Battle of the Neighborhoods in Vilnius – the Case of New Gas Station

Aistė Kisielienė

January 10, 2020

1. Introduction

1.1. Background

Vilnius is the capital of Lithuania and its largest city, with a population of 580,020. The population of Vilnius functional urban area, that stretches beyond the city limits, is estimated at 697,691. Vilnius is in the southeast part of Lithuania and is the second largest city in the Baltic states. Vilnius is fast growing city where setting up a business takes just 3 days, after which companies can tap into everything the Vilnius business environment has to offer.

1.2. Business problem

A company that owns a network of gas stations wants to expand and build a new gas station in the city of Vilnius. The problem is - how to find the best place for their new gas station in Vilnius. First, they need to decide what neighborhood in Vilnius would be most suitable for building a new gas station.

There are 21 neighborhoods in the city of Vilnius. In order to make the right decision they must examine how many gas stations there are already in each neighborhood, what is the population in each neighborhood and what is the traffic volume in each neighborhood. These are the main aspects to which they must pay attention when choosing a neighborhood. Company should choose to build a new gas station in place where the traffic volume is the highest and where number of gas stations per resident is the lowest.

1.3. Interest

Company’s which own gas stations in Vilnius and are considering expansion should be interested in this analysis of Vilnius neighborhoods.

2. Data collection and preparation

2.1. Data sources

Gas stations in Vilnius location data can be gathered from https://foursquare.com/. Having this data, we can calculate gas stations number in each neighborhood in the city of Vilnius.

Data of neighborhoods in Vilnius – number and names of neighborhoods, the coordinates of each neighborhood, population in each neighborhood can be found in Wikipedia page (https://lt.wikipedia.org/wiki/Vilnius). Number of the residents in each neighborhood will be used to calculate the number of gas stations per resident.

Data about traffic volume in each neighborhood in Vilnius can be found in Github https://github.com/vilnius/traffic.

Having all this data about each neighborhood in Vilnius I will group and compare neighborhoods in Vilnius and find the most suitable neighborhood for a new gas station.

2.2. Data collection and cleaning

In order to segment the neighborhoods and explore them, we will essentially need a dataset that contains all the neighborhoods that exist as well as the latitude and longitude coordinates of each neighborhood and population in each neighborhood. The data of all 21 neighborhoods of Vilnius was scrapped from Wikipedia page https://lt.wikipedia.org/wiki/S%C4%85ra%C5%A1as:Vilniaus_seni%C5%ABnijos.

Gas station venues location data was scrapped from https://foursquare.com/ and these gas stations results was merged to existing Vilnius neighborhoods dataset.

After scrapping this data and putting it into one dataset I calculated how many gas stations per resident there are in each neighborhood - number of gas stations in each neighborhood was divided by population in each neighborhood.

Table 1. Neighborhoods of Vilnius data: population, coordinates and gas station number.

	Neighborhoods	Population	Latitude	Longitude	Count_gas_stations	Gas_station_per_person
0	Verkiai	30856	54.708707	25.284686	14	0.000454
1	Antakalnis	39697	54.701126	25.308957	4	0.000101
2	Pašilaičiai	25674	54.725942	25.231328	8	0.000312
3	Fabijoniškės	36644	54.723397	25.249529	8	0.000218
4	Pilaitė	15996	54.708126	25.175803	2	0.000125
5	Justiniškės	30958	54.717905	25.220236	4	0.000129
6	Viršuliškės	16250	54.717867	25.220222	4	0.000246
7	Šeškinė	36604	54.715694	25.244574	9	0.000246
8	Šnipiškės	19321	54.692956	25.285007	3	0.000155
9	Žirmūnai	47410	54.723249	25.297213	13	0.000274
10	Karoliniškės	31175	54.685131	25.205156	2	0.000064
11	Žvėrynas	12188	54.693353	25.250209	6	0.000492
12	Grigiškės	11617	54.676897	25.081592	1	0.000086
13	Lazdynai	32164	54.674992	25.202004	4	0.000124
14	Vilkpėdė	24749	54.671995	25.243926	3	0.000121
15	Naujamiestis	27892	54.685457	25.284650	2	0.000072
16	Senamiestis	21022	54.685443	25.284621	2	0.000095
17	Naujoji Vilnia	32775	54.690446	25.412790	1	0.000031
18	Paneriai	8909	54.630048	25.105357	4	0.000449
19	Naujininkai	33457	54.660416	25.272415	6	0.000179
20	Rasos	13054	54.677718	25.281702	5	0.000383

There are 245 traffic junctions in Vilnius where the traffic volume is measured: how many cars have crossed the junction in one hour, average of cars that have crossed the junction in month and so on. In order to measure the traffic volume in each neighborhood I used traffic junctions' data. There are 50 traffic junctions in Vilnius where the traffic volume is very high. Traffic volume is measured by average amount of cars that have crossed the junction per day. This data was used to identify which neighborhood has the highest traffic volume.

After gathering traffic volume data from 21 neighborhood in Vilnius only in 17 there were traffic junctions where traffic volume is very high. There were 4 neighborhoods with no intensive traffic junctions and no data. In order to be able to cluster all neighborhoods of Vilnius the decision was made to add these neighborhoods to dataset with value lowest value of traffic volume – 1. Neighborhoods in which data was added are Paneriai, Naujoji Vilnia, Lazdynai and Grigiškės.

Next, I counted how many gas stations there are per one junction where traffic volume is high. If there were few gas stations and a lot of intensive junctions this neighborhood should be considered as a good neighborhood for new gas station. I divided gas stations number in each

neighborhood from junctions with high traffic volume to get the gas stations number per one high traffic volume junction.

Table 2. Neighborhoods of Vilnius data: population, coordinates, gas station number and traffic junctions.

	Neighborhoods	Population	Latitude	Longitude	Count_gas_stations	Gas_station_per_person	Count_junctions	Gas_station_per_junction
0	Verkiai	30856	54.708707	25.284686	14	0.000454	3	4.666667
1	Antakalnis	39697	54.701126	25.308957	4	0.000101	1	4.000000
2	Pašilaičiai	25674	54.725942	25.231328	8	0.000312	1	8.000000
3	Fabijoniškės	36644	54.723397	25.249529	8	0.000218	4	2.000000
4	Pilaitė	15996	54.708126	25.175803	2	0.000125	1	2.000000
5	Justiniškės	30958	54.717905	25.220236	4	0.000129	2	2.000000
6	Viršuliškės	16250	54.717867	25.220222	4	0.000246	2	2.000000
7	Šeškinė	36604	54.715694	25.244574	9	0.000246	4	2.250000
8	Šnipiškės	19321	54.692956	25.285007	3	0.000155	6	0.500000
9	Žirmūnai	47410	54.723249	25.297213	13	0.000274	3	4.333333
10	Karoliniškės	31175	54.685131	25.205156	2	0.000064	1	2.000000
11	Žvėrynas	12188	54.693353	25.250209	6	0.000492	2	3.000000
12	Grigiškės	11617	54.676897	25.081592	1	0.000086	1	1.000000
13	Lazdynai	32164	54.674992	25.202004	4	0.000124	1	4.000000
14	Vilkpėdė	24749	54.671995	25.243926	3	0.000121	6	0.500000
15	Naujamiestis	27892	54.685457	25.284650	2	0.000072	5	0.400000
16	Senamiestis	21022	54.685443	25.284621	2	0.000095	2	1.000000
17	Naujoji Vilnia	32775	54.690446	25.412790	1	0.000031	1	1.000000
18	Paneriai	8909	54.630048	25.105357	4	0.000449	1	4.000000
19	Naujininkai	33457	54.660416	25.272415	6	0.000179	3	2.000000
20	Rasos	13054	54.677718	25.281702	5	0.000383	4	1.250000

2.3. Data feature selection

After gathering and preparing data, there were dataset with 21 neighborhood of Vilnius and 7 features. Upon examining the meaning of each feature and having in mind that we need to find the neighborhood with low number of gas stations per resident and with low number of gas stations per intensive traffic junctions, in further modeling stage I decided to keep only gas station per person data and gas station per junction data.

3. Exploratory Data Analysis

Before building a model, we should look into data gathered.

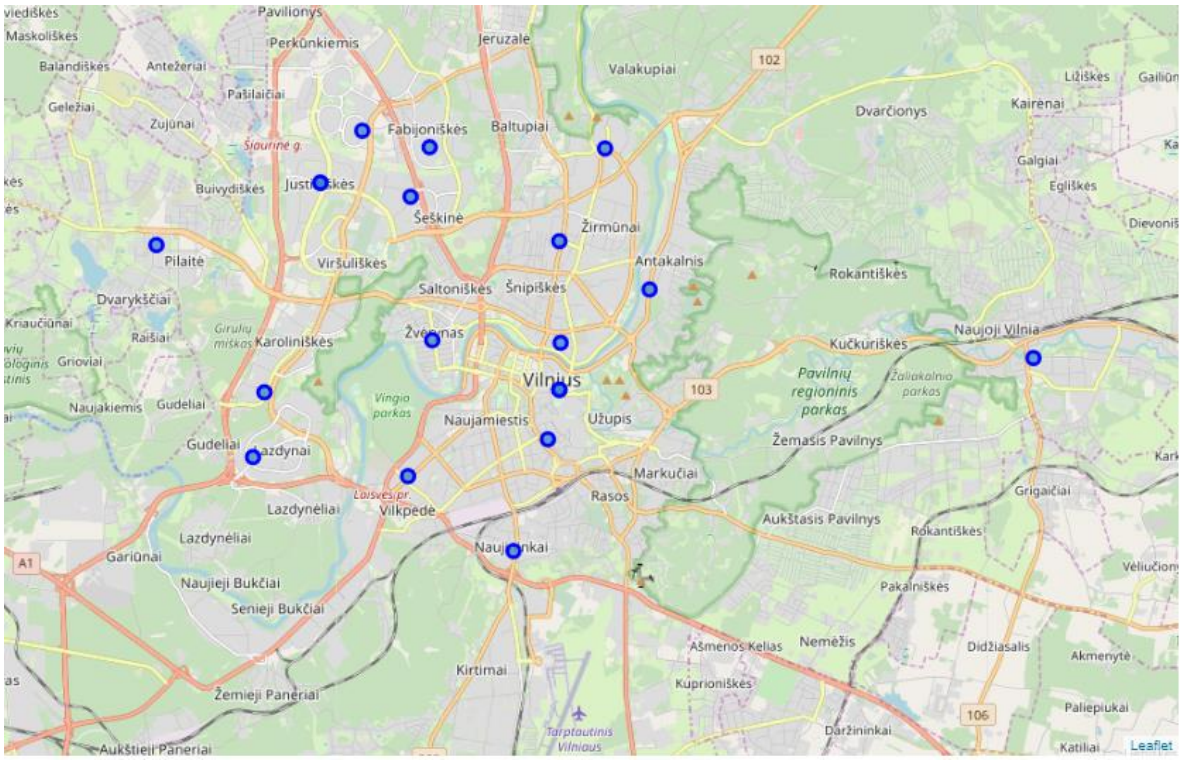

Figure 1. The map of Vilnius neighborhood.

Three neighborhoods in which number of gas stations per resident is smallest are: Naujoji Vilnia, Karoliniškės and Naujamiestis. Given only this one aspect - gas stations per resident number - these three neighborhoods should be considered as a good place to build new gas station.

Table 3. Top three neighborhoods with lowest gas station per resident number.

	Neighborhoods	Population	Latitude	Longitude	Count_gas_stations	Gas_station_per_person	Count_junctions	Gas_station_per_junction
17	Naujoji Vilnia	32775	54.690446	25.412790	1	0.000031	1	1.0
10	Karoliniškės	31175	54.685131	25.205156	2	0.000064	1	2.0
15	Naujamiestis	27892	54.685457	25.284650	2	0.000072	5	0.4

Three neighborhoods in which number of gas stations per junction is smallest are: Naujamiestis, Šnipiškės and Vilkpėdė. Given only this one aspect - gas stations per junction number - these three neighborhoods should be considered as a good place to build new gas station.

Table 4. Top three neighborhoods with lowest gas station per junction.

	Neighborhoods	Population	Latitude	Longitude	Count_gas_stations	Gas_station_per_person	Count_junctions	Gas_station_per_junction
15	Naujamiestis	27892	54.685457	25.284650	2	0.000072	5	0.4
8	Šnipiškės	19321	54.692956	25.285007	3	0.000155	6	0.5
14	Vilkpėdė	24749	54.671995	25.243926	3	0.000121	6	0.5

From first sight it looks like only one neighborhood Naujamiestis is between top three smallest values in both cases. We could think of this neighborhood as our first choice. But for more accurate results clustering model was applied to cluster Vilnius neighborhoods to see maybe there are similar neighborhoods to Naujamiestis.

4. Modeling and evaluation

Considering that data is with different magnitudes and this could result an incorrect clustering results, data normalization was implemented. StandardScaler() was used to normalize Vilnius neighborhoods dataset.

Table 5. Dataset for clustering model before normalization.

	Gas_station_per_person	Gas_station_per_junction
0	0.000454	4.666667
1	0.000101	4.000000
2	0.000312	8.000000
3	0.000218	2.000000
4	0.000125	2.000000
5	0.000129	2.000000
6	0.000246	2.000000
7	0.000246	2.250000
8	0.000155	0.500000
9	0.000274	4.333333
10	0.000064	2.000000
11	0.000492	3.000000
12	0.000086	1.000000
13	0.000124	4.000000
14	0.000121	0.500000
15	0.000072	0.400000
16	0.000095	1.000000
17	0.000031	1.000000
18	0.000449	4.000000
19	0.000179	2.000000
20	0.000383	1.250000

After having data normalized, clustering model was carried out - kmeans in this case.

Problem was that it was unclear what number of clusters to use. To find the best k in kmeans the Silhouette coefficient was calculated for different models. This index was chosen because the ground truth labels are not known and evaluation must be performed using the model itself.

The Silhouette Coefficient is an example of such an evaluation, where a higher Silhouette Coefficient score relates to a model with better defined clusters. The Silhouette Coefficient is defined for each sample and is composed of two scores:

a. The mean distance between a sample and all other points in the same class.
b. The mean distance between a sample and all other points in the next nearest cluster.

Table 6. Top three neighborhoods with lowest gas station per junction.

Number of clusters	**Value of Silhouette Coefficient**
2	0.5313
3	0.4893
4	0.3659
5	0.4757
6	0.4242
7	0.4370
8	**0.4637**
9	0.4431
10	0.4343
11	0.3950

As we can see the best score we get when clusters number is set to 2. But we get similar results when using clusters number 3, 5 and 8. Because our purpose is to narrow down the neighborhoods number, for further analysis the higher clusters number 8 was used.

After clustering the neighborhoods of Vilnius, the average values of each clusters were calculated to better understanding of each cluster.

Table 7. Clustering results.

	Population	Latitude	Longitude	Count_gas_stations	Gas_station_per_person	Count_junctions	Gas_station_per_junction
Cluster Labels							
0	22896.000000	54.683866	25.265431	2.000000	0.000093	3.500000	0.733333
1	17317.666667	54.677369	25.213417	8.000000	0.000465	2.000000	3.888889
2	30738.750000	54.704344	25.246685	6.750000	0.000222	3.250000	2.062500
3	25674.000000	54.725942	25.231328	8.000000	0.000312	1.000000	8.000000
4	35930.500000	54.688059	25.255481	4.000000	0.000113	1.000000	4.000000
5	13054.000000	54.677718	25.281702	5.000000	0.000383	4.000000	1.250000
6	47410.000000	54.723249	25.297213	13.000000	0.000274	3.000000	4.333333
7	26043.000000	54.703721	25.200398	2.666667	0.000106	1.333333	2.000000

Also, clusters where visualized in the map of Vilnius.

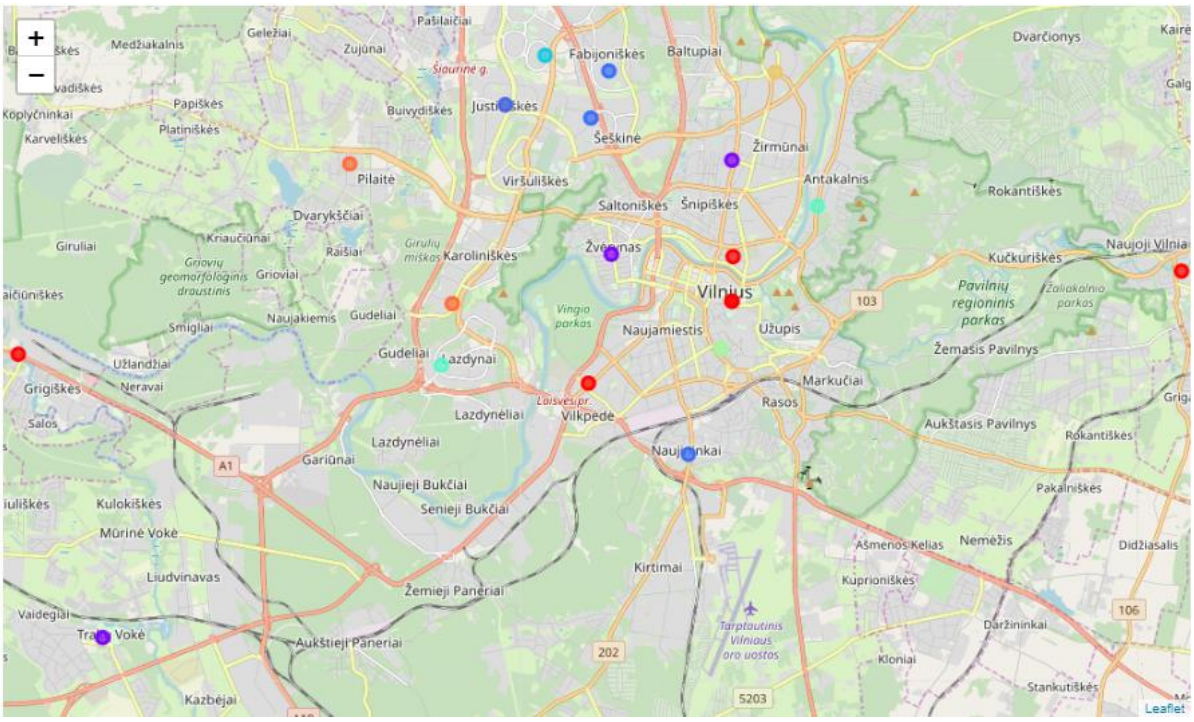

Figure 2. The map of clustered Vilnius neighborhood.

The results show that the best neighborhoods to build new gas station is in cluster 0, where average value of gas stations per person is 0.000093 and average value of gas stations per junction is 0.733333. These average values are lowest.

Table 8. Neighborhoods in cluster 0.

	Neighborhoods	Count_gas_stations	Gas_station_per_person	Count_junctions	Gas_station_per_junction
8	Šnipiškės	3	0.000155	6	0.5
12	Grigiškės	1	0.000086	1	1.0
14	Vilkpėdė	3	0.000121	6	0.5
15	Naujamiestis	2	0.000072	5	0.4
16	Senamiestis	2	0.000095	2	1.0
17	Naujoji Vilnia	1	0.000031	1	1.0

5. Conclusions

As it was indicated in business problem stage company should choose to build a new gas station in place where the traffic volume is the highest and where number of gas stations per resident is the lowest. These neighborhoods were included in cluster 0.

As we can see there are 6 neighborhoods in cluster 0: Šnipiškės, Grigiškės, Vilkpedė, Naujamiestis, Senamiestis and Naujoji Vilnia.

There are some insights from our analysis:

1. During data collection stage there were 4 neighborhoods with no intensive traffic junctions' data and data was added to these neighborhoods with value 1 (lowest value of traffic) in order to have all neighborhoods in dataset. Naujoji Vilnia and Grigiškės were among these 4 neighborhoods. So, offer should be to omit these two neighborhoods, because they did not have any high traffic volume junctions, despite the fact that these neighborhoods have low gas stations per person number.
2. This leaves us with 4 neighborhoods: Šnipiškės, Vilkpedė, Naujamiestis and Senamiestis. As we can see Naujamiestis has the lowest number of gas stations per person and the lowest number of gas stations per junction. So, our first-choice neighborhood should be Naujamiestis.
3. Two neighborhoods Šnipiškės and Vilkpedė are also very similar and should be also considered as good choice for new venue.
4. In Senamiestis neighborhood as in an old town of Vilnius there are a lot of restrictions for building anything and a lot of protected heritage. This neighborhood would not be a good choice despite it has low gas stations per person number and low gas station per junction number.