## Report of Capstone Data Science project

## Introduction

## Background

## Business Problem

## Data acquisition

## Data requirements

## Data sources

## Data cleaning

## Methodology

## Exploratory Data Analysis

## Clustering

## Result

## Discussion

## Conclusion

### 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 km² with an average density of inheritance of 4924.96 people / km².

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

Moscow has a very uneven population density from 30429 people / km² for the "Zyablikovo" borough, to 560 people / km² for the "Molzhaninovskij" borough.

The average cost of real estate varies from 68,768 rubles / m² for the "Klenovskoe" borough to 438,568 rubles / m² for the "Arbat" borough.

### 1.2 Business Problem

Owners of cafes, resturants, gym, bank department 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 bank department in Moscow boroughs, taking into account the number of people and the cost of real estate.

The key criteria for selecting suitable locations for new bank department will be:

* High population of the borough
* Low cost of real estate in the borough
* Venues in borough

The main stakeholders of my research will be bank owner or key investors, interested in expanding the geography of the representative office.

## 2. Data acquisition and cleaning

### 2.1. 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

### 2.2. Data sources

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 Borough 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 Borough Population Density
* Information about housing price of the each Moscow Borough were downloaded from the page Moscow Borough Housing Price

A special Python function was 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

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. Shape of the each Moscow Borough in GEOJSON format was downloaded from the page [Moscow Boroughs GEOJSON](http://gis-lab.info/data/mos-adm/mo.geojson)

#### 2.2.1 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.

### 2.3. Data 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 97 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 thousands of 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 10 boroughs from my analysis. Fortunately, they all had a low population density, which meat criteria of my research and did not reduce it quality.

#### 2.3.2. Moscow Boroughs geographical coordinates cleansing

Nominatim service not only quite unstable.

It also have an occasionally problem with russian letter **ё**. So I have to manually obtain coordinates for such boroughs as:

* Дес**ё**новское, Поселение, Новомосковский
* Сав**ё**лки, Муниципальный округ, ЗелАО
* Кл**ё**новское, Поселение, Троицкий
* And some others.

Another problem with Nominatim service is that it return not very accurate coordinate of some Boroughs(.

Because of this, need to adjust they data 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

#### 2.3.3. Moscow Boroughs shape in GEOJSON format cleansing

GEOJSON file downloaded from the page [Moscow Boroughs GEOJSON](http://gis-lab.info/data/mos-adm/mo.geojson) was quite good and not required any addition clearing.

#### 2.3.4. Moscow Boroughs venues cleansing

Using **Forsquare API** I obtained 1549 venues

First was a need to remove duplicates venues.

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.

## 3. Methodology

The key criteria for my research are:

* high population of the boroughs
* low cost of real estate in the boroughs area
* the absence in the immediate vicinity of the other fitness facilities

So I need to perform at least two tasks during analysis:

* first is to find boroughs with highest population and smallest housing price
* second is to provide a tool or methodology for determining vicinity of other fitness facilities in the boroughs

For the first task I try to use some approaches and methods of machine learning. And found out, what of the approaches suits my tasks best. I will use:

* exploratory data analysis, including descriptive statistical analysis, categorical variables analysis and сorrelation analysis
* segmentation with K-Means clustering

For the second task I decided to use visualization approach to mapping fitness facilities on to the interactive choropleth map and heatmap.

This approach can be easily used by stakeholders of my research to identify vicinity of other fitness facilities in the eache Boroughs.

### 3.1. Exploratory Data Analysis

We have following 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:

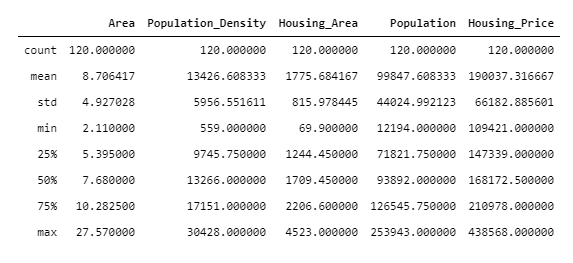
* descriptive statistical analysis
* categorical variables analysis

#### 3.1.1. 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.



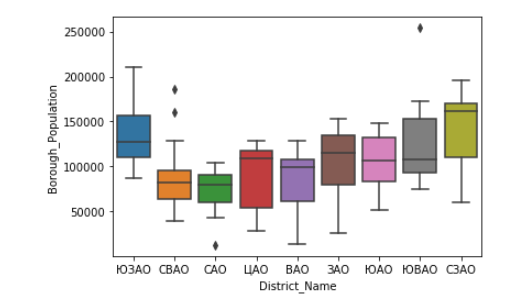
#### 3.1.2. Categorical variables analysis

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'.

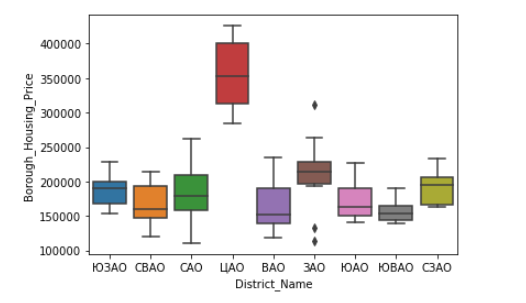
We can see that the distributions of Population between Boroughs in the different Districts have aт overlap, but we can estimate, that the most populated Boroughs are placed in 'ЮЗАО', 'ЮАО', 'СЗАО' and 'ЗАО' Districts.



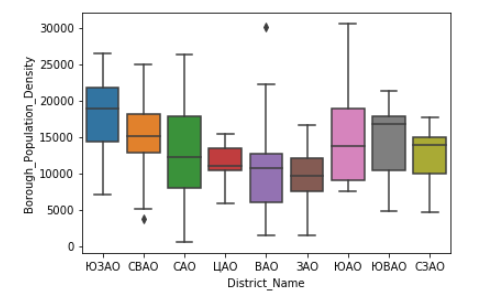
The next picture shows relationship between 'District' and 'Housing Price'.

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 predictor only of Housing Price.



Final picture show relationship between “District” and “Density.



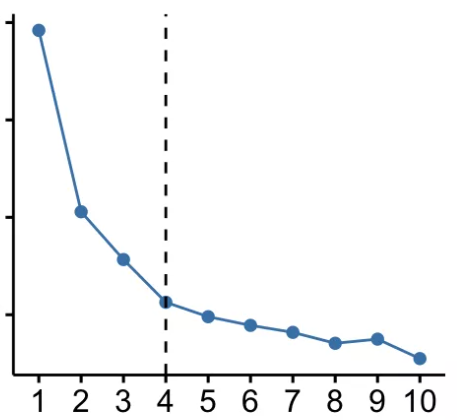
### 3.2. Clustering

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

#### 3.2.1. K-Means Clustering with elbow method

To determine right number of clusters, I used elbow method. According elbow method 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.



Best option will be using 4 centroid in my research.

"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)

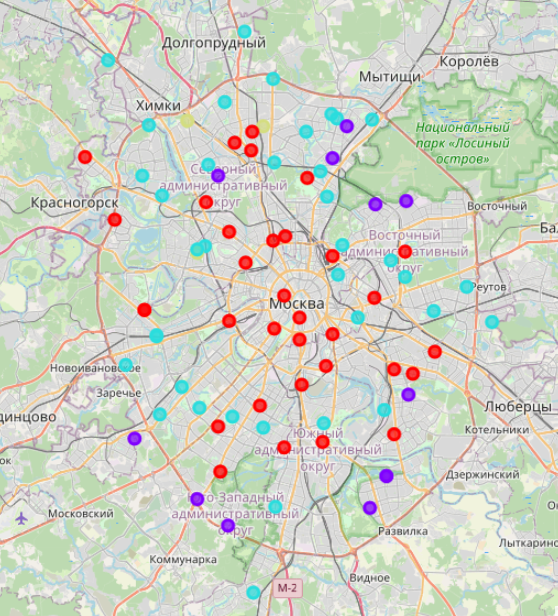
"2" Cluster - characterized by low mean population (79805 people per Borough), highest mean housing price (333794 rubles/m²) and low population density (10533 people/km²)

“4” Cluster – characterized by relatively high mean population (123020 people per Borough), relatively small mean housing price (165438 rubles/m²) and medium population density (11486 people/km²)

Cluster №2 perfectly fits 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

3.3 Vizualize clusters



## 4. Result

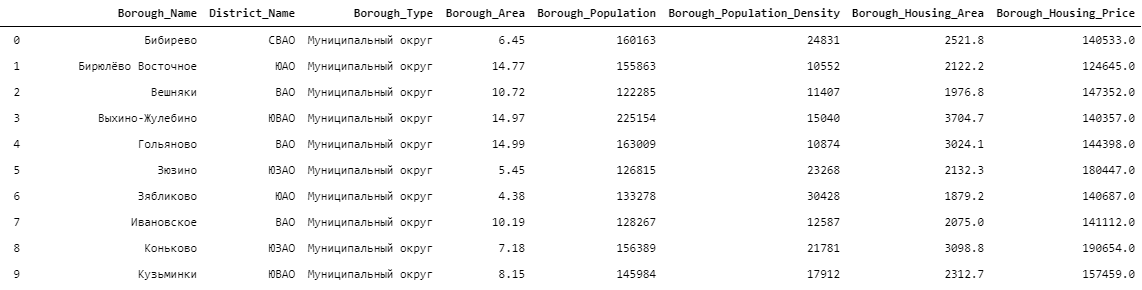
The result of my research consisted of:

* List of the optimal Boroughs for the location of bank department, according to the main criterias
  + high population of the borough
  + low cost of real estate in the borough
  + Venues in borough

Result dataset contains columns:

* **Borough\_Name** - name of the Moscow Borough
* **District\_Name** - name of the Moscow District in which Borough is belong to
* **Borough\_Type** - type of the Moscow Borough
* **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 thousands of square meters
* **Borough\_Housing\_Price** - average housing price of the Moscow Borough

The picture below shows a part of this dataset.



## 5. Discussion

In the course of my research I gathered a lot of informations about Moscow Boroughs, such as:

* 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
* geographical coordinates of the each Moscow Borough
* list of venues placed in the each Moscow Borough with their geographical coordinates and categories

All this information’s was cleared and hosted on GitHub and can be accessed in future research.

I have used segmentation with K-Means clustering to detect Boroughs that have highest population and smallest housing price. When I tested the elbow method, I set the optimum k value to 4, but there are another elbow at 5 centroid. Additional analysis can be done with 5 clusters, that can present slightly another set of optimal Boroughs for the location of fitness centers. So recommended boroughs list should therefore be considered only as a starting point for more detailed analysis.

In the future, research can be done additional analysis using categorical segmentations of the fitness facilities and calculate grid of location candidates taking into account bank department density.

## 5. Conclusion

Purpose of my project was to identify the optimal places for the location of bank department in Moscow boroughs, taking into account the dencity of population, the cost of real estate and the density of other fitness facilities in order to aid stakeholders in narrowing down the search for optimal location for a new bank department.

Results provided by research can ease new investors to found out potentially best place for new a new bank department.

Thanks for attention,

Marichev Ivan