IBM Data Science Professional Certificate

Capstone Project

Analysis of Top-5 Volga Federal District Cities in Russian Federation Using Foursquare API

author:
Marat S. Mukhametzhanov

January 2021

Table of Contents

Table of Contents

1 Introduction
- 1.1 Background
- 1.2 Business Problem
- 1.3 Interest
- 1.4 Software used for solving the problem

2 Data
- 2.1 Data wrangling
  - 2.1.1 Postal codes and their coordinates.
  - 2.1.2 Venues available from each neighborhood
- 2.2 Data Cleaning and Feature Selection

3 Methodology
- 3.1 Exploratory data analysis
- 3.2 Clustering of the cities

4 Results

5 Discussion
- 5.1 Battle of the cities
- 5.2 Placing a restaurant in the cities

6 Conclusion

Acknowledgements

References

1 Introduction

1.1 Background

Russian Federation consists of the following 8 large federal districts:

Federal district	Area (km2)	Population(2010 census)	Population density(per km2)	HDI(2017)	Federal subjects	Administrative Centre
Central	650200	38438600	59.1	0.838	18	Moscow
Volga	1037000	29900400	28.8	0.797	14	Nizhny Novgorod
Siberian	4361800	17178298	3.9	0.788	10	Novosibirsk
Southern	427800	16141100	37.7	0.793	8	Rostov-on-Don
Northwestern	1687000	13583800	8.1	0.827	11	Saint Petersburg
Ural	1818500	12082700	6.6	0.833	6	Yekaterinburg
North Caucasian	170400	9496800	55.7	0.785	7	Pyatigorsk
Far Eastern	6952600	8371257	1.2	0.801	11	Vladivostok

Table 1.1. Federal districts of Russian Federation and their description (source: https://en.wikipedia.org/wiki/Federal_districts_of_Russia)

The Volga Federal District is one of the biggest districts in Russia consisting of 14 federal subjects with almost 30 million people in total. It is mainly placed along Volga, being the longest river in Europe. It has 5 cities with a population over 1 million: Nizhny Novgorod (rus. Нижний Новгород, the administrative centre of the district, population ~1.252 million people), Kazan (rus. Казань, population ~1.257 million people), Ufa (rus. Уфа, population ~1.128 million people), Samara (rus. Самара, population ~1.156 million people), Perm (rus. Пермь, population ~1.055 million people).

1.2 Business Problem

A series of restaurants in Nizhny Novgorod has become successful, so its stakeholders are interested to extend their business into other cities. However, Russian Federation is a country with large distances between different cities and each region has its own cultural, economic and even natural or weather specifics. So, the stakeholders first consider opening a restaurant in a city of the same district. It is reasonable to consider only cities similar to Nizhny Novgorod, where their business has already become successful. One can see that there are 4 cities similar to Nizhny Novgorod in Volga federal district: Kazan, Ufa, Samara, and Perm. These cities are administrative centres of their regions and each of them has more than one million residents.

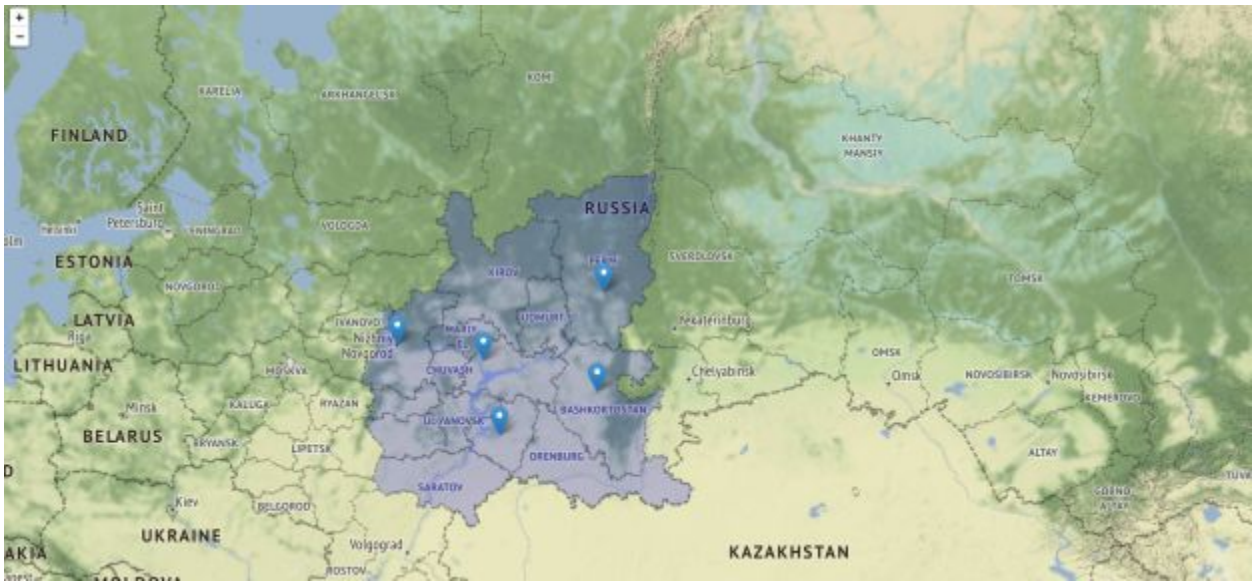

Figure 1.1. Volga Federal District and its top-5 cities (indicated by markers)

The main problem considered in this study can be briefly described as follows.

1. Find the city of the Volga Federal District closest to Nizhny Novgorod in terms of public venues availability and their categories.
2. In the chosen city, find the best places, where it is reasonable to open a restaurant based on known locations of the restaurants in the original city (Nizhny Novgorod).

1.3 Interest

Since Nizhny Novgorod is a big, grown and economically developed city, then the present study can be interesting to different businessmen who consider extending their business to other cities. In particular, there are several series of food restaurants, sushi bars, pizzerias, bars, etc. (they are not listed here only for privacy reasons). Moreover, the methodologies presented in this study are easily extendable to other regions and districts not only in Russia, but in other countries as well (e.g., in Europe, USA or Canada). In this case, only the Data section of the Jupyter notebook related to this study will differ.

1.4 Software used for solving the problem

All the methods used in the present study have been implemented in Python 3.7 using Jupyter notebooks in IBM Watson Studio. The complete notebook is presented in github. The following standard libraries have been mainly used during the work: *pandas* and *numpy* for all handlings with the data frames and data series, *seaborn, matplotlib,* and *folium* for

visualizations, *geopy* and *geocoder* for obtaining the coordinates of the neighborhoods, *json* for handling json-files, *scikit-learn* for clustering, and *requests* for working with the Foursquare API.

2 Data

2.1 Data wrangling

2.1.1 Postal codes and their coordinates.

The main methodology used in the present research is clustering of a city based on the information of all venues available for a neighborhood in a fixed radius. Since all chosen cities are well divided by neighborhoods based on their postal codes, which is a unique common criterion for different cities in Russian Federation, then we will use postal codes of each city as places around which the venues are searched.

In this case, first, we need the following data for each chosen city:

1. List of all postal codes
2. Latitude and longitude associated with each postal code in order to make queries using Foursquare API.

The list of all postal codes in Russia is available from the following link:
http://download.geonames.org/export/zip/RU.zip

The information available in this zip-archive is well described in the respective readme file. In particular, for each postal code (obviously, unique), its place name (mostly consisting of the city's name), estimated latitude and longitude are provided:

	country code	postal code	place name	admin name1	admin code1	admin name2	admin code2	admin name3	admin code3	latitude	longitude	accuracy
0	RU	385000	Майкоп	Адыгея Республика	1.0	NaN	NaN	NaN	NaN	44.8802	40.2166	1.0
1	RU	385001	Майкоп 1	Адыгея Республика	1.0	NaN	NaN	NaN	NaN	44.8802	40.2166	1.0
2	RU	385002	Майкоп 2	Адыгея Республика	1.0	NaN	NaN	NaN	NaN	44.8802	40.2166	1.0
3	RU	385003	Майкоп 3	Адыгея Республика	1.0	NaN	NaN	NaN	NaN	44.8802	40.2166	1.0
4	RU	385006	Майкоп 6	Адыгея Республика	1.0	NaN	NaN	NaN	NaN	44.8802	40.2166	1.0

Table 2.1. Data set of all postal codes in Russian Federation

However, the latitude and longitude for each postal code are estimated very roughly, because they coincide for many postal codes in each city. To solve this issue, let us use the

Python's Nominatim tool in geopy.geocoders library for retrieving the real coordinates for each postal code (and for each city location as well). An example for the postal code '603076' in Nizhny Novgorod is presented below:

```
geolocator = Nominatim(user_agent="PFO_explorer")
pcode = '603076'
address = pcode+', '+'Nizhny Novgorod'+', Russia'
try:
        location = geolocator.geocode(address)
except:
        print('error in '+address+', coordinates have not been changed')
```

Since the *geolocator.geocode(address)* for several postal codes returned *None,* then for these postal codes, the original roughly measured values have been used (for each city, there were no more than 5% postal codes with this issue, so the obtained accuracy is acceptable). After that, all postal codes with non unique coordinates are excluded from our dataset in order to reduce the biases possible to neighborhoods with the same coordinates.

The folium map of Nizhny Novgorod covered by the neighborhoods defined through the postal code is presented below. Here, each neighborhood is represented by a circle of a fixed radius (equal to 500 metres). However, one can see that the radius of 500 metres is not sufficient for covering the city. For this reason, the radius of 1000 metres has been chosen for each neighborhood (this radius will be used in the Foursquare API for searching the venues), since, on the one hand, it becomes sufficient to cover the city, and, on the other hand, a venue located in 1000 metres is accessible even by walking.

Figure 2.1. The folium maps of Nizhny Novgorod covered by the neighborhoods defined by the postal codes. Each neighborhood is represented by a circle of a fixed radius. (a) The map of the whole city using the radius = 500m. (b) The map of a part of the city that includes a park, cinema, metro station, mall, etc. showing that the radius of 500m. is not sufficient (this part of the city is not covered even it is easily accessible). (c) The map of the whole city using the radius of 1000m. In this case, all parts of the city are covered well.

2.1.2 Venues available from each neighborhood

Once the coordinates of each neighborhood have been obtained, the Foursquare API can be used for retrieving the available nearby venues for each neighborhood by the following commands:

```
url = 
'https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll={},{}&radius={}&limit
={}'.format(CLIENT_ID, CLIENT_SECRET, VERSION,lat, lng, radius, LIMIT)
results = requests.get(url).json()["response"]['groups'][0]['items']
```

By default, the Foursquare API returns the following information for each neighborhood (other fields are omitted): Neighborhood, Neighborhood Latitude, Neighborhood Longitude, Venue, Venue Latitude, Venue Longitude, Venue Category. However, using the default venue categories can be redundant: e.g., in our case, the venue categories “chinese restaurant” and “italian restaurant” are similar, we are not interested in detailed categories of each venue. For this reason, we will use macro categories for each venue:

- Arts & Entertainment (id 4d4b7104d754a06370d81259),
- College & University (id 4d4b7105d754a06372d81259),
- Event (id 4d4b7105d754a06373d81259),
- Food (id 4d4b7105d754a06374d81259),
- Nightlife Spot (id 4d4b7105d754a06376d81259),
- Outdoors & Recreation (id 4d4b7105d754a06377d81259),
- Professional & Other Places (id 4d4b7105d754a06375d81259),
- Residence (id 4e67e38e036454776db1fb3a),
- Shop & Service (id 4d4b7105d754a06378d81259),
- Travel & Transport (id 4d4b7105d754a06379d81259),

First, for each macro category, we will retrieve the list of all venue categories belonging to it:

```
url='https://api.foursquare.com/v2/venues/categories?&client_id={}&client_secret={}&v={}'.format(
        CLIENT_ID,CLIENT_SECRET, VERSION)
results = requests.get(url).json()['response']['categories']
```

The *results* variable contains now the list of all macro categories. Each macro category is represented by a dictionary with the key *‘categories’* containing the list of all its sub-categories (which in their turn are represented by the dictionaries in the same way).

After retrieving the list of all sub-categories for each macro-category, it is easy to assign a macro category for each venue (let us denote 'venue category' as 'micro category' and 'venue macro category' simply as 'venue category', hereinafter, just for simplicity, but the data about the micro categories will be also used later in order to determine the best places in the chosen city):

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Micro Category	Venue Category
0	603000	56.317042	43.994228	Burrito Family	56.316153	43.992785	Burrito Place	Food
1	603000	56.317042	43.994228	Franky Bar	56.316330	43.994536	Cocktail Bar	Nightlife Spot
2	603000	56.317042	43.994228	Surf Coffee	56.317038	43.994256	Coffee Shop	Food
3	603000	56.317042	43.994228	Большая Покровская улица	56.320291	43.998442	Road	Travel & Transport
4	603000	56.317042	43.994228	Бикрам-Йога | Нижний Новгород	56.319637	43.994256	Yoga Studio	Outdoors & Recreation

Table 2.2. An example of the main dataset containing the information about venues for each postal code (indicated by the column 'Neighborhood')

2.2 Data Cleaning and Feature Selection

Let us count the number of all venues for each category in each city:

Category	Nizhny Novgorod	Kazan	Ufa	Samara	Perm
Arts & Entertainment	209	203	261	157	125
College & University	5	3	4	3	1
Event	0	0	0	0	0
Food	685	813	654	882	697
Nightlife Spot	143	148	182	123	129
Outdoors & Recreation	288	282	308	311	201
Professional & Other Places	9	30	20	12	11
Residence	1	0	3	0	3
Shop & Service	586	603	877	966	409
Travel & Transport	300	180	235	214	100
Total	2226	2262	2544	2668	1676

Table 2.3. Total number of venues for each category in each city.

We can see that there is almost no data about the categories 'Event' and 'Residence' for all the cities. Moreover, even if there is some data about 'College & University' in Nizhny Novgorod, it is almost unavailable for the other cities, so this category cannot be used to estimate similarity of the cities. For the same reason, the category 'Professional & Other Places' cannot be used to estimate similarity between these cities, because the data about it is almost unavailable for Nizhny Novgorod. The data about the other categories is widely

available for all the cities, so we will use these 6 categories to estimate similarity between the cities: Arts & Entertainment (id 4d4b7104d754a06370d81259), Food (id 4d4b7105d754a06374d81259), Nightlife Spot (id 4d4b7105d754a06376d81259), Outdoors & Recreation (id 4d4b7105d754a06377d81259), Shop & Service (id 4d4b7105d754a06378d81259), Travel & Transport (id 4d4b7105d754a06379d81259)

In order to visualize better the distribution of all venues between categories for each city, let us calculate the relative number of venues in each category dividing the number of venues from the previous table by the total number of venues for each city (the categories excluded previously are not considered here). The results in % are visualized in the following table and the respective barplot (see Fig. 2.2):

Category\City	Nizhny Novgorod	Kazan	Ufa	Samara	Perm
Arts & Entertainment	9.5	9.1	10.4	5.9	7.5
Food	31	36.5	26	33.2	42
Nightlife Spot	6.5	6.6	7.2	4.6	7.8
Outdoors & Recreation	13	12.7	12.2	11.7	12.1
Shop & Service	26.5	27.1	34.8	36.4	24.6
Travel & Transport	13.6	8.1	9.3	8.1	6

Table 2.4. Relative number of venues (in %) for each category in each city.

Let us now calculate the relative number of venues for each neighborhood in each city, dividing the number of nearby venues for each neighborhood in each category by the total number of venues in each category: e.g., for Nizhny Novgorod, the following data is assigned for the first 5 neighborhoods:

	Neighborhood	Arts & Entertainment	Food	Nightlife Spot	Outdoors & Recreation	Shop & Service	Travel & Transport
0	603000	0.010160	0.026172	0.017893	0.016981	0.024422	0.024775
1	603001	0.023222	0.021302	0.029821	0.018868	0.007712	0.018018
2	603002	0.005806	0.004869	0.000000	0.005660	0.017995	0.013514
3	603003	0.004354	0.006086	0.001988	0.009434	0.007712	0.002252
4	603004	0.002903	0.004869	0.000000	0.009434	0.006427	0.006757

Table 2.5. An example from the dataset containing the relative number of venues for each neighborhood (the first 5 rows for Nizhny Novgorod)

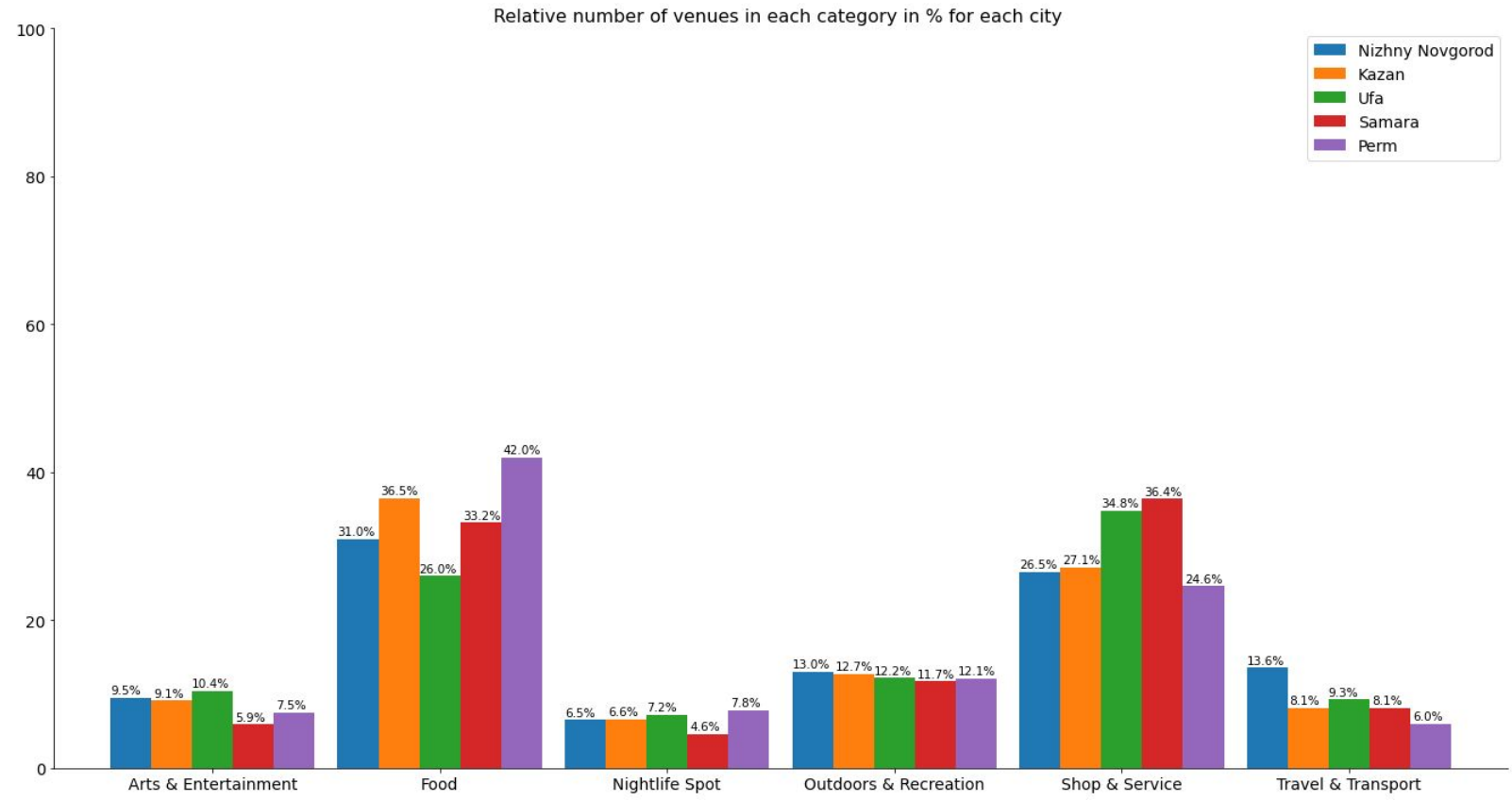

Figure 2.2. Barplots of the relative number of venues (in %) for each category in each city.

The obtained data will be used for clustering the cities and determining the closest city to Nizhny Novgorod in terms of available venue categories. After that, the number of venues in each micro category for each neighborhood obtained previously will be used to determine the best places to locate the restaurants in the chosen city.

3 Methodology

3.1 Exploratory data analysis

Let us first study the data obtained at the previous step. In particular, one can see from Fig.2.2 that the distribution of the categories is comparable in each city: e.g., in each city, the mostly presented categories are Food and Shop&Service, while other categories are presented much less. Since the first main aim of this research is to find the city similar to Nizhny Novgorod, then let us calculate root mean square deviations between Nizhny Novgorod and all other cities for the values given in Table 2.4:

$$RMSD_A = \sqrt{\frac{1}{Number\ of\ categories} \sum_{category\ \in\ Categories} (X_{category,\ A} - X_{category,\ Nizhny\ Novgorod})^2}\ ,$$

Where A is one of the cities except Nizhny Novgorod (Kazan, Ufa, Samara, and Perm), Categories is the set of all categories presented in Table 2.4 (Arts & Entertainment, Food, Nightlife Spot, Outdoors & Recreation, Shop & Service, and Travel & Transport). The results are presented in the following table:

City	RMS w.r.t. Nizhny Novgorod
Kazan	3.19166
Ufa	4.36501
Samara	5.02262
Perm	5.61056

Table 3.1. Root Mean Square Deviations w.r.t. Nizhny Novgorod for each city.

We can see at this moment that Kazan and Perm are closer to Nizhny Novgorod in terms of the distribution of the categories within each city. However, comparisons based only on these results for the whole cities can be biased and non significant. To illustrate this reason, let us build the following boxplots of the relative number of venues for each neighborhood in, e.g., Nizhny Novgorod (an example of the dataset is given in Table2.5).

One can see in Fig. 3.1 that the distributions of venues in each category are biased: e.g. the quartile Q1 is much closer to the median than the quartile Q3 for the Food or Arts&Entertainment category. In order to visualize better this issue, let us construct scatter plots for each category. In Fig. 3.2, scatter plots for each venue category versus each other category are presented for the Nizhny Novgorod city. Only unique scatter plots are presented avoiding so redundant plots for the same categories, e.g., the scatter plot 'Food vs. Food' is not presented, since its information can be immediately obtained from the presented plots.

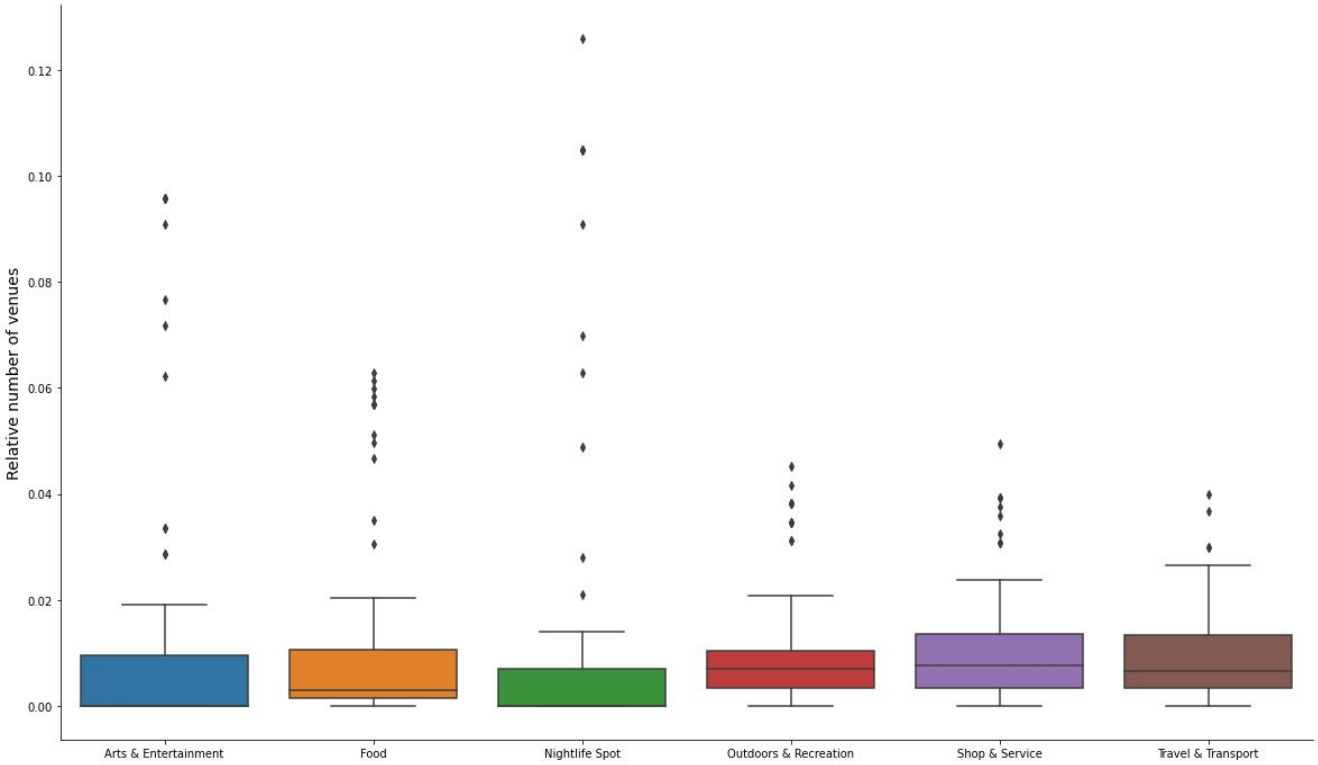

Figure 3.1. Boxplots of a relative number of venues in each category for Nizhny Novgorod.

One can see from Fig. 3.2 that the distribution of the venue categories within the city are not homogeneous: e.g., the plot 'Outdoors&Recreation vs. Shop&Service' shows that there are at least three groups of neighborhoods. In the first one, there are a lot of venues in both the categories for each neighborhood. In the second group, there are a lot of 'Outdoor&Recreation' venues, but less 'Shop&Service' venues. Finally, the third group contains neighborhoods with less venues in both the categories. This issue can be explained by a heterogeneity structure of the city. For example, in each city, there are areas (e.g., city centre), where there are a lot of monuments and/or parks, while in other areas there are almost none parks, but a lot of shops. But the stakeholders can be not interested in such areas, considering only more balanced places in quiet and calm neighborhoods. In this case, it can be more reasonable to compare the distributions only in such areas and not in the whole city. For this reason, first, we will subdivide each city into different clusters (determining the clusters in the same way for each city). Then, we will study the distributions of all categories within each cluster determining the most similar city with

respect to Nizhny Novgorod both visually and numerically for each cluster separately and for the clusters in average.

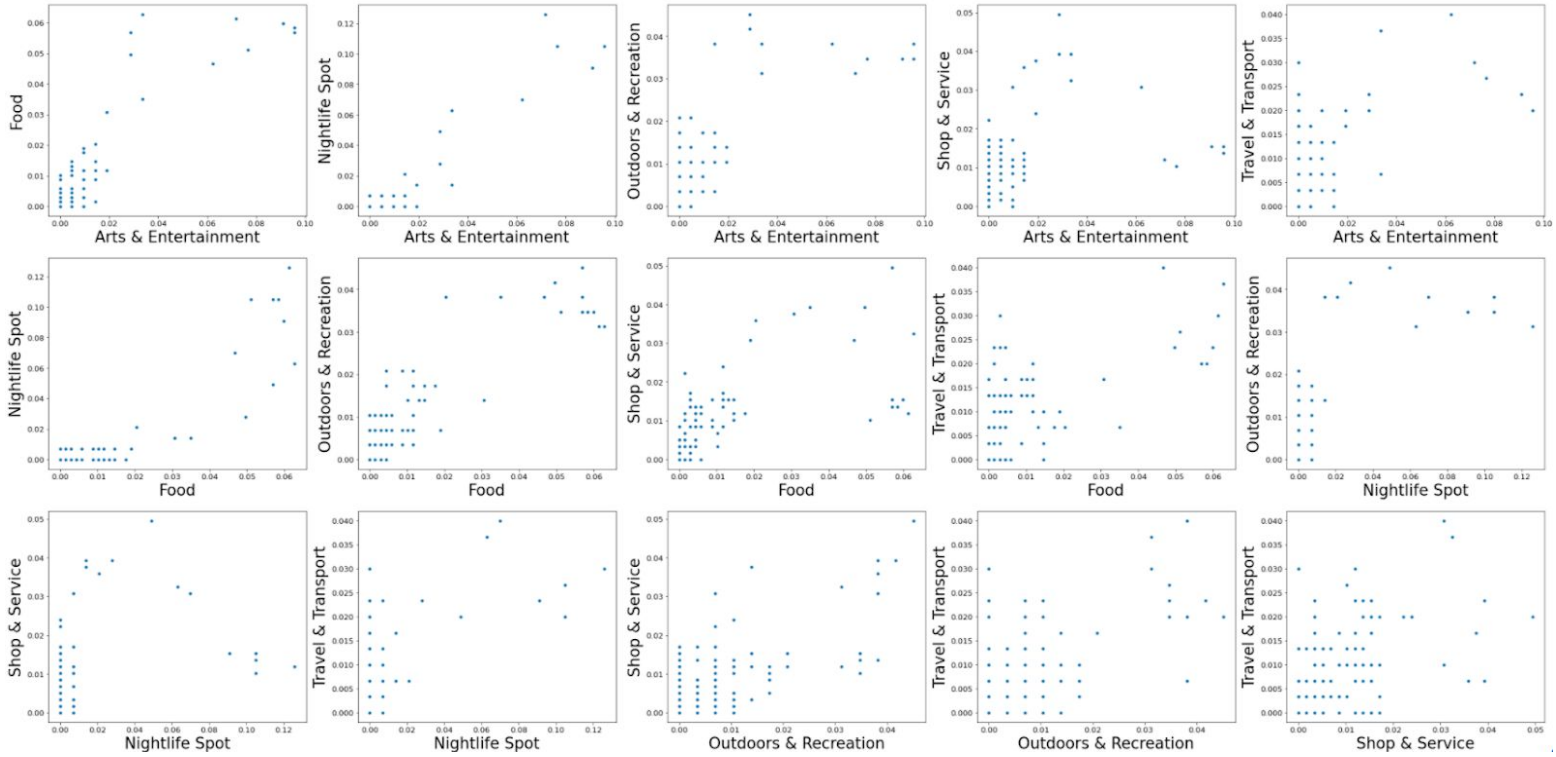

Figure 3.2. Scatter plots of a relative number of venues in each category versus each other category for Nizhny Novgorod.

3.2 Clustering of the cities

The clustering is performed in 6-dimensional space, since there are 6 venues categories in our dataset (see Table 2.5 for an example of the dataset). Each point (or a row in our dataset) is represented by the postal code and 6 numeric values of the range (0,1) being the relative number of venues in each category around the neighborhood defined through the postal code. As we see from Fig. 3.2 for Nizhny Novgorod, there is no evidence for special structures or non-convex patterns in the data, so the standard K-means method can be performed for the clustering. There is only one control parameter in this method: the number of clusters k. In order to determine an optimal value of k, let us use the Elbow method, which consists of clustering each city with different k from 1 to 10 and then choosing an optimal k for each city, visually showing the obtained inertia for each value of k. The respective graphs are presented in Fig. 3.3.

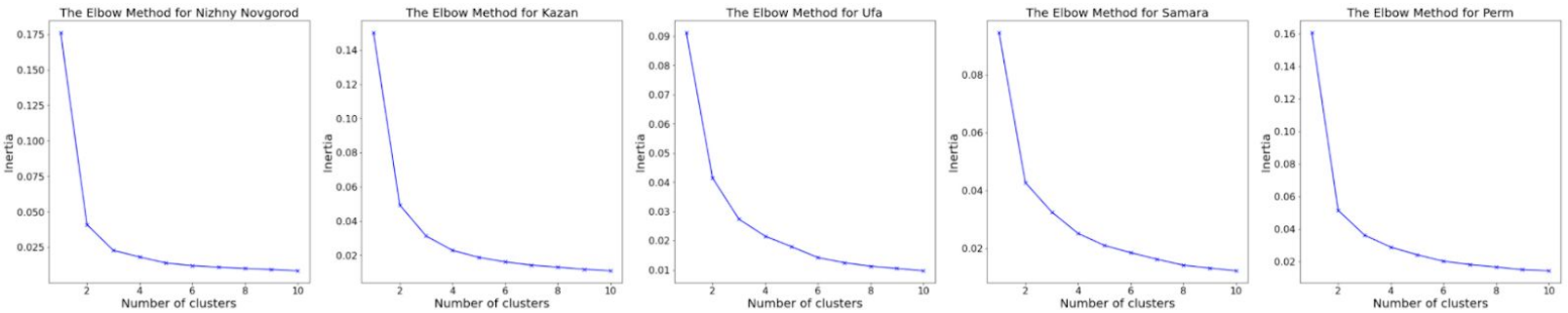

Figure 3.3. The graphs of inertia vs. number of clusters for the Elbow method for finding the optimal number of clusters for each city.

One can see from Fig. 3.3 that 3 clusters should be sufficient for an efficient clustering of Nizhny Novgorod and Kazan, while for the remaining cities 3 clusters might not be sufficient. Since the original city in our research is Nizhny Novgorod, then, in order to substantiate the final choice of the number of clusters, let us first cluster the city of Nizhny Novgorod using up to 5 clusters. After that, the best choice will be determined visually.

In Fig. 3.4, the results of the clustering of Nizhny Novgorod using the dataset from Table 2.5 are presented. Here, the clusters are ordered as follows. First, in each cluster, the average number of venues in each category is calculated. Then, the average value for all categories is also calculated. Finally, the clusters are ordered in a descending order with respect to the calculated values. Such the ordering of the clusters can be useful for determining similar clusters in different cities both visually and numerically.

One can see from Fig. 3.4 that the most efficient number of clusters is 4, since the difference between 4 and 5 clusters is not significant (only a few neighborhoods located at the same part of the city change their cluster), while 3 clusters are not sufficient, since the biggest part of the city becomes unclustered (only the city center was classified adequately). Since the difference in inertia between the clustering using more than 5 clusters becomes non-significant, then we will use 4 clusters for all the cities in order to find the closest one to Nizhny Novgorod.

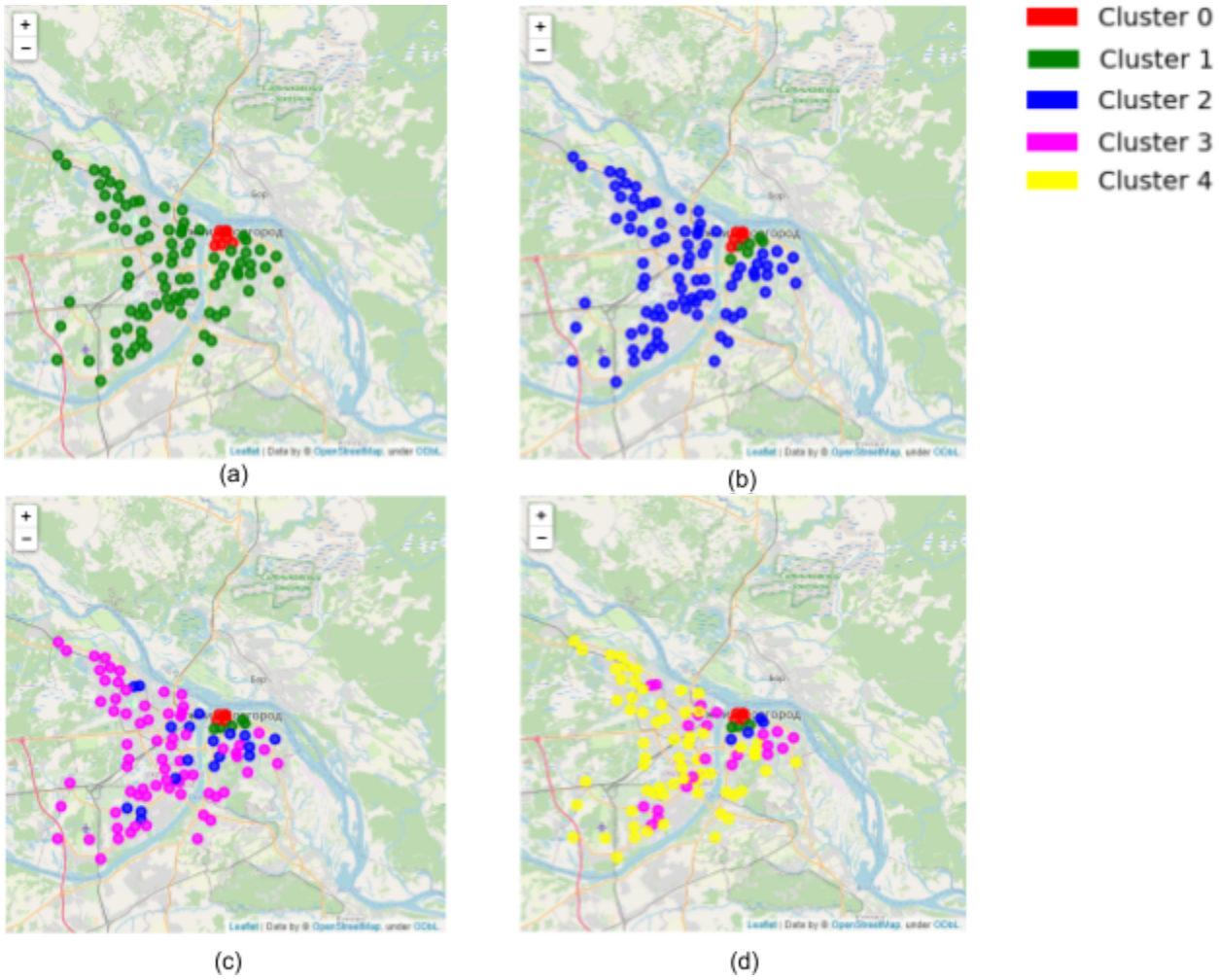

Figure 3.4. Clustered neighborhoods in Nizhny Novgorod using (a) 2 clusters, (b) 3 clusters (c) 4 clusters, and (d) 5 clusters. Each cluster is indicated by the respective color (cluster 0 - by red, cluster 1 - by green, cluster 2 - by blue, cluster 3 - by magenta, cluster 4 - by yellow).

4 Results

The results of the clustering of the cities Kazan, Ufa, Samara, and Perm are presented in Fig. 4.1. One can see from this figure that all the cities have been clustered in a similar way: Neighborhoods in Cluster 0 are located at the historical city center, they have the highest mean number of venues and a lot of arts & entertainment places. Cluster 1 consists of neighborhoods close to the city center, but having more food and shopping places. Cluster 2 consists of neighborhoods located outside of the city center with a smaller or medium

number of venues in each neighborhood. Finally, cluster 3 mainly consists of outskirts with a very small number of venues in each neighborhood.

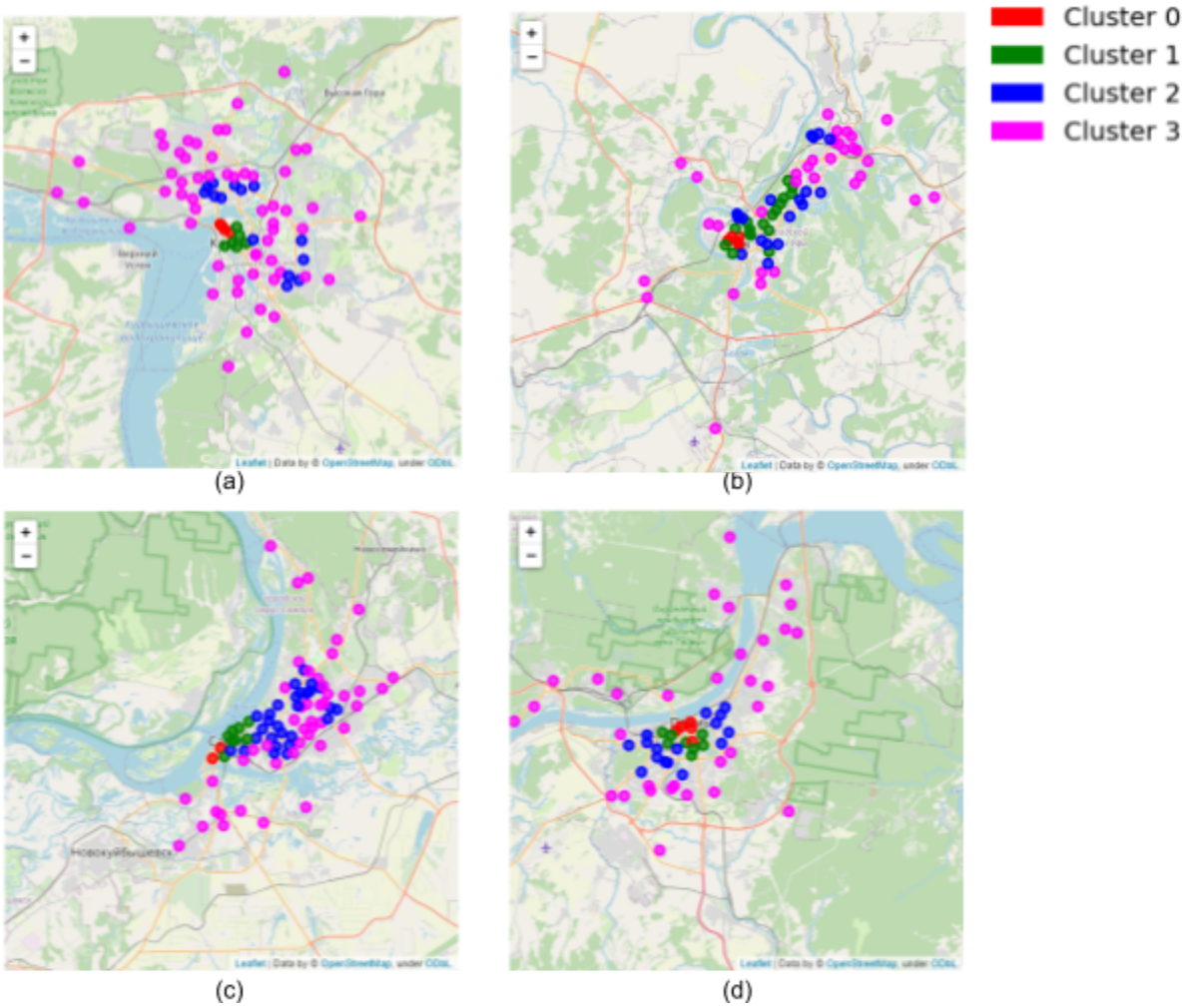

Figure 4.1. Neighborhoods divided into 4 clusters in (a) Kazan, (b) Ufa, (c) Samara, and (d) Perm. Each cluster is indicated by the respective color (cluster 0 - by red, cluster 1 - by green, cluster 2 - by blue, cluster 3 - by magenta).

In order to visualize better the obtained clusters, let us describe them visually using different diagrams:

1. Box plots for studying the neighborhoods within each cluster (see Figures 4.2 - 4.6).
2. Pie charts for studying the distributions of the venue categories within each cluster (see Fig. 4.7).

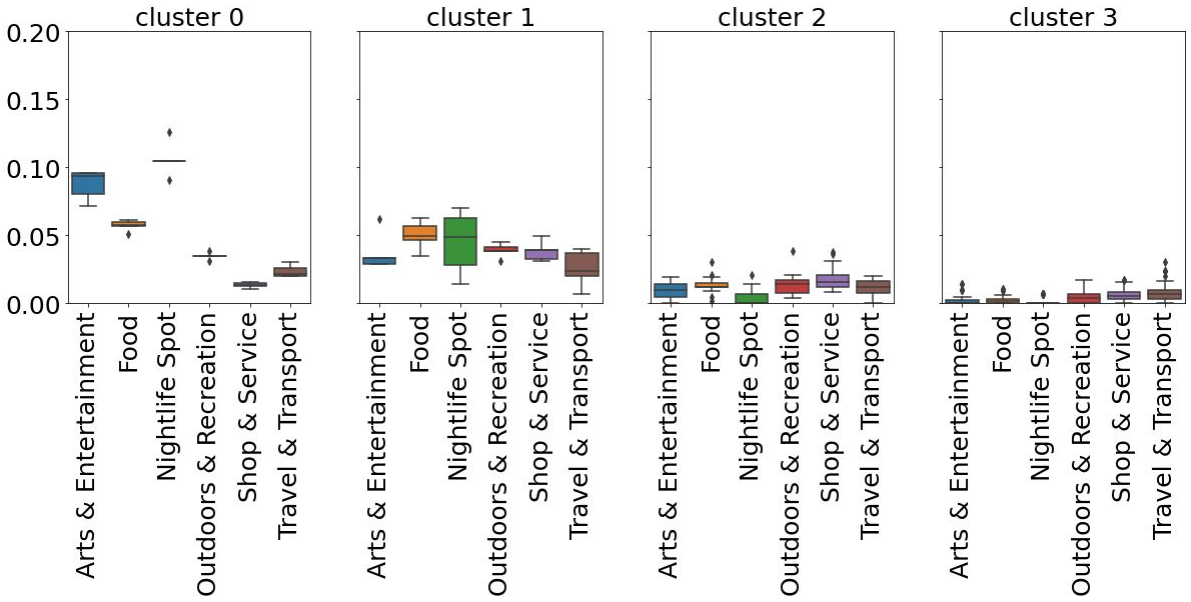

Figure 4.2. Box plots of the venue categories within each cluster for Nizhny Novgorod.

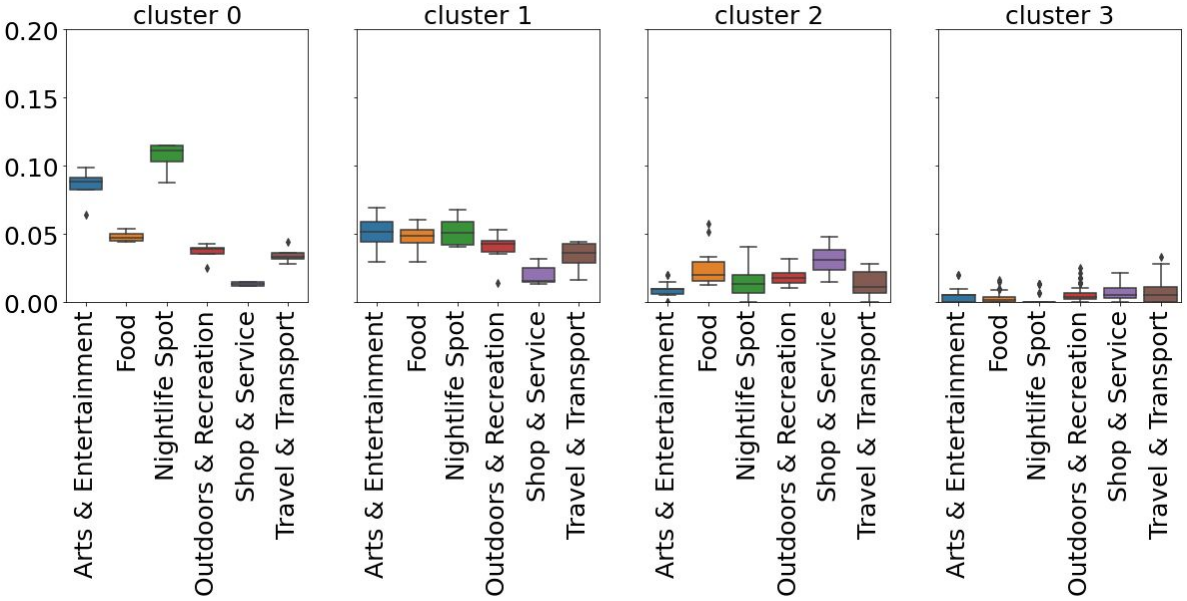

Figure 4.3. Box plots of the venue categories within each cluster for Kazan.

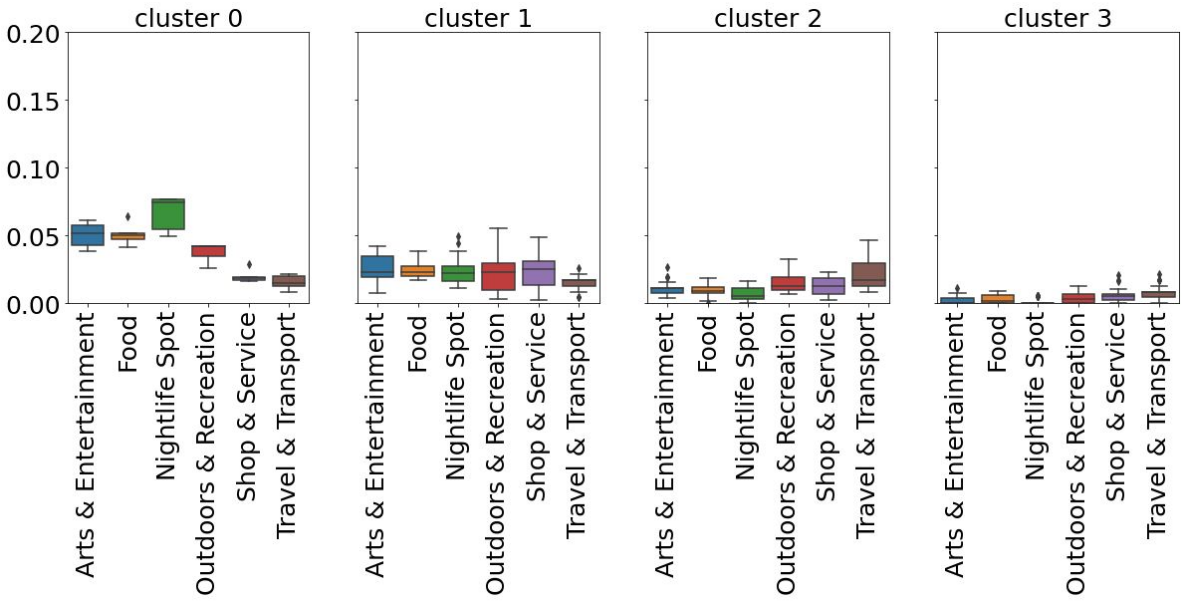

Figure 4.4. Box plots of the venue categories within each cluster for Ufa.

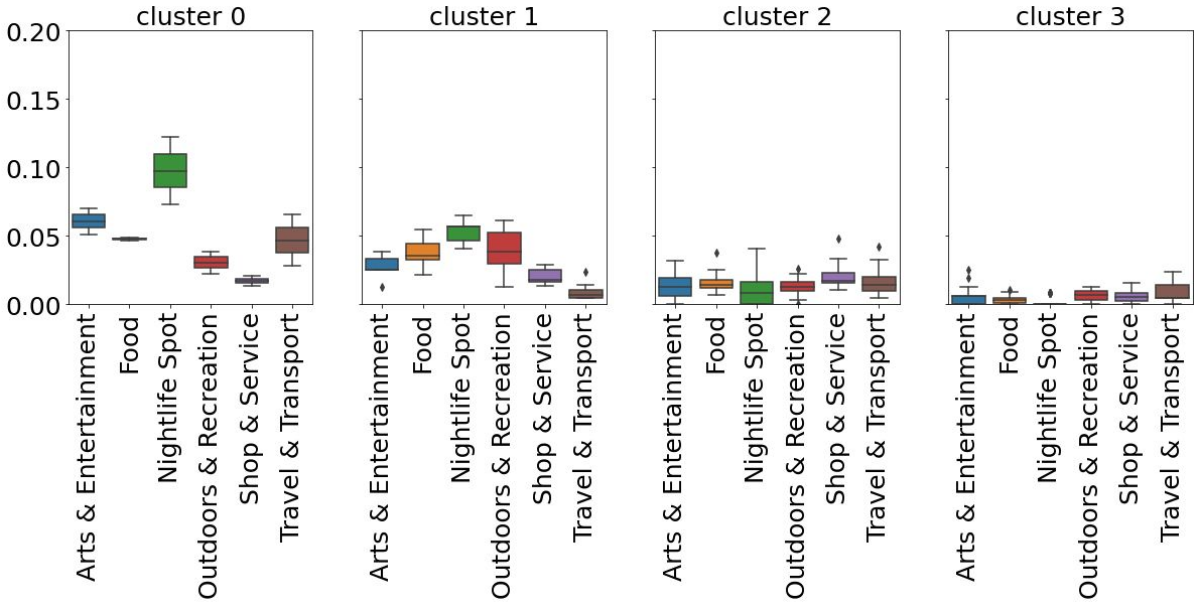

Figure 4.5. Box plots of the venue categories within each cluster for Samara.

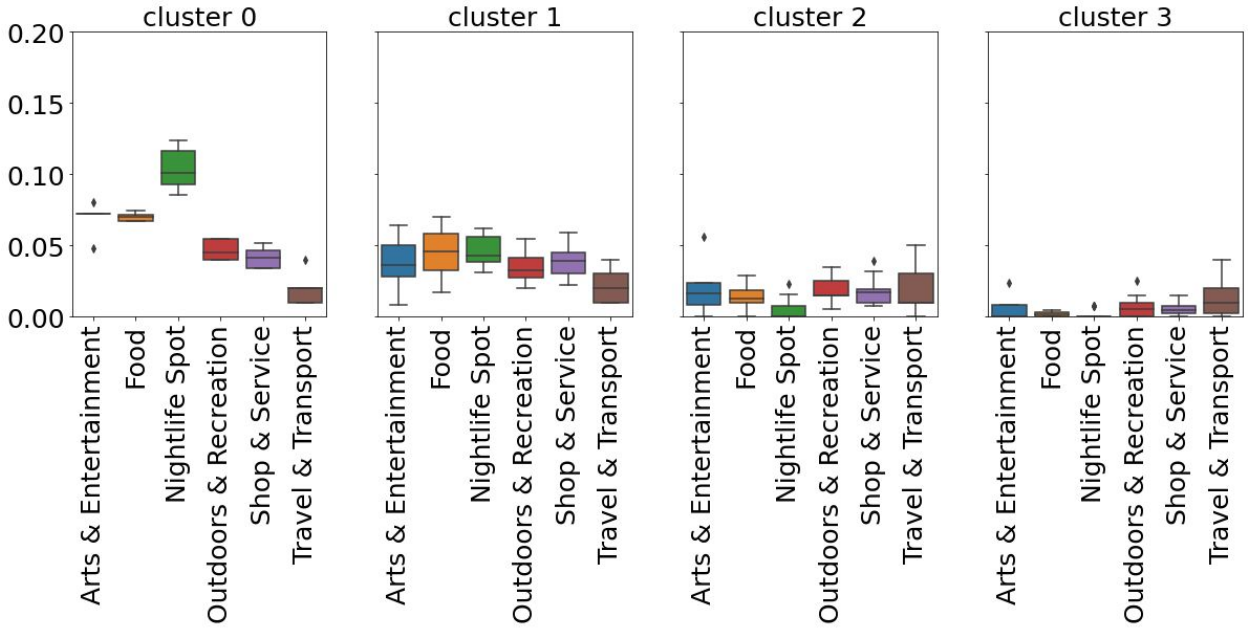

Figure 4.6. Box plots of the venue categories within each cluster for Perm.

One can see that the clusters are significantly different in all the cities. One can decide if the clusters 0 and 1 should be unified into one cluster, since they both are related to the city centers: cluster 0 is related to the old historical center, e.g., to the kremlins in Nizhny Novgorod or Kazan, while cluster 1 is related to a more modern, but always central part of the city. However, since in the original city (Nizhny Novgorod) these two parts of the city are significantly different, then I decide to save them as two different clusters (even if Cluster 0 has only a few neighborhoods in each city): e.g., the neighborhoods in cluster 0 have more arts & entertainment venues and much less shops & service venues, with respect to the cluster 1.

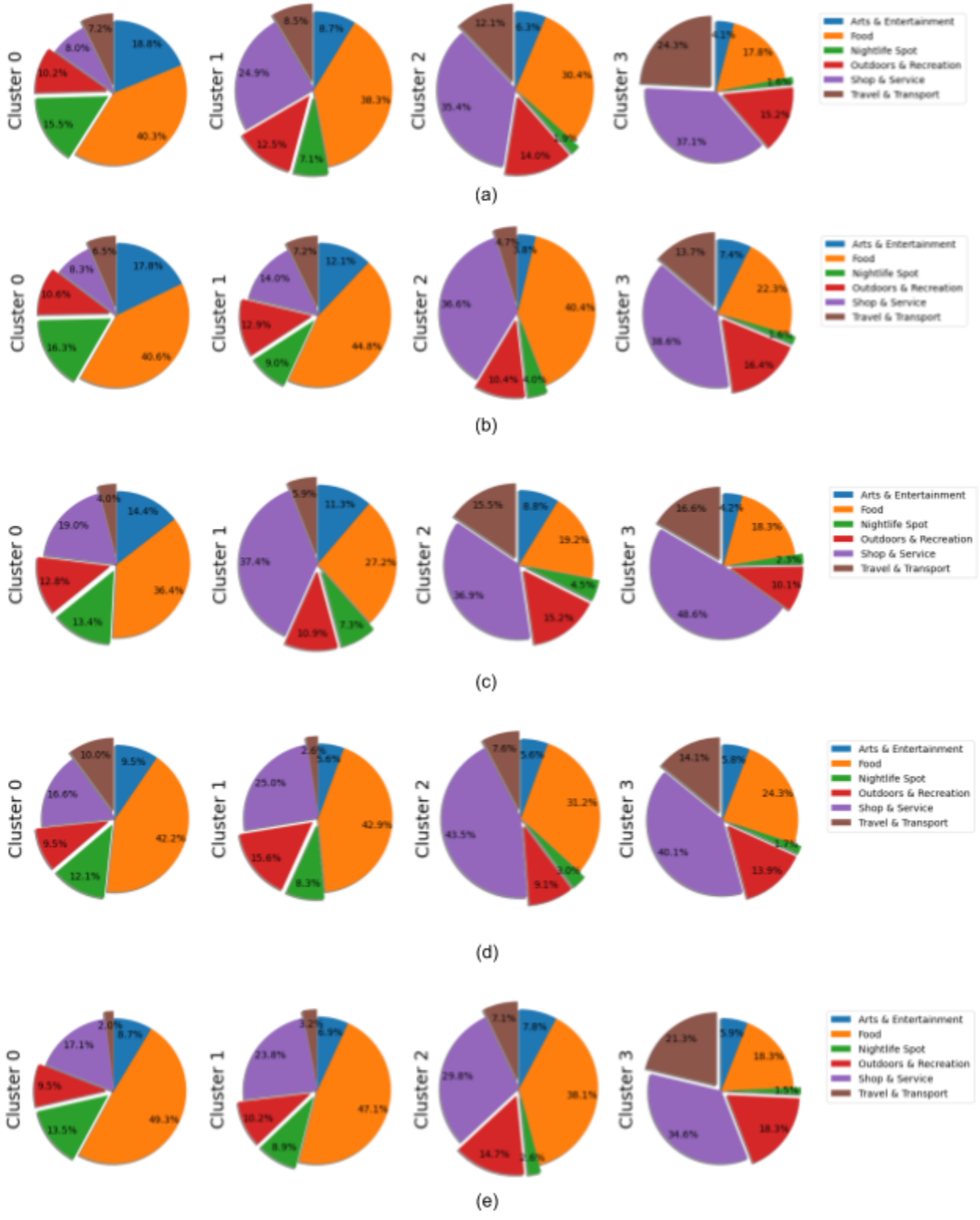

Figure 4.7. Pie charts of the venue categories within each cluster for (a) Nizhny Novgorod, (b) Kazan, (c) Ufa, (d) Samara, and (e) Perm.

5 Discussion

5.1 Battle of the cities

In order to decide, which city is the closest one to Nizhny Novgorod in terms of venues, let us calculate the root mean square deviations for each cluster and each city with respect to Nizhny Novgorod. First, let us use the percentage distributions of each category within each cluster (see Fig. 4.7):

City	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Average
Kazan	0.608988	5.446536	5.481505	4.909425	4.111613
Ufa	5.414327	7.002333	5.053364	6.027620	5.874411
Samara	5.520522	3.557337	4.331435	5.140835	4.637533
Perm	7.045545	4.451563	4.453031	2.174331	4.531118

Table 5.1. Root mean square deviations of the percentage distributions of each category within each cluster. The smallest RMS value is indicated in red for each cluster.

We can see that, according to this criterion, we cannot choose the closest city to Nizhny Novgorod, since there are different “best” cities for each cluster. In particular, if the stakeholders are interested in locating a restaurant in a neighborhood in the city center, then it can be reasonable to choose Kazan, since it is the most similar to Nizhny Novgorod within the Cluster 0 (or in average within the Cluster 0 and Cluster 1, if we would decide to join them). The same is correct for Cluster 3, i.e., outskirts. However, the distributions of the categories in Kazan within the Cluster 2 are very different w.r.t. Nizhny Novgorod. In the case the stakeholders are interested in Cluster 2, i.e., neighborhoods located outside of the city center with medium number of venues, then the reasonable choice is Ufa.

Another possible criterion for the comparison of the cities is to use the average relative number of venues within each cluster for each category. In this case, we can also calculate the RMS deviations in the same way and then rank the cities.

City	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Average
Kazan	0.609262	1.047903	0.866304	0.133041	0.664128
Ufa	2.269787	1.741474	0.480833	0.060139	1.138058
Samara	1.574283	1.215271	0.445047	0.155724	0.847581
Perm	1.571093	0.326216	0.382862	0.226016	0.626547

Table 5.2. Root mean square deviations of the average relative number of venues for each category within each cluster (in percents). The smallest RMS value is indicated in red for each cluster.

We can see that, according to this criterion, Perm is the most similar city to Nizhny Novgorod for Clusters 1 and 2, while Kazan and Ufa are the most similar to Nizhny Novgorod for Clusters 0 and 3, respectively.

Finally, another possible criterion is to choose the best city based on the most common venue categories for each cluster. We can see from the Fig. 4.7, which categories are the most common for each cluster:

City	Cluster 0	Cluster 1	Cluster 2	Cluster 3
Nizhny Novgorod	[Food, Arts & Entertainment, Nightlife Spot]	[Food, Shop & Service, Outdoors & Recreation]	[Shop & Service, Food, Outdoors & Recreation]	[Shop & Service, Travel & Transport, Food]
Kazan	[Food, Arts & Entertainment, Nightlife Spot]	[Food, Shop & Service, Outdoors & Recreation]	[Food, Shop & Service, Outdoors & Recreation]	[Shop & Service, Food, Outdoors & Recreation]
Ufa	[Food, Shop & Service, Arts & Entertainment]	[Shop & Service, Food, Arts & Entertainment]	[Shop & Service, Food, Travel & Transport]	[Shop & Service, Food, Travel & Transport]
Samara	[Food, Shop & Service, Nightlife Spot]	[Food, Shop & Service, Outdoors & Recreation]	[Shop & Service, Food, Outdoors & Recreation]	[Shop & Service, Food, Travel & Transport]
Perm	[Food, Shop & Service, Nightlife Spot]	[Food, Shop & Service, Outdoors & Recreation]	[Food, Shop & Service, Outdoors & Recreation]	[Shop & Service, Travel & Transport, Food]

Table 5.3. Top-3 most common venue categories within each cluster for each city.

We can see from Table 5.3 that Kazan is the most similar to Nizhny Novgorod according to this criterion for the central part of the cities, while Perm is more similar for Clusters 2 and 3. However, if the order of the categories within top-3 for each cluster is not significant, then the Jaccard index can be used to estimate the similarity between the cities:

City	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Average
Kazan	1.000000	1.000000	1.000000	0.500000	0.875000
Ufa	0.500000	0.500000	0.500000	1.000000	0.625000
Samara	0.500000	1.000000	1.000000	1.000000	0.875000
Perm	0.500000	1.000000	1.000000	1.000000	0.875000

Table 5.4. Jaccard indices for the cities for each cluster w.r.t. Nizhny Novgorod according to Table 5.3.

However, the Jaccard index does not provide significant information in this case, since there are a small number of categories in each city (only 6 venue categories). Moreover, no one of the previous criteria did not provide the best city in the overall competition (that can be done only in average for the previous criteria). In order to compare the cities more efficiently using different numeric metrics for estimation of the similarities, let us apply the following metrics from Wolda(1981):

1. Bray and Curtis similarity index.
2. Bray and Curtis similarity index after the logarithmic transformation.
3. Canberra Metric.
4. Squared Euclidean distance.
5. Morisita index.
6. Simplified Morisita index.

Each of these metrics estimate the similarity (higher the metric means higher similarity) of two samples. Each sample i (i = 1,2) consists of $n_{i,j}$ individuals of a type j and $\sum_j n_{i,j} = N_i$ individuals in total. In our case, we have a sample of neighborhoods in each city, where each neighborhood can belong to one of 4 clusters. In this case, for each city i (i=0,1,2,3,4), the sample consists of $n_{i,j}$ individuals of the cluster j (j = 0,1,2,3). The results of the metrics are presented in Table 5.5. One can see from Table 5.5 that for each similarity index, the best city is Kazan: it has the highest similarity indices in all the cases.

City	Bray and Curtis	Bray and Curtis (log)	Canberra Metric	Squared Euclidean distance	Morisita index	Simplified Morisita index
Kazan	0.877778	0.954826	0.866495	0.999150	1.009178	0.999207
Ufa	0.715909	0.916262	0.768784	0.895599	0.893019	0.878816
Samara	0.765027	0.900998	0.706925	0.943633	0.952931	0.941298
Perm	0.743902	0.937598	0.824342	0.954763	0.965999	0.950675

Table 5.5 Similarity metrics for the cities. The best value for each metric is indicated in red.

5.2 Placing a restaurant in the cities

After the decision in which city to open a restaurant based on the criteria from the previous subsection, an efficient place for the restaurant should be chosen. There can be qualitatively different neighborhoods even within one cluster: e.g., in one neighborhood there can be already more food restaurants, than in another, but in the second one there are more non-food venues. In this case, it can be reasonable to choose the second neighborhood, since this allows to reduce the concurrency and to have more clients due to a high traffic.

For this reason, let us first determine the number of food venues and non-food venues for each neighborhood. After that, let us order them within each cluster by the number of food venues in a descending order and take only unique numbers of food venues. Choose the first quartile of the obtained unique values and then take only neighborhoods, which have a number of food venues less or equal to this value. After that, order the obtained neighborhoods in an ascending order by the number of non-food venues and take the top-3 neighborhoods for each cluster. The obtained neighborhoods are visualized in the map and described by the respective table below for Kazan being the closest city to Nizhny Novgorod according to the previous section:

Neighborhood	Sum of Food Venues	Sum of Non-Food Venues	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
420014	36	64	55.796878	49.108822	0	Food	Arts & Entertainment	Nightlife Spot
420017	37	63	55.798551	49.106324	0	Food	Arts & Entertainment	Nightlife Spot
420111	40	60	55.793351	49.112408	0	Food	Arts & Entertainment	Nightlife Spot
420107	34	54	55.779228	49.131082	1	Food	Shop & Service	Arts & Entertainment
420021	24	43	55.779342	49.112822	1	Food	Outdoors & Recreation	Arts & Entertainment
420043	39	39	55.782064	49.144244	1	Food	Shop & Service	Nightlife Spot
420133	12	31	55.830835	49.157022	2	Shop & Service	Food	Travel & Transport
420140	14	31	55.767897	49.231726	2	Shop & Service	Food	Travel & Transport
420044	12	28	55.832822	49.095629	2	Shop & Service	Food	Outdoors & Recreation
420029	0	18	55.813868	49.186202	3	Shop & Service	Travel & Transport	Arts & Entertainment
420104	1	15	55.757109	49.196832	3	Outdoors & Recreation	Shop & Service	Travel & Transport
420032	2	13	55.811061	49.068375	3	Shop & Service	Outdoors & Recreation	Travel & Transport

Table 5.6. 3 best found neighborhoods within each cluster in Kazan ordered by the sum of non-food venues in the descending order within each cluster.

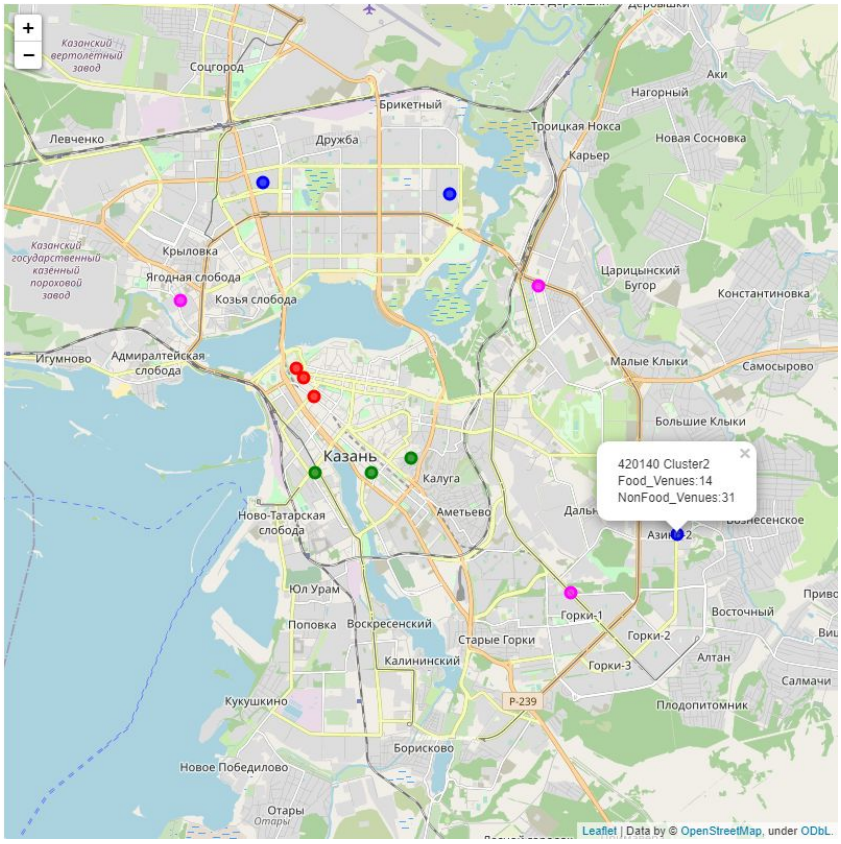

Figure 5.1. 3 best found neighborhoods within each cluster in Kazan. For each neighborhood, the number of food venues and non-food venues are indicated as well as the cluster (the cluster is indicated by both the text and the color).

6 Conclusion

Five biggest cities of the Volga Federal District have been studied in this research. For each city, the neighborhoods based on the postal code have been subdivided into 4 clusters: 1) neighborhoods in the old city center, 2) neighborhoods in the modern city center, 3) neighborhoods outside of the center, 4) outskirts of the city. There have been identified several criteria for the comparison of the cities based on the obtained results. In particular, several numeric metrics used in literature for estimation of the similarity between samples have been used.

According to the used similarity indices, the closest city to Nizhny Novgorod in terms of the number of venues within each cluster is Kazan. This result is reasonable, because both the cities are very similar even structurally. First, each of them has a kremlin as a core of the old historical center of the city. Second, the main pedestrian streets proceed right to the kremlins and a lot of venues are placed around them. Both of the cities are placed at the Volga river, they are administrative centers of two large regions within the Volga Federal District, and Kazan is the closest (in terms of the distance) city to Nizhny Novgorod with respect to other cities studied in this research. For all these reasons, we can substantiate that the most similar city w.r.t. Nizhny Novgorod between the studied cities is Kazan.

A future study is possible with a close collaboration with the stakeholders restricting and understanding their requirements: e.g., if the stakeholders are interested in a particular particular part of the city (e.g., neighborhoods outside of the city center, but not outskirts) and the similarity of distributions of the venue categories is the most useful criterion for them, then a particular city can become more appropriate than Kazan. A stronger collaboration with the stakeholders will improve the obtained accuracy and restrict the obtained results.

Other important conclusions can be also made observing the obtained results. First, the studied cities are very similar: they have a strict old city center, a modern city center, a developing part of the city with more transport and service venues, and, finally, outskirts with a low number of venues. Second, it has been obtained that the Foursquare API does not provide all necessary information about the cities. Moreover, some important venues are not presented in the Foursquare, since the cities are not touristy, being distant from the

main touristic venues or cities. Foursquare API can be more reasonable to use for studying bigger or more touristic cities, e.g., Moscow or Saint Petersburg. For a more precise analysis, it can be more reasonable to use other APIs (possibly, integrating their results with the ones obtained by Foursquare), for example, Google or Yandex. Yandex is a russian company providing Internet-related products and services, including transportation, search and information services. It is very similar to Google, but provides the most actual and full information about venues in russian cities. However, at this moment, it does not provide any free instrument for study of venues in russian cities as well as Google, so I decided to use Foursquare only for now. But, since Yandex grows very fast, being the largest IT company in Russia (see, e.g., the following link) with a rapidly growing service Yandex.map and particular interest to Machine Learning and Data Science, it is already clear that it will be used for the analogous study in the near future, probably, together with Foursquare and/or Google.

Acknowledgements

The author is grateful to Coursera and IBM for an opportunity to obtain/improve all necessary basic skills in the field of Data Science. Many thanks, in particular, to all instructors and moderators of the courses included to the IBM Data Science Professional Certificate.

References

1. Github repository containing the present study, https://github.com/maratsmuk/Coursera_Capstone
2. Federal districts in Russian Federation, Wikipedia, https://en.wikipedia.org/wiki/Federal_districts_of_Russia
3. IBM Watson Studio, https://www.ibm.com/cloud/watson-studio
4. Dataset of all russian postal codes, http://download.geonames.org/export/zip/RU.zip
5. Foursquare API, developer page, https://developer.foursquare.com/developer/
6. K-means clustering using scikit-learn, https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html
7. Wolda, H. Similarity indices, sample size and diversity. Oecologia 50, 296–302 (1981). https://doi.org/10.1007/BF00344966