues in the neighborhoods.

Foursquare API was used in the next step of the project analysis. After defining access credentials, algorithm that fetches limited number of venues in every neighborhood was defined and used. It is important to notice, that only venues that are in the radius of 500 meters from neighborhood centers were retrieved into a new dataframe. 725 venues were retrieved from Foursquare API after running the algorithm in total.

Of course, some Pizza places are already open in Vilnius city. All opened spots may have a significant impact on the opening of a new local Pizzeria. If there are many competitors in the area, people might not be interested in another Pizza restaurant and those regions should be out of our focus. To better understand local business prospects in Vilnius areas, all retrieved pizzerias in every neighborhood were counted. As an assumption, pizzas are served in Italian restaurants as well, therefore all Italian restaurants in Vilnius neighborhoods were added. 42 local competitors in Vilnius were counted in total. Some neighborhoods in Vilnius has no Pizza place venues. Missing values were replaced with 0 as a Pizza place count of the neighborhood.

3.2. Cluster Analysis

Some data ambiguity was noticed in the data set. Neighborhood density can be calculated dividing population by area. To avoid correlated features, it was decided to drop neighborhood density feature in the data cleaning step. Neighborhood Area and Population features were left in the dataset as some useful insights might be retrieved after the cluster analysis is conducted by visually exploring clustered relations of features.

As neighborhood area, population and no. of Pizza places features have different variations and scales, these features were Standardized to better fit clustering model. For standartisation purposes StandardScaler tool was used from Sci-Kit Learn library.

After features of interest were standardized, optimal number of clusters for neighborhood clustering was identified. While fitting the k-Means clustering model with a cluster counts from 1 to 9, model distortions were calculated on every iterative step. Distortions were plotted to identify the "elbow" – the point in plotted line that represents the optimal number of clusters.

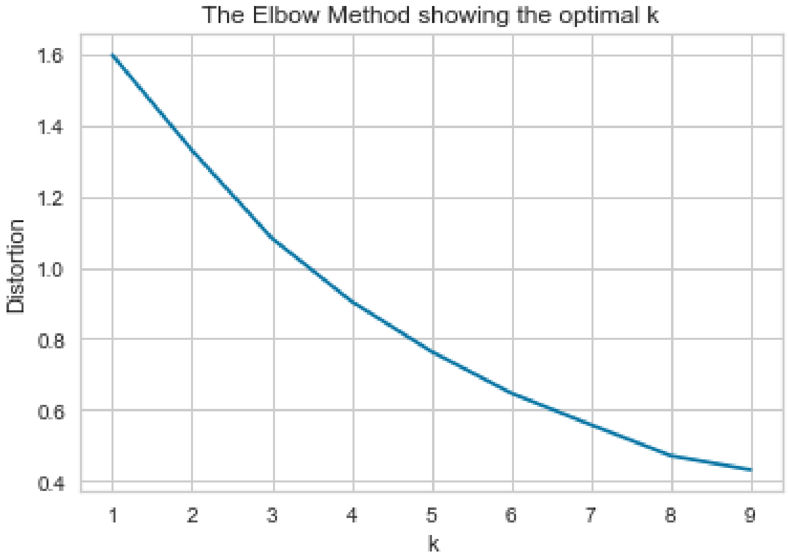


Figure 3 The Elbow method to identify optimal number of clusters

However, after visually inspecting the plot, a clear elbow point was not identified. To find optimal cluster number of our investigation, Silhouette score of k-Means model was used in the next step of modeling.

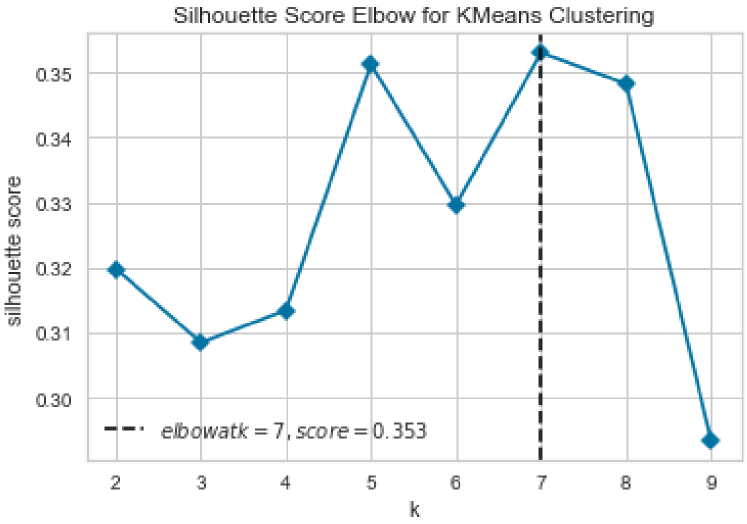


Figure 4 The Silhouette method to identify optimal number of clusters

This time, highest Silhouette was achieved with 7 clusters of the k-Means model. Identified optimal number of clusters is used in next steps of clustering the Vilnius neighborhoods.

k-Means model was one more time fitted with standardized feature dataset. After this final clustering step, cluster labels were modeled for Vilnius neighborhoods. Cluster Labels were added to neighborhood dataset and clustered neighborhoods were plotted on map of Vilnius. To better understand neighborhood specifics, neighborhood center points were visualized as bubbles with a population size based radius of the buble, i.e. bigger bubbles represent areas with higher populated neighborhoods.

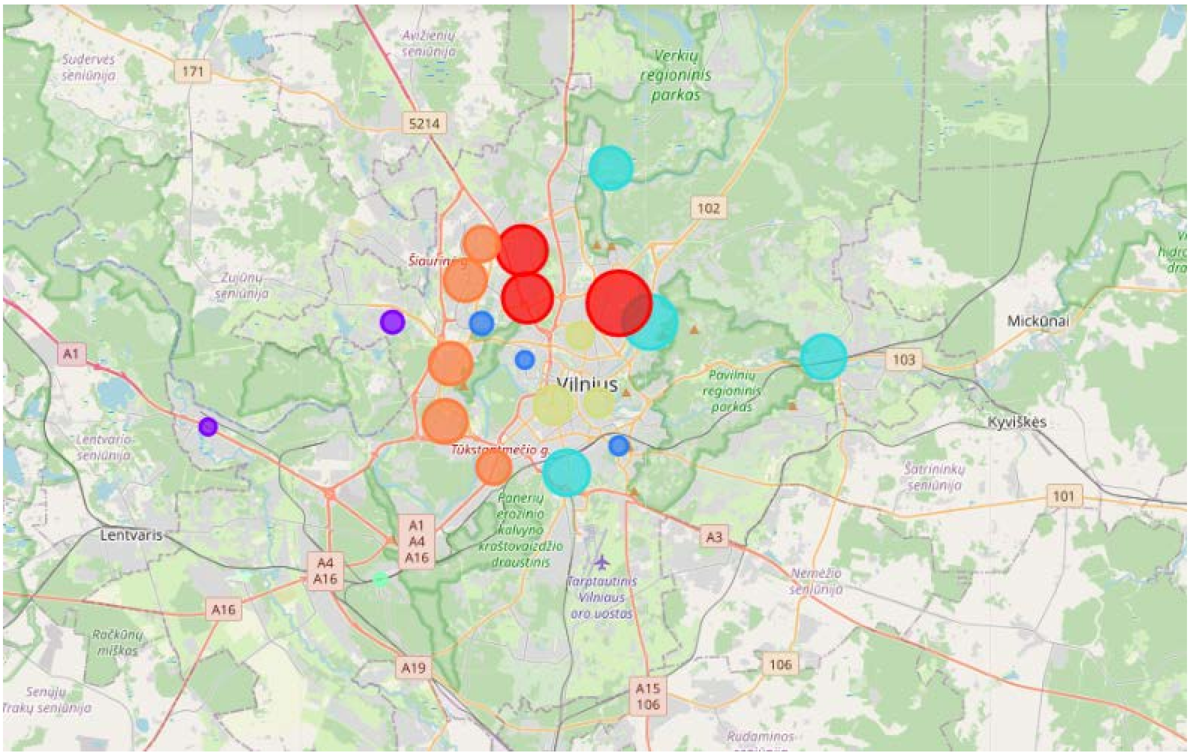


Figure 5 Clustered neighborhoods of Vilnius

First glance at a generated Vilnius map reveals some clear clusters of neighborhoods. k-Means clustering segments the observations but not defines labels for clusters. For this reason, more bar plots were generated for possibly identifying label names for generated clusters.

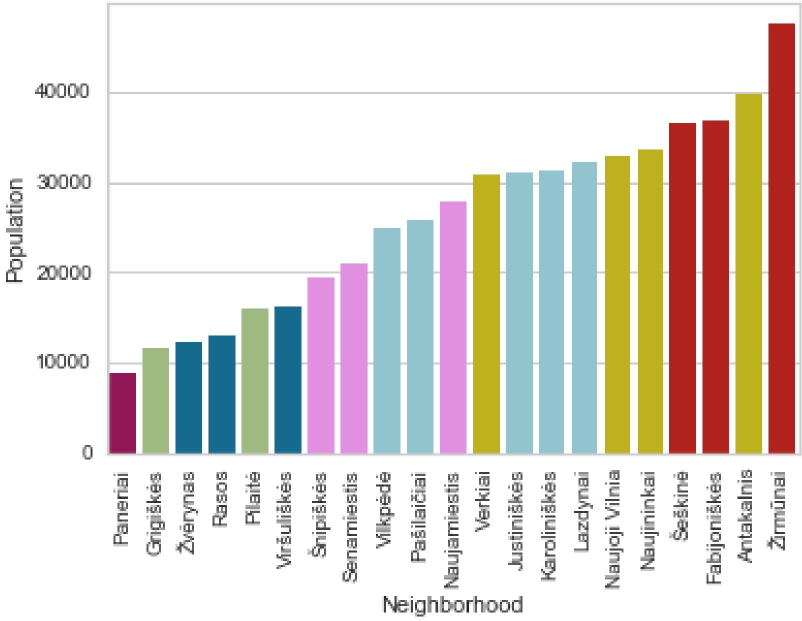


Figure 6 Cluster coloured neighborhoods by Population

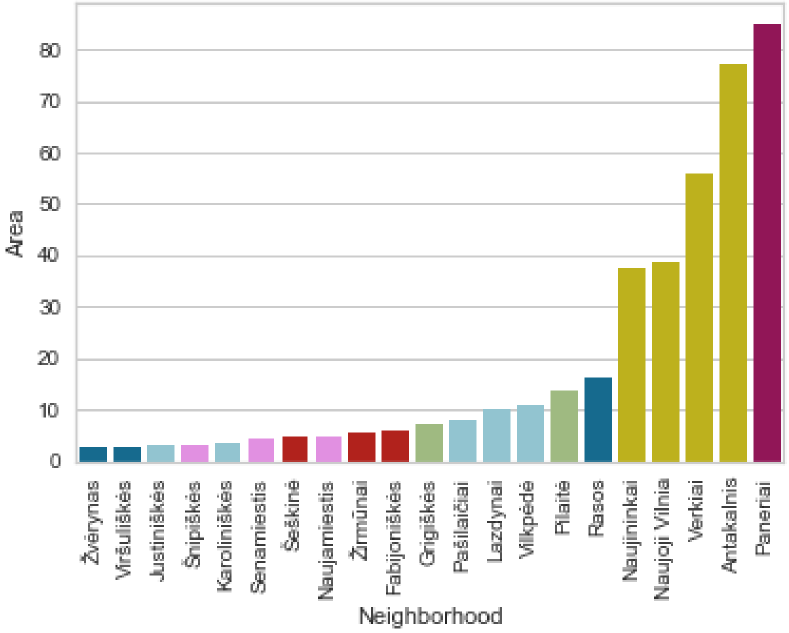


Figure 7 Cluster coloured neighborhoods by Area

Visual inspection of bar plots of neighborhoods suggests the following names for clusters:

- Cluster 0 - Small area, highly populated area with many Pizzerias
- Cluster 1 - Small area, small population area with no Pizzerias
- Cluster 2 - Small area, small population area with many Pizzerias
- Cluster 3 - Large area, highly populated area some Pizzerias
- Cluster 4 - Very large area, small population area with no Pizzerias
- Cluster 5 – Small area, medium population area with many Pizzerias
- Cluster 6 - Small area, medium population area with some Pizzerias

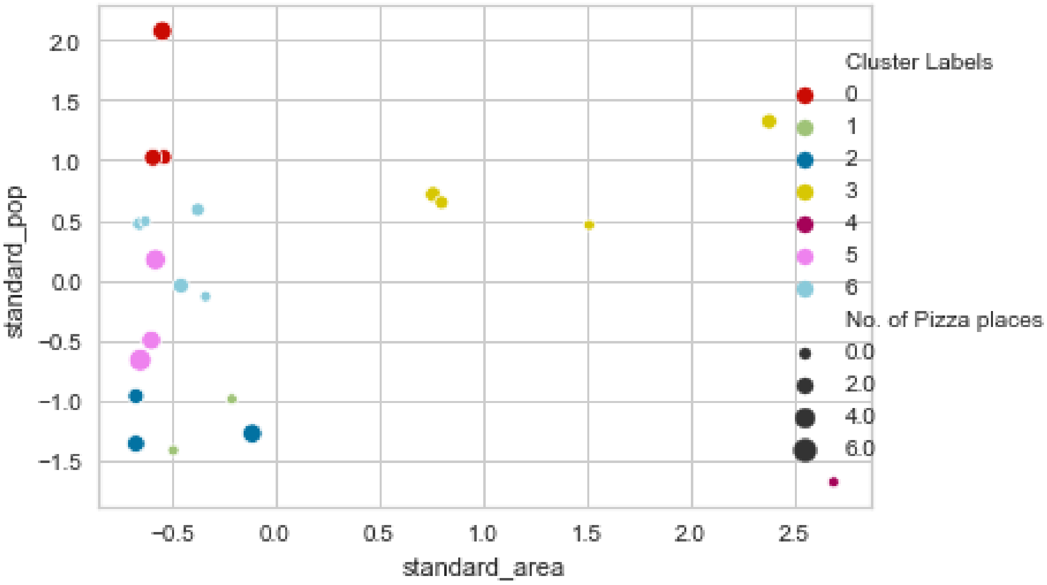


Figure 8Scatter plot of neighborhoods based on their area and population

Clustering analysis revealed neighborhoods which are relatively similar based on their population density and number of Pizza places in the area. The results of analysis can provide some answers to the defined problem of the project. Of course, the decision should be made based on many other factors such as what risk and strategy the business want to take when opening a new restaurant.

4. Results and discussion

Data cleaning steps of the analysis provided the insights that many data sets that may be found on internet might be outdated. Location such as Vilnius has not so many information about its venues and it might be relatively hard to retrieve other trusted statistics for the area. Many other features or models can be included in further analysis to create more precise/more relevant model for business decision making.

Visual inspection of clustered Vilnius neighborhoods identified that Paneriai neighborhood is only one observation in its cluster. It is a large area, low populated area which has no pizzerias. People may not be interested to have a new pizzeria in this area and going with a decision to open a restaurant there is risky.

Cluster 5 area neighborhoods such as Senamiestis, Naujamiestis, Šnipiškės are in the central part of Vilnius. It consists of relatively small and averagely populated neighborhoods. Opening new pizzeria might be risky move as many operating competitors are operating in the cluster. Of course, central part of the city is attraction to many city visitors and new restaurant might get some interest. The decision to open a restaurant in Cluster 5 neighborhoods depends on the strategy and risk assessment of the business.

Clustering model precisely identified the “sleeping” neighborhoods of Vilnius – Cluster 6 (Pašilaičiai, Justiniškės, Karoliniškės, Lazdynai, Vilkpėdė). These neighborhoods are highly populated areas with many blocks of flats, where many people homes are located. These neighborhoods have 0-2 pizzerias. While containing relatively small competition and high population the neighborhoods of cluster might be of high interest for the business to open a new Pizzeria in the area. Karoliniškės and Vilkpėdė are neighborhoods with 0 count venues of interest and might be a potential neighborhood to open a new pizza spot.

5. Conclusion

This project consists of basic data analysis that was conducted to identify a potential location for a new Pizza restaurant in Vilnius. Many other factors, such as purchasing power, other competitor strategy, work specifics, etc. are needed to be evaluated in the deeper analysis. Clustering analysis performed in this project revealed the possibility to suggest the user an optimal location for a new business based on some selected features and statistics.