Venues Data Analysis of Moscow City

1. Introduction

1.1 Background

Moscow, one of the largest metropolises in the world with a population of more than 12 million people, covers an area of more than 2561.5 km^2 with an average density of inheritance of 4924.96 people / km^2 1.

Moscow is divided into 12 districts (125 boroughs, 2 urban boroughs, 19 settlement boroughs).

Moscow has a very uneven population density from 30429 people / km^2 for the "Зябликово" borough, to 560 people / km^2 for the "Молжаниновский" borough 2.

The average cost of real estate varies from 68,768 rubles / m^2 for the "Кленовское" borough to 438,568 rubles / m^2 for the "Арбат" borough 3.

1.2 Business Problem

Owners of cafes, fitness centers and other social facilities are expected to prefer boroughs with a high population density. Investors will prefer areas with low housing costs and low competitiveness.

On the part of residents, the preference is expected for a boroughs with a low cost of housing and good accessibility of social places.

In my research, I will try to determine the optimal places for the location of fitness centers in Moscow boroughs, taking into account the number of people, the cost of real estate and the density of other fitness facilities.

The key criteria for selecting suitable locations for fitness centers will be:

- High population of the borough
- Low cost of real estate in the area
- The absence in the immediate vicinity of other fitness facilities of a similar profile

I will use the approaches and methods of machine learning to determine the location of fitness centers in accordance with the specified criteria.

The main stakeholders of my research will be investors interested in opening new fitness centers.

2. Data acquisition and cleaning

2.2. Data requirements

Based on the problem and the established selection criteria, to conduct the research, I will need the following information:

1. main dataset with the list of Moscow Borough, containing the following attributes:
   - name of the each Moscow Borough
   - type of the each Moscow Borough
   - name of the each Moscow District in which Borough is belong to
   - area of the each Moscow Borough in square kilometers
   - the population of the each Moscow Borough
   - housing area of the each Moscow Borough in square meters
   - average housing price of the each Moscow Borough
2. geographical coordinates of the each Moscow Borough
3. shape of the each Moscow Borough in GEOJSON format
4. list of venues placed in the each Moscow Borough with their geographical coordinates and categories

2.3. Describe data sources

2.3.1. Moscow Boroughs dataset

Data for Moscow Boroughs dataset were downloaded from multiple HTTP page combined into one pandas dataframe.

- List of Moscow District and they Boroughs were downloaded from the page Moscow Boroughs
- Information about area of the each Moscow Borough in square kilometers, their population and housing area in square meters were downloaded from the page Moscow Boroughs Population Density
- Information about housing price of the each Moscow Borough were downloaded from the page Moscow Boroughs Housing Price

A special Python function has been developed for HTML table parse. This function help me:

- to find number of rows and columns in a HTML table
- to get columns titles, if possible
- to convert string to float, if possible
- return result in form of the Pandas dataframe

2.3.2. Moscow Boroughs geographical coordinates

Geographical coordinates of the each Moscow Borough were queried through Nominatim service. As the Nominatim service are quite unstable it was quite a challenge to request coordinate in several iterations.

2.3.3. Moscow Boroughs shape in GEOJSON format

Shape of the each Moscow Borough in GEOJSON format was downloaded from the page Moscow Boroughs GEOJSON

2.3.4. Moscow Boroughs venues

To determine **venues** the service **Forsquare API** was used.
The API of **Forsquare** service have the restriction of 100 **venues**, which it can return in one request.
To obtain list of all **venues** I used the following approach:

- present Moscow area in the form of a regular grid of circles of quite small diameter, no more than 100 **venues** in each circle
- perform exploration using **Forsquare API** with quite bigger radius than circle of a grid to make sure it overlaps/full coverage to don't miss any venues
- cleaning list of venues from duplicates.

This approach and some of the Python code was taken from the work presented here.
https://cocl.us/coursera_capstone_notebook

Circle of 28 000 meter in radius cover all Moscow Boroughs.
In my research grid of circles contains 7899 cells with radius 300 meter.
Foursquare API have a certain limitation for API call in one day to explore venues.
In my case it was about 2000 calls per day.
So in addition I have to divide grid dataset into subset and call Foursquare API for several days.

2.4. Describe data cleansing

2.4.1. Moscow Boroughs dataset cleansing

As data for Moscow Boroughs dataset were downloaded from multiple HTTP page it was necessary to perform a data cleaning. Such as:

- remove some unused colums
- strip text columns from additional information like ' \n\t'
- replace some Borough_Name as of russian letters "е" and "ё"
- change places of some words in Borough_Name
- clear Borough Name from additional information, such as ', поселение ', ', городской округ '
- replace '\n', ' ↗' and '↘' in some columns
- delete extra spaces in numeric columns
- replace ',' to '.' for float columns
- convert from float to int for integer columns

- convert from string to float for numeric columns

As the result I, had a dataset with all 146 Moscow Boroughs. Result dataset contains columns:

- **Borough_Name** - name of the Moscow Borough - is a unique key of the dataset
- **District_Name** - name of the Moscow District in which Borough is belong to
- **Borough_Type** - type of the Moscow Borough
- **OKATO_Borough_Code** - numeric code of the Moscow Borough
- **OKTMO_District_Code** - numeric code of the Moscow District
- **Borough_Area** - area of the Moscow Borough in square kilometers
- **Borough_Population** - population of the Moscow Borough
- **Borough_Population_Density** - population density of the Moscow Borough
- **Borough_Housing_Area** - housing area of the Moscow Borough in square meters
- **Borough_Housing_Area_Per_Person** - housing area per person of the Moscow Borough in square meters
- **Borough_Housing_Price** - average housing price of the Moscow Borough

I had a problem to found proper statistics about “housing prices” and “housing area” for some Moscow boroughs, so I had to exclude 26 boroughs from my analysis. Fortunately, they all had a low population density, which meat criteria of my research and did not reduce it quality.

2.4.2. Moscow Boroughs geographical coordinates cleansing

Nominatim service not only quite unstable.
It also have an occasionally problem with russian leter **ё**. So I have to manyaly obtain coordinates for such boroughs as:

- Десёновское, Поселение, Новомосковский
- Савёлки, Муниципальный округ, ЗелАО
- Клёновское, Поселение, Троицкий
- And some others.

Another problem with Nominatim service is that it return not very accurate coordinate of some Boroughs.
So I needed to adjust they manually in the map.

As the result I, had a dataset with all 146 Moscow Boroughs geographical coordinates:

- **Borough_Name** - name of the Moscow Borough
- **Latitude** - geographical Latitude of the Moscow Borough
- **Longitude** - geographical Longitude of the Moscow Borough

2.4.3. Moscow Boroughs shape in GEOJSON format cleansing

GEOJSON file downloaded from the page Moscow Boroughs GEOJSON was quite good and not required any addition clearing.

2.4.4. Moscow Boroughs venues cleansing

Using **Forsquare API** I obtained 34460 venues in 7899 cells.
As I used a quite bigger radius (350 meters) for venue explorations than circle of a grid (300 meters), there was a need to remove duplicates venues.
After duplicates removal I had 27622 unique venues in the circle radius of 28 000 meters around the Moscow City.

The second task was to bind each venue to Moscow Boroughs in which borders they were placed. To perform this task I created a polygon for each Moscow Borough from GEOJSON file and found which venues coordinate included into each polygon.

The third task was to remove all the venues that placed outside of the Moscow boroughs.

The fourth task was to get main category from the category list for each venue.

As the result, I had list of 20864 venues placed in the Moscow Boroughs with their geographical coordinates and categories

2.5. Example of the resulting datasets

2.5.1. The result Moscow Boroughs dataset

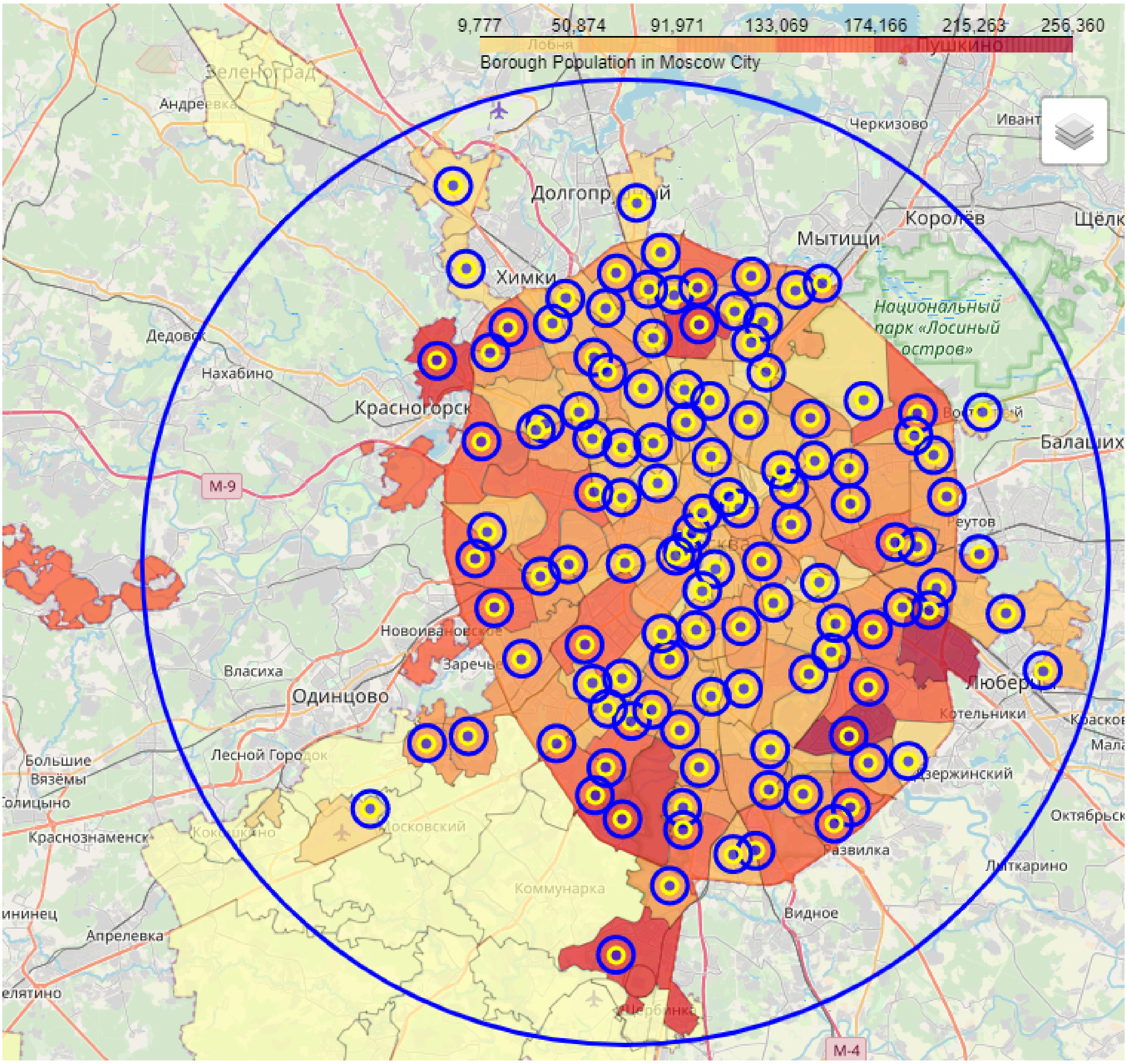
The prepared and cleared Moscow Boroughs dataset can be downloaded by link Moscow Boroughs dataset

The picture below shows a small part of the Moscow Boroughs dataset

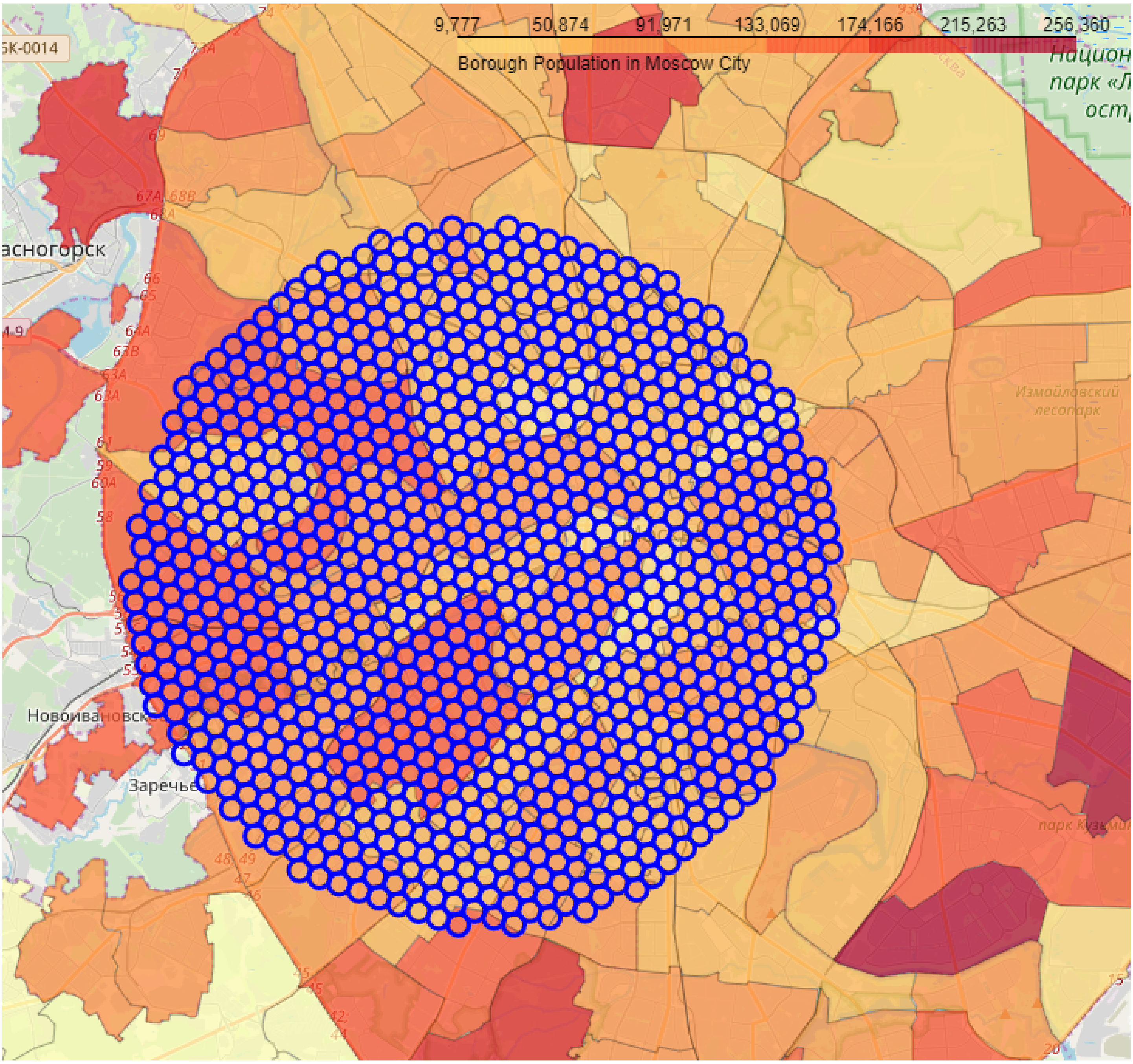
Index	Borough_Name	District_Name	Borough_Type	ATO_Borough_Cc	TMO_District_C	Borough_Area	orough_Populatic	Populatic	rough_Housing_Ar	using_Are	Latitude	Longitude	orough_Housing_Pric
0	Академический	ЮЗАО	Муниципальный округ	45293554	45397000	5.83	109387	18762	2467.00	22.70	55.69	37.58	199999.00
1	Алексеевский	СВАО	Муниципальный округ	45280552	45349000	5.29	80534	15223	1607.90	20.50	55.81	37.65	199474.00
2	Алтуфьевский	СВАО	Муниципальный округ	45280554	45350000	3.25	57596	17721	839.30	15.50	55.88	37.58	138021.00
3	Арбат	ЦАО	Муниципальный округ	45286552	45374000	2.11	36125	17120	731.00	26.00	55.75	37.59	438568.00
4	Аэропорт	САО	Муниципальный округ	45277553	45333000	4.58	79486	17355	1939.70	25.90	55.80	37.53	234544.00
5	Бабушкинский	СВАО	Муниципальный округ	45280556	45351000	5.07	88537	17462	1586.30	18.50	55.87	37.66	164324.00
6	Басманный	ЦАО	Муниципальный округ	45286555	45375000	8.37	110694	13225	1991.80	18.40	55.78	37.69	302021.00
7	Беговой	САО	Муниципальный округ	45277556	45334000	5.56	42781	7694	791.10	18.80	55.78	37.57	261402.00
8	Бескудниковский	САО	Муниципальный округ	45277559	45335000	3.30	79603	24122	1391.70	18.40	55.86	37.56	158398.00
9	Бибирево	СВАО	Муниципальный округ	45280558	45352000	6.45	160163	24831	2521.80	15.80	55.88	37.60	140533.00
10	Бирюлёво Восточное	ЮАО	Муниципальный округ	45296553	45911000	14.77	155863	10552	2122.20	14.70	55.59	37.66	124645.00
11	Бирюлёво Западное	ЮАО	Муниципальный округ	45296555	45912000	8.51	88672	10419	1183.20	13.20	55.59	37.64	109421.00
12	Богородское	ВАО	Муниципальный округ	45263552	45301000	10.24	109324	10676	1744.10	16.90	55.82	37.71	178577.00
13	Братеево	ЮАО	Муниципальный округ	45296557	45913000	7.63	110021	14419	1585.40	15.50	55.64	37.76	136300.00
14	Бутырский	СВАО	Муниципальный округ	45280561	45353000	5.04	71458	14178	1236.20	18.30	55.81	37.59	182641.00
15	Вешняки	ВАО	Муниципальный округ	45263555	45302000	10.72	122285	11407	1976.80	16.20	55.73	37.82	147352.00
16	Внуково	ЗАО	Муниципальный округ	45268552	45317000	17.42	25471	1462	416.60	17.80	55.61	37.30	113399.00
17	Войковский	САО	Муниципальный округ	45277565	45336000	6.61	70729	10700	1531.00	23.10	55.82	37.49	207242.00
18	Восточное Дегунино	САО	Муниципальный округ	45277568	45337000	3.77	98923	26239	1592.50	16.70	55.88	37.56	146300.00

2.5.2. Boroughs population in Moscow City map

The picture below shows a choropleth map of the Moscow Boroughs population and the center of each boroughs.
As we can see, use center of the boroughs for searching venues is quite useless as each borough have very sophisticated shape.
So I needed to present Moscow area in the form of a regular grid of circles of quite small diameter inside the circle of 28 000 meter in radius, which cover all the Moscow Boroughs in my research.



The picture below shows an Example of such hexagonal grid of area candidates



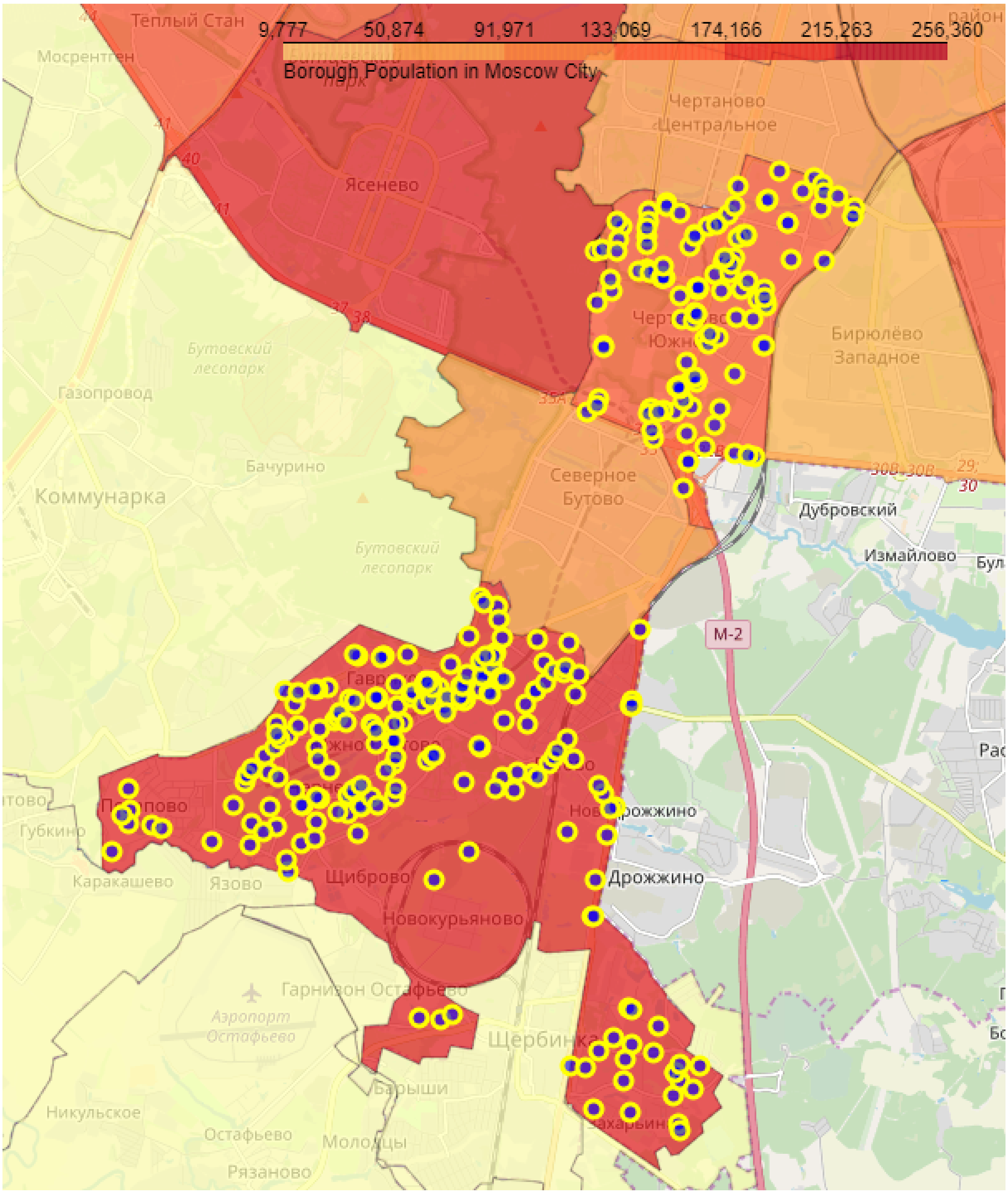
2.5.3. The result Moscow venues dataset

The prepared and cleared Moscow venues dataset can be downloaded by link Moscow venues dataset

The picture below shows a small part of the Moscow Boroughs dataset

Cell_id	Venue_Id	Borough_Name	Venue_Name	Venue_Latitude	Venue_Longitude	Venue_Category_Name
55.7020821...	511629f5e4b051a081439bf5	Очаково-Матвеевское	"Aminevskoe hotel" restaurant	55.703032	37.454590	Hotel
55.8350558...	5023841de4b0e6fe1a411c7d	Ростокино	"Cosmos 2" Hotel	55.836780	37.665548	Hotel
55.8277624...	505f30d2e4b0d9a2f19a319d	Покровское-Стрешнево	"Karaoke&Bar G-Voice"	55.827876	37.409241	Karaoke Bar
55.6864545...	4efb158da17cdc15b40b98fc	Очаково-Матвеевское	"MOON"	55.686766	37.414477	Furniture / Home Store
55.7213688...	5905a5870123587260ffe1d5	Южнопортовый	"Mime" Film Company (Мим Кинокомпания)	55.722946	37.679820	Film Studio
55.7488985...	5083dcc4e4b0ba1a3249d19f	Вешняки	"Red House" Клуб-Сауна	55.746088	37.838734	Sauna / Steam Room
55.7454108...	50eadc9de4b02662c430d51c	Новокосино	"Александр"	55.744217	37.877648	Department Store
55.7366957...	4eb12a04b63434fc86fa3310	Дорогомилово	"Аргумент - кафе"	55.738145	37.532077	Restaurant
55.7143244...	53a02544498e62c556da1f3f	Хамовники	"Банкет Холл" Лужники	55.715131	37.547142	Russian Restaurant
55.8692166...	5299878d11d2d1319ecea89f	Северное Тушино	"Бегемотики"	55.870727	37.440701	Kids Store
55.7623045...	50162ce6e4b01bcdb30b45e0	Крылатское	"Беговая дорожка" в Крылатском	55.762294	37.416648	Athletics & Sports
55.6249294...	4d877bec99b78cfaf7f5f91f	Орехово-Борисово Севе...	"Борисовский" билиардная	55.624427	37.709809	Bar
55.7949991...	503ccbf9e4b0708fceeb8ad1	Строгино	"Веселуха"	55.795756	37.405038	Dance Studio
55.8866119...	50420be2e4b0b5223de4c8a5	Дмитровский	"Волчий лес" / "Wolf Wood"	55.885273	37.528364	Café
55.6367977...	4f2c1f33e4b0ecad92a8352c	Коньково	"Гермес"	55.639274	37.544578	Convenience Store
55.6645507...	4f6a1b18e4b0ed0504f11293	Марьино	"Городская аптека"	55.662385	37.773821	Pharmacy
55.8777268...	50fbfea6e4b09f8ff7c27c93	Куркино	"Золотые Дуги"	55.880515	37.396922	American Restaurant
55.7902398...	4d43cae40349224b7365f34e	Восточное Измайлово	"Измайловский СДС" Филиал ГУП "Мосзеленх...	55.793075	37.823913	Flower Shop
55.7110205...	56b5e6ed498e16a72e900561	Даниловский	"Комус"	55.709422	37.657847	Paper / Office Supplies Store
55.8952978...	5558da32498ed73c64236d90	Лианозово	"Лавочки"	55.896766	37.580660	Park
55.8951981...	4ead5cf729c2a9bb97952c9e	Дмитровский	"Левый Берег" торговый центр	55.895344	37.503386	Shopping Mall
55.6521319...	4ea54de79adff6343ad6ff45	Тропарёво-Никулино	"Леди & Бродяга"	55.651273	37.470040	Pet Store
55.6833684...	51f7c3b0498e305d9ef6b5b2	Некрасовка	"Магнит"	55.683751	37.928274	Supermarket
55.8798507...	541c4831498e76f1b432ffee	Ярославский	"Магнит"	55.878228	37.729744	Supermarket
55.6628188...	51bea6bf498ea7d17efe1403	Люблино	"Мекона" Сервис	55.661802	37.807258	Auto Workshop

The picture below shows a example of the some Moscow Boroughs and their venues



Exploratory Data Analysis

The key criteria for my research are:

- high population of the borough
- low cost of real estate in the area

We have theese key features in Moscow Boroughs dataset:

- District - name of the Moscow District in which Borough is belong to
- Area - area of the Moscow Borough in square kilometers
- Population_Density - population density of the Moscow Borough

- Housing_Area - housing area of the Moscow Borough in square meters

Let's analyze features and key criteria using:

- descriptive statistical analysis
- categorical variables analysis
- correlation analysis

Descriptive statistical analysis

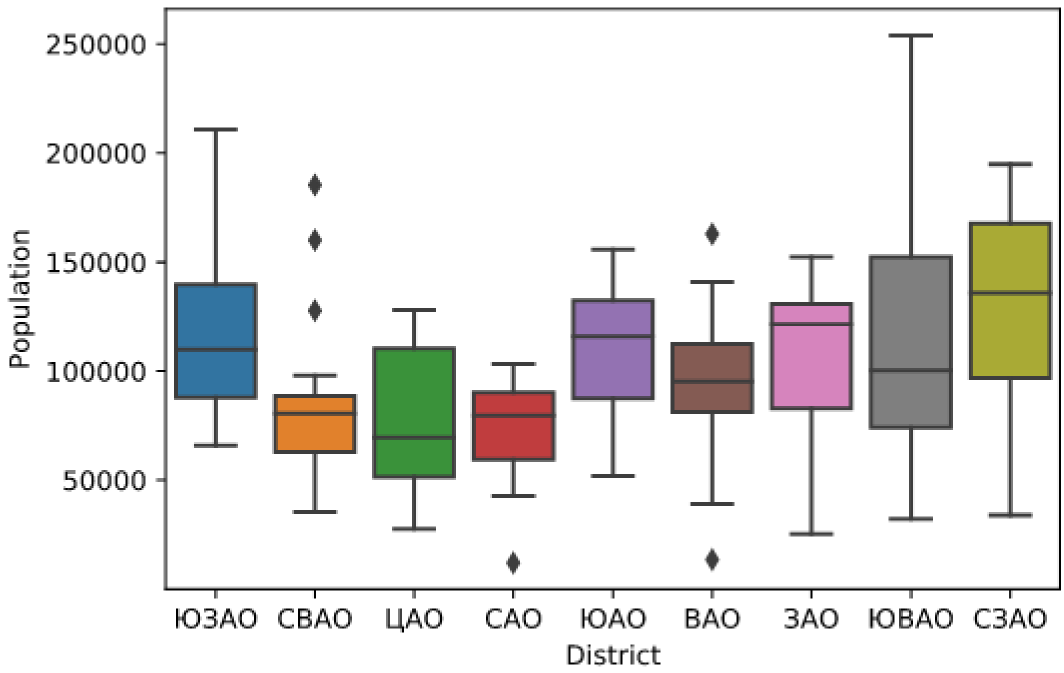
The picture below shows basic statistics for all features.
As we can see, Moscow Boroughs has a very uneven population from 12 194 people to 253 943 people.
The average cost of real estate varies from 109 421 rubles/m² to 438 568 rubles/m².

	Area	Population_Density	Housing_Area	Population	Housing_Price
count	120.000000	120.000000	120.000000	120.000000	120.000000
mean	8.706417	13426.608333	1775.684167	99847.608333	190037.316667
std	4.927028	5956.551611	815.978445	44024.992123	66182.885601
min	2.110000	559.000000	69.900000	12194.000000	109421.000000
25%	5.395000	9745.750000	1244.450000	71821.750000	147339.000000
50%	7.680000	13266.000000	1709.450000	93892.000000	168172.500000
75%	10.282500	17151.000000	2206.600000	126545.750000	210978.000000
max	27.570000	30428.000000	4523.000000	253943.000000	438568.000000

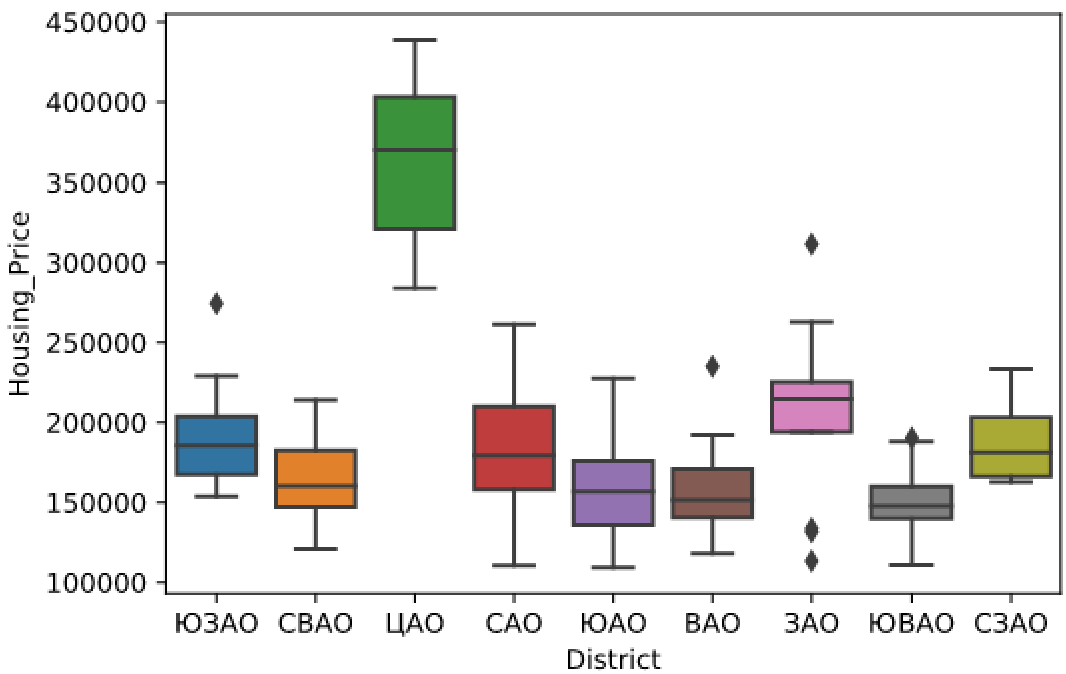
Categorical variables analysis

I have one categorical variable - name of the Moscow District in which Borough is belong to.
Let's analize relationship between categorical feature 'District' and key criteria using boxplots visualization.

The picture below shows relationship between 'District' and 'Population'.
We can see that the distributions of Population between Boroughs in the different Districts have aт overlap, but we can estimate, that the most populated Boroughs are placed in 'ЮЗАО', 'ЮАО', 'СЗАО' and 'ЗАО' Districts.



The next picture shows relationship between 'District' and 'Housing Price'.
We can see that the distributions of Housing Price between Boroughs in the different Districts are distinct enough.
As the result of boxplots visualization, categorical feature 'District' would be a good potential redictor only of Housing Price.



Correlation analysis

The picture below shows correlation matrix.
Correlation between 'Area', 'Population_Density' and 'Population' is statistically significant, although the linear relationship isn't extremely strong.
Correlation between 'Housing_Are' and 'Population' is statistically hughly significant, and the linear

relationship is extremely strong.
Correlation between 'Area', 'Population_Density', 'Housing_Area' and 'Housing_Price' is not statistically significant, although the linear relationship isn't strong.
Correlation between 'Area' to 'Population_Density' is statistically hughly significant, and the linear relationship is extremely strong.
So we can exclude 'Population_Density' from our considerations.

	Area	Population_Density	Housing_Area	Population	Housing_Price
Area	1.000000	-0.585991	0.344188	0.380587	-0.154996
Population_Density	-0.585991	1.000000	0.289456	0.338621	-0.101348
Housing_Area	0.344188	0.289456	1.000000	0.887856	-0.016971
Population	0.380587	0.338621	0.887856	1.000000	-0.195774
Housing_Price	-0.154996	-0.101348	-0.016971	-0.195774	1.000000

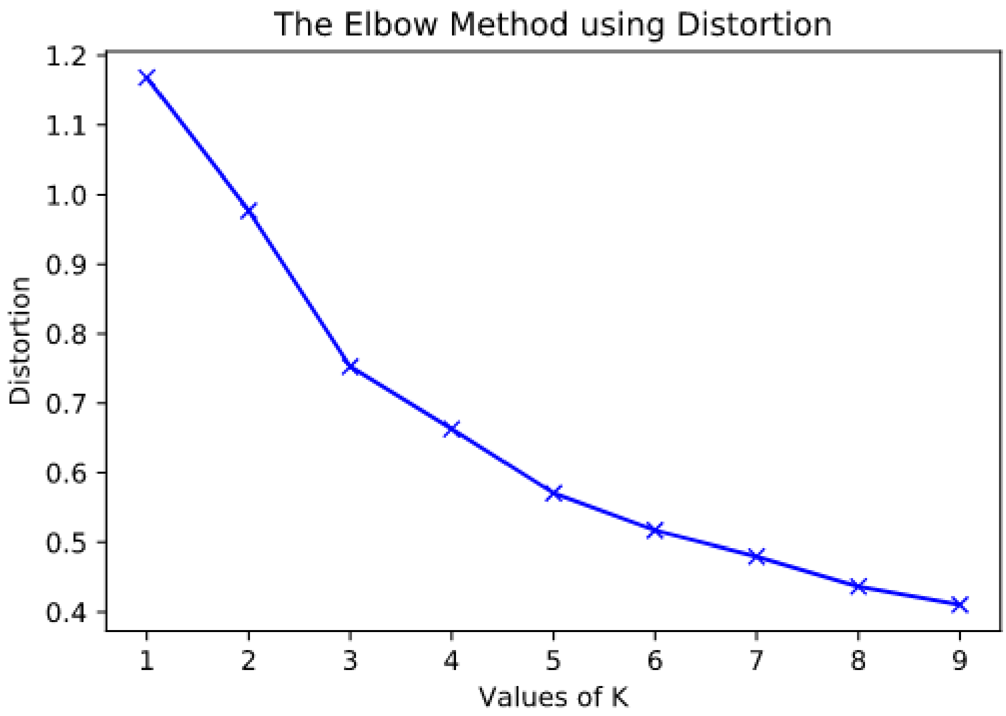
Clustering

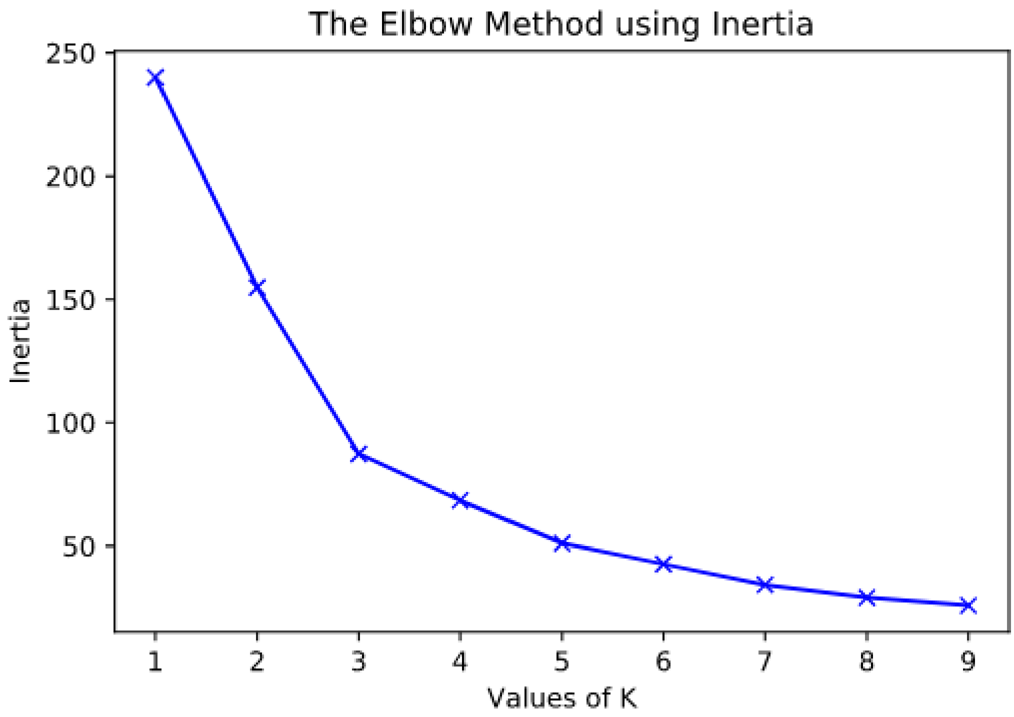
In my research, I decided to perform Moscow Boroughs segmentation with K-Means to detect Boroughs that have highest population and smallest housing price.

K-Means Clustering with elbow method

To determine right number of clusters, I used elbow method. According elbow method I implemented K-Means clustering from 1 to 10 centroids and calculate distortion and inertia for each variant.

The next pictures show elbow method using Distortion and Inertia. We can see that there are elbows at 3 and 5 centroid.
I decided to use 3 centroid In my research.





Analyze K-Means clusters

To analyze K-Means clusters I calculated some statistics:

- count boroughs in the cluster
- sum population in the cluster
- sum area of the cluster
- mean population in the boroughs in the cluster
- mean housing price in the boroughs in the cluster
- % population in the cluster to all Moscow City population
- % area of the cluster to all Moscow City area
- population density in the cluster

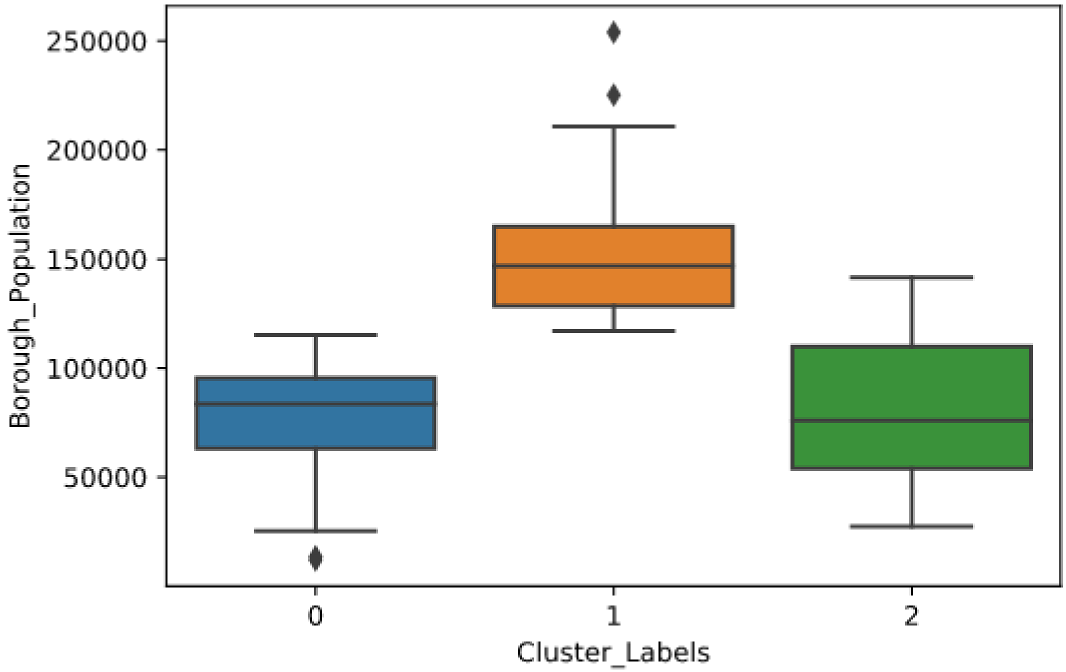
The next pictures show these statistics

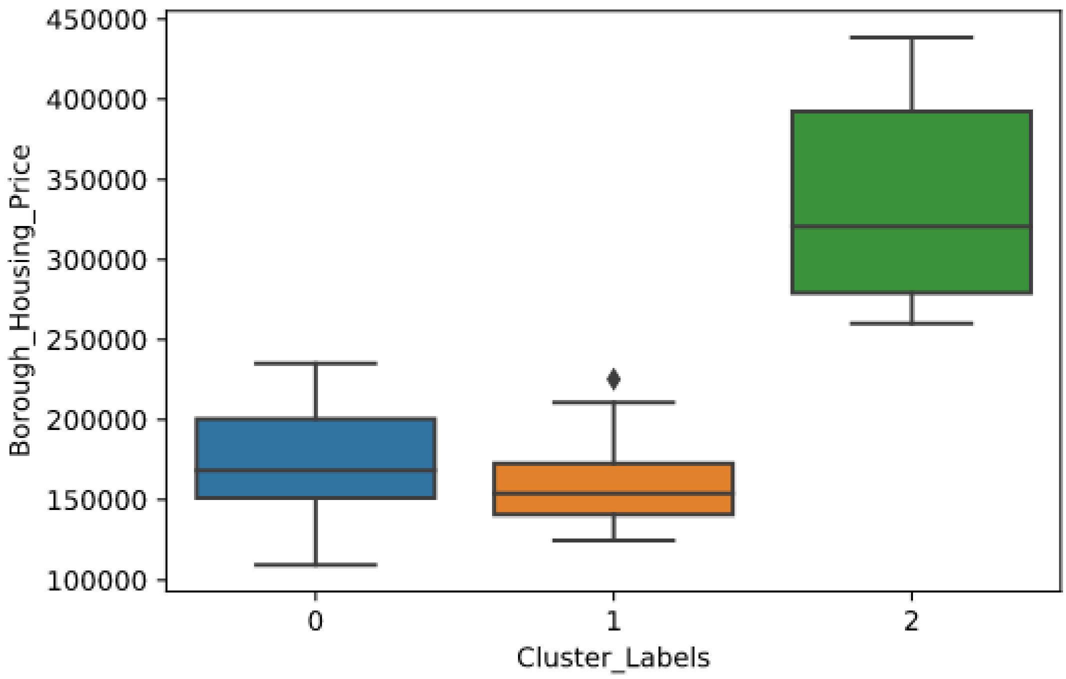
	Cluster_Labels	Population_Mean	Housing_Price_Mean	Population_Sum	Population_%	Borough_Count	Area_Sum	Area_%	Population_Density
0	0	78538.901408	173695.070423	5576262	46.539773	71	539.87	51.673574	10328.897698
1	1	153187.235294	160741.323529	5208366	43.469294	34	391.25	37.448434	13312.117572
2	2	79805.666667	333794.866667	1197085	9.990934	15	113.65	10.877992	10533.084030

As we can see, there are 3 clusters

- "0" Cluster - characterized by low mean population (78538 people per Borough), relatively high mean housing price (173695 rubles/m²) and low population density (10328 people/km²)
- "1" Cluster - characterized by highest mean population (153187 people per Borough), smallest mean housing price (160741 rubles/m²) and highest population density (13312 people/km²)
- "2" Cluster - characterized by low mean population (79805 people per Borough), highest mean housing price (333794 rubles/m²) and low population density (10533 people/km²)

The next pictures show these clusters using boxplots visualization.





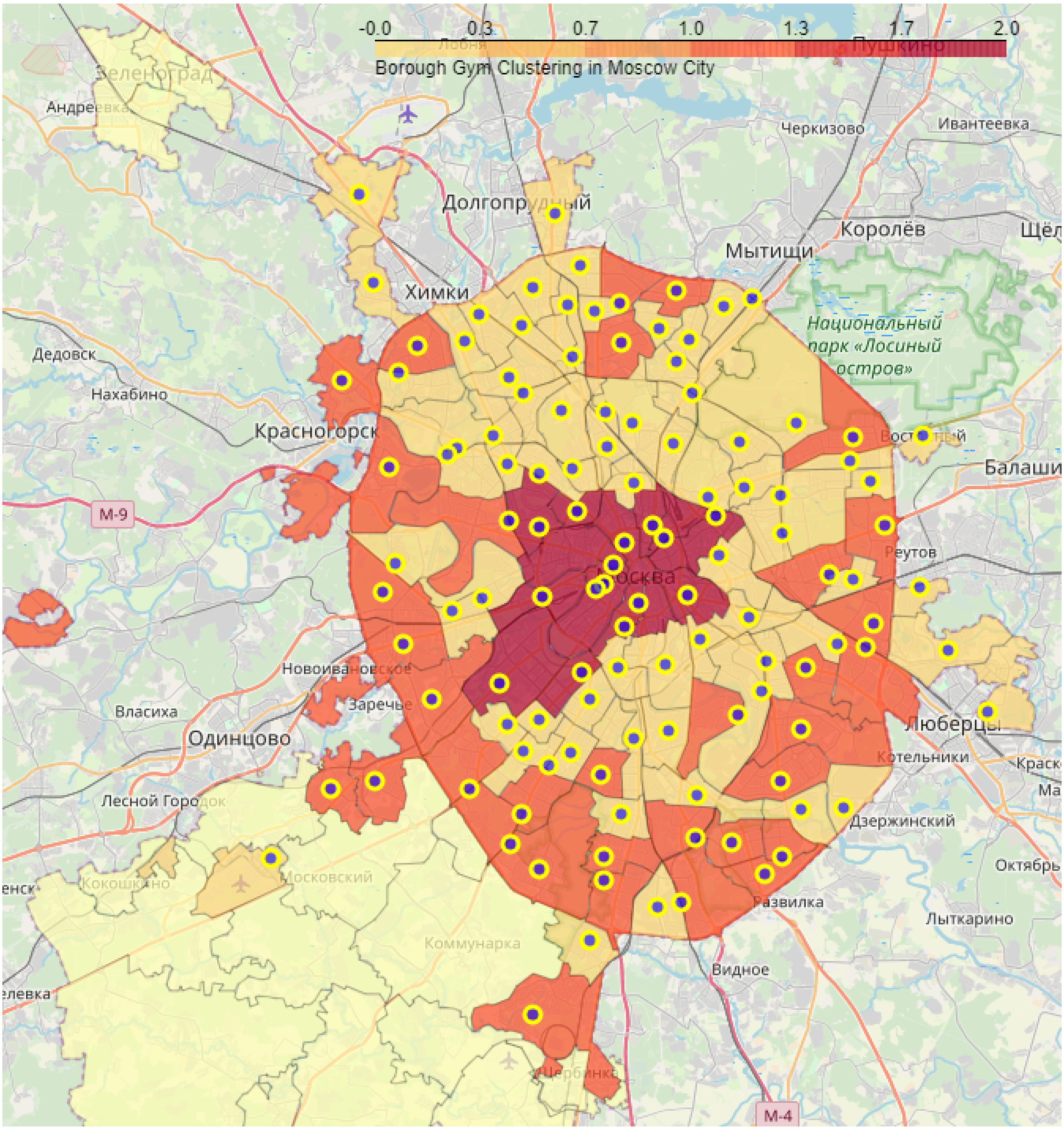
Very good result of the KMean clustering.

"1" Cluster perfectly fits my research criteria:

- boroughs from this cluster have highest mean population and smallest mean housing price
- in 34 boroughs about 43% of the Moscow population occupied only 37% of the Moscow City area, that mean the highest population density

Vizualize clusters on choropleth map

The next picture shows clusters on choropleth map.


-0.0 0.3 0.7 1.0 1.3 1.7 2.0
Borough Gym Clustering in Moscow City