IBM Data Science Professional Certificate

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

Marat S. Mukhametzhanov

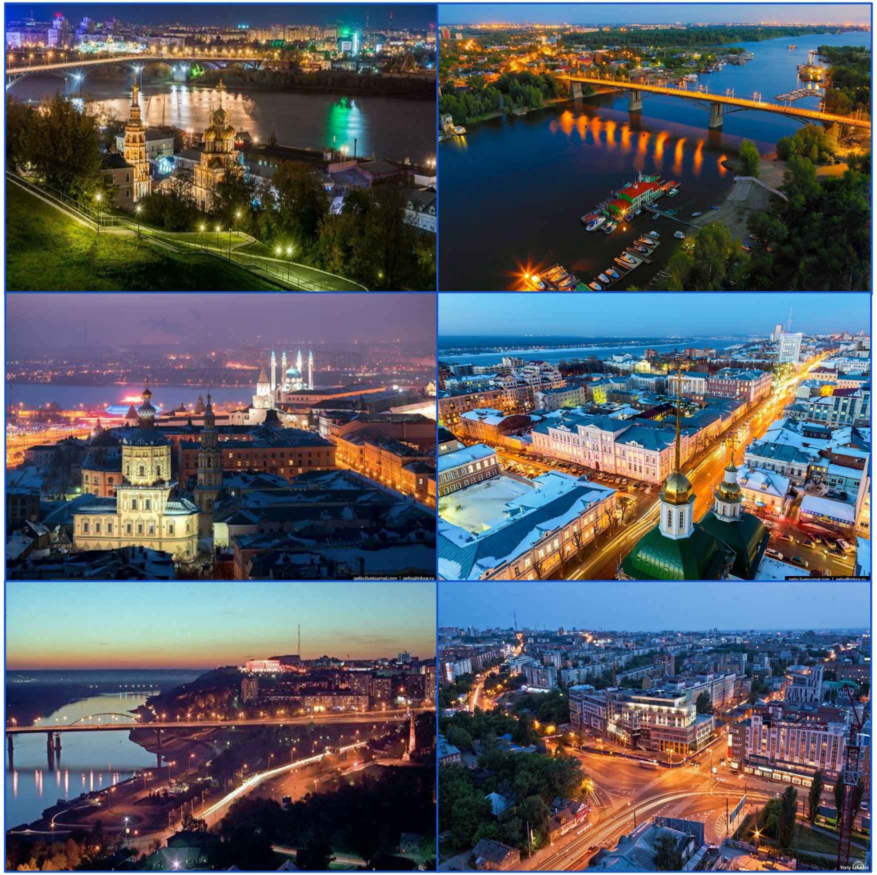
# Background

The Volga Federal District is one of the federal districts in Russian Federation.

It consists 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:**

1. **Nizhny Novgorod** (rus. *Нижний Новгород*, administrative center, population ~1.252 million people),
2. **Kazan** (rus. *Казань*, population ~1.257 million people),
3. **Ufa** (rus. *Уфа*, population ~1.128 million people ,
4. **Samara** (rus. *Самара*, population ~1.156 million people),
5. **Perm** (rus. *Пермь*, population ~1.055 million people)



# Volga Federal District
# Capital: Nizhny Novgorod


# Business Problem

- A series of restaurants in Nizhny Novgorod has become successful, so its stakeholders are interested to extend their business into other cities.
- 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.
- 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.
- 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.

**The main problem in this research:**

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 successful restaurants in the original city (Nizhny Novgorod).

# Interest

- Different businessmen who consider extending their business from one to other cities
  - E.g., there are several series of food restaurants, sushi bars, pizzerias, bars, etc.
- 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).
- Only the Data section of the Jupyter notebook related to this study should be changed in the latter case.

# Used Software

- All methods have been implemented in Python 3.7 using Jupyter notebooks in IBM Watson Studio.
- 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;
  - *requests* for working with the Foursquare API.

# Data wrangling

**Two main datasets have been used:**

1. Neighborhoods associated with postal codes and their coordinates:
   a. List of all postal codes in the studied cities from geonames.

| | 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 |

   b. Coordinates of each neighborhood obtained using Nominatim in geopy.geocoders library.
   For example:

*geolocator = Nominatim(user_agent="PFO_explorer")*
*pcode = ‘603076’*
*address = pcode+', '+’Nizhny Novgorod’+', Russia'*
*location = geolocator.geocode(address)*

| | Postal Code | Latitude | Longitude |
|---|---|---|---|
| 0 | 603000 | 56.317042 | 43.994228 |
| 1 | 603001 | 56.328599 | 43.991916 |
| 2 | 603002 | 56.317991 | 43.946463 |
| 3 | 603003 | 56.353209 | 43.859960 |
| 4 | 603004 | 56.244075 | 43.871575 |

2. Venues available from each neighborhood obtained using Foursquare API (within the radius of 1000 m.)

| | 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 |

# Data Cleaning

## Number of all venues for each category in each city

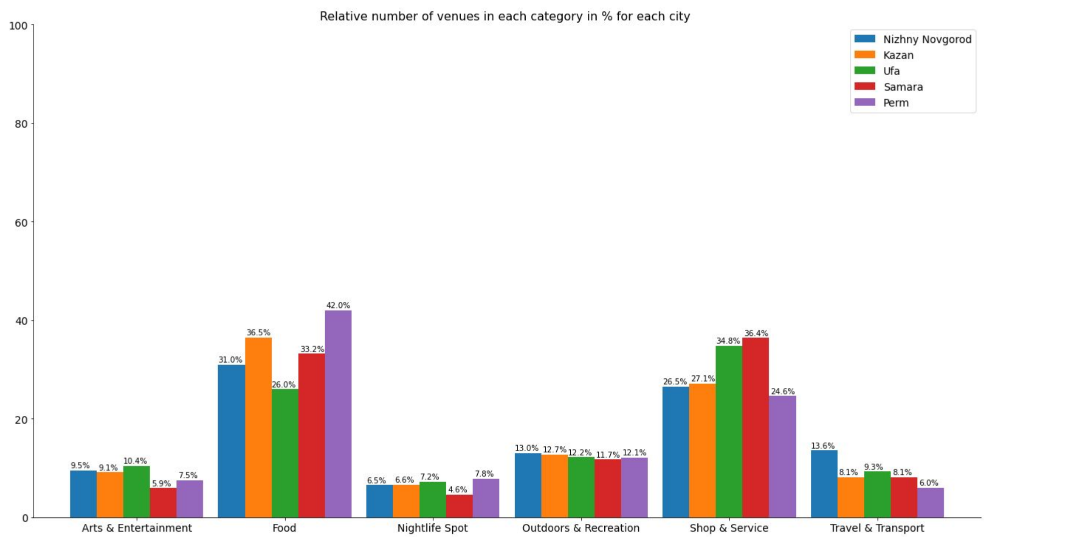
### Before cleaning:

| 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 |

### After cleaning:

| Category\City | Nizhny Novgorod | Kazan | Ufa | Samara | Perm |
|---|---|---|---|---|---|
| Arts & Entertainment | 209 | 203 | 261 | 157 | 125 |
| Food | 685 | 813 | 654 | 882 | 697 |
| Nightlife Spot | 143 | 148 | 182 | 123 | 129 |
| Outdoors & Recreation | 288 | 282 | 308 | 311 | 201 |
| Shop & Service | 586 | 603 | 877 | 966 | 409 |
| Travel & Transport | 300 | 180 | 235 | 214 | 100 |
| Total | 2211 | 2229 | 2517 | 2653 | 1661 |

## Bar plots of relative number of venues in each category (in %)



# Feature Selection

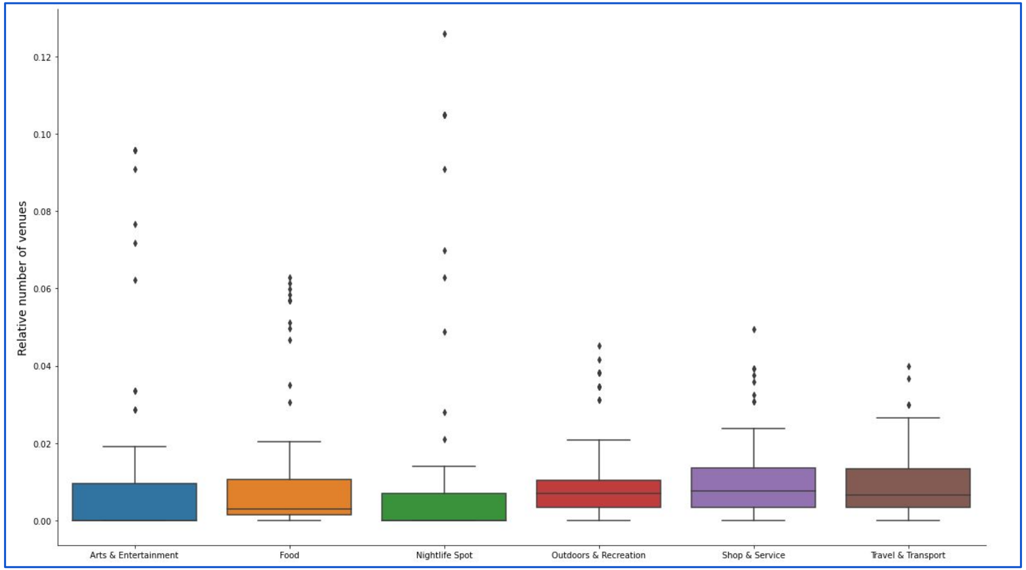
- Relative number of venues for each neighborhood  (the first 5 rows for Nizhny Novgorod)

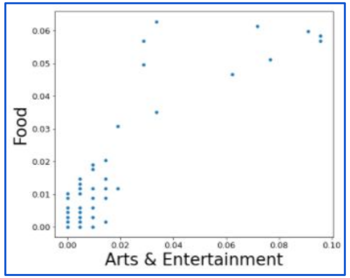
| | Neighborhood | Arts & Entertainment | Food | Nightlife Spot | Outdoors & Recreation | Shop & Service | Travel & Transport |
|---|---|---|---|---|---|---|---|
| 0 | 603000 | 0.010160 | 0.026172 | 0.017893 | 0.016981 | 0.024422 | 0.024775 |
| 1 | 603001 | 0.023222 | 0.021302 | 0.029821 | 0.018868 | 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 |

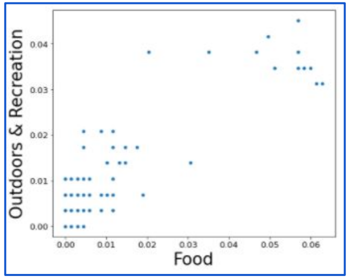
- Dimension of the search space: 6 (there are 6 venues categories in the dataset)
- Each point (or a row in the dataset) is represented by the postal code and 6 numeric values of the range (0,1) being the relative numbers of venues in each category around the neighborhood defined through the postal code.

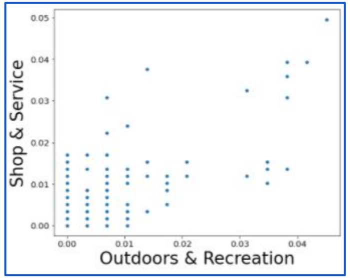
# Exploratory Data Analysis

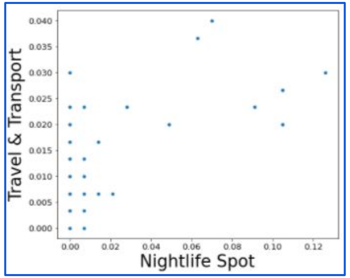
Box plots/scatter plots of relative number of venues in each category in Nizhny Novgorod:



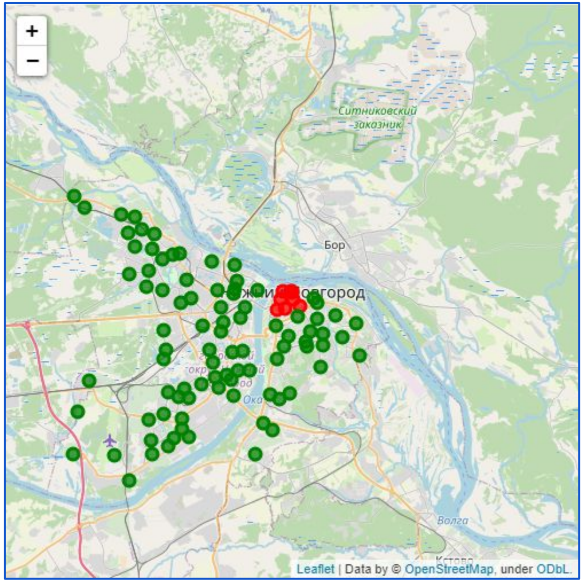



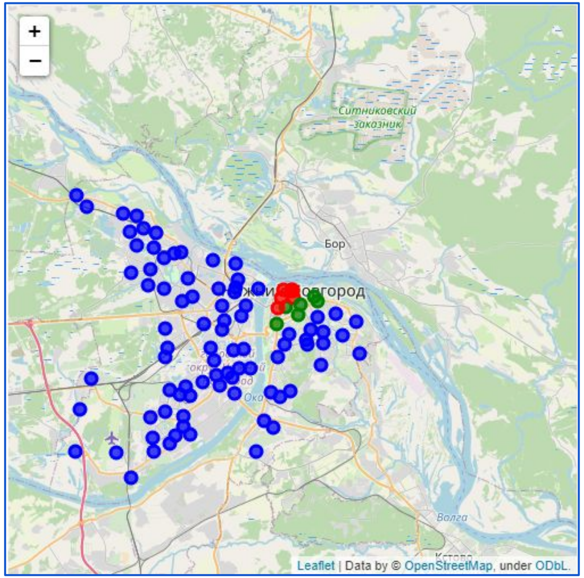

- Box plots show a bias in several categories, scatter charts show a heterogeneous structure of the city.
- Clustering is necessary.
- Distributions of the categories within each cluster should be studied separately.

# Optimal k for k-means clustering, Nizhny Novgorod

Results using 2 clusters:



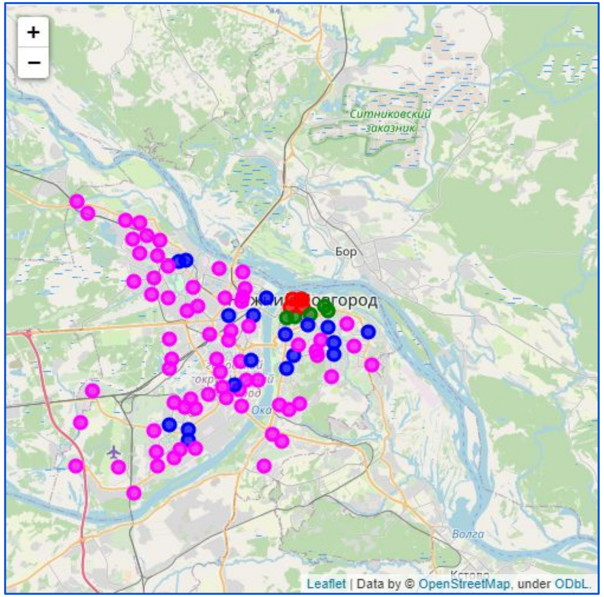

Results using 3 clusters:




- Clusters are ordered by the average number of mean venues in each category within each cluster (in descending order)
- 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.

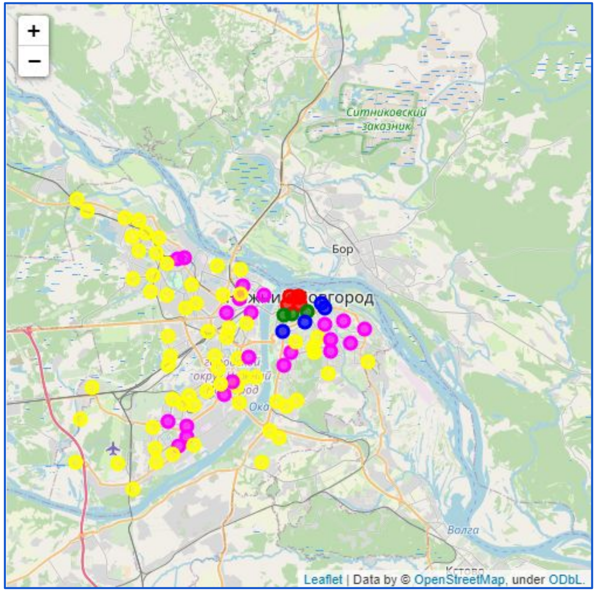
# Optimal k for k-means clustering, Nizhny Novgorod

Results using 4 clusters:




Results using 5 clusters:

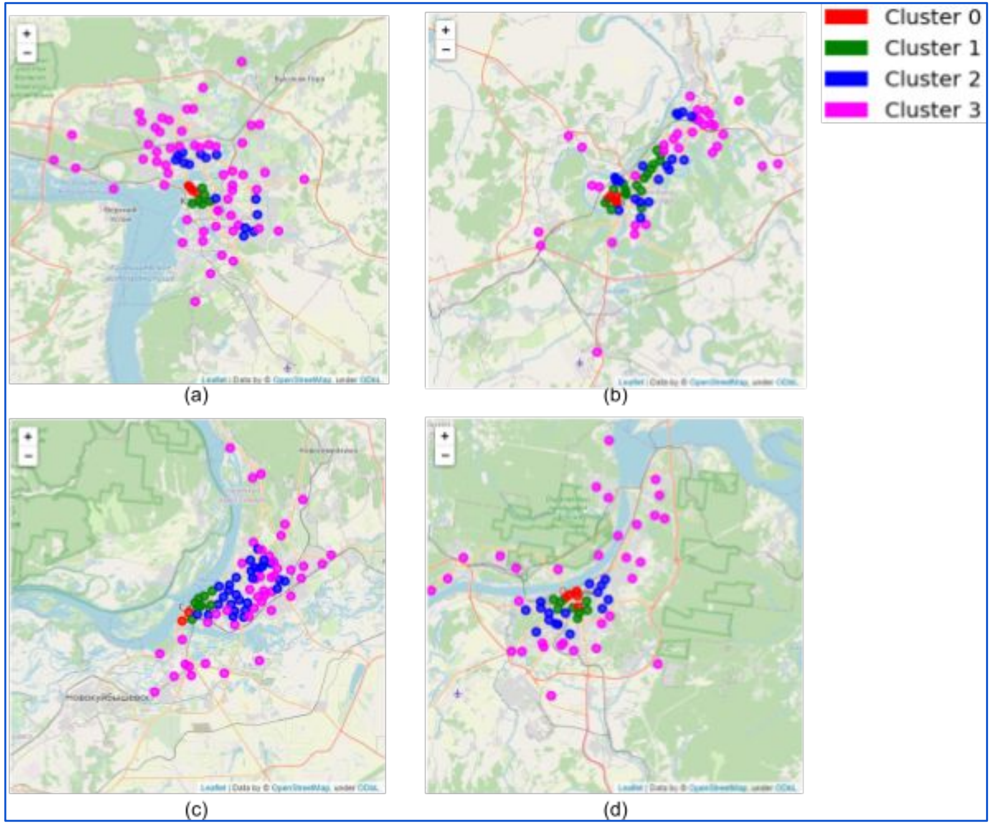
Best value: k=4




- Clusters are ordered by the average number of mean venues in each category within each cluster (in descending order)
- 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.

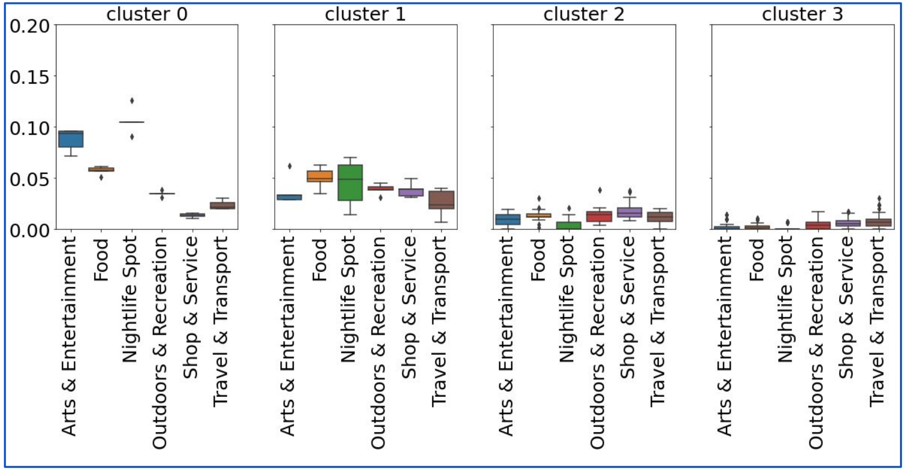
# Results

- Clusters are ordered by the average number of mean venues in each category within each cluster (in descending order):
  - **Cluster 0:** historical city center.
    The highest mean number of venues.
    Top venues:
    Arts&Entertainment, Food, and Nightlife.
  - **Cluster 1:** modern city center.
    Top venues:
    Food, Shopping&Service,
    Outdoors&Recreation/Arts&Entertainment.
  - **Cluster 2:** neighborhoods outside the center.
    Top venues:
    Shop&Service, Food,
    Outdoors&Recreation/Travel&Transport.
  - **Cluster 3:** outskirts, the smallest number of venues.
    Top venues:
    Shop&Service, Food, Travel&Transport.

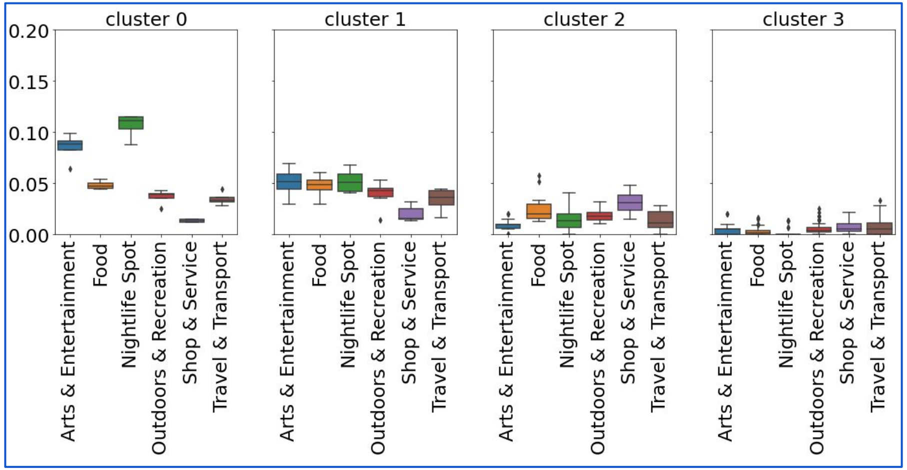


# Boxplots within clusters: N. Novgorod vs. other cities

## Nizhny Novgorod:

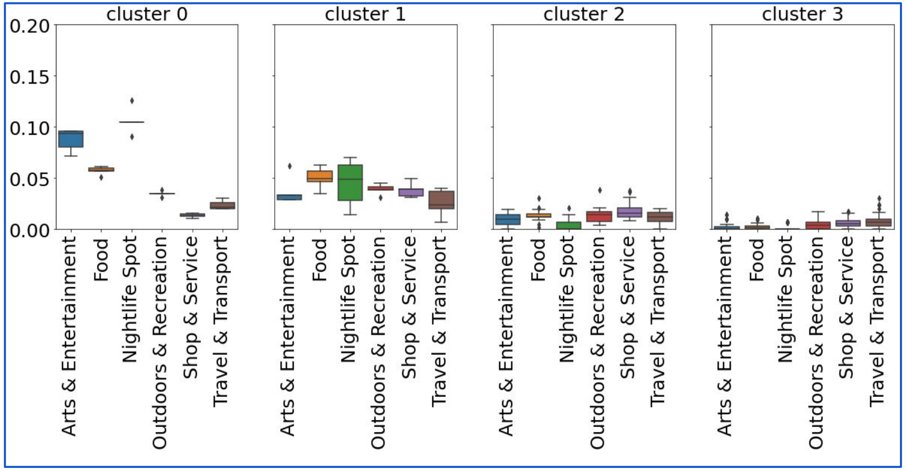



## Kazan:

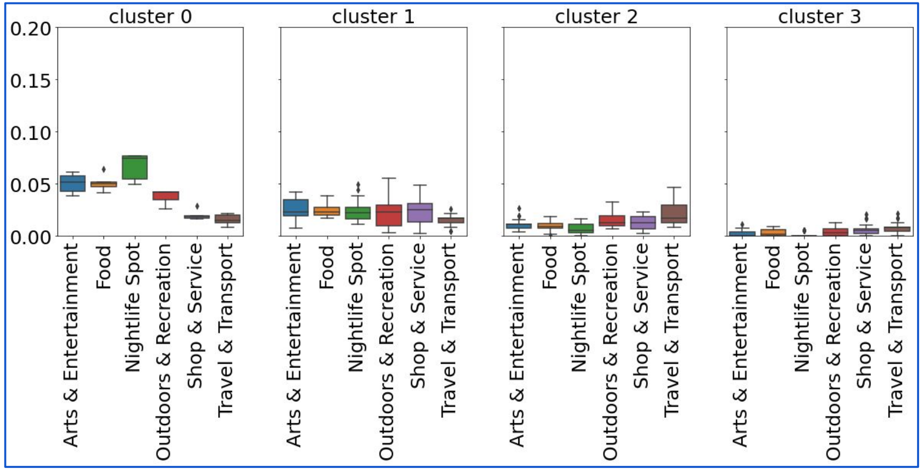


# Boxplots within clusters: N. Novgorod vs. other cities

## Nizhny Novgorod:

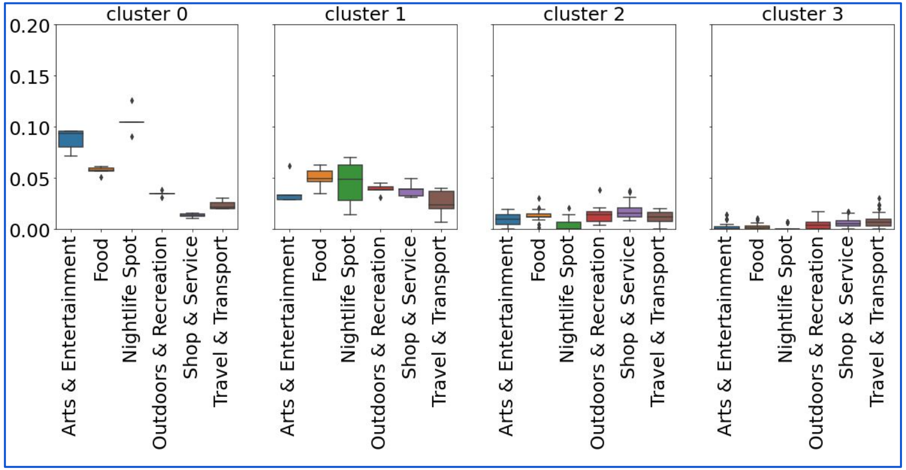



## Ufa:

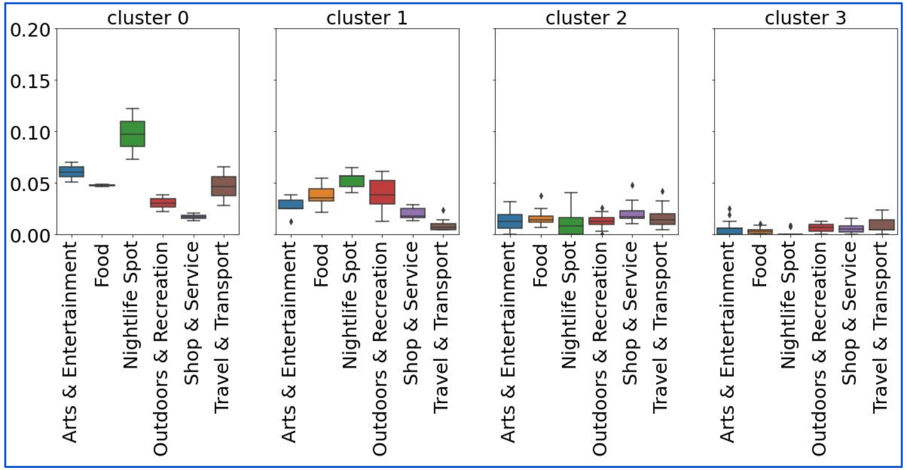


# Boxplots within clusters: N. Novgorod vs. other cities

## Nizhny Novgorod:

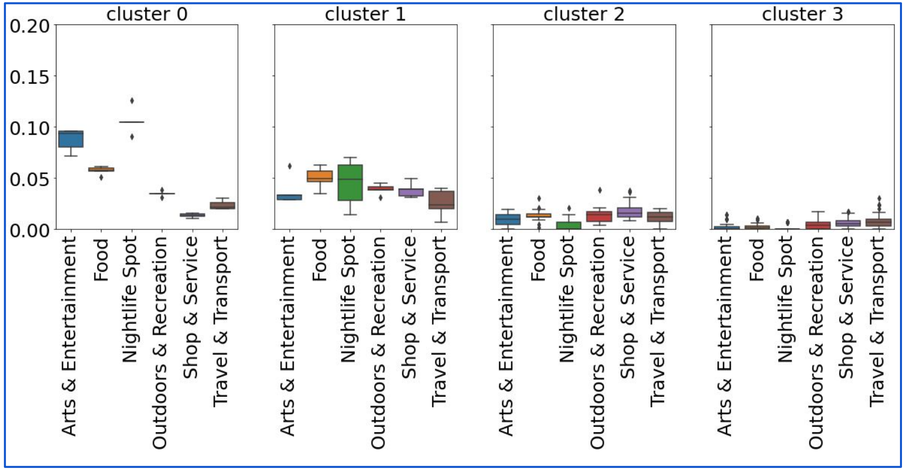



## Samara:

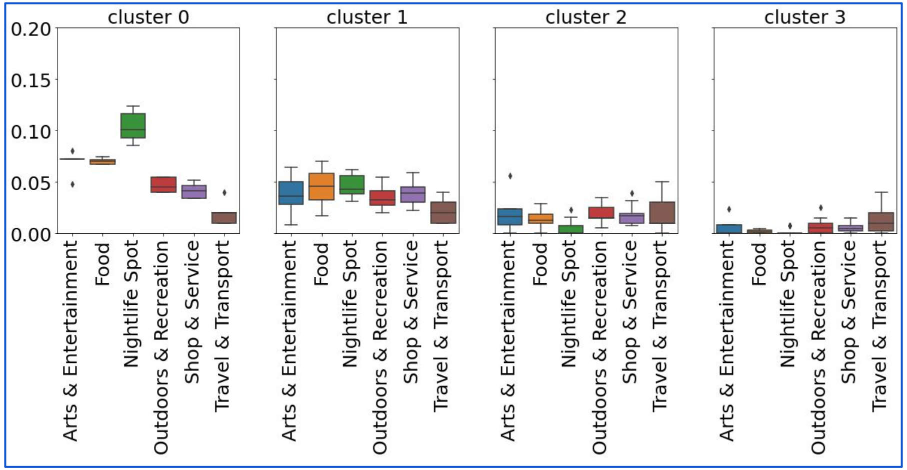


# Boxplots within clusters: N. Novgorod vs. other cities

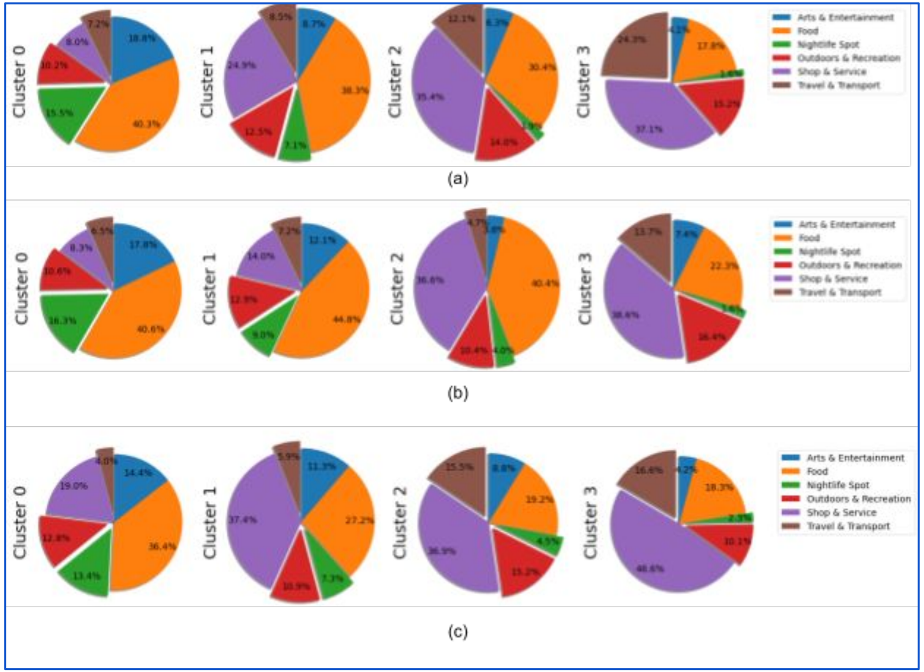
## Nizhny Novgorod:

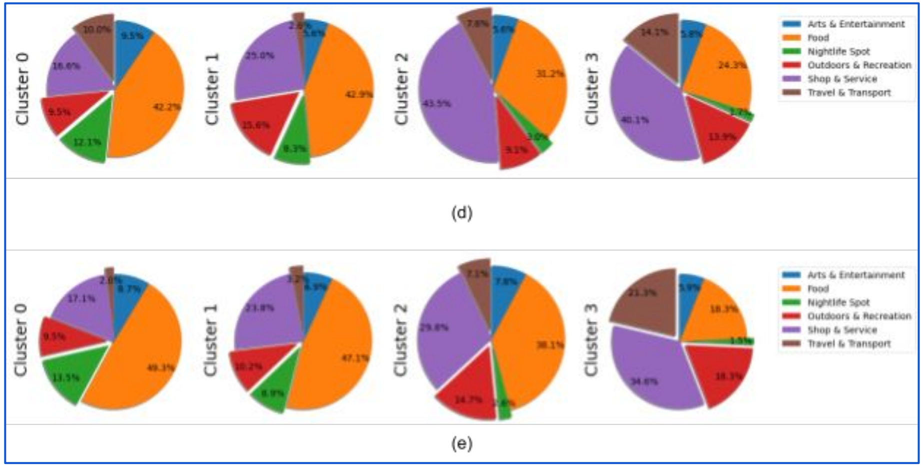



## Perm:



# Pie charts of the venue categories within each cluster




(a) Nizhny Novgorod, (b) Kazan, (c) Ufa, (d) Samara, and (e) Perm.

# Discussion

- Similarity indices from Wolda, 1981:

| 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 |

- Each of these metrics estimate the similarity (a higher metric means a higher similarity) of two samples:
  - Each sample i (i = 1,2) consists of n_ij individuals of the type j and 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_ij individuals of the cluster j (j = 0,1,2,3).
  - The similarities are estimated between the city 0 (Nizhny Novgorod) and cities i (i=1 for Kazan, i=2 for Ufa, i=3 for Samara, and i=4 for Perm) one-by-one.
- For each similarity index, the best city is Kazan (indicated by red in the table): it has the highest similarity indices in all the cases.

# Placing a restaurant in Kazan

- First, the number of food venues and non-food venues for each neighborhood in Kazan is determined.
- After that, the neighborhoods are ordered within each cluster by the number of food venues in a descending order.
- For each cluster, take only the neighborhoods, where the number of food venues are less than a predefined threshold.
- Re-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 following table.

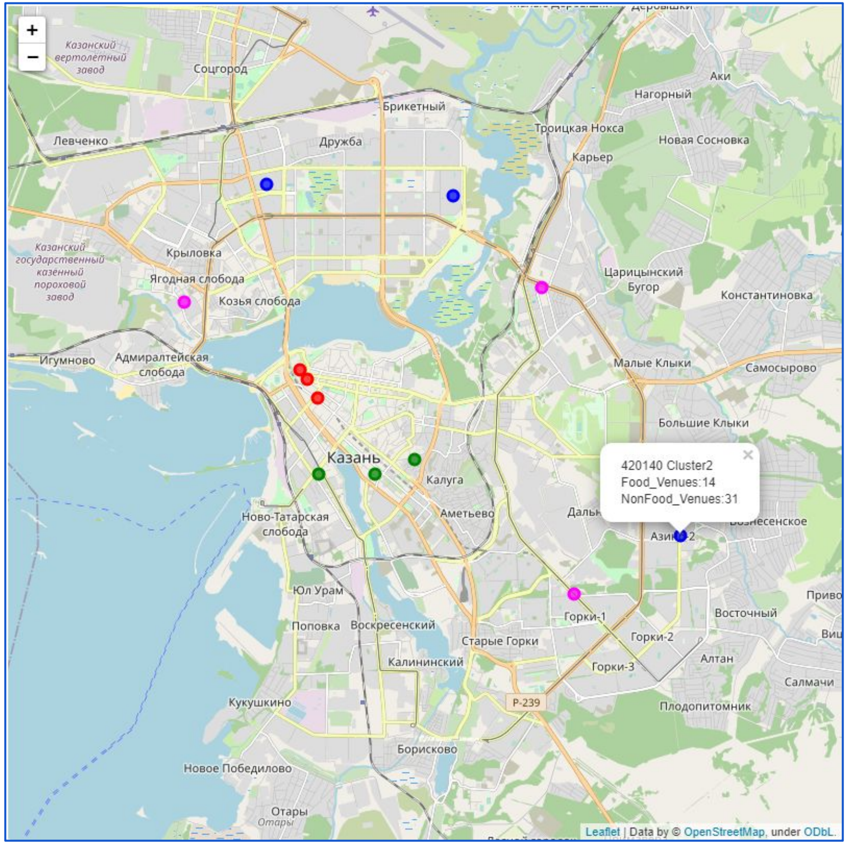
| Neighborhood | Sum of Food Venues | Sum of Non-Food Venues | Latitude | Longitude | Cluster Labels | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue |
|---|---|---|---|---|---|---|---|---|
| 420014 | 36 | 64 | 55.796878 | 49.108822 | 0 | Food | Arts & Entertainment | Nightlife 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 |

# Best places for a restaurant in *Kazan*

The clustering of Kazan is implemented in the same way as of Nizhny Novgorod: obtained clusters are similar for 2 cities.

Based on the information of the placement of successful restaurants in Nizhny Novgorod, **an optimal place for a restaurant in Kazan can be determined:**

- E.g., if a restaurant in Nizhny Novgorod placed in a neighborhood of the cluster 1 is successful, then it could be useful to consider only neighborhoods of the same cluster in Kazan.
- 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.
- Finally, 3 optimal places for each cluster are presented in order to determine the best one during a further analysis, e.g., already in the city.



# Conclusion

- 5 biggest cities of the Volga Federal District have been studied in this research.
- For each city, neighborhoods based on the postal code have been subdivided into 4 clusters:
  - neighborhoods in the old city center,
  - neighborhoods in the modern city center,
  - neighborhoods outside of the center
  - 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 these metrics, the most similar city to Nizhny Novgorod 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 really Kazan.
- Top-3 optimal places for a restaurant within each cluster in Kazan have been identified. The best location can be identified based on the information of the location of the most successful restaurants in Nizhny Novgorod.
- A stronger collaboration with the stakeholders will improve the obtained accuracy and restrict the obtained results.
- Foursquare is not well-adapted for use in Russian Federation, especially, in non-touristy cities: e.g., some important venues haven't been identified correctly (e.g., the multiplex cinema "Karo Film Rossiya" around the postal code '87036' wasn't found by Foursquare in Nizhny Novgorod)
- Using of locally adapted APIs is more preferable: e.g., Google or Yandex could provide a more detailed information about the neighborhoods and, as a consequence, a more precise analysis.