IBM – Coursera

Data Science Specialization

Capstone project - Final report

Examining Restaurants in Belgrade Neighborhoods

Jovan Njegic - 2019

Table of content:

Introduction

The objective of the research is to explore which neighborhoods in the city of Belgrade have similar preferences regarding restaurants. The results of the research could be beneficial both for restaurant customers, as well as for the potential future owners.

Data

Data for the research were obtained from the multiple sources:

- Wikipedia (https://en.wikipedia.org/wiki/List_of_Belgrade_neighbourhoods_and_suburbs)
- Geocoder API,
- FourSquare API

The names of neighborhoods of the city of Belgrade were originally scrapped from Wikipedia source. The data was filtered in order only the urban neighborhoods, and saved in csv format. In the code, data were loaded from csv by using Pandas library. This way, we managed to gather total of 225 different neighborhoods.

The geographical coordinates were obtained by using Geocoder API, and these data were combined with previously gathered neighborhoods data.

The neighborhoods of the city of Belgrade can be seen in the Figure 1.

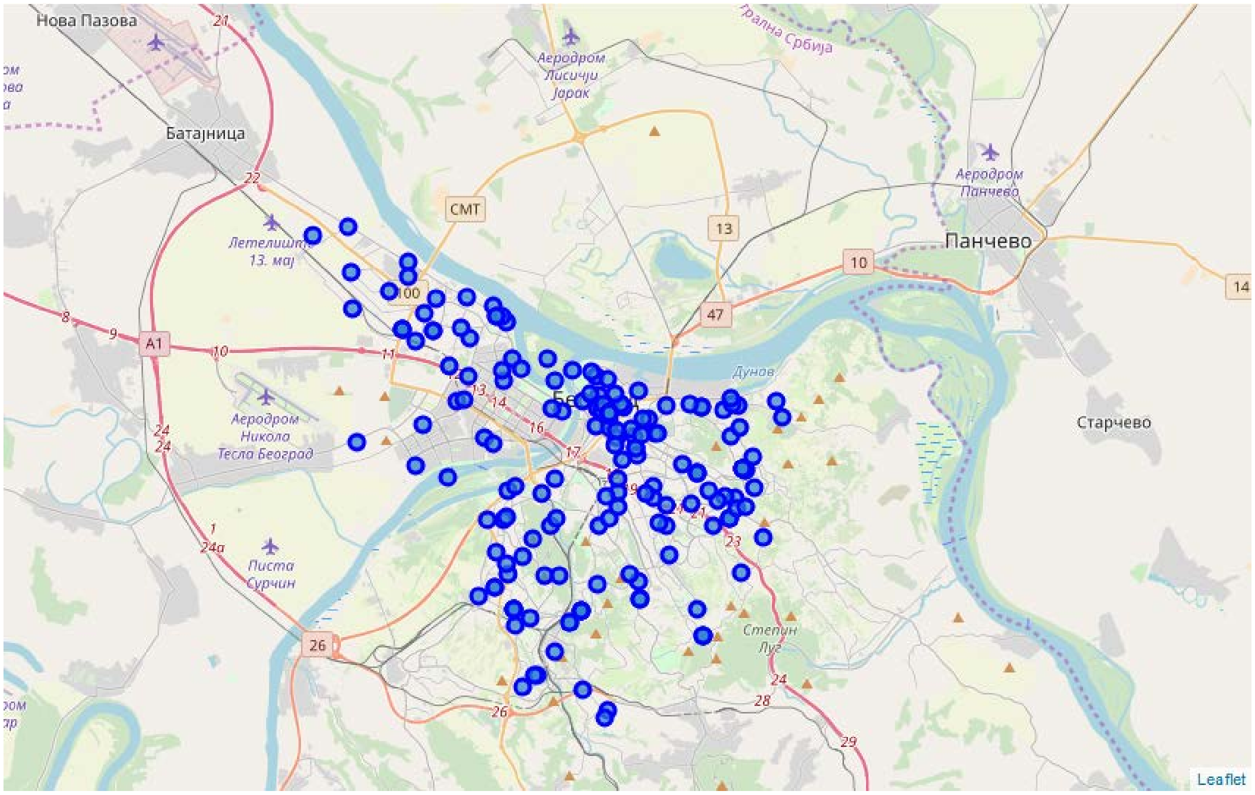


Figure 1. The neighborhoods of the city of Belgrade

Source: Wikipedia, Geocoder

Finally, we used FourSquare API in order to gather the data regarding all available venues in the observing neighborhoods. The venues data were collected with dynamically determined range, which represented a half of the Euclidean distance between the observed neighborhood and its closest neighborhood, expressed in meters. The formula for the range is:

$$R = 111120 \times \frac{1}{2} \times \sqrt{(X_n - X_c)^2 + (Y_n - Y_c)^2},$$

Where R represents Range, 111120 is approximate number of meters for one degree of latitude/longitude, X represents the latitude coordinate, and Y represents the longitude coordinate.

This way, we obtained the data regarding 1705 Belgrade venues from 206 unique categories. The data was then filtered in order to extract only venues which contain word ‘restaurant’ in a venue category name, which resulted in dataset with 21 different restaurant type.

Methodology

Since all of our data was categorical without non-available values, there was no need to use any of the usual methodologies, such as data standardization, outlier detection, non-available data handling, etc. The only preprocessing technique we used was one-hot encoding. This is one of the well-known and standard techniques for categorical data preprocessing, which represents categorical data in numerical way, in order to be applicable for the usage in machine learning models. After described preprocessing, we calculated the frequency of every restaurant type in each neighborhood.

After that, the neighborhoods were grouped by most common restaurant type, and only neighborhoods with at least one restaurant were selected.

For the cluster analysis, we applied K-means clustering method, which aims to partition neighborhoods into k clusters in which each observation belongs to the cluster with the nearest mean. Formally, the objective can be represented as:

$$\underset{S}{\operatorname{argmin}} \sum_{i=1}^{k} \sum_{x \in S_i} ||x - \mu_i||^2 = \underset{S}{\operatorname{argmin}} \sum_{i=1}^{k} |S_i| Var S_i$$

The optimal hyperparameter k for the algorithm was obtained by Silhouette score, which refers to a method of interpretation and validation of consistency within clusters of data. The highest score was for the k value of 3, as presented on the Figure 2.

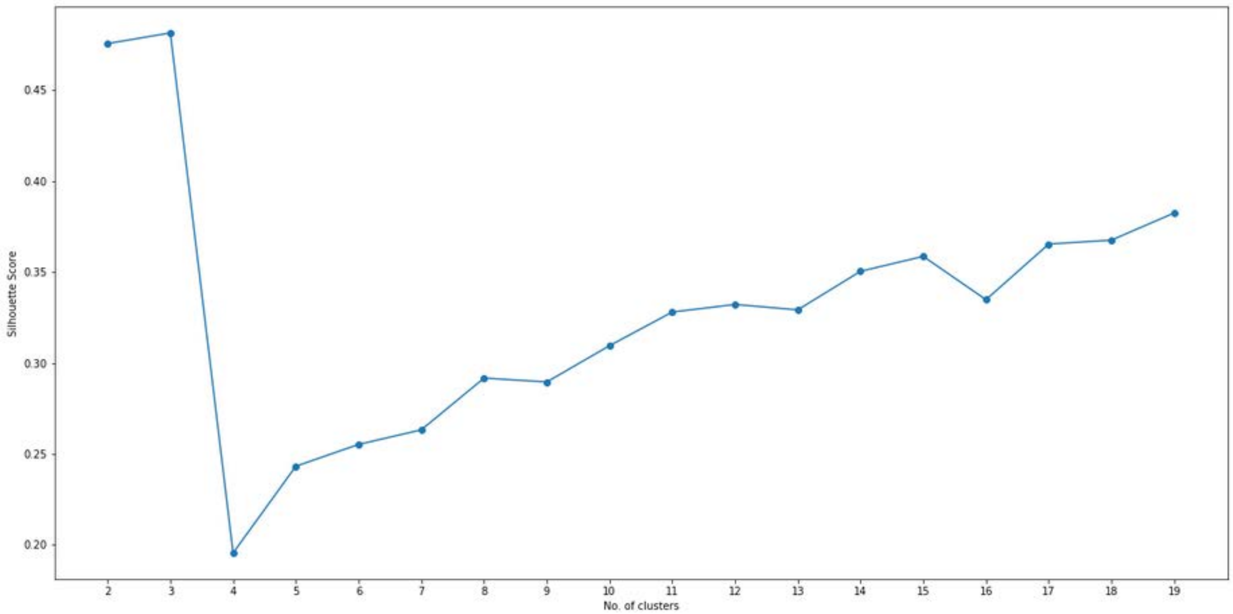
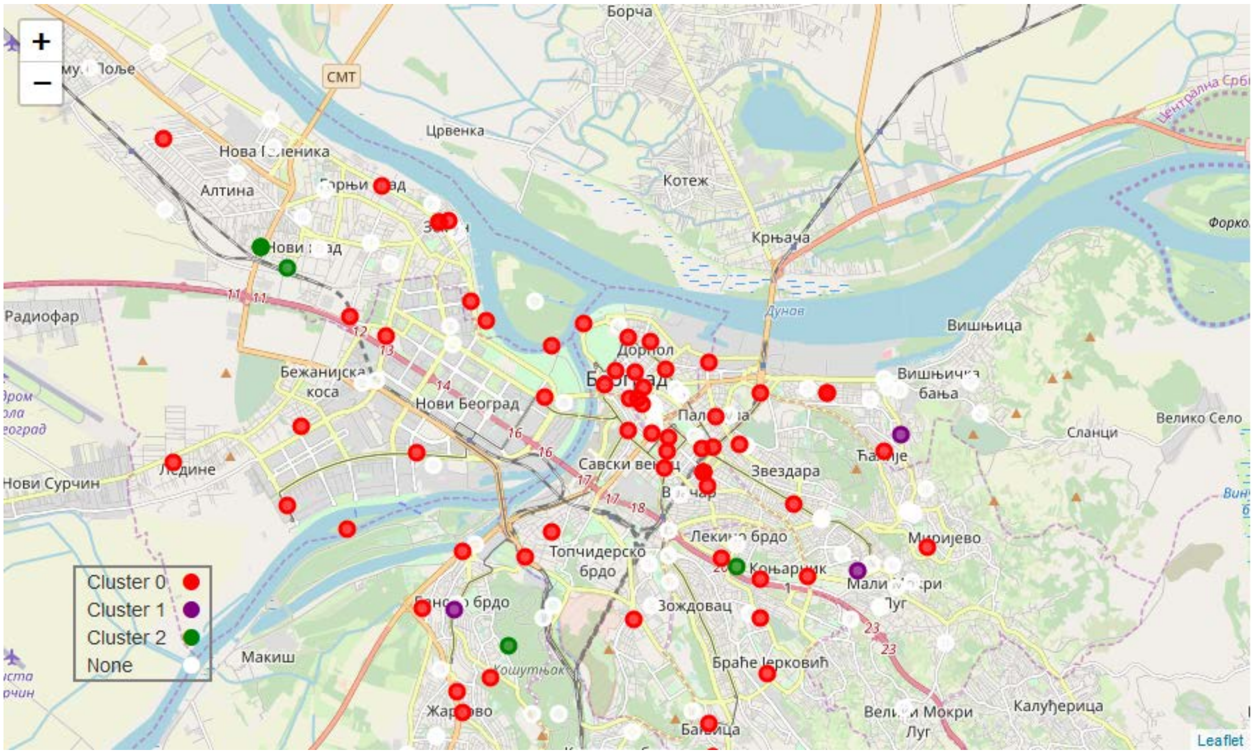


Figure 2. Silhouette score for different values of hyperparameter k

Source: Author

Results

By applying above described methodology, we managed to cluster data in three different clusters. The map of the clustered neighborhoods is presented on the Figure 3.



The obtained clusters can be seen in Figure 4.

CLUSTER 1

•••

	Neighbourhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
1	Ada Ciganlija	Seafood Restaurant	Vegetarian / Vegan Restaurant	Fast Food Restaurant
2	Ada Medjica	Italian Restaurant	Seafood Restaurant	Eastern European Restaurant
9	Banjica	Greek Restaurant	Vegetarian / Vegan Restaurant	Fast Food Restaurant
15	Bezanija	Fast Food Restaurant	Chinese Restaurant	Falafel Restaurant
23	Bogoslovija	Comfort Food Restaurant	Theme Restaurant	Vegetarian / Vegan Restaurant
...	...	...	...	...
198	Zapadni Vracar	Seafood Restaurant	Italian Restaurant	Vegetarian / Vegan Restaurant
201	Zeleni Venac	Chinese Restaurant	Theme Restaurant	Vegetarian / Vegan Restaurant
206	Zemun	Eastern European Restaurant	Theme Restaurant	Vegetarian / Vegan Restaurant
207	Zemun Backa	Eastern European Restaurant	Theme Restaurant	Vegetarian / Vegan Restaurant
209	Zemunski Kej	Mexican Restaurant	Theme Restaurant	Vegetarian / Vegan Restaurant

68 rows × 4 columns

CLUSTER 2

•••

	Neighbourhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
11	Banovo Brdo	Fast Food Restaurant	Theme Restaurant	Vegetarian / Vegan Restaurant
76	Karaburma	Fast Food Restaurant	Eastern European Restaurant	Theme Restaurant
146	Rudo	Fast Food Restaurant	Theme Restaurant	Vegetarian / Vegan Restaurant

CLUSTER 3

•••

	Neighbourhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
57	Golf Naselje	Eastern European Restaurant	Theme Restaurant	Vegetarian / Vegan Restaurant
136	Petlovo Brdo	Eastern European Restaurant	Theme Restaurant	Vegetarian / Vegan Restaurant
168	Sumice	Eastern European Restaurant	Theme Restaurant	Vegetarian / Vegan Restaurant
192	Vojni Put	Eastern European Restaurant	Theme Restaurant	Vegetarian / Vegan Restaurant
193	Vojni Put I	Eastern European Restaurant	Theme Restaurant	Vegetarian / Vegan Restaurant
194	Vojni Put II	Eastern European Restaurant	Theme Restaurant	Vegetarian / Vegan Restaurant
203	Zeleznicka Kolonija	Eastern European Restaurant	Theme Restaurant	Vegetarian / Vegan Restaurant

Figure 4. Clustered neighborhoods by most frequent restaurant types

Discussions

As it can be seen from the obtained results, most of the neighborhoods in the center of the Belgrade are populated with the variety of restaurants. There are three neighborhoods in the cluster 2, which are similar to each other by the preferences to fast food and theme restaurants. Finally, we can see seven neighborhoods in the cluster 3 which prefer eastern European cuisine.

Data presented with white circles are neighborhoods which don’t have any restaurant in the area. These spots could be interpreted as possible opportunities for the potential future restaurant owners.

Conclusion

In this project, we investigated the preferences of the neighborhoods of the city of Belgrade regarding different restaurant types. We managed to identify three different clusters of neighborhoods. First cluster indicate that central neighborhoods are populated with the wide variety of different restaurants. Second cluster managed to identify three neighborhoods with the preferences to fast food/theme restaurants combination, while the last cluster showed the neighborhoods which prefer Eastern European cuisine.

We hope that the findings of this project could help interested parties to understand restaurant preferences of different Belgrade neighborhoods. Also, these results could hopefully be beneficial for the potential future restaurant owners.

References

https://en.wikipedia.org/wiki/List_of_Belgrade_neighbourhoods_and_suburbs

https://foursquare.com/

https://pypi.org/project/geocoder/

https://en.wikipedia.org/wiki/K-means_clustering

https://en.wikipedia.org/wiki/Silhouette_(clustering)

https://www.coursera.org/learn/applied-data-science-capstone/home