Lomonosov Moscow State University

Economics Department

Research topic:

«Moscow real estate: pricing analysis through the prism of statistics and machine learning»

Authors:

Stepan Smirnov (student) ssm1508@mail.ru

Vladimir Tlostanov (student) pichpik@yandex.ru

Moscow 2019

Table of contents

Abstract

This paper provides an extensive statistical analysis of Moscow's real estate market. The main feature of the work – representation of the main variables on the map of Moscow, which allows a better understanding of the situation on the market. Moreover, it provides a correlation and cluster analysis, which show the main drivers of the cost of an apartment and types of flats in Moscow. In the section of decreasing the dimensions, it is shown which main factors influence the cost of apartments in Moscow and which of them make the greatest contribution to the cost of apartments. In the last section, 4 price forecasting models were built, and it turned out that the "random forest" model has the best predictive ability in the Moscow real estate market

Keyword: real estate, property market, city economy, housing market

JEL Codes: C02, C55, R31, O18,

Introduction

The purpose of this work is to study pricing in the real estate market in Moscow. Demand for residential space in the capital of Russia is now growing and hence the prices on the real estate market are rising as well. However, this growth can hardly be called linear or uniform. This paper sets its objective in identifying and studying factors affecting pricing on the Moscow apartment market.

Thus, the objective of this study is Moscow real estate market. The subject – pricing on this market.

Challenges of this work are:

- To examine the preliminary descriptive statistics of the available data
- Carry out a correlation analysis of the real estate market and identify the most influencing factors on the price
- Perform cluster analysis
- Conduct a dimensionality reduction procedure using factor analysis as well as PCM
- Predict housing prices in Moscow

The data used in this work is an array of data of Sberbank: "Sberbank Russian Housing Market", designed just for studying and predicting the price of apartments in the capital. This dataset is incredibly comprehensive because it contains 292 variables for more than 30,000 observations. It takes into account a variety of factors that can affect the price of the apartment, such as, for example, the distance to the 3 transport rings of Moscow, to the Kremlin, the area of green and industrial zones around, the year of construction of the house, floor, the presence of nearby cultural sites, universities, the presence of schools/kindergartens and their workload, as well as many others. Despite a certain number of typos and omissions, this array is a very exhaustive and accurate source of information to study the issue of our interest.

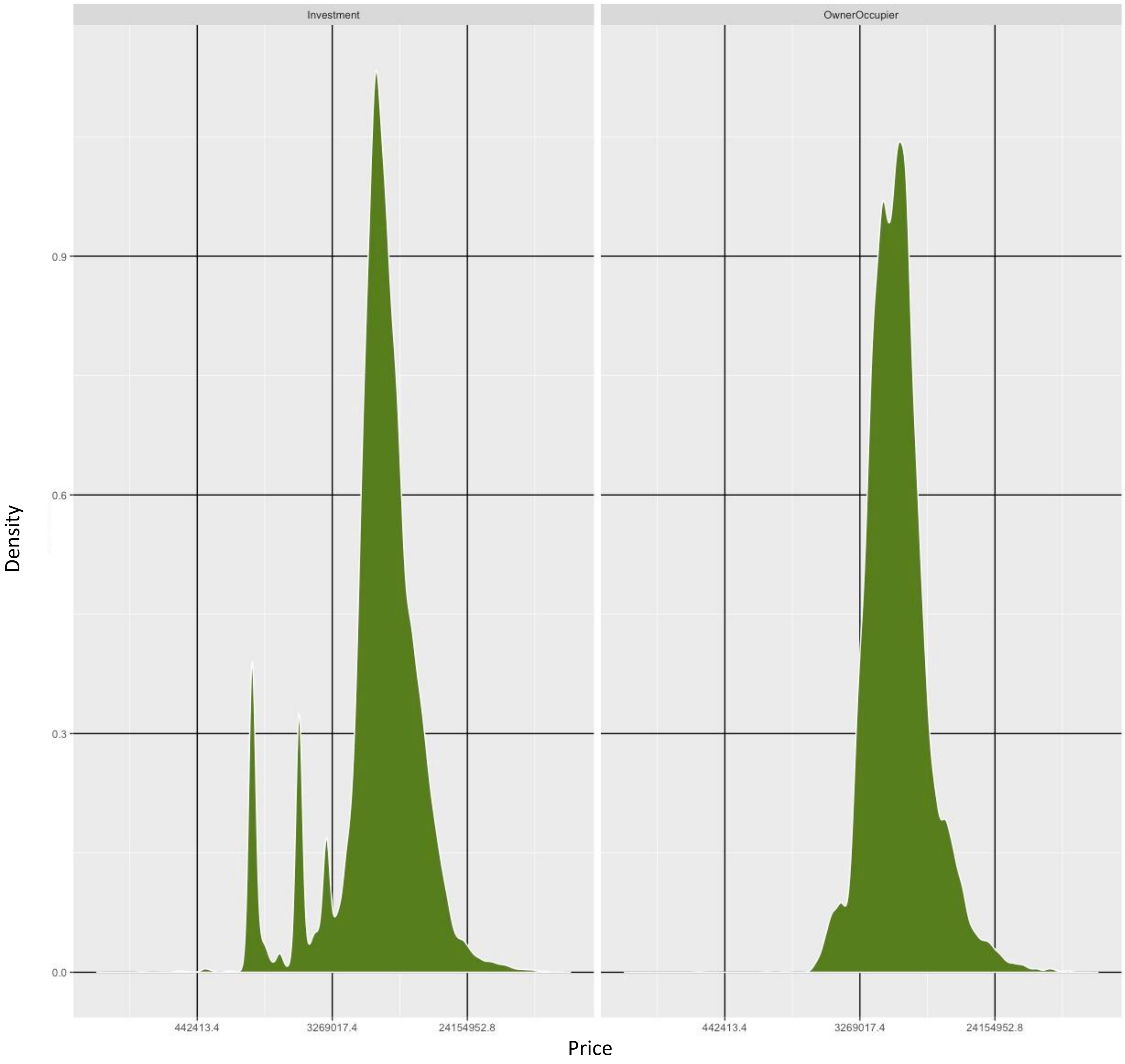
We should make a note that before the very beginning the available data was slightly fixed: first of all, all the conspicuous typo (such as the year of 2210, etc.) were removed; secondly, some custom variables that are of particular importance for the study were constructed and added to the entire array; also every factor variable were transformed into numeric format; moreover all the gaps were fulfilled using the EM-algorithm from “Amelia” package.

Descriptive statistics

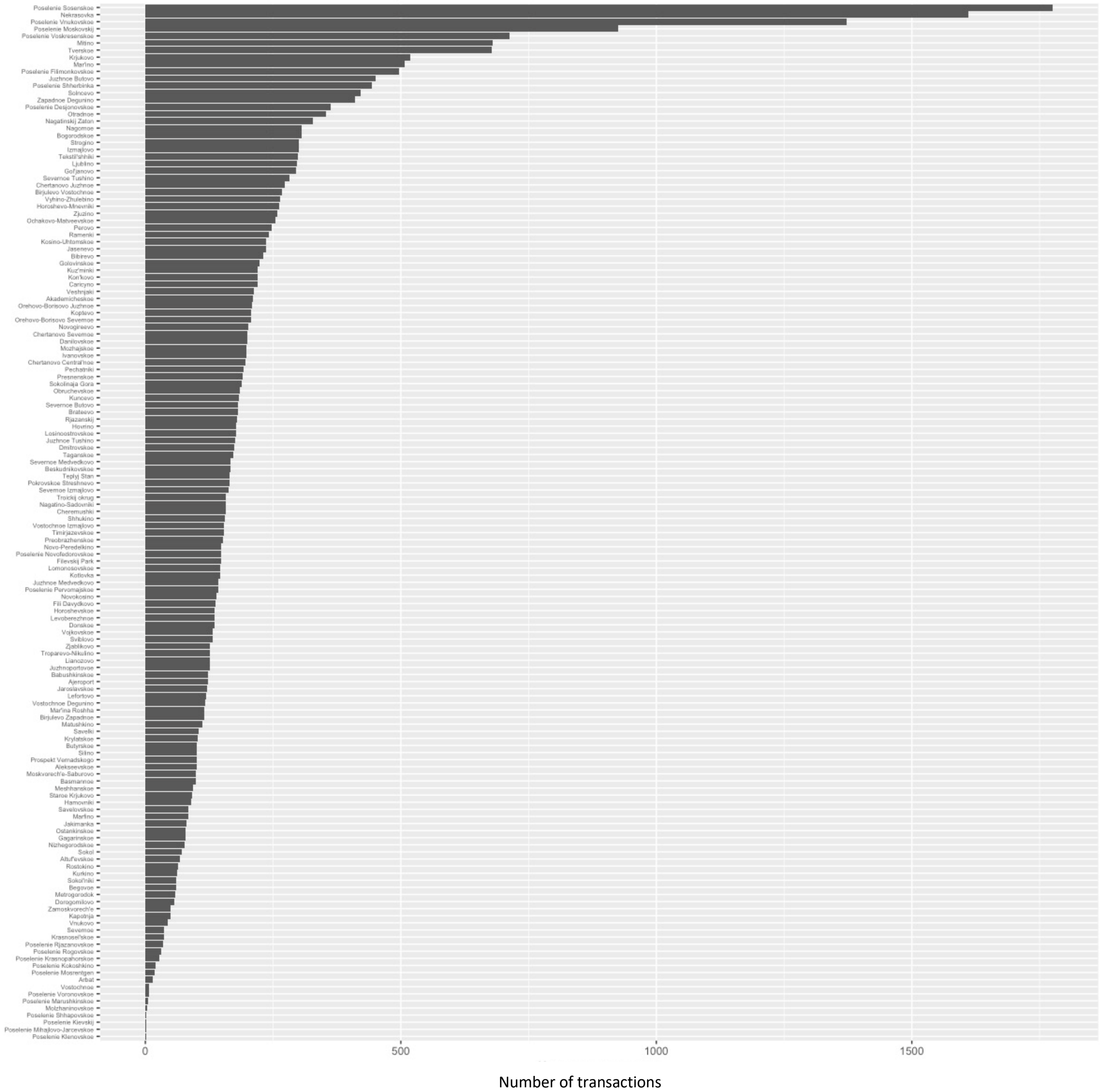
In the very beginning of this study, we examine a couple of graphs describing the data available. Firstly, we want to check whether the prices in private and in commercial sector are different. Let's have a look at the distribution of flats prices in the "Investment" and "Own possession" sectors.

As we can see the distributions are quite similar, however in the commercial sector the maximum price is higher and there're also some local peaks. At the same time the distribution of private sector is much smoother.

Distribution of prices



Number of transactions by the areas

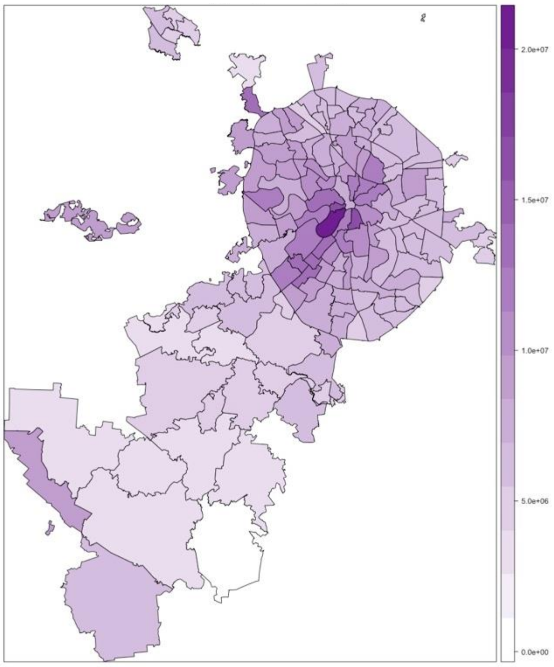


Secondly, we can glimpse at the distribution of the number of purchasing contracts by the areas of the city. We can see that the definite leaders here are the areas of so-called “New Moscow”. However, the absolute outsiders of this rating are also recently attached territories. Among such territories, one can even see an area with 0 transaction. The rest of the areas are related to the “Old Moscow” and has a market volume of no more than 500 transactions.

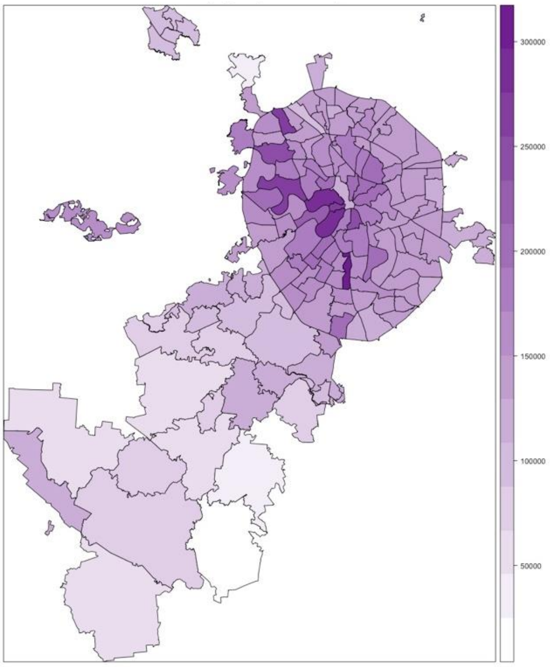
Also, before we start the main part of this research, we would like to give a visual descriptive statistic. These statistics are for some variables of interest for us and also some variables which describe the situation in the city and can influence apartment prices.

We start with the distribution of average prices of entire flats and prices of square meter lived by the areas of Moscow. The charts clearly show that the most valuable property in the center of Moscow. We can also see a relatively high price in the West and South-West of the city.

Average price per flat



Average price per square



In the North and East, as well as in the areas of "New Moscow", there're the lowest prices.

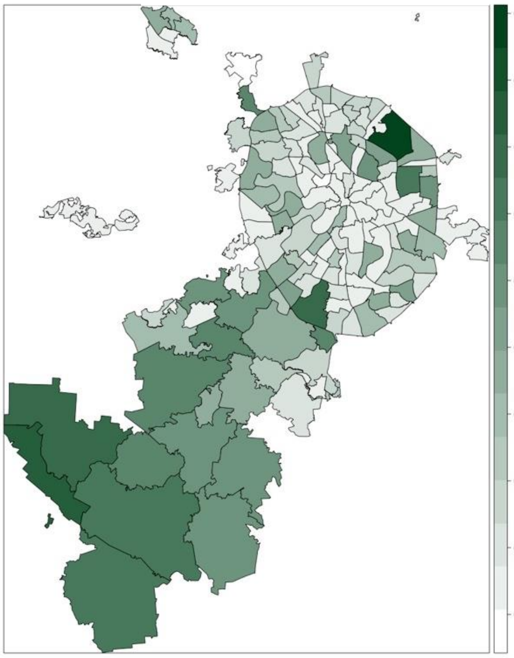
If one takes a look at the distribution of green and industrial zones around the city, one can see that concentration of green spaces is highest in most remote from "MKAD"[1] areas and also on the territory of the reserve "Losinyy Ostrov"[2]. The rest of the areas have the proportion of the green spaces of the total area between 10% and 40%. One can also see that almost every area inside Sodovoe ring[3] has the lowest green rating.

[1] "MKAD" stands for Moscow Circular Auto Road in English

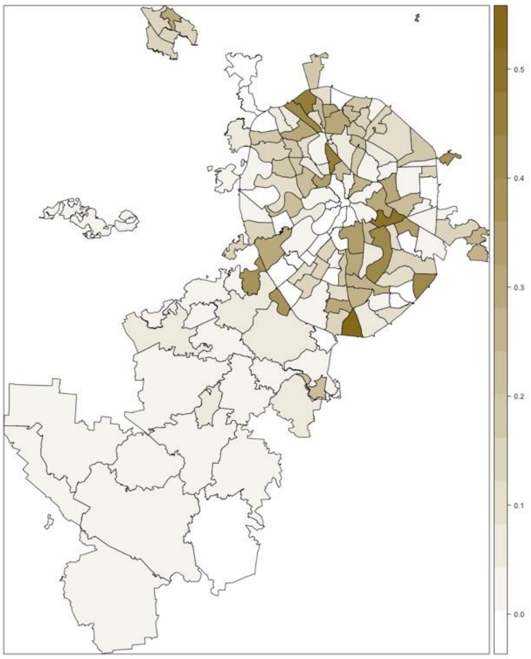
[2] "Losinyy ostrov" stands for The Moose Island in English

[3] Here means the circular road inside the city, Garden ring in English

Green spaces

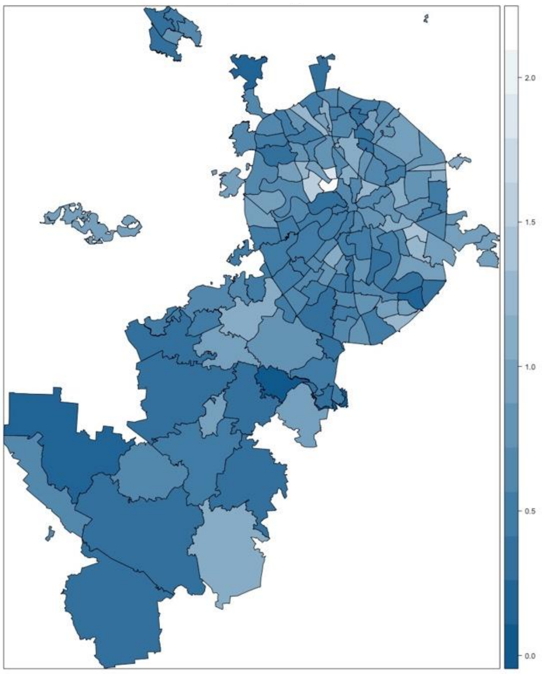


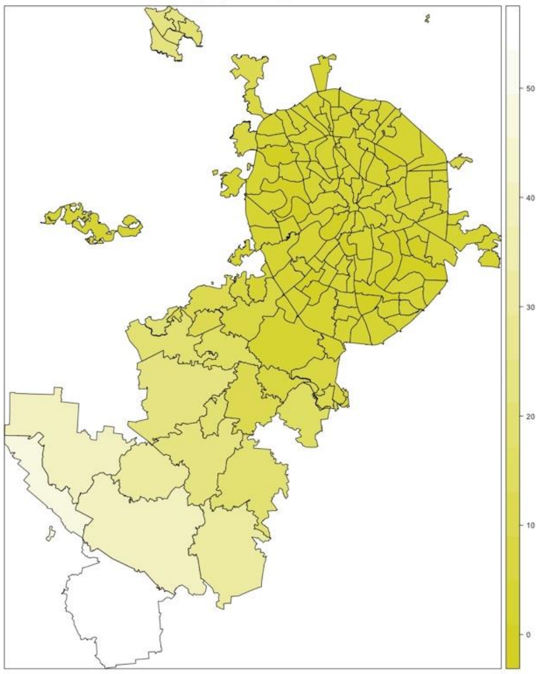
Industrial areas



Now look at the distribution of industrial zones. In New Moscow, their concentration is minimal. In the very center there are no such ones at all. In the remaining areas, a uniformly distributed variation of this indicator can be observed, with values ranging from 0 to 50%.

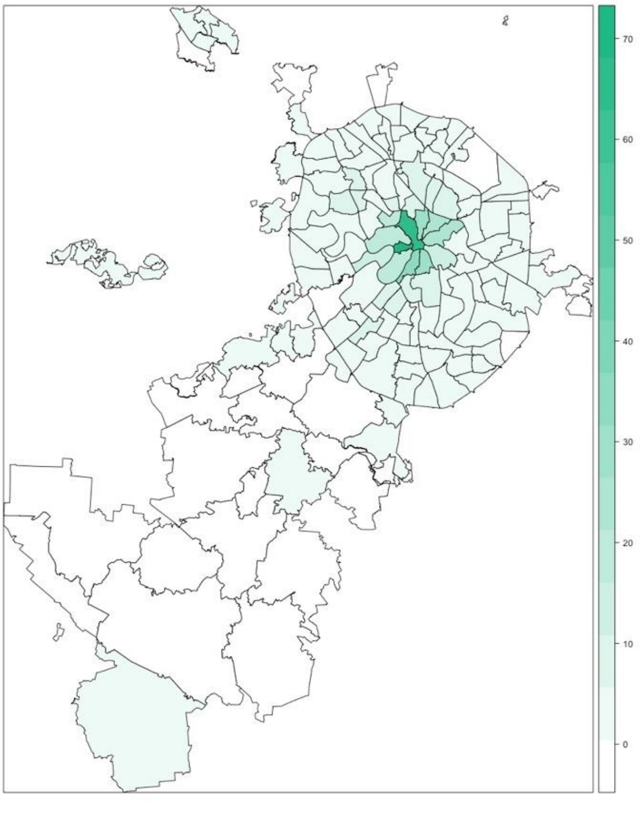
Now we give an illustration of the saturation of areas with water bodies and sources of radiation. When calculating these indicators, proxies were used - the distance to the nearest water body and the source of radiation, respectively. So, you can see that the radiation level is distributed throughout Moscow as homogeneous as possible and, on average, in each region the nearest source of radiation is within a radius of several kilometers. This indicator weakens only at a very significant distance from the Moscow Ring Road.


Water treatment


Radiation

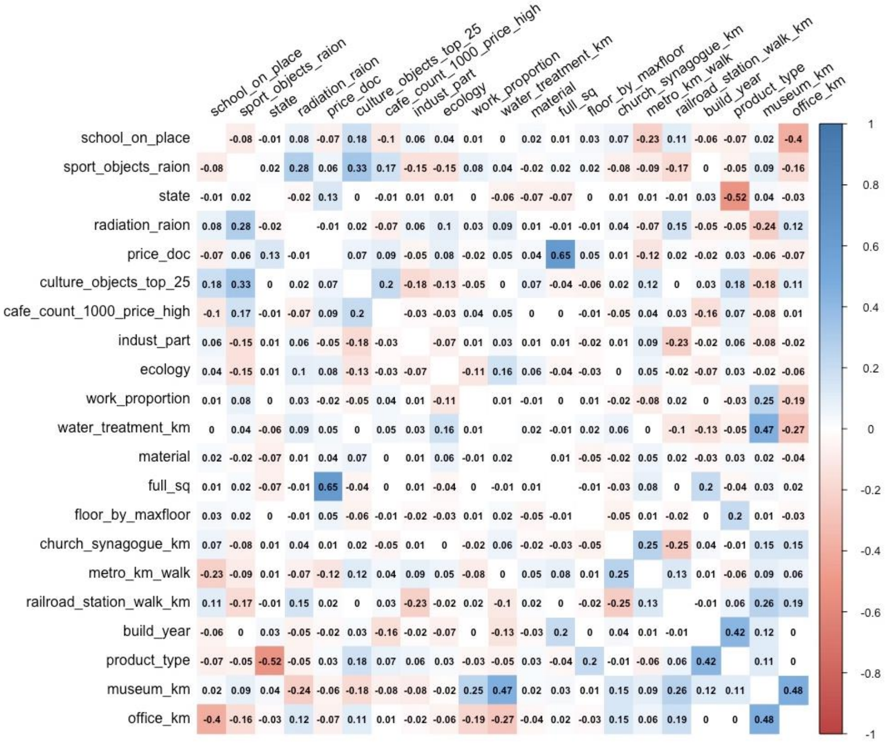
As for saturation with water sources, in almost every area the reservoir is on average no more than 1.5 kilometers from the residential sector.

At the end of this section, we present the saturation of cafe areas within walking distance. You can see that in the city there are certain problems with the presence of a cafe near the house. Virtually the entire city, except for the center itself, does not abound with such leisure facilities.


Cafes

Correlation analysis

On the graphs below you can observe correlograms constructed by partial correlation coefficients. Correlation analysis was carried out on a limited set of variables, consisting of only 21 units, including the apartment price itself. Such a decision was made based primarily on the consideration that it would be extremely difficult to consider and interpret 292 variables, and quite time consuming. Therefore, we selected two sets of variables - the base one, consisting of 21 variables (their description you can see in Appendix), and the extended one, consisting of 55. These sets were compiled by excluding from the initial list of variables those that were most correlated with those similar in meaning, and, based on logical considerations. That is, a kind of dimension reduction procedure was performed. For the correlation analysis, a basic set is used because such a quantity is much easier to display on the correlogram and interpret.

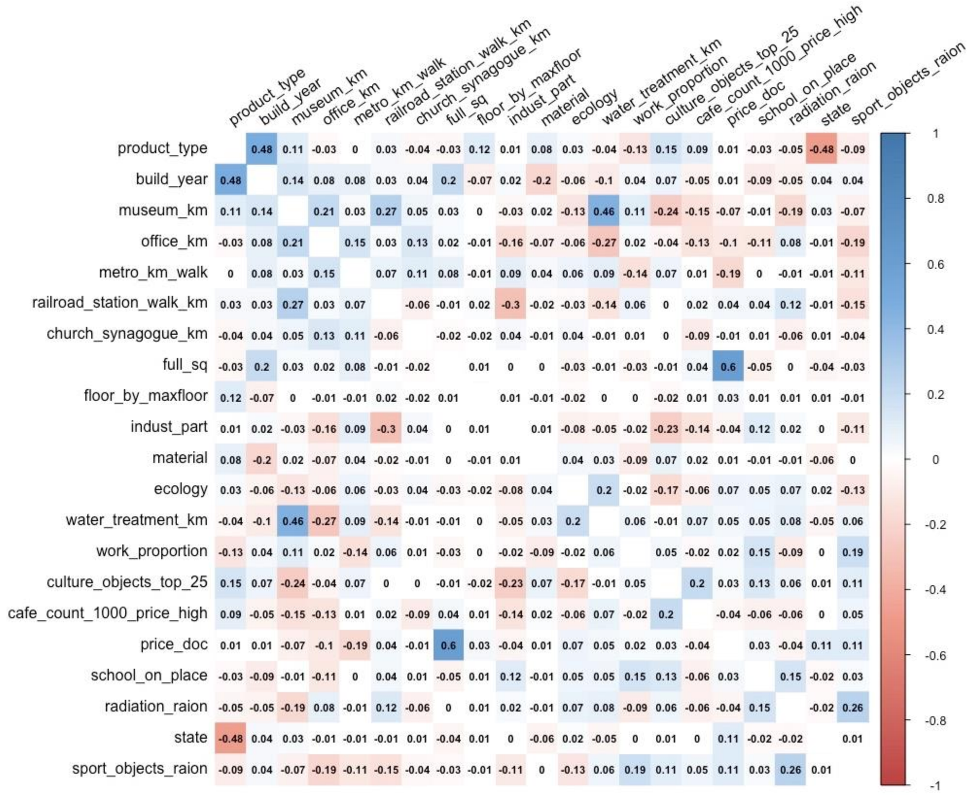


	school_on_place	sport_objects_raion	state	radiation_raion	price_doc	culture_objects_top_25	cafe_count_1000_price_high	indust_part	ecology	work_proportion	water_treatment_km	material	full_sq	floor_by_maxfloor	church_synagogue_km	metro_km_walk	railroad_station_walk_km	build_year	product_type	museum_km	office_km
school_on_place		-0.08	-0.01	0.08	-0.07	0.18	-0.1	0.06	0.04	0.01	0	0.02	0.01	0.03	0.07	-0.23	0.11	-0.06	-0.07	0.02	-0.4
sport_objects_raion	-0.08		0.02	0.28	0.06	0.33	0.17	-0.15	-0.15	0.08	0.04	-0.02	0.02	0.02	-0.08	-0.09	-0.17	0	-0.05	0.09	-0.16
state	-0.01	0.02		-0.02	0.13	0	-0.01	0.01	0.01	0	-0.06	-0.07	-0.07	0	0.01	0.01	-0.01	0.03	-0.52	0.04	-0.03
radiation_raion	0.08	0.28	-0.02		-0.01	0.02	-0.07	0.06	0.1	0.03	0.09	0.01	-0.01	-0.01	0.04	-0.07	0.15	-0.05	-0.05	-0.24	0.12
price_doc	-0.07	0.06	0.13	-0.01		0.07	0.09	-0.05	0.08	-0.02	0.05	0.04	0.65	0.05	0.01	-0.12	0.02	-0.02	0.03	-0.06	-0.07
culture_objects_top_25	0.18	0.33	0	0.02	0.07		0.2	-0.18	-0.13	-0.05	0	0.07	-0.04	-0.06	0.02	0.12	0	0.03	0.18	-0.18	0.11
cafe_count_1000_price_high	-0.1	0.17	-0.01	-0.07	0.09	0.2		-0.03	-0.03	0.04	0.05	0	0	-0.01	-0.05	0.04	0.03	-0.16	0.07	-0.08	0.01
indust_part	0.06	-0.15	0.01	0.06	-0.05	-0.18	-0.03		-0.07	0.01	0.03	0.01	0.01	-0.02	0.01	0.09	-0.23	-0.02	0.06	-0.08	-0.02
ecology	0.04	-0.15	0.01	0.1	0.08	-0.13	-0.03	-0.07		-0.11	0.16	0.06	-0.04	-0.03	0	0.05	-0.02	-0.07	0.03	-0.02	-0.06
work_proportion	0.01	0.08	0	0.03	-0.02	-0.05	0.04	0.01	-0.11		0.01	-0.01	0	0.01	-0.02	-0.08	0.02	0	-0.03	0.25	-0.19
water_treatment_km	0	0.04	-0.06	0.09	0.05	0	0.05	0.03	0.16	0.01		0.02	-0.01	0.02	0.06	0	-0.1	-0.13	-0.05	0.47	-0.27
material	0.02	-0.02	-0.07	0.01	0.04	0.07	0	0.01	0.06	-0.01	0.02		0.01	-0.05	-0.02	0.05	0.02	-0.03	0.03	0.02	-0.04
full_sq	0.01	0.02	-0.07	-0.01	0.65	-0.04	0	0.01	-0.04	0	-0.01	0.01		-0.01	-0.03	0.08	0	0.2	-0.04	0.03	0.02
floor_by_maxfloor	0.03	0.02	0	-0.01	0.05	-0.06	-0.01	-0.02	-0.03	0.01	0.02	-0.05	-0.01		-0.05	0.01	-0.02	0	0.2	0.01	-0.03
church_synagogue_km	0.07	-0.08	0.01	0.04	0.01	0.02	-0.05	0.01	0	-0.02	0.06	-0.02	-0.03	-0.05		0.25	-0.25	0.04	-0.01	0.15	0.15
metro_km_walk	-0.23	-0.09	0.01	-0.07	-0.12	0.12	0.04	0.09	0.05	-0.08	0	0.05	0.08	0.01	0.25		0.13	0.01	-0.06	0.09	0.06
railroad_station_walk_km	0.11	-0.17	-0.01	0.15	0.02	0	0.03	-0.23	-0.02	0.02	-0.1	0.02	0	-0.02	-0.25	0.13		-0.01	0.06	0.26	0.19
build_year	-0.06	0	0.03	-0.05	-0.02	0.03	-0.16	-0.02	-0.07	0	-0.13	-0.03	0.2	0	0.04	0.01	-0.01		0.42	0.12	0
product_type	-0.07	-0.05	-0.52	-0.05	0.03	0.18	0.07	0.06	0.03	-0.03	-0.05	0.03	-0.04	0.2	-0.01	-0.06	0.06	0.42		0.11	0
museum_km	0.02	0.09	0.04	-0.24	-0.06	-0.18	-0.08	-0.08	-0.02	0.25	0.47	0.02	0.03	0.01	0.15	0.09	0.26	0.12	0.11		0.48
office_km	-0.4	-0.16	-0.03	0.12	-0.07	0.11	0.01	-0.02	-0.06	-0.19	-0.27	-0.04	0.02	-0.03	0.15	0.06	0.19	0	0	0.48	

Partial Pearson Correlation Coefficients

So, based on the method of calculating the Pearson, it can be argued that the area of the apartment has the strongest influence on the price - their pair correlation is 65%. Also, a noticeable and significant positive correlation is present between the price and the condition of the apartment, as well as the environment and the number of cafes, cultural and sports facilities nearby. Among the factors that have a negative impact on the price, the distance to the metro is the strongest - the price correlates with this indicator - (-12%). Other distance factors have a negative connection with the price - the distance to the nearest museum and office. Also, the area of industrial zones and the downtroddenness of schools have a negative impact.

If we look at the Spearman correlation coefficients, we will see a practically similar picture - the strongest positive correlation with the area of the apartment, and the most negative - with the distance to the metro. Among the significant positive correlation is also the correlation with ecology, the state of housing and the presence of sports facilities. At the same time, there is a much lower correlation with the number of cultural objects, but there is a significant correlation with the distance to the recreation area near the water. The presence of a nearby café has a negative influence on the price now. Such things are not quite explainable from the point of view of everyday logic, and this is an alarm. If we look at the negative dependencies, here we can also distinguish the distance to the museum and the industrial zone. However, the absolute values of the coefficients are even lower here, than in the case of the Pearson technique. In general, the results obtained in this case appear much less logical and meaningful, so we tend to trust the Pearson correlations and assume the presence of a linear relationship.



Partial Spearman Correlation Coefficients

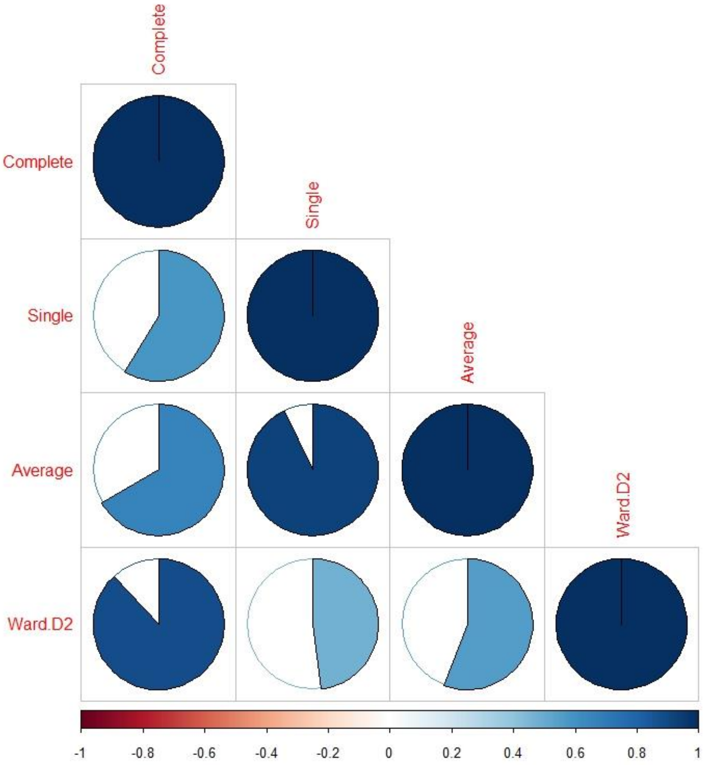
Cluster analysis

In our study, we tried to sort out all the apartments available on the date set for a certain number of classes in order to systematize the data and have an idea about their structure. The cluster analysis was conducted on an extended sample (out of 55 variables) and included two main methods: hierarchical clustering and the EM algorithm. In the beginning we consider the first one.

For hierarchical clustering, we used 4 different methodologies:

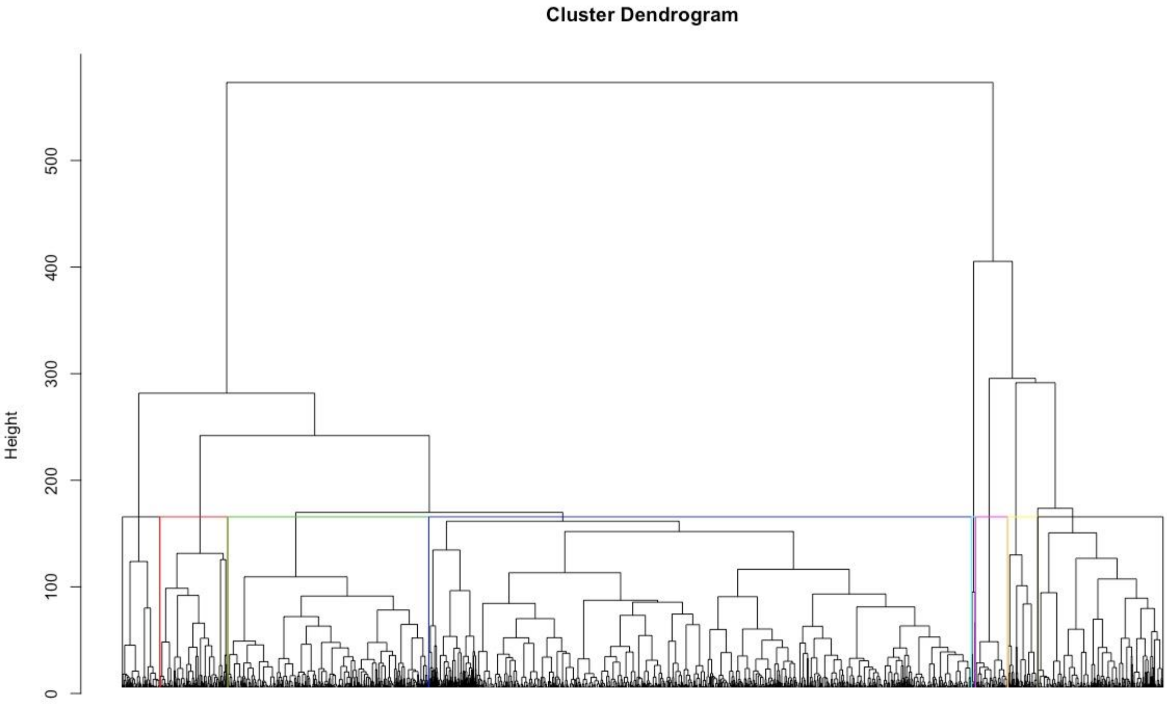
- WardD2
- Average
- Single
- Complete

To compare the obtained results, we constructed a special correlogram showing the similarity of the results obtained by each of the methods, which you can see below.



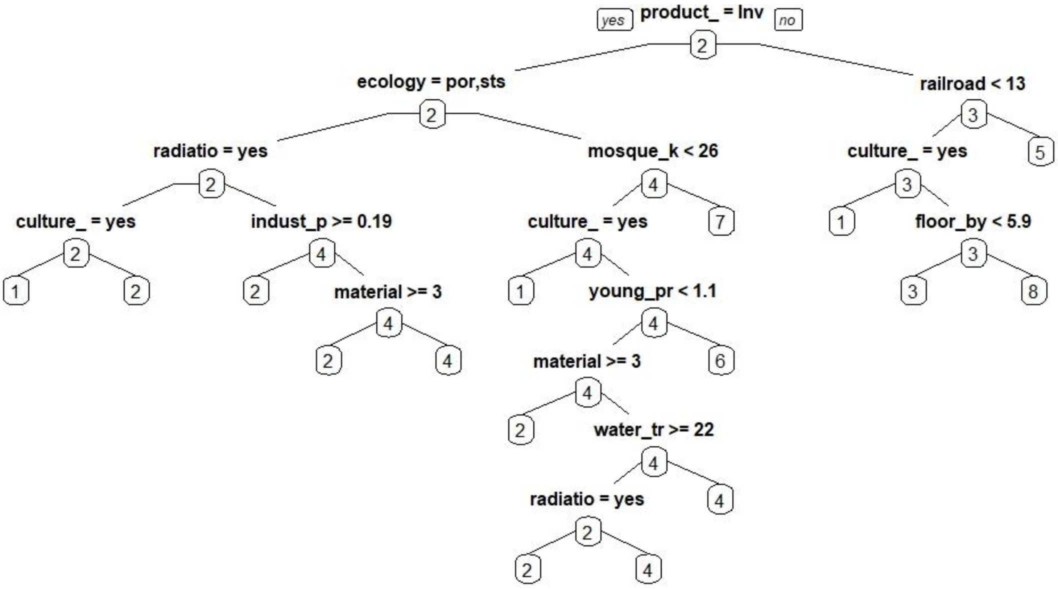
The results obtained by all 4 methods are similar by more than 40%. The results of the WardD2 and Complete methods, as well as Average and Single, are especially close. This result indicates that when the array is divided into clusters using the hierarchical clustering method, the type of procedure carried out does not matter too much.

On the dendogram below, you can see the resulting hierarchical structure with 9 selected groups. This choice of the number of clusters is due to the following logic: in Moscow there are 3 options for the location of housing - in the center, on the outskirts and between them, as well as 3 price categories - expensive, low-cost housing and housing of an average price category.



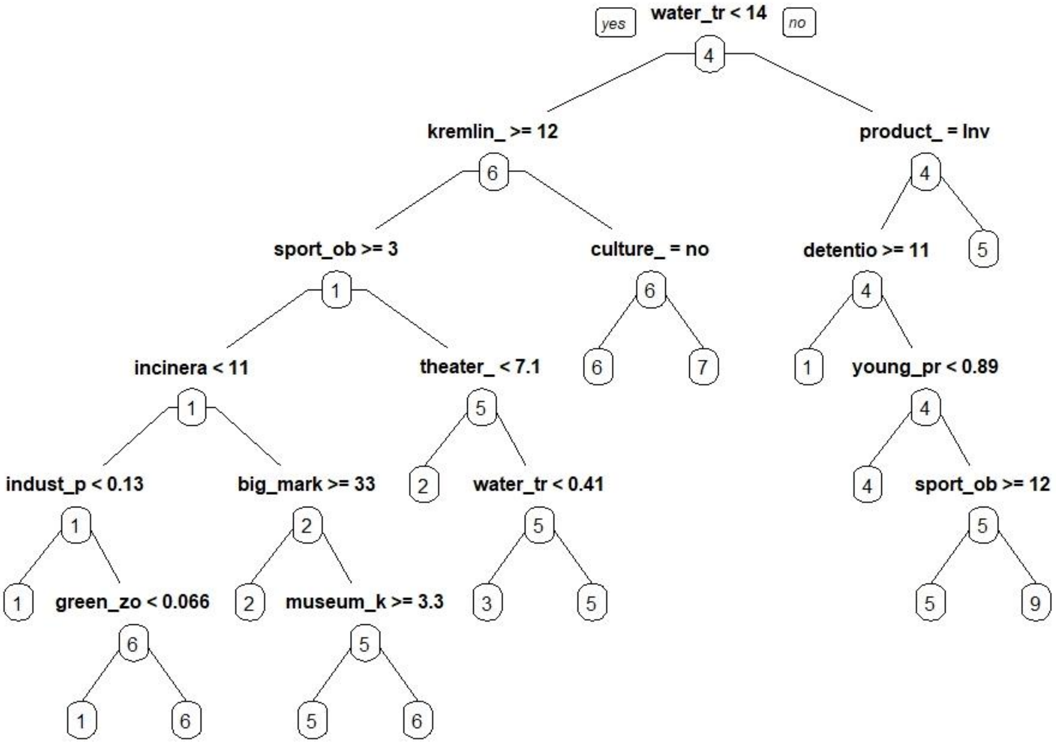
On the graph below there is a decision tree, describing the principle of the distribution of apartments into groups.

Decision tree for hierarchical clustering



As for EM clustering, this algorithm chooses the optimal number of clusters by itself, and in this case, there are also 9 of them, which indicates the validity of the chosen logic. Below you can see the decision tree, which helps with group interpretation. In the tree you can also see that groups are formed using very likely criteria.

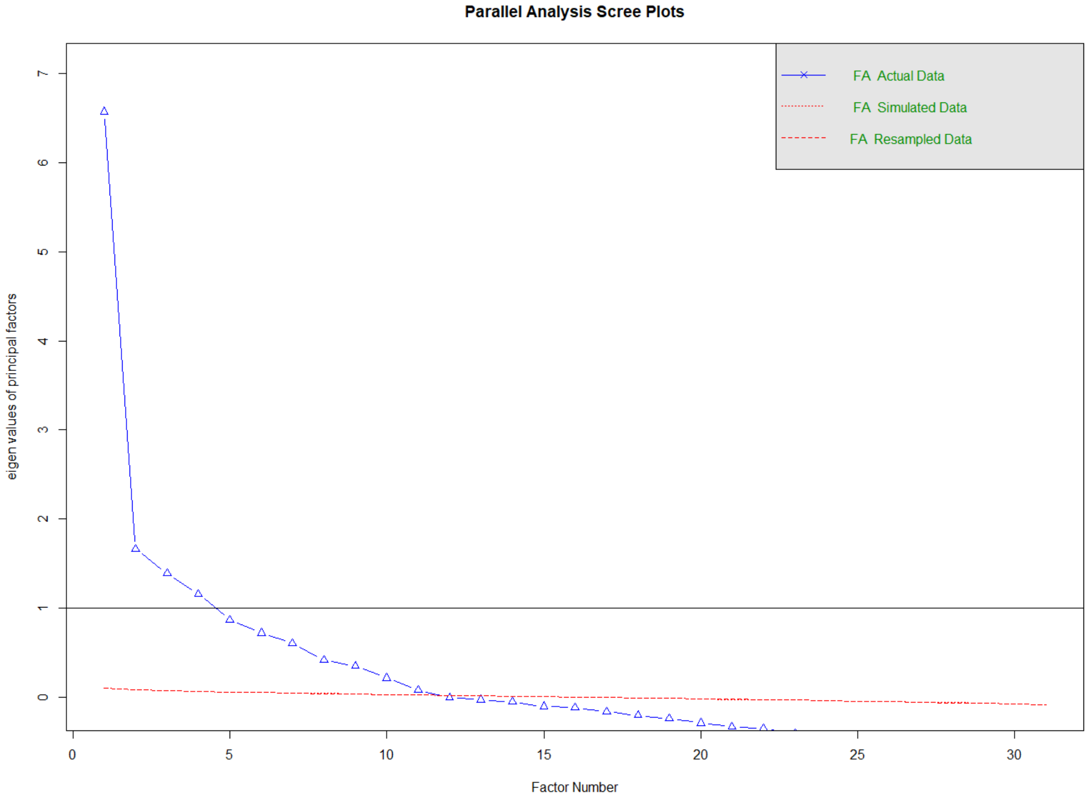
Decision Tree for EM Algorithm



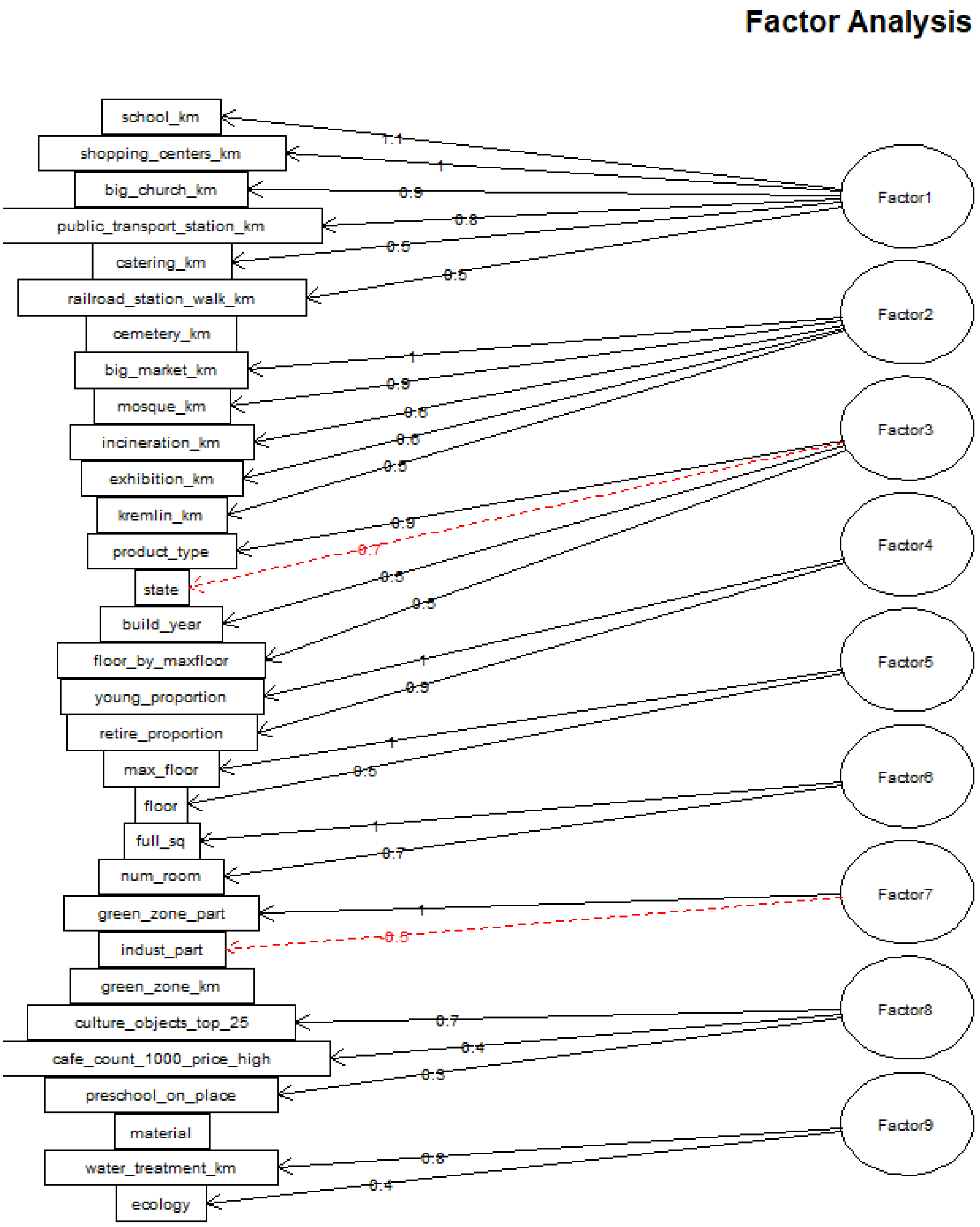
Dimension reduction

Factor analysis

Before we conduct factor analysis itself, we have to identify the optimal cluster number using the *fa.parallel* function. The result is presented in the graph below. Thus, we can clearly see that the optimal number fluctuates around 10, so we can safely choose 9 as we used in cluster analysis. It is necessary to note that factor analysis is conducted on the extended set of variables, though it is slightly adjusted. Thus, in purposes of calculation ease only poorly correlated variables (correlation below 65% threshold) were left.



After conducting the factor analysis, we obtained the following results:



Let's interpret them:

Factor 1 can be called as “The distance to objects of daily use” or as “The distance to the most important places”. Indeed, the factor includes variables such as distance to the nearest: subway station, public transport stop, railroad station, catering facilities, shopping center and even Orthodox church (this is probably because of the large number of Orthodox believers).

Factor 2 also includes variables of distance, but at this time it is the distance to much more rarely used places, we can even say symbolic places. Thus, this factor contains the following variables: distance to the Kremlin, distance to an exhibition complex, distance to a mosque (this is due to a comparatively small number of Muslims, we believe) and the distance to an incineration plant.

Factor 3 is “characteristics of the apartment”. It contains condition of the apartment, type of the property (commercial or own possession) and the house construction year.

Factor 4 is “Demography”. Proportion of the younger generation residents, as well as proportion of the elderly, residents are included in this factor.

Factor 5 is “The height of the apartment”. It includes the number of floors in the building and floor on which the flat is located itself.

Factor 6 is “Apartment size”. This factor contains the area of the apartment as well as the number of rooms.

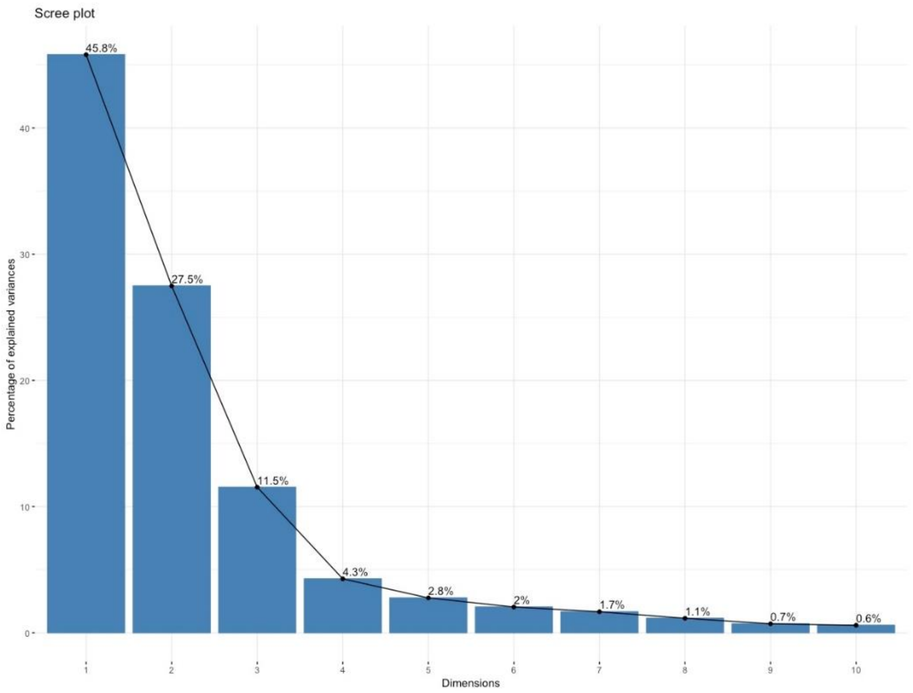
Factor 7 is “Living environment”. It includes the proportion of residential as well as industrial areas around the flat.

Factor 8 is “Recreation”. It contains the number of cafes within a kilometer, as well as the availability of popular cultural sites and attractions nearby.

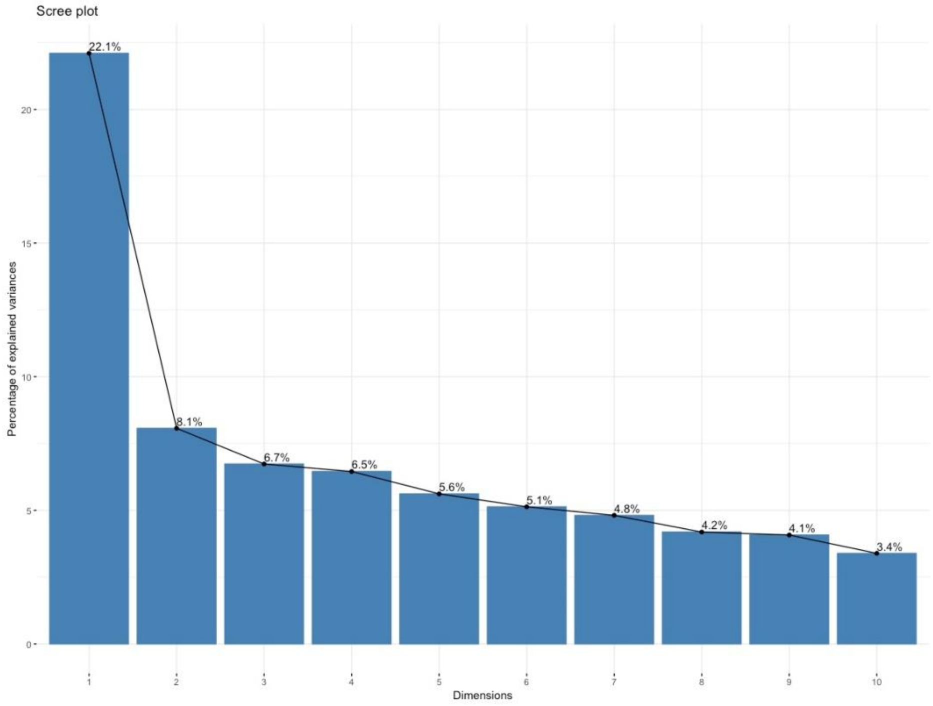
Factor 9 - "Natural factor". It consists of the ecological state in the area of the apartment and the distance to the nearest pond with the possibility of recreation.

Principal component method

By reducing the dimension by the method of principal components, we obtained 32 (as well as the number of explanatory variables) main components. The graph below shows the shares of the explanatory dispersion of each component. So, one can see that even the first 3 components explain 85% of the variation of the initial variables, and the first 7 explain 95%.

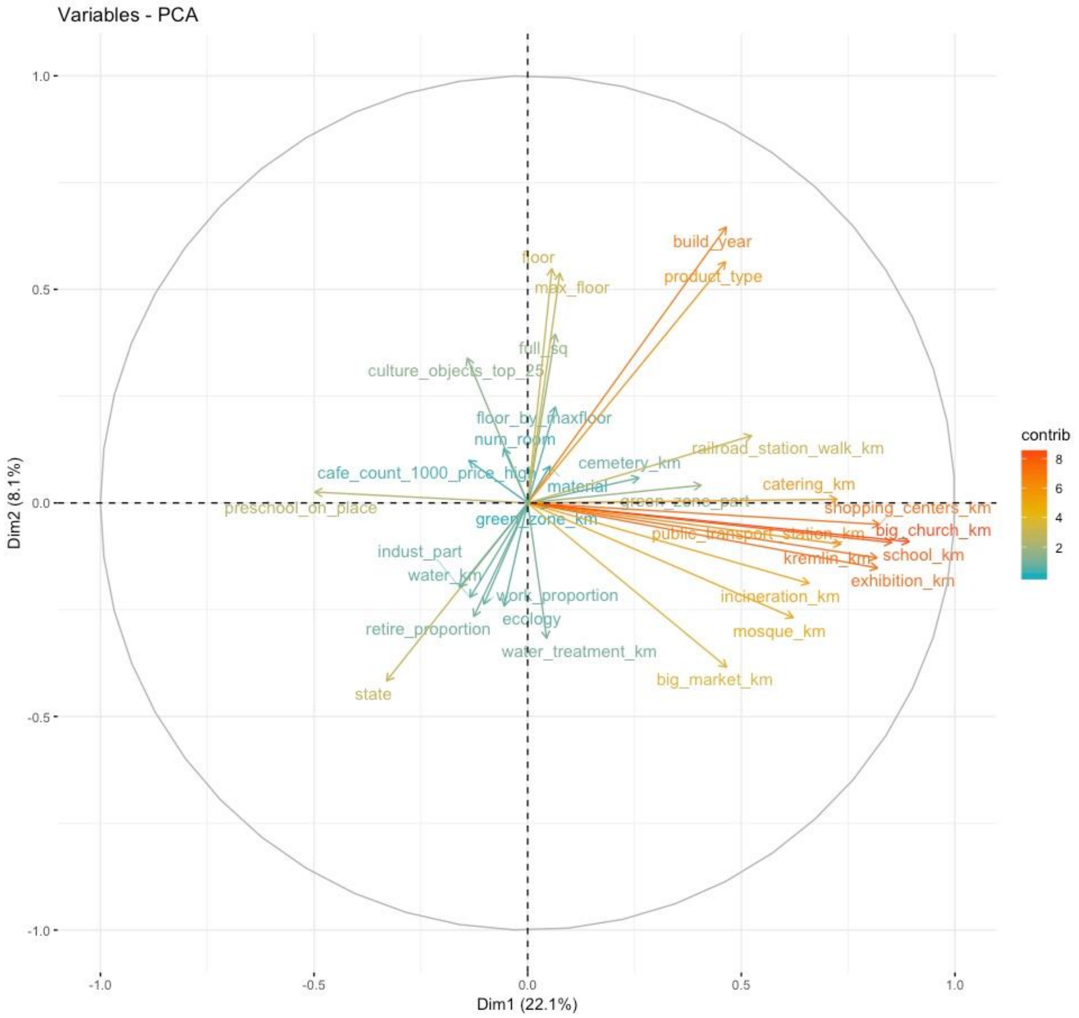


However, by building the principal components using a different function, we obtained different results. The variables still aggregate in 32 (which is a good sign) principal components, however the share of explanatory variance is distributed quite differently.



As we can see from the graph: in this case 85% of the variance explains as many as 16 principal components, and 95% - 22. Considering the heterogeneity and volume of data, we assume that results obtained from the second method are more logical, and therefore we conduct further analysis using them.

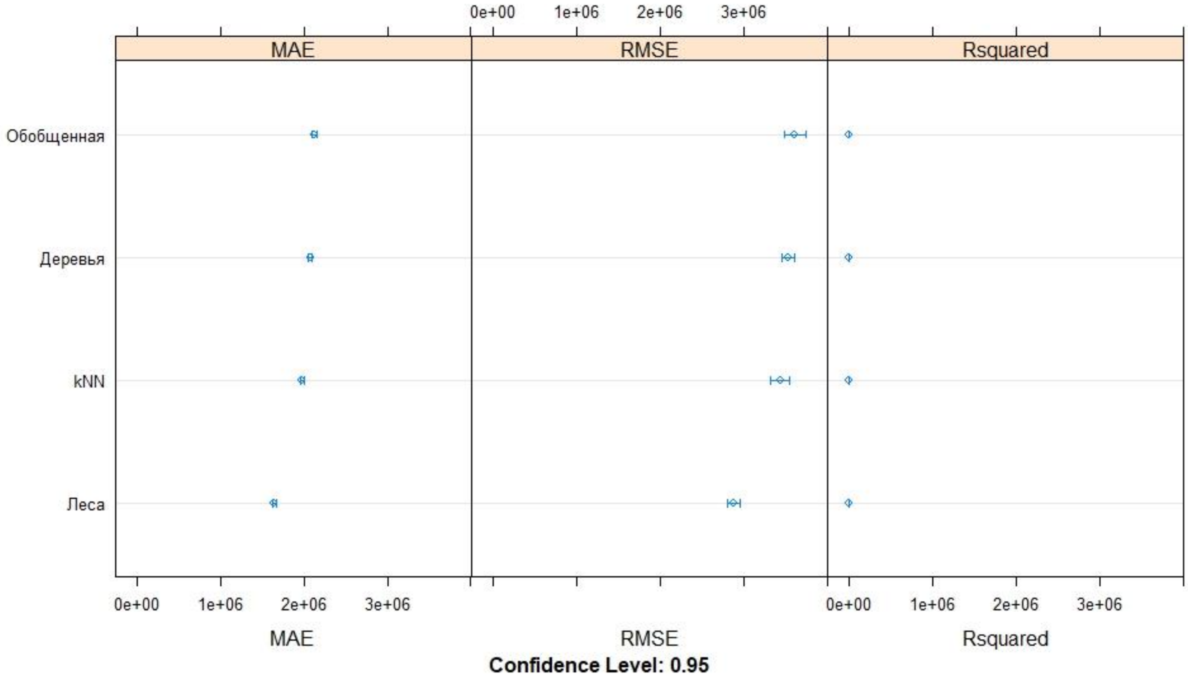
Now let's turn to the graph showing the contribution of various variables to the first 2 components - so called "biplot" and interpret it. The color of the variable vector indicates the amount of contribution in forming the principal component while the direction of the vector indicates which component it contributes in. At the same, time the angle between vectors indicates the correlation degree between corresponding variables. Thus, we can see that the biggest contribution have distance variables (partly matching factor 1 and factor 2 from the factor analysis). These variables form mainly the first principal component and highly correlated. The least contribution have variables such as the number of cafes around, the floor, the distance to the nearest cemetery and the proportion of working residents. These variables are also poorly correlated. Someplace between these two groups occupies the house construction year, type of the apartment, its condition and workload of neighboring kindergartens. These variables make a moderate contribution and not highly correlated.



Forecasting

Now let's move on to the methods price predicting. The generalized least squares method, trees, regression using the "K" method of nearest neighbors and random forests were used as the main models. Each model used 56 regressors for 15963 observations (since the initial sample was subdivided into training and test).

To improve accuracy in the first three methods, the data were centered and scaled, as well as subjected to the Box-Coke transformation. For random forests, the data were only transformed by the Box-Cox method. Also, for the purpose of improving the accuracy of prediction, the technique of repeated cross-validation with ten blocks of three repetitions was applied. In order to speed up the process a bit (random forests, for example, are calculating around half an hour on 6GB of RAM), the test sample was cut by 20%, and highly dependent regressors were removed (more than 65% of the correlation). Thus, according to the obtained results, as shown in the picture, the forest cope best in all respects. Such large errors are likely to be associated with a very large number of observations, of which we have almost half a million. In General, we believe that the forest model can be used for prediction. As for R^2, for every model it fluctuates around 0,6, however, this metric can’t be trusted due to the possible congestion of the model.



Conclusion

This work is a quite comprehensive analysis of data by methods of multivariate statistical analysis. Having at our disposal an incredibly rich set of data, including more than 30,000 observations on 292 variables, we were able to successfully apply all the basic methods of Multivariate statistical analysis and even a little more.

Thus, we have built a colorful and visual descriptive statistic, giving the reader an idea of the real estate markets of Moscow and the city. Then we conducted a correlation analysis of the main variables and identified the factors that have the closest relationship with the price of housing. In the next section, we clustered the available observations in five different ways and compared them. In addition, we have provided illustrations of the algorithm. In the next section of this work, we have carried out procedures to reduce the dimension of the existing dataset. First, this task was performed by using factor analysis, where we obtained very well interpreted factors, the number of which matched with the number of clusters from the previous section. We then used the principal components method, which also allowed us to compress many available variables into a smaller number of aggregated ones; we studied which derived principal components carry the most information and which variables they consist of. In the last part of our work, we built apartment price forecasts in four different ways and compared the quality of the forecasts, choosing the most effective.

Thus, it can be argued that this paper provides a qualitative and illustrative example of multivariate statistical analysis practical use in the analysis of real data.

Appendix

Variable name	Description
school_on_place	Number of children per place at school
sport_object_raion	The number of sport facilities in raion
state	Apartment condition
radiation_raion	Presence of radioactive waste disposal
price_doc	sale price
culture_objects_top_25	Presence of the key objects of cultural heritage
cafe_count_1000_price_high	Cafes and restaurant bill, average over 4000 in 1000 meters zone
indust_part	Share of industrial zones in area of the total area
ecology	Ecological zone where the house is located
work_proportion	Share of the working population of this zone
water_treatment_km	Distance to water treatment
material	Wall material
full_sq	Wotal area in square meters, including loggias, balconies and other non-residential areas
floor_by_maxfloor	Relative height indicator
church_synagogue_km	Distance to Christian chirches and Synagogues
metro_km_walk	Distance to the metro, km
railroad_station_walk_km	Distance to the railroad station (walk)
build_year	year built
product_type	owner-occupier purchase or investment
museum_km	Distance to museums
office_km	Distance to business centers/ offices