Bakery shop: how to move out from Berlin and not lost yourself in Sofia, Bulgaria.

Dmitry Pankov

January 24, 2020

1. Introduction

1.1 Background

After 6 years living in Berlin my friend decided to move home to Bulgaria and open in Sofia bakery shop with national breakfast kitchen. To meet this decision were taken different factors in consideration.

Sofia is the biggest city in Bulgaria, and the 15th largest city in the European Union [1]

In the recent time the city became fast developing tourist destination, since 2009 the city has been experiencing a stable growth in tourism - approximately 13% in both 2016 and 2017, that is expected to continue in the coming years, though at a lower rate. [2]

Sofia Airport launched 42 new routes between 2015-2018, including new direct flights to such destinations as Baku, Nice, Malaga, etc. A new line to Beijing is planned for 2020. [2]

Most of the visitors (87% in 2016) coming to Sofia for leisure purpose.

Sofia is among the top 3 cities in Europe with the highest growth of international visitors. The average increase for the period between 2009 and 2016 was 9.4% according to the Mastercard Global Destination Cities Index report published in September 2017. [2]

The Daily Backpacker Index for Sofia in 2018 is 53.05 BGN or around 27.11 Euro and includes a night stay at a centrally located hostel with good reviews, transportation, and food and entertainment costs. [2]

Therefore, it is advantageous for my friend to accurately predict in which areas of the city may occurs high demand on the cafe breakfast service and in which areas the target group may be found.

1.2 Problem

The analyses of available data source can help us to determine the places where the potential consumer may be founded, and especially:

1. Tourists who wants to start a day with affordable and on good quality national breakfast

2. Freelancer who loves to have their own cafe as work place

1.3 Objective

The aim of this report is to analyses different locations in Sofia and group them into similar clusters, which can help us to:

1. Find ideal locations for cafe close to main touristic accommodations with low competition from other players

2. Find ideal locations for advertising of new café

1.4 Target Audience

The provided information would be interesting for everyone who wants to open own restaurant or cafe business, or other tourism relevant services, which would benefit of knowing the main touristic allocations and saturation of those places with relevant services.

2. Data Sources

Taking into consideration main goal of the project we can use following data's:

1. Sofia Neighborhood Data: The Wikipedia page 'Districts of Sofia' was scraped to pull out the necessary information. [3]
   The information obtained i.e. the table of postal codes was transformed into a pandas data frame for further analysis.

2. Coordinate data for each Neighborhood in Sofia. I used Nominatim to get the center coordinates of the each Neighborhood and coordinates for every required adress [4]

3. Foursquare API to get the information of restaurants, cafe and hotels venues of given neighborhoods and borough of Sofia [5]

3.Methodology

3.1. Getting names and coordinates of districts in Sofia

Name	Idle	Population	Type
Bankya, Sofia	10.4	9,186	Town
Vitosha, Sofia	3.5	42,953	Suburb
Vrabnitsa, Sofia	4.6	47,417	Urban
Vazrazhdane, Sofia	5.3	47,794	Urban
Izgrev, Sofia	3.1	33,611	Urban
Ilinden, Sofia	4.5	37,256	Urban
Iskar, Sofia	3.9	69,896	Urban
Krasna polyana, Sofia	9.2	65,442	Urban
Krasno selo, Sofia	3.7	72,302	Urban
Kremikovtsi, Sofia	5.8	23,599	Suburb
Lozenets, Sofia	3.3	45,630	Urban
Lyulin, Sofia	5.4	120,897	Urban
Mladost, Sofia	4.2	110,852	Urban
Nadezhda, Sofia	3.8	77,000	Urban
Novi Iskar, Sofia	4.5	26,544	Town
Ovcha kupel, Sofia	3.8	47,380	Urban
Oborishte, Sofia	2.8	36,000	Urban
Pancharevo, Sofia	5.3	24,342	Suburb
Poduene, Sofia	4.5	85,996	Urban
Serdika, Sofia	3.6	52,918	Urban
Slatina, Sofia	4.1	65,772	Urban
Studentski, Sofia	2.9	50,368	Urban
Sredets, Sofia	4.0	41,000	Urban
Triaditsa, Sofia	3.7	65,000	Urban
TOTAL, Sofia	4.5	1,299,155	

The data about Sofia districts and their population was processed from Wikipedia with help of Beautiful soup library, which helped to get the following data frame:

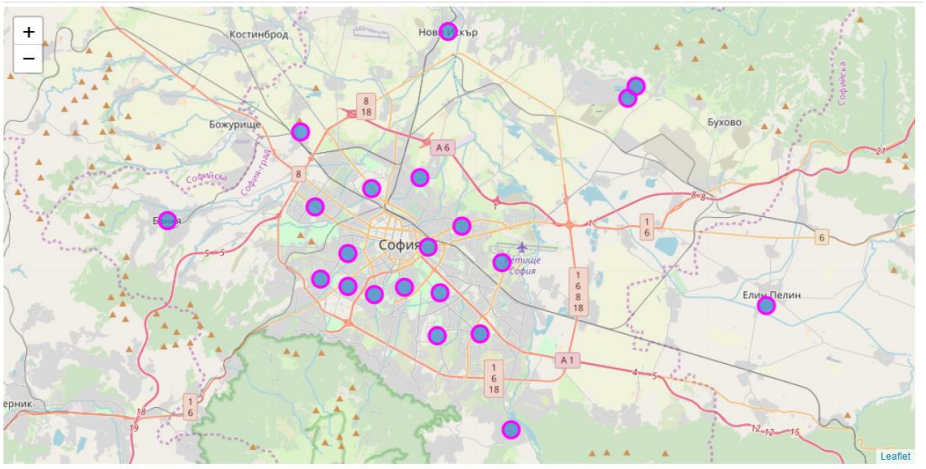
With help of Nominatim Library were founded coordinates of all Sofia Districts

Finding coordinates of Sofia districts

```
from geopy.geocoders import Nominatim
geolocator = Nominatim(user_agent="Sofia_explorer")

df['Latitude']= df['Name'].apply(geolocator.geocode, timeout=15).apply(lambda x: (x.latitude))

df['Longitude']= df['Name'].apply(geolocator.geocode, timeout=15).apply(lambda x: (x.longitude))

sofia_city_data = df.head()

sofia_city_data = df.head()
```

	Name	Idle	Population	Type	Latitude	Longitude
0	Bankya, Sofia	10.4	9,186	Town	42.710125	23.146497
1	Vitosha, Sofia	3.5	42,953	Suburb	42.560000	23.280000
2	Vrabnitsa, Sofia	4.6	47,417	Urban	42.759103	23.247337
3	Vazrazhdane, Sofia	5.3	47,794	Urban	42.777602	23.493854
4	Izgrev, Sofia	3.1	33,611	Urban	42.670481	23.351794

It helped us with leaflet to visualize them on a map:



3.2 Finding the most touristic cluster in Sofia

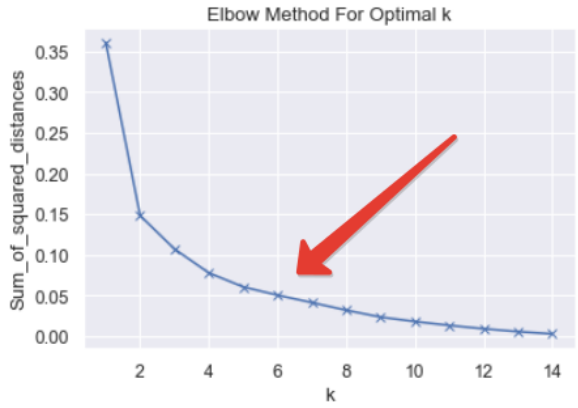
To find the area with most touristic infrastructure and similar and probably less expensive in terms of the rental prices areas we proceed cluster analysis based on the venues data's, which we have collected via Foursquare API.

The initial dataset after primary Venue extraction:

District	Dist_Latitude	Dist_Longitude	Venue	Venue_Lat	Venue_Long	Venue_Category
Bankya, Sofia	42.710125	23.146497	Тенис Клуб Банкя (Bankia Tennis Club)	42.710281	23.148693	Tennis Stadium
Bankya, Sofia	42.710125	23.146497	Централен Парк Банкя	42.707522	23.144989	Park
Bankya, Sofia	42.710125	23.146497	Плувен басейн "Здраве"	42.709876	23.146219	Pool
Bankya, Sofia	42.710125	23.146497	Ресторанта На Кортовете	42.710298	23.148785	Restaurant
Bankya, Sofia	42.710125	23.146497	Aqualand Bankya	42.706415	23.146734	Pool

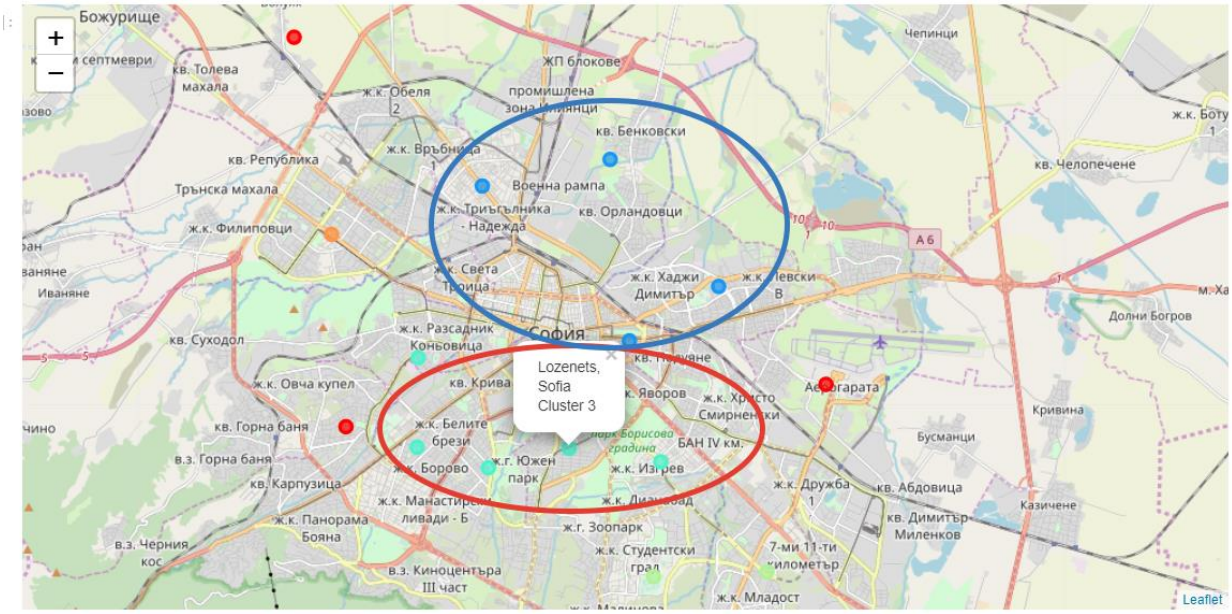
The venue data were grouped and converted in binominal format. To see the similarities between different district was selected K-Means clustering as most suitable method for simple cluster aggregation.

To find the right number of clusters was used Elbow method:



By finding compromise between high sum of squared distance and high granularity were selected 6 cluster.

K-means Algorithm had allowed to see the differences between two big central areas in Sofia, which we can call North and South, here is map visulaization:



Cluster 2 examination helped us to see that in the North Theaters, Cafes and Hotels are one of the most frequent venues, which give us right to call this cluster – Touristic North.

Cluster 3 showed different picture, here the most frequent venues were Gyms and Supermarkets, which gave us the opportunity to identify this area as living one.

Examine main clusters

```
sofia_merged.loc[sofia_merged['Cluster Labels'] == 2, sofia_merged.columns[[0] + list(range(6, sofia_merged.shape[1]))]]
```

	District	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
13	Nadezhda, Sofia	2	Bar	Park	Bulgarian Restaurant	Italian Restaurant	Dessert Shop	Hotel	Coffee Shop	Restaurant	Gym / Fitness Center	Burger Joint
16	Oborishte, Sofia	2	Park	Bar	Restaurant	Bakery	Theater	Italian Restaurant	Cupcake Shop	Coffee Shop	Café	Hotel
18	Poduene, Sofia	2	Park	Theater	Bar	Bakery	Restaurant	Café	Cupcake Shop	Yoga Studio	Cocktail Bar	Hotel
19	Serdika, Sofia	2	Bar	Park	Theater	Bakery	Hotel	Cocktail Bar	Restaurant	Café	Italian Restaurant	Dessert Shop

Cluster 3

```
sofia_merged.loc[sofia_merged['Cluster Labels'] == 3, sofia_merged.columns[[0] + list(range(6, sofia_merged.shape[1]))]]
```

	District	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
4	Izgrev, Sofia	3	Park	Restaurant	Bakery	Coffee Shop	Gym	Vegetarian / Vegan Restaurant	Bulgarian Restaurant	Cupcake Shop	Clothing Store	Pizza Place
7	Krasna polyana, Sofia	3	Park	Bakery	Coffee Shop	Italian Restaurant	Gym / Fitness Center	Bar	Vegetarian / Vegan Restaurant	Dessert Shop	Bookstore	Restaurant
8	Krasno selo, Sofia	3	Coffee Shop	Gym / Fitness Center	Bakery	Park	Italian Restaurant	Café	Gym	Dessert Shop	Supermarket	Cosmetics Shop
10	Lozenets, Sofia	3	Park	Bakery	Italian Restaurant	Restaurant	Coffee Shop	Theater	Dessert Shop	Ice Cream Shop	Vegetarian / Vegan Restaurant	Gym / Fitness Center
23	Triaditsa, Sofia	3	Bakery	Restaurant	Coffee Shop	Park	Gym / Fitness Center	Italian Restaurant	Cheese Shop	Dessert Shop	Vegetarian / Vegan Restaurant	Modern European Restaurant

3.3 Finding the best district in the cluster 2 for café

The last question in our research was – finding right district for our bakery shop, as the target group should be tourists, we check in which district of North cluster is highest number of cafes:

And it was Nadezhda, analyses of competitions showed us that in Nadezhda the number of Cafes – competitors the lowest one, which gave us answer on the question where to place our Café – in Nadezhda, which means Hope in Bulgarian!

4. Results

The result of the exploratory data analysis and clustering are summarized below:

- The number of Hotels and Hostels in the Sofia is pretty low, followed with low saturation of cafes it gives the good opportunities to start a business
- Cluster Analyses shows that the central area of Sofia can be divided on the art of venues on two parts - "Touristic Nord", which has high saturation on hotels and theaters, and "Living South", where are concentrated recreation activities like Gym, Cafes, Supermarkets
- For cafe is the best idea to find location in cluster North.
- Analyzing the competitions of cafes in different districts of North cluster, most profitable looks Nadezhda district, where we can find 5 Hotels with only one Cafe.
- Thanks analyze we got a list of all touristic venues, which can be suitable for online advertising.

5. Discussion

According to this analysis, Nadezhda district seems to be one of the most promising area, with high frequency on typical touristic attractions and low number of breakfast infrastructure.

But we need to take into consideration that the clustering was completely based on the most common venues obtained from Foursquare data, which numbers seems to be a bit outdated and unrealistically low. The additional verification via 3th party sources would be beneficial.

Also I didn't took into analysis the rental prices in different districts, which can play a role.

The analysis itself lacks on the granularity, the better idea seems the dividing the city with the artificial grid like squares 200*200 meters and proceeding cluster analysis on their base, it can give better understanding of areas. But with number of valuable data's in Foursquare it doesn't make so much sense, pother sources are here required.

6. Conclusion

Finally finishing this project.

I got the idea how data-science projects should look like, got in my hand great data science toolkit with powerful libraries and technic.

The project changed my perception how the obvious questions could be solved more efficient way, and the results can drastically vary from initial expectations.

[1] https://en.wikipedia.org/wiki/List_of_cities_in_the_European_Union_by_population_within_city_limits
[2] https://investsofia.com/wp-content/uploads/2018/09/Sofia-Tourism-and-Air-Transport-Market-September-2018.pdf
[3] https://en.wikipedia.org/wiki/Districts_of_Sofia
[4] http://nominatim.org/
[5] Forsquare API - https://developer.foursquare.com/docs/api