Moscow Metro Stations Analysis

Anatoly Anpilogov

February 7, 2020

1. Introduction

1.1 Background

Moscow is the capital and most populous city of Russia, with approximately 15 million residents within the city limits and 20 million within the metropolitan area. Moscow has the lowest unemployment rate of all federal subjects of Russia. The average gross monthly wage in the city is ₽60,000 (US$2,500 in Purchasing Power), which is almost twice the national average of ₽34,000 (US$1,400 in Purchasing Power). Moscow has one of the largest metro in the world consisting of 263 stations. On average, approximately 5 million people use the Moscow metro per day. Therefore, it is a very reasonable idea to open a business in the vicinity of a metro station, but it would be useful to know which stations or types of businesses are more suitable for this.

1.2 Problem

This project aims to select metro stations in the vicinity of which it is better to open a new business, as well as to select the type of business. Data that might contribute to determining this might include the Foursquare location data and passenger traffic data.

1.3 Interest

This analysis will be useful for businessmen or managers of large businesses for a more informed decision-making on location to open a new sales point.

2. Data acquisition and cleaning

2.1 Data sources

A list of Moscow Metro stations with coordinates can be found on Wikipedia here. Passenger traffic data can be found here. I also used the service Foursquare to list venues near each metro station.

2.2 Data cleaning

Data downloaded or scraped from multiple sources was combined into one table. First, I took the table with coordinates of metro stations and data with the values of average daily traffic at metro stations for 2017 from Wikipedia. Then I corrected some inaccuracies and typos and merged into one data frame. The stations with the type ‘surface’ and ‘elevated, open’ are discarded. These are the stations of Moscow Monorail and Moscow Central Circle, which are very different from regular stations.

Stations, located closer than 250 meters to each other are essentially one station. They just have different names on different lines. Therefore, I combine them into one station, and the passenger traffic values for these stations are added up (for several stations the traffic values coincide completely, in this case the value is taken).

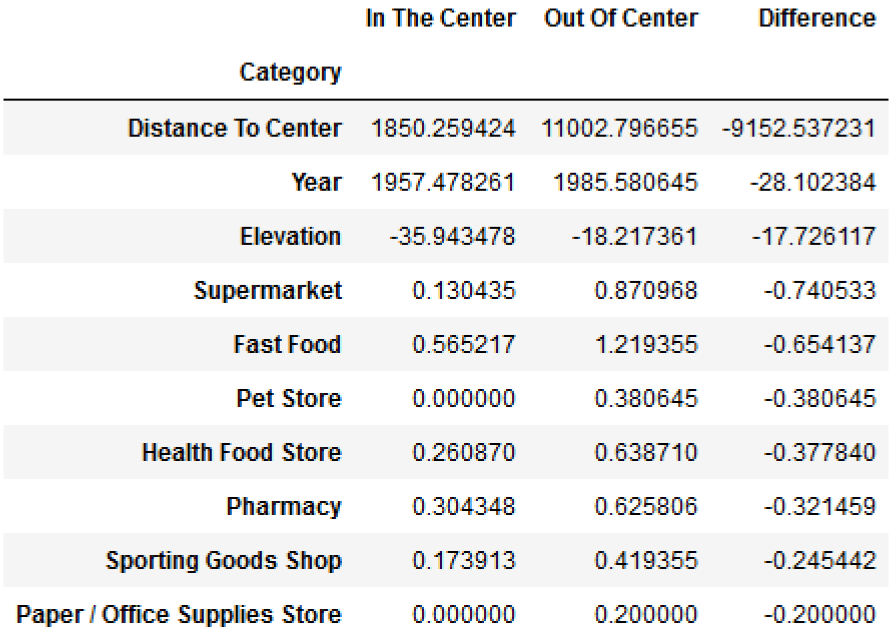
Using the Foursquare service, I get a list of venues located within the radius of 500 meters for each metro station. Then, I group some categories. All categories ending in “Restaurant” and “Pizza Place” are renamed to “Restaurant” to analyze the restaurants as a whole piece. I also group categories ending in ‘Bar’ or ‘Pub' into one category 'Bar / Pub'.

Thus, I get the final table consisting of 178 entries for metro stations. It contains information on the coordinates, passenger traffic (for 119 stations) and the number of venues of each category near each station.

3. Methodology

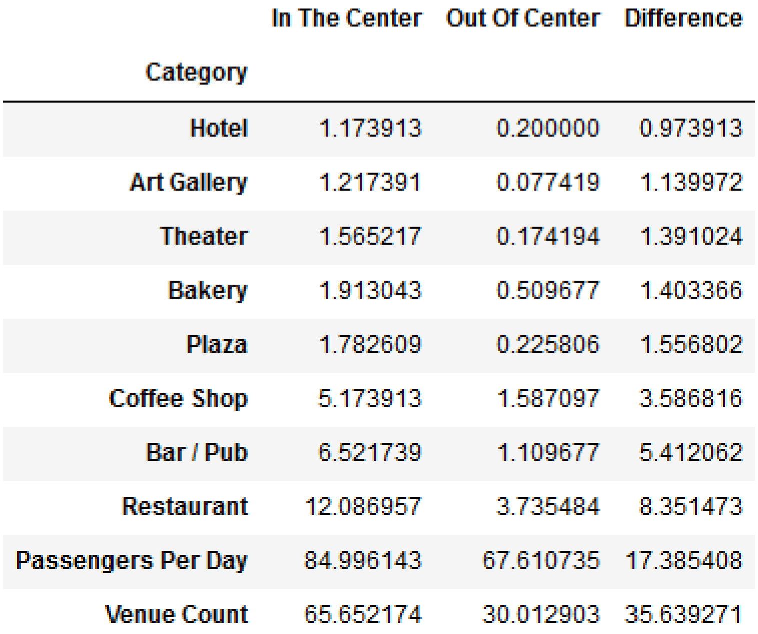
3.1 Exploratory Data Analysis

Historically, Moscow has a circular structure with the Kremlin in the center. Therefore, I decided to add parameter “**Distance To Center**” to the final table which shows the distance from the metro station to the center of Moscow and split the metro stations into two groups. Those that are within a radius of **3 km** to the center and the rest. Then I calculated the average values of all columns of the final table for each group and created a column “**Difference**”, that shows the difference between the average values of the two groups. Thereafter I sorted the table by field **'Difference'**. Let’s take a look at the head of the result table.



	In The Center	Out Of Center	Difference
Category			
Distance To Center	1850.259424	11002.796655	-9152.537231
Year	1957.478261	1985.580645	-28.102384
Elevation	-35.943478	-18.217361	-17.726117
Supermarket	0.130435	0.870968	-0.740533
Fast Food	0.565217	1.219355	-0.654137
Pet Store	0.000000	0.380645	-0.380645
Health Food Store	0.260870	0.638710	-0.377840
Pharmacy	0.304348	0.625806	-0.321459
Sporting Goods Shop	0.173913	0.419355	-0.245442
Paper / Office Supplies Store	0.000000	0.200000	-0.200000

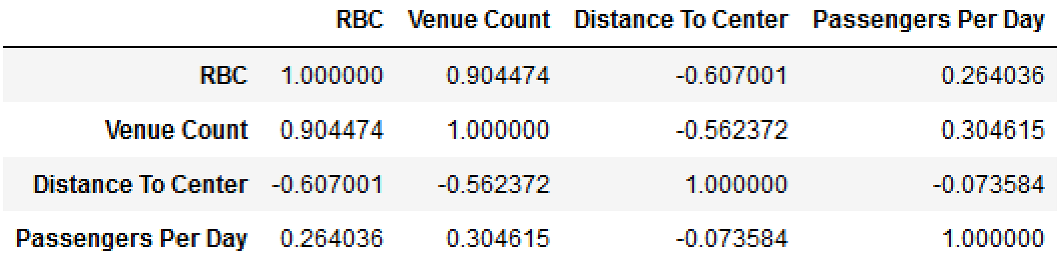
It's obvious that value for **Distance To Center** is smaller for the group in the center. The same is true for **Year** and **Elevation**. We also see that outside the center there are more venues of categories: **Supermarket, Fast Food, Pet Store, Health Food Store, Pharmacy**. These categories are really typical for residential areas. Now let's look at the tail of the table.



	In The Center	Out Of Center	Difference
Category			
Hotel	1.173913	0.200000	0.973913
Art Gallery	1.217391	0.077419	1.139972
Theater	1.565217	0.174194	1.391024
Bakery	1.913043	0.509677	1.403366
Plaza	1.782609	0.225806	1.556802
Coffee Shop	5.173913	1.587097	3.586816
Bar / Pub	6.521739	1.109677	5.412062
Restaurant	12.086957	3.735484	8.351473
Passengers Per Day	84.996143	67.610735	17.385408
Venue Count	65.652174	30.012903	35.639271

We see the obvious that **Venue Count** and **Passengers Per Day** are greater for a group in the center. Of the categories of venues, there are 3 that are very dominant in the center: **Restaurant**, **Bar / Pub**, **Coffee Shop**. These three categories are similar in meaning and I decided to combine these three categories into one and name it **RBC** by the first letters of each category.

Let's take a look at correlations between **RBC, Venue Count, Distance To Center, Passengers Per Day**.

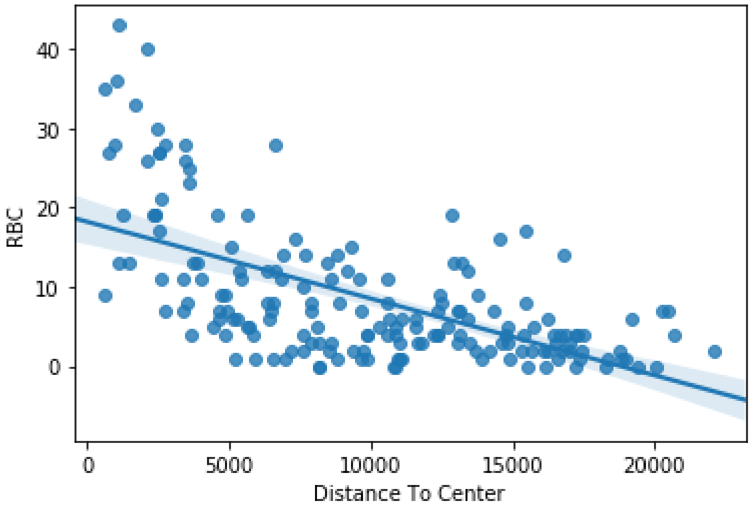


	RBC	Venue Count	Distance To Center	Passengers Per Day
RBC	1.000000	0.904474	-0.607001	0.264036
Venue Count	0.904474	1.000000	-0.562372	0.304615
Distance To Center	-0.607001	-0.562372	1.000000	-0.073584
Passengers Per Day	0.264036	0.304615	-0.073584	1.000000

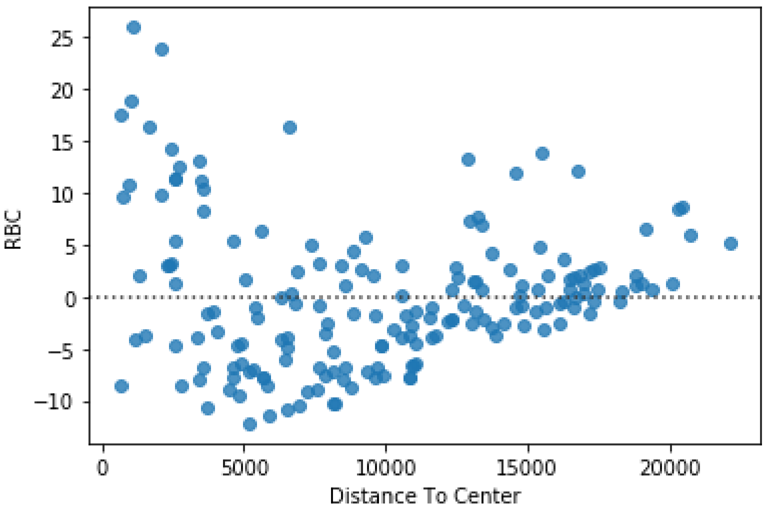
The strong correlation between **RBC** and **Venue Count** is obvious and of no interest to us. We can see a negative correlation (-0.6) between **RBC** and **Distance To Center**, which requires our further investigation. **Passengers Per Day** variable is most correlated with **Venue Count**, but the correlation is weak (0.3). I think this relationship also requires our further research.

3.2 Regression between “RBC” and “Distance To Center”

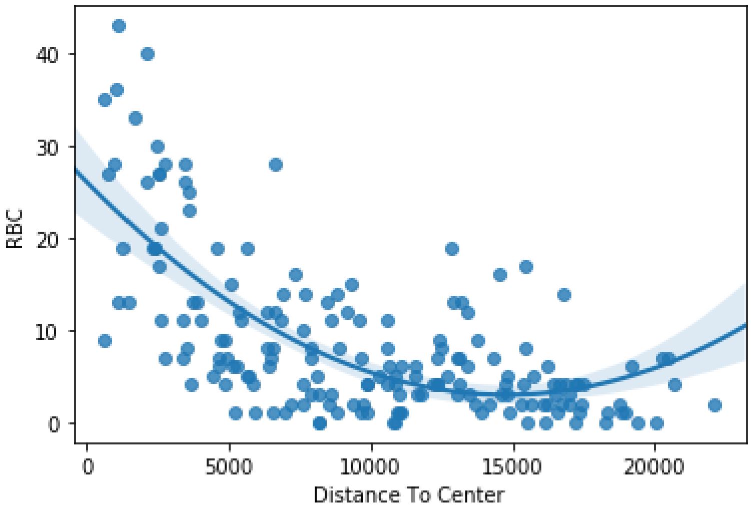
Let's try to predict the number of venue categories **RBC** (**R**estaurant, **B**ar / Pub, **C**offee Shop) by the distance of the metro station to center. Firstly, look at linear regression.



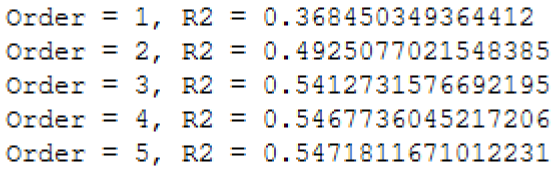
Now, let's take a look at the residual plot.



As we can see the residual is especially noticeable on the left and on the right, most likely the dependence is non-linear. Let's see the second order relationship.

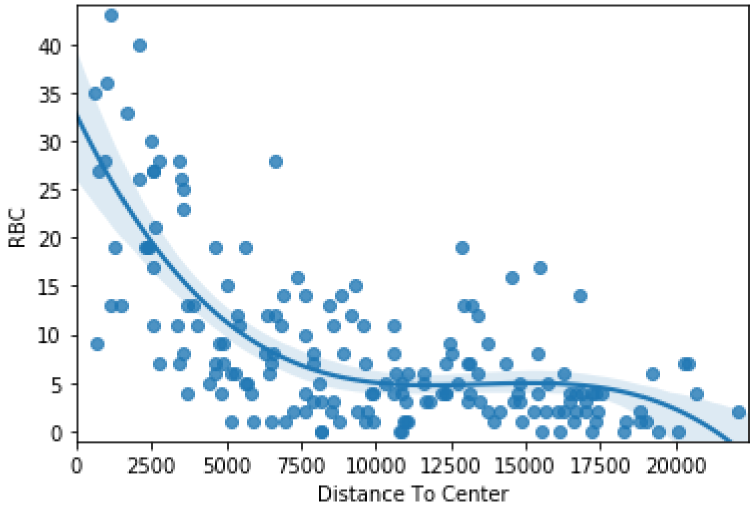


It looks much better, but let's make calculations and check.



```
Order = 1, R2 = 0.368450349364412
Order = 2, R2 = 0.4925077021548385
Order = 3, R2 = 0.5412731576692195
Order = 4, R2 = 0.5467736045217206
Order = 5, R2 = 0.5471811671012231
```

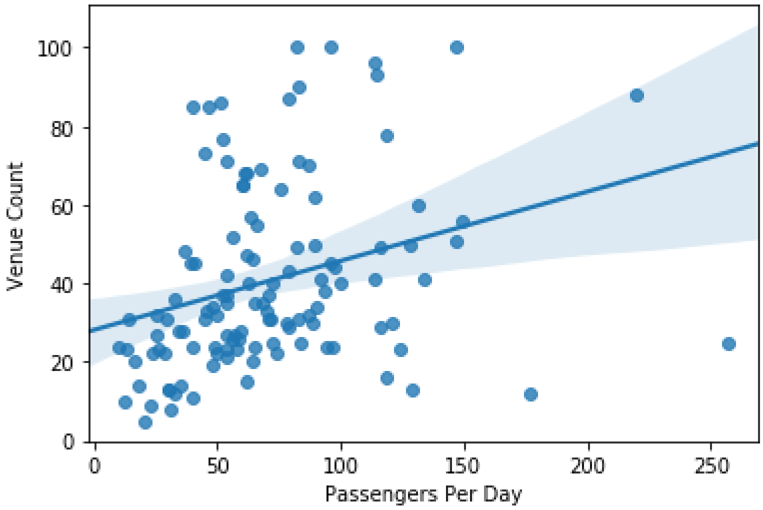
As you can see, order 3 is best suited. Further order increasing is not helpful.



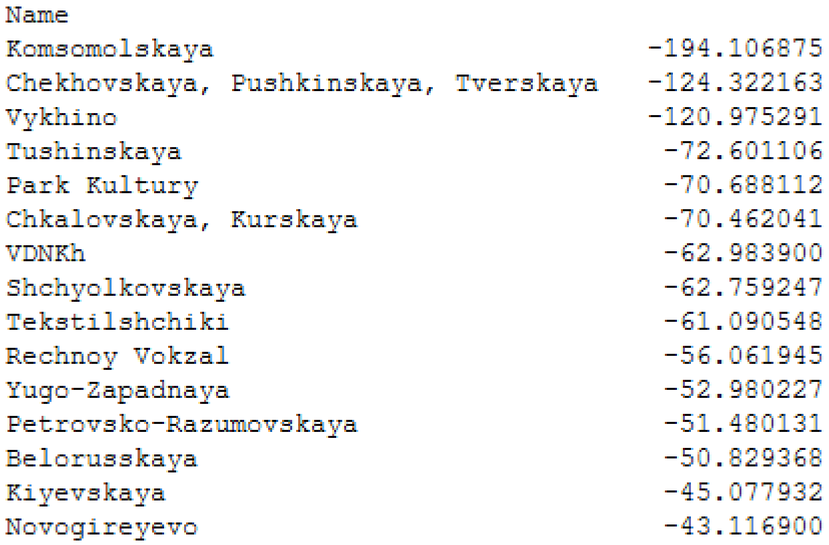
Now, we can get the list of stations near which the number of venues of the RBC category is much smaller than what the model predicts.

3.3 Regression between Venue Count and Passengers Per Day

Consider the graph of linear regression between **Passengers Per Day** and **Venue Count.**

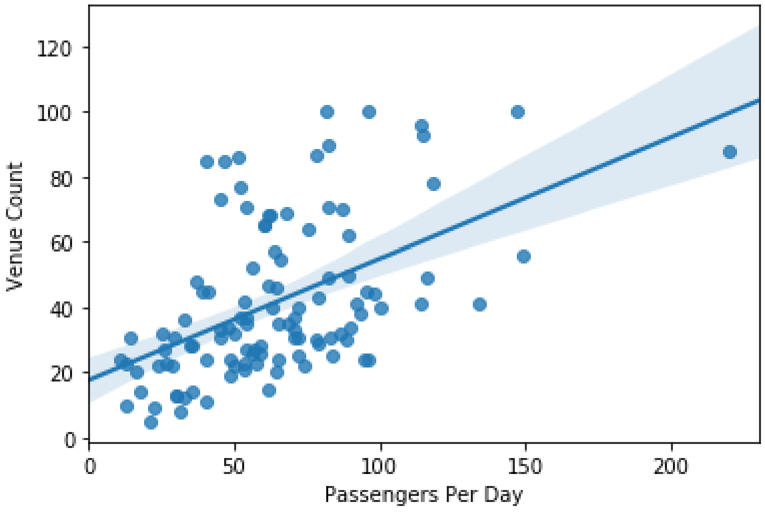


Correlation = 0.3, which is very small. There is a number of outliers. Let's calculate them.



```
Name
Komsomolskaya                                -194.106875
Chekhovskaya, Pushkinskaya, Tverskaya        -124.322163
Vykhino                                      -120.975291
Tushinskaya                                   -72.601106
Park Kultury                                  -70.688112
Chkalovskaya, Kurskaya                        -70.462041
VDNKh                                         -62.983900
Shchyolkovskaya                               -62.759247
Tekstilshchiki                                -61.090548
Rechnoy Vokzal                                -56.061945
Yugo-Zapadnaya                                -52.980227
Petrovsko-Razumovskaya                        -51.480131
Belorusskaya                                  -50.829368
Kiyevskaya                                    -45.077932
Novogireyevo                                  -43.116900
```

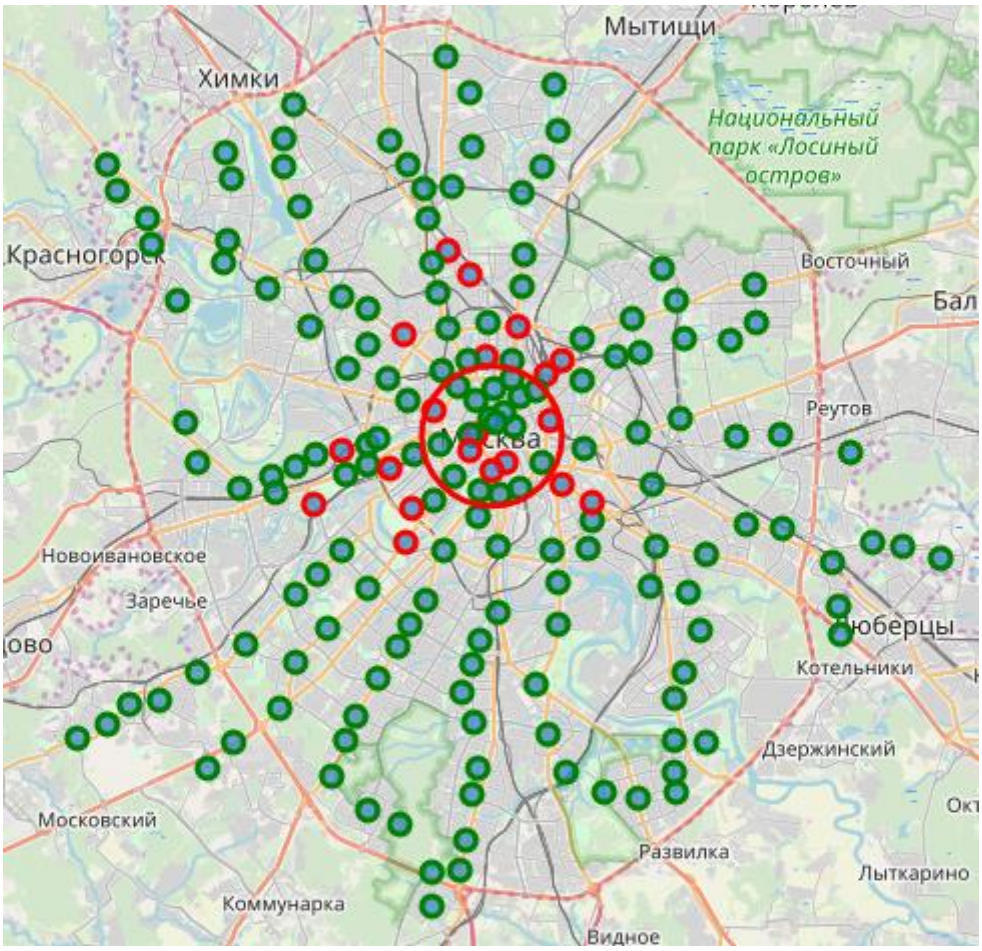
In this list we can see three stations near the railway stations of Moscow - **Komsomolskaya, Chkalovskaya, Kurskaya, Belorusskaya.** Obviously, the huge traffic at these stations is connected with the railway stations, and not with the fact that the people live or work there, which means that the number of venue should be less than for ordinary stations. The list also has the largest transport hubs. Moscow has a huge working intraday migration. Every day, millions of people from Moscow suburbs come to Moscow to work and return back in the evening. Therefore, a number of stations that are located on the outskirts of Moscow, or are located next to electric train platforms, are have huge passenger traffic. At the same time, people at these stations turn out to be in transit and do not create demand for venues in their vicinity. So I decided to drop these stations.



Now the correlation is **0.51**, which is better than the initial **0.3**, but still weak. I hypothesized that these two variables are more closely related, but this is not so.

4. Results and Discussion

In this study, I managed to find only one strong dependence of the metro station parameters on the distance to the center of Moscow. Since Moscow has a circular structure, the distance to the center is a good characteristic for metro stations. Especially good the distance to the center correlates with the number of venues of category **RBC** (**R**estaurant, **B**ar / Pub, **C**offee Shop). The cubic regression between these variables is the main result of this study. If a businessman plans to open a new business of this category in Moscow near the metro station, he should first look at the stations from the **goodForRbc** list. Because near these stations, the number of venue RBC categories is much less than the model predicts, which means that maybe they are not enough there. Let's look at the map. Stations from the **goodForRbc** list are marked in **red**.



In addition, you can see that the stations in the center are much deeper and older than those outside the center. This is logical, because initially the metro was built in the center and only then grew as the city grew. And the depth of the original stations in the center is due to the fact that they still had the function of bomb shelters.

5. Conclusion

Purpose of this project was to identify metro stations in the vicinity of which it is better to open a new business in order to aid stakeholders in narrowing down the search for optimal location. I managed to find the cubic regression between the number of venues of category RBC (Restaurant, Bar / Pub, Coffee Shop) and the distance to the center. Stakeholders should use the list of metro stations based on this dependence.

Final decision on optimal location will be made by stakeholders based on specific characteristics of metro station, taking into consideration additional factors like attractiveness of each location (proximity to park or water), levels of noise / proximity to major roads, real estate availability, prices, social and economic dynamics etc. It also seems to me to be useful to use the ratio between "Venue Count" and "Passengers Per Day".