FOTO: Silbaski, Old Belgrade Panorama From The Building "Ušće"

Where to open new business in Belgrade?

Data Science show-case project: investigating Belgrade neighborhoods and venues

The Battle of Neighborhoods Capstone Project
by Goran D. Stevanovic
11-May-2020

Table of Contents

Executive summary

Belgrade is 1.6 million city, capital of Serbia, with multiple urban municipalities (boroughs) and suburban municipalities. Each urban borough has multiple neighborhoods.
In recent years Belgrade is developing very rapidly, where some new neighborhoods are emerging and gaining popularity, while some traditional neighborhoods are being re-shaped. This obviously creates potential gaps, but also an opportunity. Inadequate accessibility of certain venues in specific neighborhoods is potential opportunity for investors to consider opening new, or extending existing business into those neighborhoods.

This Data Science show-case project will inspect current status of Belgrade neighborhoods and venues, and make recommendation for potential investors.

Introduction

Business Understanding

Business Goal is to define in which neighborhoods there is a potential for opening new business, i.e. new venue. For this purposes, combination of statistical data (i.e. average income, real-estate values) and geolocation and venue data (i.e. Foursquare data) will be used to develop the analysis of the most promising locations for opening new business.

Business Objective of the business owner, opening new, or extending existing network of the business, is to locate the most suitable location for opening its business venue. Equipped with the info about potential neighborhoods of interests, stakeholders will be able to plan the expansion of their business in Belgrade.

The focus is businesses/venues primarily relevant to neighborhood residents i.e. gyms, fitness, kids playgrounds, pharmacies, beauty salons, grocery stores etc. However, for evaluating potential for other venue categories, like shop, shopping malls, restaurants, café bars etc. other input data would need to be used for analysis – instead of leaving desirability other aspects of business desirability, accessibility and commuting traffic should be considered, which is not in scope of this particular analysis.
Please note, this analysis is Showcase Data Science Project, built with Open Source tools, and built only on feely available data. Communalization of this (or similar) project, should include more accurate and up-to-date data sources, which could imply certain costs.

Analytical Approach

To address Business Problem, classification model will be used.
Neighborhoods will be classified and outcome of those classifications will be discussed.

- Publicly available statistical data about employees' gross income and real-estate average price will be to determine current living desirability of each neighborhoods, classifying neighborhoods accordingly.
- Beside the statistical data, this classification will also be supported with Foursquare data, listing most common venue categories available in each neighborhoods. Evaluation will be used with options to use only statistical data, to use only venue categories data, or to both statistical and venue categories data in each neighborhood. Outcome and precision of each approach will be discussed.
- After delivering the satisfactory classification, Foursquare data will be used to count the specific venue groups for each neighborhood – enabling an identification of neighborhoods, within the same classification, with bellow average access to specific venue categories
- If residents of the neighborhood have below average access to specific venues categories, this is a potential opportunity to start such business in that neighborhood.

Constraints

This assignment is created as part of the Capstone Project – The Battle of Neighborhoods, and as such it must:

1. Leverage the Foursquare location data,
2. Explore or compare neighborhoods or cities or come up with a problem that can use the Foursquare location data to solve

Data

Data Requirements

The current living desirability will be primarily determined based on financial and market data.

1. Belgrade employees' average gross income per municipalities, in RSD. This information is publicly available either from the State Statistical Office (https://www.stat.gov.rs/sr-Latn/oblasti/trziste-rada/zarade) or from the official City of Belgrade publication web portal (https://zis.beograd.gov.rs/index.php/2013-12-03-10-50-11/2013-11-04-10-15-34/finish/4-z-p-sl-n-s-i-z-r-d/113-pr-s-cn-z-r-d-p-z-p-sl-n-s-p-b-r-2019.html). The State Statistical office makes regular vast number of publications and/or provide statistical data on demand, but neither there is public (free) available API, nor is this information available on web pages for html stripping. The City of Belgrade provide necessary employees' average gross income per Belgrade municipalities. This data be used. However this data is publicly available only in PDF format, there is public (free) available API, so data will be manually parsed into CSV table.
2. Belgrade real-estate average price per municipalities and respective neighborhoods, in EUR per square meter. This information is publicly available on several web portals. The most comprehensive list is available on Imovina.net (https://imovina.net/statistiika_cena_nekretnina/?viewType=table&category=1&city=1&year=2020&stats=699&load=PRIPREMI). This data will be used. However data is provided as query result in JavaScript, no html tripping of data is possible, there is no public (free) available API to use, so data will be manually parsed into CSV table. </br>

Data about venues categories is collected from Foursquare.

1. In order to utilize Foursquare, geolocation (Latitude and Longitude) for Belgrade neighborhoods are extracted using google. Note: considering there is no unified list of Belgrade neighborhoods, instead of using geopy library, geolocation (Latitude and Longitude) is extracted manually from google and prepared in CSV table. This is one time activity which can be re-used for all future analysis.
2. Foursquare application API is used to gather venues and venues categories which will be used both in the classification, and discussion of the outcomes to count the specific venue groups for each neighborhood – enabling an identification of neighborhoods, within the same classification, with bellow average access to specific venue categories It should be noted, there is no official list of all neighborhoods. Various real-estate web portals provides different lists, which included additional neighborhoods, other than those traditional neighborhoods.

Data Collection

Data is collected from the following Data Sources

- Belgrade employees' average gross income per municipalities is provided as CSV table on the level of municipalities.
- Belgrade real-estate average price is available per municipalities and their neighborhoods is provided as CSV table on the level of neighborhoods.
- Geolocation (Latitude and Longitude) is provided on the level of traditional neighborhoods.
- Foursquare venue data is collected for each geolocation of neighborhoods

Additional, data sources to be considered.

- For determining the living desirability of the particular neighborhood. The State Tax Administration data can be used as complementary data. Taxation is based on several criteria, including calculation of real-estate average price, as well as city zone predefined by city regulations. At the initial stage of this exercise, this data will not be used, as data is not refreshed frequently and cannot caught market trends.
- In order to support better classification of neighborhoods, population density can be used. At the initial stage of this exercise, this data will not be used, as data is available only per municipalities where some municipalities are urban, some suburban, and some mixture. Determining relevance and normalizing population density for the mixed urban/suburban municipality would require additional data sources.

Data Understanding

All collected data will be prepared and cleaned:

- Out of 17 municipalities in Belgrade only 6 are fully urban/integrated municipalities; 4 are partially urban/integrated and partially suburban; 6 are fully suburban with both urban (incl. municipal seats) and rural settlements; and 1 is fully suburban only rural. Using real-estate data per neighborhood will support more accurate neighborhood classifications, as obviously not all neighborhoods in one municipality should have the same treatment. Combined neighborhoods DataFrame with employees' average gross income, real-estate average price and geolocation will be created from the three data sources:
    - Employees' average gross income is available only per municipalities, municipal average will be applied for all neighborhoods within the municipality.
    - Suburban municipalities may not have associated neighborhoods
        - If in the combined neighborhoods DataFrame row, municipality does not have neighborhood, then neighborhood = municipality.
    - Geolocation is available only for traditional neighborhoods (and/or larger landmarks). Data for neighborhoods from real-estate data source is matched and paired with geolocation of neighborhoods. Rows with micro-locations (neighborhoods from real-estate data source) which do not have geolocations, are omitted.
        - If in the combined neighborhoods DataFrame row, there is no geolocation, then row is removed.
    - Foursquare venue data is matched with the combined neighborhoods DataFrame.
- Venue categories are "cleaned" and simplified:
    - All venue categories with expression like "restaurant" are combined into single venue category "restaurant". Similar approach is used for "coffie" and "café", "gym" and "fitness" etc.
    - Venue categories are used to calculate most common venue categories for each neighborhood. This data will be used to support the classification of the neighborhoods.
    - Venues categories count for each neighborhood will be used for the discussion

Methodology

Tools

Project is delivered on the IBM Watson Studio on the IBM Cloud platform. Main working tool is Jupyter Notebook with Python 3.6. Kernel.

For data manipulation, analysis, wrangling and vectored calculation, *Pandas* (incl. *Json*) and *NumPy* and *Requests* libraries, are used. For geolocations *Nominatim* from *GeoPy Geocoders* is used. For maps and plotting, *Folium* and *Matpltlib* libraries are used.

Fur unsupervised Machine Learning *K-Means* from *Sklearn Cluster* library is used.

Since IBM Cloud data storage is used, specific libraries for loading source data files are used as well; like *IBM_Boto3* and *Config* from *Botocore*.

Access to the Foursquare venue data is provided through free developer's API.

Data files used are in EASTERN EUROPEAN encoding, formatted in UTF-8 encoding.

Data processing

First data source is Belgrade employees' average gross income per municipalities in RSD (sample data provided for February 2020).

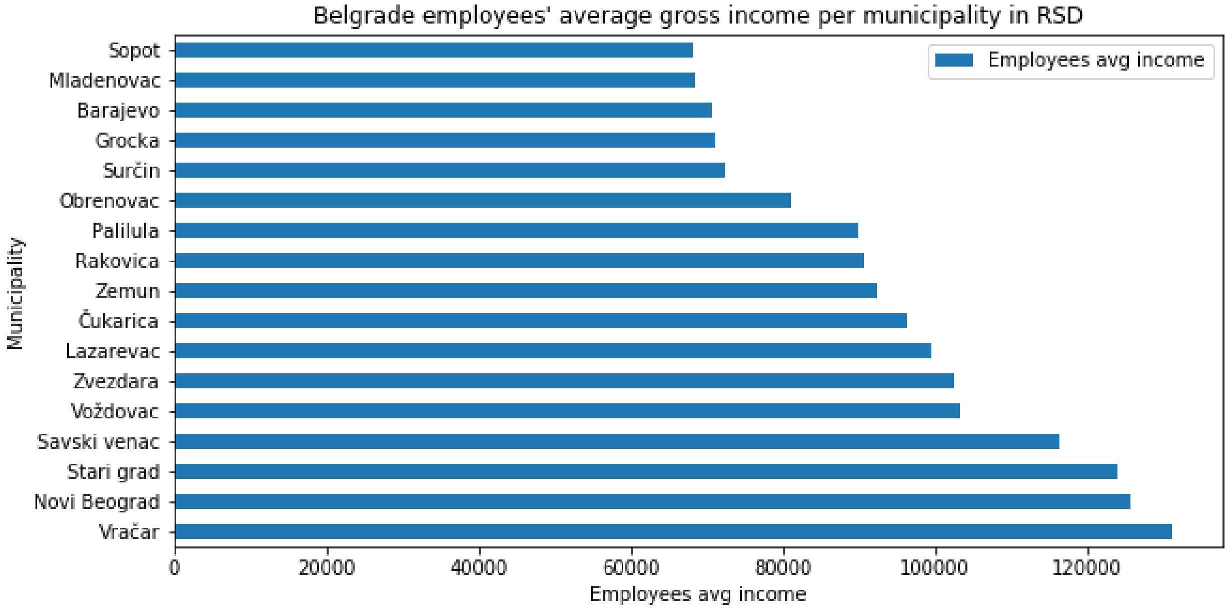
There are 17 records. Here are the first few rows:

	Municipality	Classification	Employees avg income
0	Barajevo	rural	70603
1	Voždovac	mostly urban/integrated	103300
2	Vračar	urban/integrated	131100
3	Grocka	suburban/nonintegrated	71142
4	Zvezdara	mostly urban/integrated	102383
	…		

In Belgrade, statistically there are:

- 6 municipalities which are fully urban and integrated into the Belgrade as a settlement (New Belgrade, Rakovica, Stari Grad, Savski Venac, Zvezdara, Vračar)
- 4 municipalities which are partially urban/integrated and partially suburban, including rural areas (Palilula, Zemun, Voždovac, Čukarica)
- 6 municipalities which are suburban, with both urban (usually municipal seats) and rural settlements (Obrenovac, Lazarevac, Sopot, Grocka, Mladenovac, Surčin)
- 1 municipality which is suburban and fully rural (Barajevo)

Following bar chart demonstrate clear discrepancy in the employees' average gross income throughout different municipality, Gross income is higher in the urban then in suburban municipalities, with highest income in the most central city municipalities like Vračar and Stari Grad.



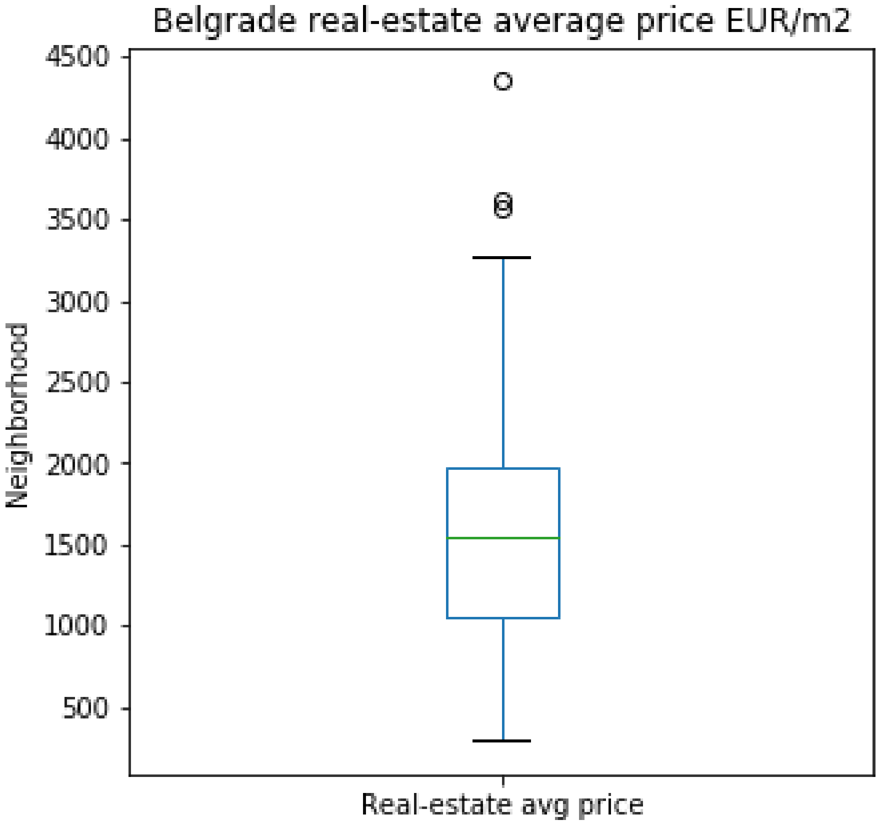
Next data source to consume is Belgrade real-estate average price per municipalities and respective neighborhoods, in EUR per square meter (sample data provided for February 2020).

There are 149 records. Here are the first few rows:

	Municipality	Neighborhood	Location size	Real-estate avg price
0	Savski venac	Belgrade Waterfront	2	3569
1	Savski venac	Zeleni venac	2	2357
2	Savski venac	Savski Trg	1	2188
3	Savski venac	Ekonomski Fakultet	1	2100
4	Savski venac	Palata pravde	1	2043
	…			

Following box plot demonstrate the overall distribution of the real-estate average price. Several observations can be immediately made from the plot bellow:

1. The minimum price is around **300** (min), maximum outlier around **4,500** (max), and median price is around **1,500** EUR/m2 (median).
2. 25% of the real-estate prices are below **1,000** (First quartile).
3. 75% of the real-estate prices are below **2,000** (Third quartile).



Geolocation (Latitude and Longitude) for Belgrade neighborhoods are extracted using *GeoPy* library.

Note: Considering there is no unified list of Belgrade neighborhoods, instead of using *GeoPy* library, geolocation (Latitude and Longitude) is extracted manually from google and prepared in CSV table. This is one time activity which can be re-used for all future analysis.

This data source has 124 records. Here are the first few rows:

	Neighborhood	Latitude	Longitude
0	A blok	44.806583	20.401885
1	Andrićev venac	44.810467	20.462996
2	Arena	44.814569	20.421348
3	Belgrade Waterfront	44.804835	20.448303
4	Belvil	44.805555	20.407605
	…		

Combined neighborhoods DataFrame with employees’ average gross income, real-estate average price and geolocation is created from the three data sources:

- Employees’ average gross income is available only per municipalities, municipal average will be applied for all neighborhoods within the municipality.
- Some municipalities may not have associated neighborhoods
    - If in the combined neighborhoods DataFrame row, municipality does not have neighborhood, then neighborhood = municipality.
- Geolocation is available only for traditional neighborhoods (and/or larger landmarks). Data for neighborhoods from real-estate data source is matched and paired with geolocation of neighborhoods. Rows with microlocations (neighborhoods from real-estate data source) which do not have geolocations, are omitted.
    - If in the combined neighborhoods DataFrame row, there is no geolocation, then row is removed.

Location size of each neighborhood is used to set radius, around which Foursquare venues will be searched.

- Standard urban neighborhood has the default neighborhood size, and radius is set to 500m.
- Microlocation is a small urban neighborhood, radius is 200m,
- Large neighborhood size is usually in the outskirts of the city area, radius is 1,000m
- Very large neighborhood size is usually suburban settlement or separate town outside of the main city area, but within metropolitan city area, radius is set to 1,500m

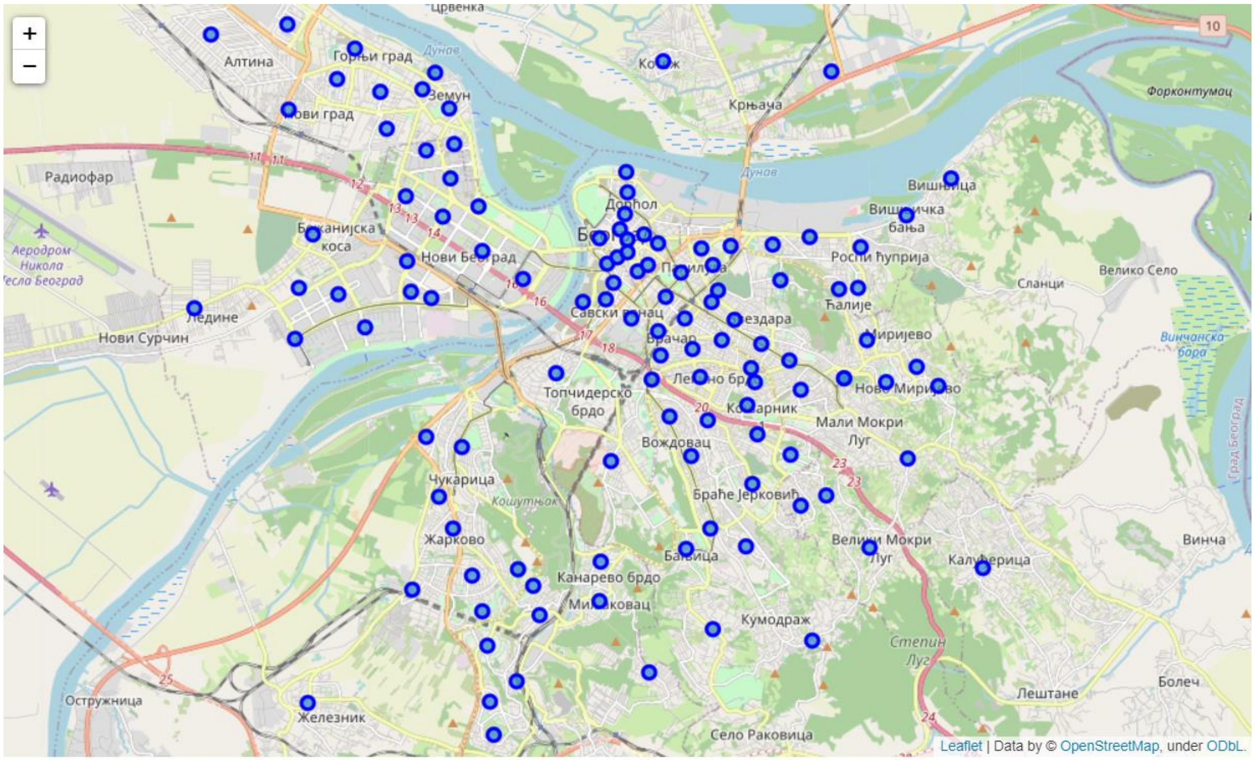
Geolocator from *GeoPy* library uses address as input parameter, and returns geolocation.

```
address = 'Belgrade, Serbia'
geolocator = Nominatim(user_agent="beograd_explorer")
location = geolocator.geocode(address)
latitude = location.latitude
longitude = location.longitude
```

The geolocation (latitude, longitude) of Belgrade is 44.8178131, 20.4568974. Using this information, we can use *Folium* to plot map of Belgrade with neighborhoods superimposed on top.

```
# create map
Beograd_Map = folium.Map(location=[latitude, longitude], zoom_start=12)

# add markers to map
for lat, lng, municipl, neigh in zip(beo2['Latitude'], beo2['Longitude'], beo2['Municipality'], b
eo2['Neighborhood']):
    label = '{}, {}'.format(neigh, municipl)
    label = folium.Popup(label, parse_html=True)
    folium.CircleMarker(
        [lat, lng],
        radius=5,
        popup=label,
        color='blue',
        fill=True,
        fill_color='#3186cc',
        fill_opacity=0.7,
        parse_html=False).add_to(Beograd_Map)
```

Beograd map with neighborhoods:



Leaflet | Data by © OpenStreetMap, under ODbL.

For each neighborhood, using its geolocation and radius, we are calling Foursquare API to return all nearby venues. Foursquare API responds in Json, which we convert into DataFrame for further processing. This an example of the user defined function used:

```
def getNearbyVenues(neighborhood, latitude, longitude, radius, LIMIT):
    venues_list=[]
    for neigh, lat, lng, rad in zip(neighborhood, latitude, longitude, radius):
        # create the API request URL
        url = 'https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&l
l={},{}&radius={}&limit={}'.format(
            CLIENT_ID,
            CLIENT_SECRET,
            VERSION,
            lat,
            lng,
            rad,
            LIMIT)

        try:
            # make the GET request
            results = requests.get(url).json()["response"]['groups'][0]['items']

            # return only relevant information for each nearby venue
            venues_list.append([(
                neigh,
                lat,
                lng,
                v['venue']['name'],
                v['venue']['location']['lat'],
                v['venue']['location']['lng'],
                v['venue']['categories'][0]['name']) for v in results])
        except:
            continue
    nearby_venues = pd.DataFrame([item for venue_list in venues_list for item in venue_list])
    nearby_venues.columns = ['Neighborhood',
                  'Neighborhood Latitude',
                  'Neighborhood Longitude',
                  'Venue',
                  'Venue Latitude',
                  'Venue Longitude',
                  'Venue Category']

    return(nearby_venues)
```

For this particular functional call, Foursquare is returning 3,115 venues, with 212 unique venue categories.

Based on the inspection of the data, for this analysis we can merge/consolidate[1] similar venue categories:

- Restaurants
- Pubs and Bars
- Coffee Shops
- Gym, Fitness, Yoga and Pilates
- Movie Theater and Multiplex
- Kids Indoor Play Areas
- etc.

After the consolidation, there are 162 unique venue categories. In order to execute classification and statistical methods, we must transposed venue categories into individual columns, where for each individual venue (total 3,115) there will be a separate row, where column of the particular venue category (in which that venue belongs) is marked `True` and all other columns are marked `False`, This is called One Hot Encoding. Transposed DataFrame is grouped by taking the mean of the frequency of occurrence of each venue category.

```
# one hot encoding
newDataFrame = pd.get_dummies(oldDataFrame['Venue Category'], prefix="", prefix_sep="")
newDataFrame = newDataFrame.groupby('Neighborhood').mean().reset_index()
```

Statistical data about income and real-estate price is also normalized using min-max normalization. Basically all numerical values are represented in a range of 0 to 1, where min numerical value in the array becomes 0, and max becomes is 1. All others are between 0 and 1, in a proportion compared to min and max value:

```
# normalize statistical data (min-max)
newDataFrame['Real-estate avg price']=(
    newDataFrame['Real-estate avg price']-newDataFrame['Real-estate avg price'].min())/(
    newDataFrame['Real-estate avg price'].max()-newDataFrame['Real-estate avg price'].min())
newDataFrame['Employees avg income']=(
    newDataFrame['Employees avg income']-newDataFrame['Employees avg income'].min())/(
    newDataFrame['Employees avg income'].max()-newDataFrame['Employees avg income'].min())
```

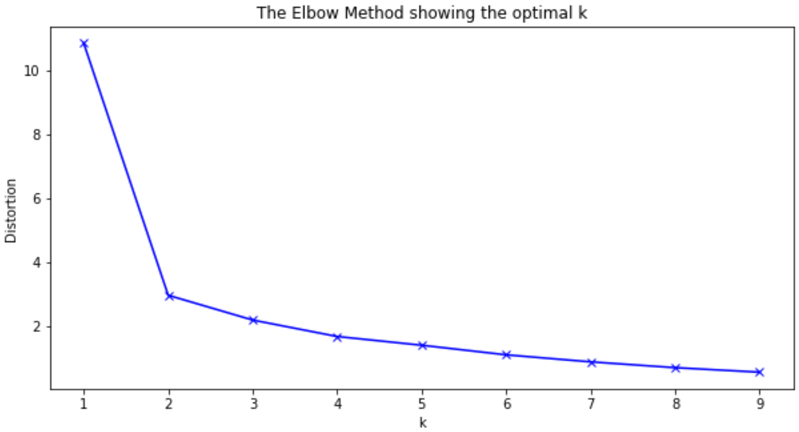
[1] More (or less) merging and consolidation can be applied, depending of the scope of the analysis.

Clustering

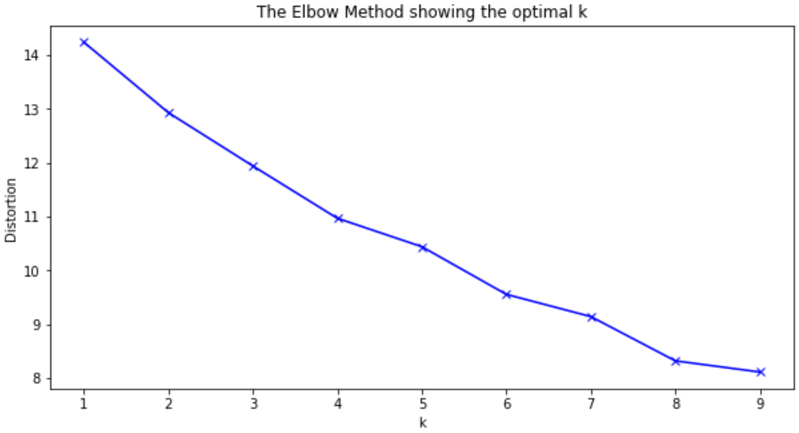
For unsupervised *K-Means* clustering, we have evaluated evaluate options to use only statistical data, to use only venue categories data, or to both statistical and venue categories data in each neighborhood.

- Running K-Means with a range of k 1 to 10.
- Plotting the distortions of K-Means

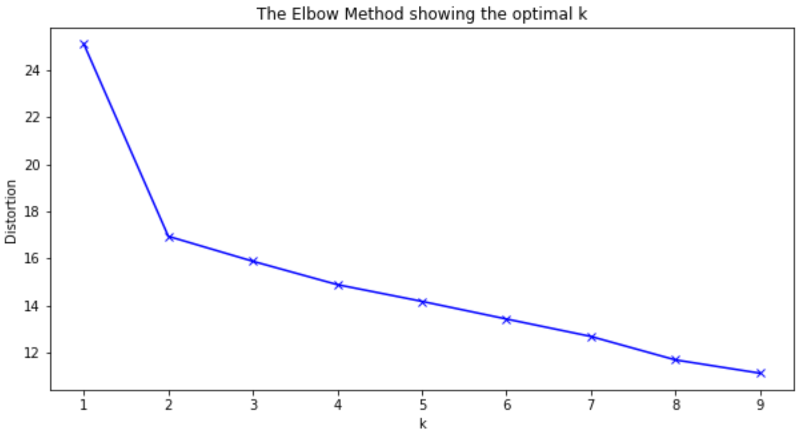
When using only statistical data, Elbow method is suggesting k=2 is the most prominent k.



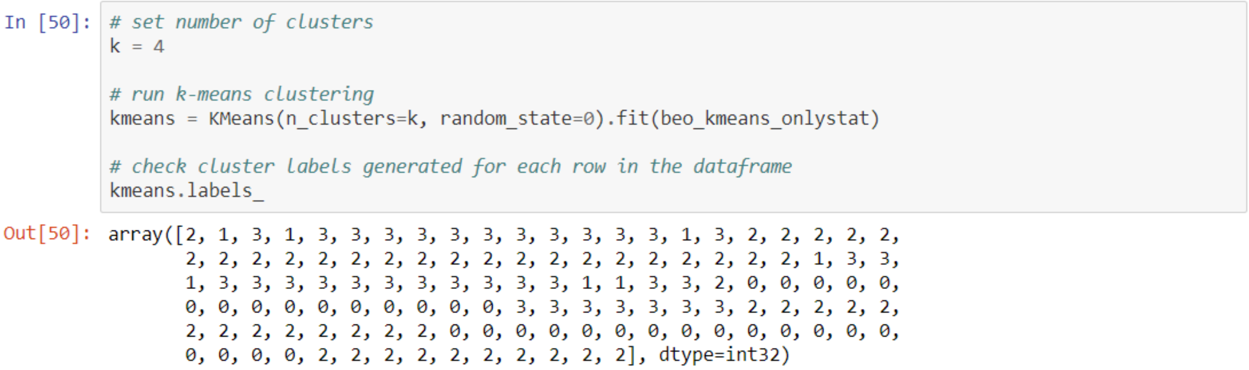
When using only venue category data, it is obvious there is no specific k with satisfactorily results. The reason is that, Foursquare is returning different venues and trending venues depending of the particular time, inquiry is being made on. Furthermore, due to the usage of the resources, the limit is set to only 100 venues per location. Therefore, depending of the time of the inquiry, and number of returned venues, this curve might look different, and in reality may support the classification.



When using both statistical and venue category data, due to non-specific Elbow in venue categories, we are not improving k.



For this analysis at this run, only statistical data is used, with k=4.



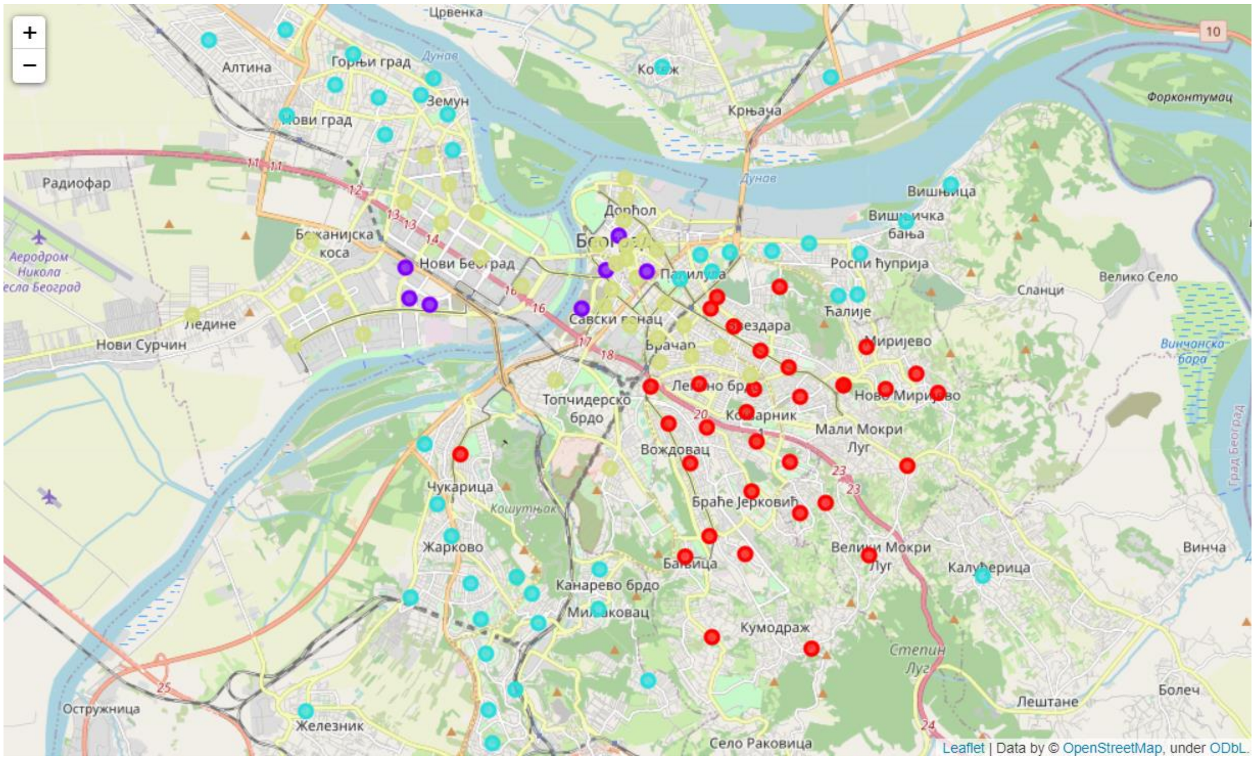
```
In [50]: # set number of clusters
         k = 4

         # run k-means clustering
         kmeans = KMeans(n_clusters=k, random_state=0).fit(beo_kmeans_onlystat)

         # check cluster labels generated for each row in the dataframe
         kmeans.labels_

Out[50]: array([2, 1, 3, 1, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 1, 3, 2, 2, 2, 2, 2,
                2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 1, 3, 3,
                1, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 1, 1, 3, 3, 2, 0, 0, 0, 0, 0,
                0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3, 3, 3, 3, 3, 3, 3, 2, 2, 2, 2, 2,
                2, 2, 2, 2, 2, 2, 2, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
                0, 0, 0, 0, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2], dtype=int32)
```

Exploring

For each neighborhood, we have identified 5 most common venue categories. Here are first few record:

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	A blok	Coffee Shop	Bistro	Restaurant	Bank	Zoo
1	Altina	Restaurant	Soccer Field	Gym and Recreation	Donut Shop	Food Court
2	Andrićev venac	Coffee Shop	Restaurant	Hotel	Hostel	Pub and Bar
3	Arena	Restaurant	Coffee Shop	Pub and Bar	Gym and Recreation	Pizza Place
4	Autokomanda	Restaurant	Indoor Play Area	Fried Chicken Joint	Butcher	Drugstore
	…					

By assigning Cluster Label to each neighborhood, we can further explore the clusters:



Results

In which neighborhoods to start new business?

In order to respond on key Business Problem; which neighborhoods has potential for additional venues, we will select several venue categories whose business is relevant to the neighborhood residents[2]:

- Gym and Recreation - incl. also Fitness, Yoga, Pilates studios facilities etc.,
- Indoor Play Area – incl. Indoor Kids Playgrounds,
- Pub and Bar - incl. all specific types of Pubs and Bars,
- Grocery and Supermarket,
- etc.

Analysis is performed on the statistically normalized venue categories counts, taking into consideration that different neighborhoods have different sizes, and taking into consideration average (mean) value of the venue category per each Cluster Label.

	Municipality	Neighborhood	Latitude	Longitude	Radius	Gym and Recreation	Indoor Play Area	Pub and Bar	Grocery and Supermarket	Coffee Shop
0	Grocka	Kaluđerica	44.753719	20.555951	1500	0.356364	0.000000	0.000000	0.312102	0.167808
1	Novi Beograd	A blok	44.806583	20.401885	200	0.000000	0.000000	0.000000	0.000000	1.647059
2	Novi Beograd	Arena	44.814569	20.421348	500	1.296296	0.000000	0.731070	0.000000	1.624473
3	Novi Beograd	Belvil	44.805555	20.407605	200	0.000000	0.000000	0.000000	3.888889	1.647059
4	Novi Beograd	Bežanijski blokovi	44.806209	20.382241	500	1.944444	0.000000	0.182768	1.944444	0.886076
	…									

Where to open a new venue?

So, what are the most prominent locations (i.e. neighborhoods) where one can start new or extend existing business in Belgrade?

- Any location (i.e. neighborhood) which have below average number of venues (< 1) is potentially good location.
- We select locations which are well below average number of venues (< 0.7)

[2] Some venue categories might be excluded, or others can be included where applicable, but in this analysis only this sample will be used.

In this analysis, we will focus and investigate potential opportunities for **Gym and Recreation**[3] venues. Similar approach can be used for other venue categories. Here is the list of records, where **Gym and Recreation** factor is below 0.7:

	Municipality	**Neighborhood**	**Latitude**	**Longitude**	**Gym and Recreation**
0	Novi Beograd	A blok	44.806583	20.401885	0.000000
1	Novi Beograd	Belvil	44.805555	20.407605	0.000000
2	Novi Beograd	Ledine	44.803522	20.343482	0.000000
3	Novi Beograd	Sava centar	44.809021	20.432033	0.000000
4	Novi Beograd	Stara Bežanija	44.807301	20.371678	0.000000
5	Novi Beograd	West 65 Belgrade	44.812632	20.401026	0.000000
6	Novi Beograd	YUBC	44.822919	20.420086	0.000000
7	Palilula	Borča	44.876808	20.430449	0.000000
8	Palilula	Karaburma II	44.807359	20.522370	0.000000
9	Palilula	Kotež	44.850879	20.469875	0.000000
10	Palilula	Krnjača	44.848694	20.515100	0.000000
11	Palilula	Višnjica	44.828398	20.547663	0.000000
12	Palilula	Višnjička banja	44.821373	20.535590	0.000000
13	Palilula	Ćalije	44.807055	20.517374	0.000000
14	Rakovica	Kneževac	44.732202	20.430429	0.000000
15	Rakovica	Labudovo Brdo	44.728267	20.423394	0.000000
16	Rakovica	Miljakovac 3	44.733781	20.466226	0.000000
17	Rakovica	Rakovica	44.744847	20.436530	0.000000
18	Rakovica	Resnik	44.710875	20.458669	0.000000
19	Savski venac	Belgrade Waterfront	44.804835	20.448303	0.000000
20	Savski venac	Dedinje	44.774353	20.455856	0.000000
21	Savski venac	Palata pravde	44.805280	20.454536	0.000000
22	Savski venac	Savamala	44.812035	20.454944	0.000000
23	Savski venac	Savski Trg	44.808369	20.456420	0.000000

[3] All above selection is subject to changes, deepening of the actual scope of the analysis.

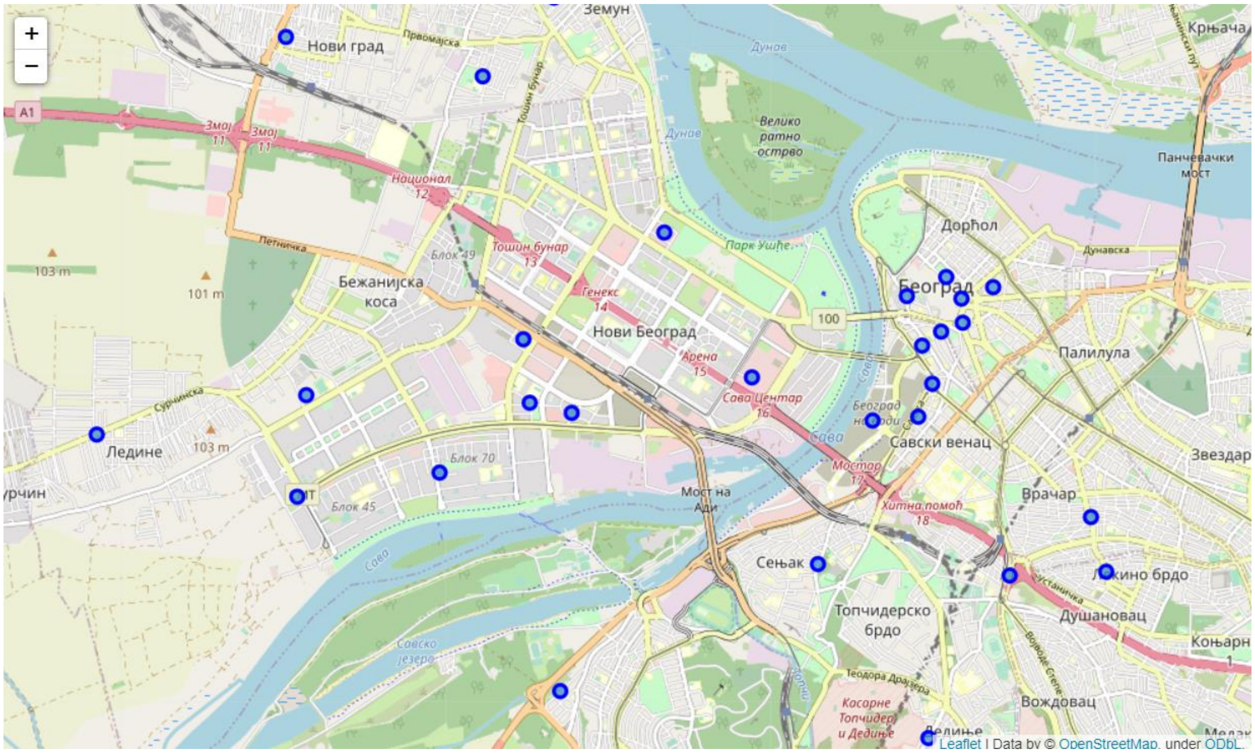
	Municipality	Neighborhood	Latitude	Longitude	Gym and Recreation
24	Savski venac	Senjak	44.791132	20.441001	0.000000
25	Stari grad	Kosančićev venac	44.816823	20.452867	0.000000
26	Stari grad	Skadarlija	44.817694	20.464684	0.000000
27	Stari grad	Studentski trg	44.818572	20.458182	0.000000
28	Stari grad	Terazije	44.814220	20.460486	0.000000
29	Stari grad	Trg republike	44.816606	20.460279	0.000000
30	Surčin	Surčin	44.790213	20.266705	0.000000
31	Voždovac	Autokomanda	44.789941	20.466910	0.000000
32	Voždovac	Dušanovac - Lekino brdo	44.790323	20.479915	0.000000
33	Voždovac	Jajinci	44.742158	20.483459	0.000000
34	Voždovac	Stepa Stepanović	44.758049	20.492183	0.000000
35	Zemun	Batajnica	44.905515	20.275730	0.000000
36	Zemun	Gardoš	44.848533	20.408333	0.000000
37	Zemun	Gornji Grad	44.853228	20.386959	0.000000
38	Zemun	Kalvarija	44.837843	20.395563	0.000000
39	Zemun	Meandri	44.847320	20.382126	0.000000
40	Zemun	Nova Galenika	44.857695	20.368797	0.000000
41	Zemun	Novi Grad	44.841583	20.369017	0.000000
42	Zemun	Zemun Polje	44.871923	20.323677	0.000000
43	Zemun	Ćukovac	44.845370	20.405212	0.000000
44	Zvezdara	Mali mokri lug	44.774852	20.535932	0.000000
45	Zvezdara	Mirijevo I	44.797329	20.524950	0.000000
46	Zvezdara	Veliki mokri lug	44.757604	20.525469	0.000000
47	Čukarica	Bele vode	44.749734	20.402341	0.000000
48	Čukarica	Julino brdo	44.767378	20.409674	0.000000
49	Čukarica	Rušanj	44.682646	20.437232	0.000000
50	Čukarica	Sremčica	44.676718	20.391154	0.000000
51	Čukarica	Čukarička padina	44.778933	20.406073	0.000000

	Municipality	Neighborhood	Latitude	Longitude	Gym and Recreation
52	Čukarica	Železnik	44.728056	20.374145	0.000000
53	Grocka	Kaluđerica	44.753719	20.555951	0.356364
54	Voždovac	Kumodraž	44.739890	20.510201	0.379310
55	Novi Beograd	Naselje Dr Ivan Ribar	44.797573	20.370549	0.648148
56	Novi Beograd	Savski blokovi	44.799847	20.389710	0.648148
57	Savski venac	Zeleni venac	44.813331	20.457530	0.648148
58	Vračar	Čubura	44.795652	20.477822	0.648148

6. Discussion

For the selected venue category **Gym and Recreation** (incl. Gym, Fitness, Yoga Studio, Pilates Studio and other Recreation venues), there are number of different locations (i.e. neighborhoods) which can be considered to start this business.

Map of Belgrade with recommended locations, illustrates the solution, for the key Business Problem addressed in this Data Science show-case project



There are **53** locations which do not have **Gym and Recreation** venues at all (*or at least not registered in Fouresquare*), out of total of **59** locations which have **well below average** number of Gyms or similar Recreation venues. It is interesting to note, that new and trendy microlocation neighborhoods in Novi Beograd like A blok, Belvil and West 65 Belgrade, don not have sufficient access to Gyms and Recreation venues, compared to similar neighborhoods.

Also, an opportunity for more Gyms and Recreation venues exists for new trendy Belgrade Waterfront neighborhood as well.

But not only new and trendy neighborhoods have potential. Older, traditional neighborhoods in Stari Grad municipality like Kosančićev venac and Skadarlija, should be considered as potential location for starting this type of new business.

An interesting observation is that residents of some large traditional neighborhoods in Novi Beograd like Savski blokovi, and some large outer neighborhoods like Veliki mokri lug and Ledine, have below average access to the Gyms and Recreation venues, compared to similar class of neighborhoods. So let's consider this as a clear business opportunity.

Limitations:

This is s show-case analysis.

In this show-case Data Science analysis project, we have used freely available data sources. However, certain limitations do apply:

- Statistical data is not available in the easy machine readable format.
- Real-estate data are subject to changes, and info may vary from portal to portal.
- Foursquare provides free API, but does not have full and comprehensive list of venues in Belgrade, compared to Google.
- In addition, Foursquare response list of venues and trendy venues depends of the time of day when query is made.

7. Conclusion

Clustering results are not surprising for Belgrade. Neighborhood cluster label is in correlation with its distance from the city downtown, with some minor exceptions - which can be empirically confirmed as a correct result for Belgrade - but these results were achieved using Data Science methods, using selected data sources, and K-Mean clustering algorithm, so one does not have to have prior knowledge of the city in order to execute this type of analysis.

- It is clear that statistical data are the most impacting factor on cluster label.
- The aim of the clustering is not to determine which neighborhoods are more or less desirable, this could have been done by simple reviewing of the market prices of the real-estates.
- The aim is to create clusters of similar neighborhoods, which can be used to answer the key Business Problem: In which neighborhoods one can start new business i.e. open new venue.

Why clustering is important?

- If one would compare neighborhoods which belongs to different clusters, resident composition and real-estate pricing (also needed for renting/acquiring business space) would not be mutually comparable.
- By comparing neighborhoods in the same cluster, it is possible to produce more accurate recommendation, in which neighborhoods residents have below average access to the specific venues.

Recommendation is based on the statistical model for each cluster (developed by K-Means clustering algorithm), where different neighborhood sizes and average (mean) value of the venue category per each cluster are taken into consideration.

Using the same model, analysis for other venue categories or for other types of business can be conducted as well.

8. References

1. Serbia State Statistical Office - https://www.stat.gov.rs/sr-Latn/oblasti/trziste-rada/zarade
2. Beograd.rs - http://www.beograd.rs/
3. Imovina.net - https://imovina.net/statistiika_cena_nekretnina/?viewType=table
4. Foursquare - https://foursquare.com/city-guide
5. Google - https://www.google.rs/maps/@44.793856,20.4668928,12z?hl=en
6. Wikipedia - https://sh.wikipedia.org/wiki/Beograd
7. Wikipedia - https://en.wikipedia.org/wiki/Subdivisions_of_Belgrade
8. Alex Aklson GitHub - https://github.com/aklson-datascientist
9. Silbaski Wikimedia Belgrade Panorama - https://sr.wikipedia.org/sr-el/%D0%94%D0%B0%D1%82%D0%BE%D1%82%D0%B5%D0%BA%D0%B0:PANORAMA_BEOGRADA_SA_PC_%22USCE%22_(Old_Belgrade_Panorama_From_The_Bilding_%22Usce%22).jpg