

**MTECH KE5107**  
**DATA MINING METHODOLOGY AND METHODS**  
**PROJECT REPORT**

---

**EXPLORATORY ANALYSIS AND PREDICTIVE MODELS**  
**ON THE RUSSIAN PROPERTY MARKET**

---

**TEAM MEMBERS**  
SIDDHARTH PANDEY  
PRANSHU RANJAN SINGH  
EDUARD ANTHONY CHAI  
NYON YAN ZHENG  
TAN KOK KENG

MASTER OF TECHNOLOGY IN  
KNOWLEDGE ENGINEERING  
BATCH KE-30(2018)

SUBMISSION DATE: 12 MARCH 2018

## INTRODUCTION

---

This project is undertaken as an academic assignment which is part of the continual assessment in the module KE5107 Data Mining Methodology and Methods of the National University of Singapore (NUS) Institute of Systems Science (ISS) Masters of Technology in Knowledge Engineering (KE) programme.

This project adopted the CRISP-DM methodology and the R programming language to mine data from a dataset describing the Property Market in Moscow, Russia.

## 1. BUSINESS UNDERSTANDING

---

### 1.1 DETERMINE BUSINESS OBJECTIVES

The dataset selected for the project comprises data on Russian properties in the vicinity of Moscow, including price. The dataset includes population, transportation, amenities related features.

The business objective is to predict the price of the property price in Moscow.

The prediction model can be used for estimating the prices of an individual property, such as:

- a new property on the market
- an existing property affected by changes in demographics or building, transportation infrastructure

It can also be used to estimate the mean prices of properties which specific features, for example, small apartments within a certain distance of transportation means in a particular municipality.

#### **Business Success Criteria**

To adopt the CRISP-DM methodology and the R programming language in executing the project, specifically:

- Perform exploratory data analysis to gain insights from the dataset.
- Build stable predictive model with  $R^2$  of more than 0.7 to predict property price.

### 1.2 ASSESS SITUATION

#### **Inventory of Resources**

- a. Project Team: The project team comprises of 5 project members.
- b. Dataset: The dataset is obtained from: <https://www.kaggle.com/c/sberbank-russian-housing-market>.
- c. Software: R Studio, an integrated development environment for the R programming language.

#### **Requirements, assumptions and constraints**

The dataset is open source and therefore there are no legal implications for its use in this project. It is assumed that the dataset provided by Sberbank, a Russian bank comprises valid data which will be useful for predicting property prices in Moscow. Sberbank is Russia's oldest and largest bank with housing investments as one of their major source of revenue.

#### **Risks and Contingencies**

There are no expected risks or contingencies identified for this project.

### 1.3 DETERMINE DATA MINING GOALS

#### Data mining goals

To predict property price (target variable) from a selected set of features (feature variables) in the dataset. In this project, feature selection will be performed using methods other than principal component analysis.

#### Data mining success criteria

To complete a predictive model with appropriate validation. No accuracy requirements are mandated.

### 1.4 PRODUCE PROJECT PLAN

#### Project plan

The following steps are planned for this project:

No.	Step	Duration (days)	Start Date	End Date
1	Selection of Dataset	7	23 Jan 18	29 Jan 18
2	Business Understanding	14	30 Jan 18	12 Feb 18
3	Data Understanding	14	13 Feb 18	26 Feb 18
4	Data Preparation	7	27 Feb 18	5 Mar 18
5	Modeling	9	1 Mar 18	9 Mar 18
6	Evaluation	4	6 Mar 18	9 Mar 18
7	Deployment (not applicable to this project)	-	-	-
8	Preparation of Project Report	4	9 Mar 18	12 Mar 18
9	Submission of Project Report	-	-	12 Mar 18

The choice of the dataset to be used in the project was approved by the module instructor on 29 Jan 18.

The initial steps of business and data understanding will be performed over a longer duration while the team is also involved in other projects.

The data preparation, modelling and evaluation steps have overlaps in planned duration as the data mining process will be performed iteratively as opposed to a linear sequential manner in a traditional waterfall model.

#### Initial assessment of tools and techniques

The use of the CRISP-DM methodology and the R programming language is a requirement of the project. The team does not expect to need any other tools to execute the project.

---

## 2. DATA UNDERSTANDING

---

### 2.1 COLLECT INITIAL DATA

The collected datasets comprised of a primary transaction dataset and supplementary dataset to help add value to the primary dataset.

#### **Primary Dataset**

The housing transaction dataset is acquired from Kaggle.

#### **Supplementary Dataset**

The sub-area and area division of Moscow is acquired from Wikipedia. URL:  
[https://en.m.wikipedia.org/wiki/Administrative\\_divisions\\_of\\_Moscow](https://en.m.wikipedia.org/wiki/Administrative_divisions_of_Moscow)

## 2.2 DESCRIBE DATA

The transaction dataset comprises of 30,473 records with 292 attributes describing the house property and local area of the each property. Also each record has a transaction price associated with it which is the sales price of each property. The dataset has transactions from August 2011 to May 2016.

Sub-area and area division of Moscow have 147 records, each record providing a mapping between sub-area and the corresponding administrative area it fall under. This data can be joined with the transaction data through its *sub\_area* column.

The following sections covers the data description in details, elucidating about various attributes of data like variable descriptions, missing values, distribution, correlation etc.

### Hypothesis-based Data Investigation

The primary dataset has large number of variables and exploring and describing each variable in the dataset will consume too much time and effort. Therefore, to understand the dataset in a prolific way, some initial hypotheses are set up to guide the course of exploration. These initial hypotheses will allow for a clustered variable exploration and cover more ground faster. Following hypotheses were assumed:

1. Price of property is affected by the property's characteristics such as size of property, total living area, number of rooms and etc.
2. The population of the area where the property is located affects the property price, i.e. there is a difference in price between properties in sub-urban and urban areas.
3. Availability of amenities such as schools, health care centres, cafes and etc. affect the price of a property. Properties with many amenities nearby are likely higher in price.

#### 2.2.1 Variables and Description

A complete description of all the variables in the primary dataset is attached in Appendix A of this report.

The sub-area & administrative-area division of Moscow data comprised of two categorical variable *sub\_area* and *admin\_area*. The former has 147 unique values mapped to 12 values of the later.

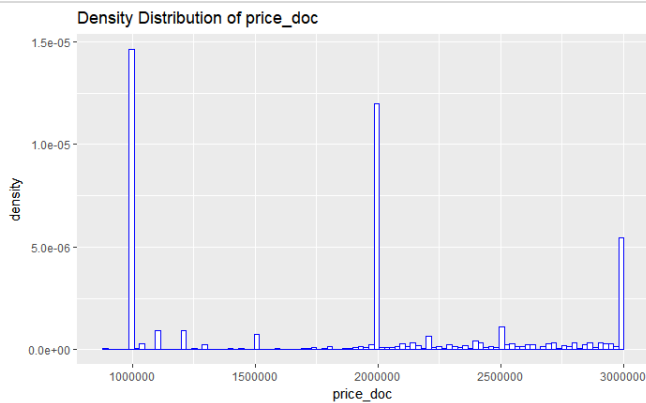
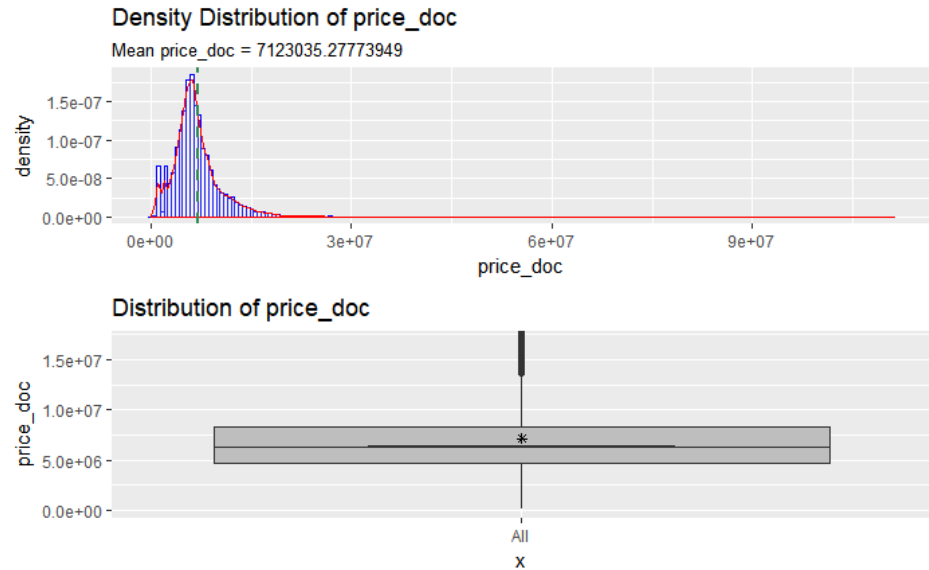
#### 2.2.2 Volumetric Analysis of Target Variable (*price\_doc*)

To have better understanding of target variable several analysis are performed. The summary statistic of the variable *price\_doc* is given below.

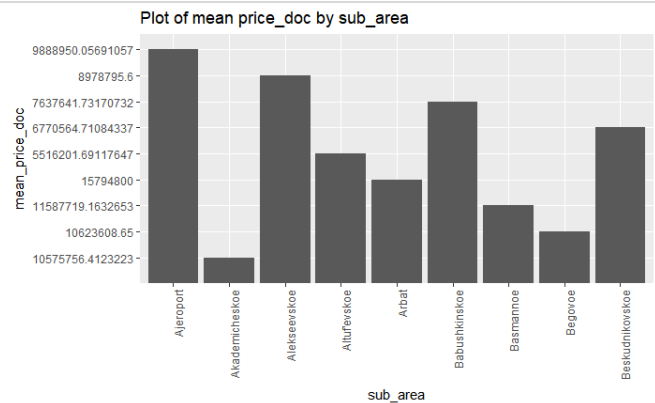
Min.	1 <sup>st</sup> Qu.	Median	Mean	3 <sup>rd</sup> Qu.	Max.
100,000	4,740,002	6,274,411	7,123,035	8,300,000	111,111,112

From the density distribution plot on the left, it was observed that the property prices are positively skewed.

From the box plot, it can be inferred that most outliers consists of abnormally large values.



In the plot above, the density distribution plot is zoomed in to prices below 3,000,000. There are some unusual peaks in probability density around  $[9 \times 10^5, 3 \times 10^6]$ . These unusual peaks are most likely due to data error or presence of duplicate transaction with different id.



9 sub\_areas are sampled and plotted against its mean price. The plot shows that the price can be very different from one area to another.

### 2.2.3 Volumetric Analysis of Variables Governing House Characteristics

This covers the volumetric analysis of variables related to the first hypothesis.

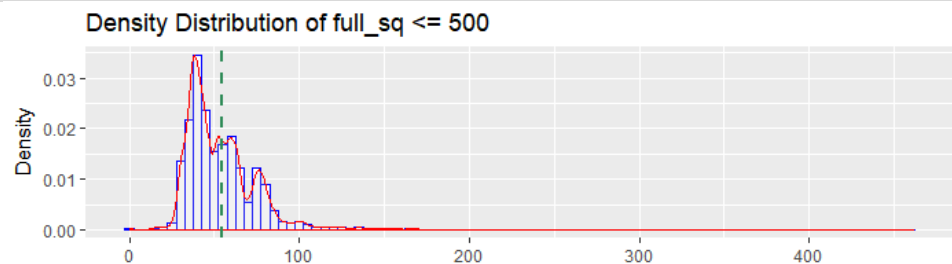
#### 1. Total Area (*full\_sq*)

Both the upper bound and lower bound of the *full\_sq* implies some data errors.

Min.	1 <sup>st</sup> Qu.	Median	Mean	3 <sup>rd</sup> Qu.	Max.
0	38	49	54.21	63	5326

The density distribution plot on the right is zoomed in to  $full\_sq \leq 500$ .

The distribution is positively skewed.



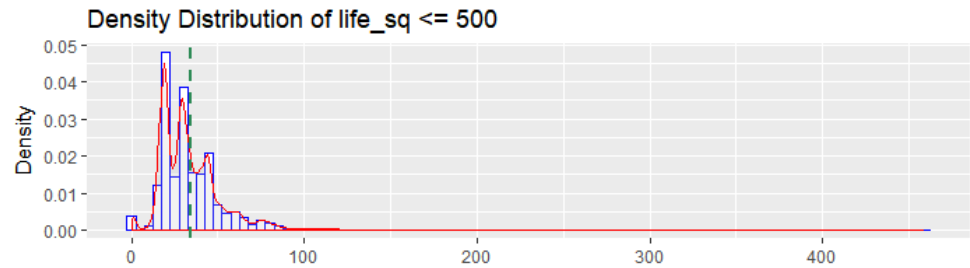
## 2. Living Area (*life\_sq*)

The maximum value of living area is very large, probably a data error. Similarly 0 is not possible as a min value.

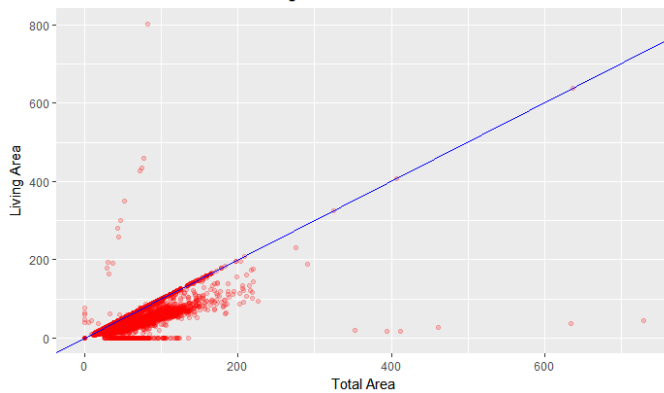
Min.	1 <sup>st</sup> Qu.	Median	Mean	3 <sup>rd</sup> Qu.	Max	NA's
0	20.0	30.0	34.4	43.0	7478.0	6383

The density distribution plot on the right is zoomed-in to  $life\_sq \leq 500$ .

The distribution is positively skewed.



Distribution of total area x living area



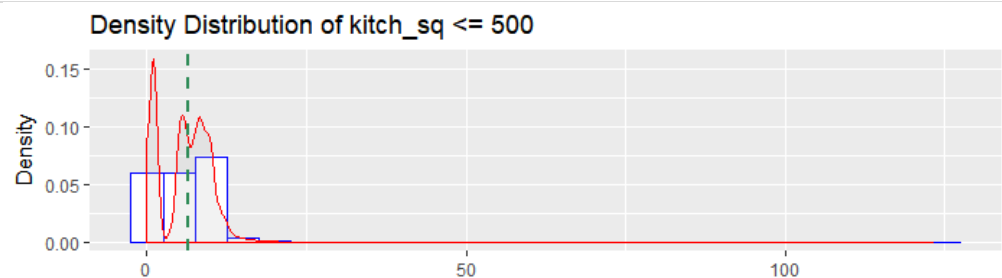
The plot on the left between *total\_area* and *life\_sq* shows a positive correlation among them. As *life\_sq* variable has a large number of missing value, this relationship can be exploited to impute those missing value using a linear model.

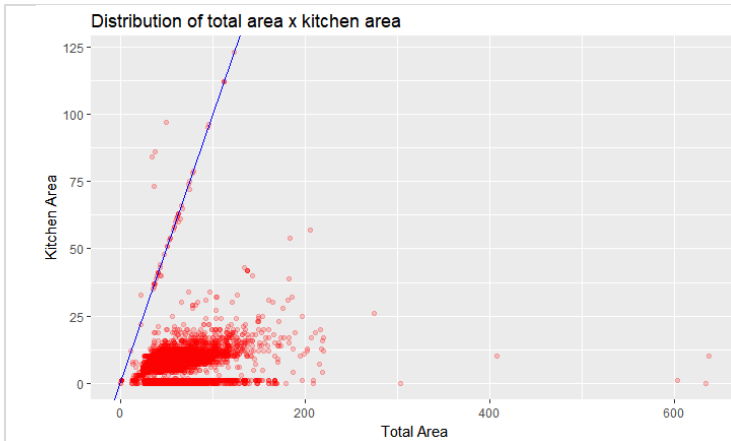
## 3. Kitchen Area (*kitch\_sq*)

Summary table of *kitch\_sq* shows that some observations has zero and large values such as 2014 which is not possible.

Min.	1 <sup>st</sup> Qu.	Median	Mean	3 <sup>rd</sup> Qu.	Max	NA's
0	1.0	6.0	6.39	9.0	2014.0	9572

The density distribution of *kitch\_sq* shows an unusual density peak for lower values. This suggests some erroneous records in the dataset. There are many missing values in this variable.



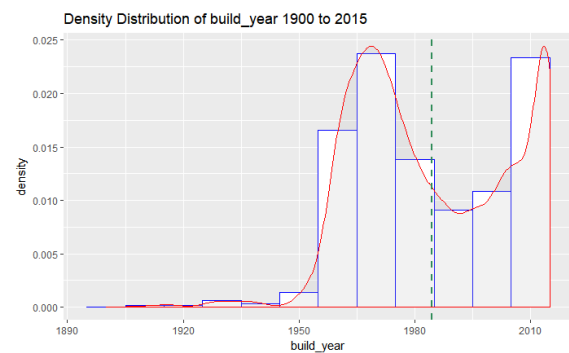
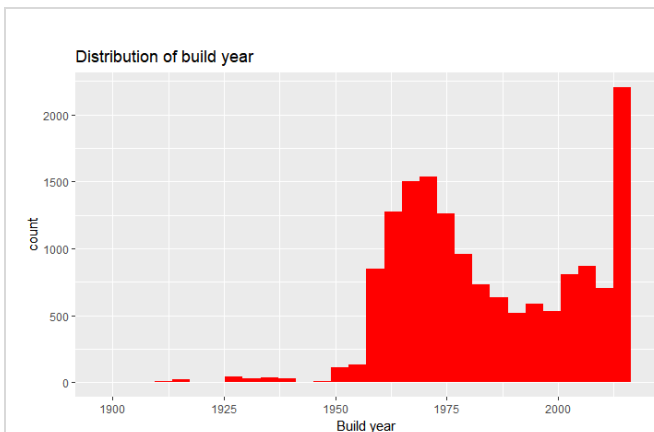


The plot on the left between total area and kitchen area shows a positive correlation. This relationship can be used to impute missing values and correct erroneous records using a linear model.

#### 4. Build Year (*build\_year*)

The table below shows that the record with the maximum build year is incorrect. It looked like a data entry error where the data could either mean year 2005 or 2009. Likewise the build year of 0 is also not logical. This variable has high number of missing values.

Min.	1 <sup>st</sup> Qu.	Median	Mean	3 <sup>rd</sup> Qu.	Max	NA's
0	1967	1979	3068	2005	20052009	13605



From the histogram and density distribution of build year, it can be inferred that there are high number of properties with build year in early 1970s and early 2000s.

#### 5. House Condition (*state*)

This variable has 13,559 missing values. The values for this variable ranged from 1 to 4. With the maximum value of 33, it is likely there is a data entry error for that particular record.

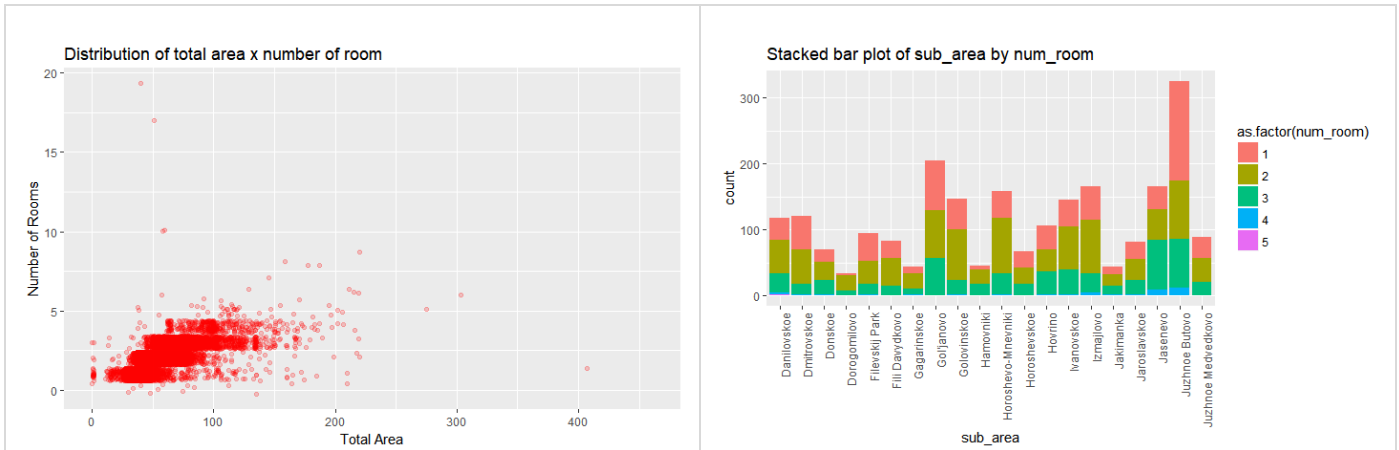
Summary of House Condition ( <i>state</i> )		
Min.	Max	NA's
1.00	33.0	13559

Count of House Condition ( <i>state</i> )			
1	2	3	4
4855	5844	5790	422

#### 6. Number of Rooms (*num\_rooms*)

The figures below hint towards incorrect records with abnormally low or high number of rooms.

Min.	1 <sup>st</sup> Qu.	Median	Mean	3 <sup>rd</sup> Qu.	Max.	NA's
0	1.00	2.00	1.91	2.00	19.00	9572



## 7. Floor and Max Floor (*floor* and *max\_floor*)

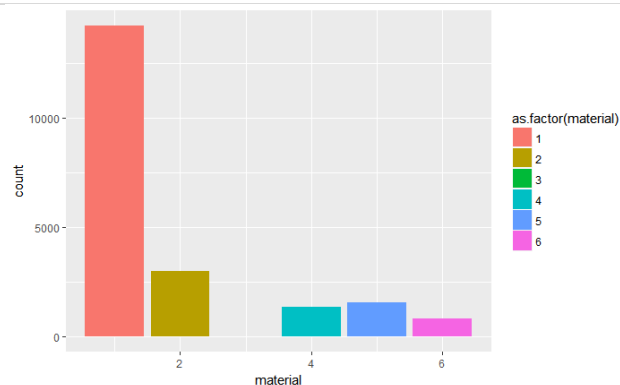
Below is the statistical summary of *floor* and *max\_floor*.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
Floor	0	3.0	6.5	7.671	11.00	77.00	167
Max Floor	0	9.0	12.0	12.56	17.00	117.00	9,572

## 8. Material (*material*)

Material is categorical variable with 6 possible values.

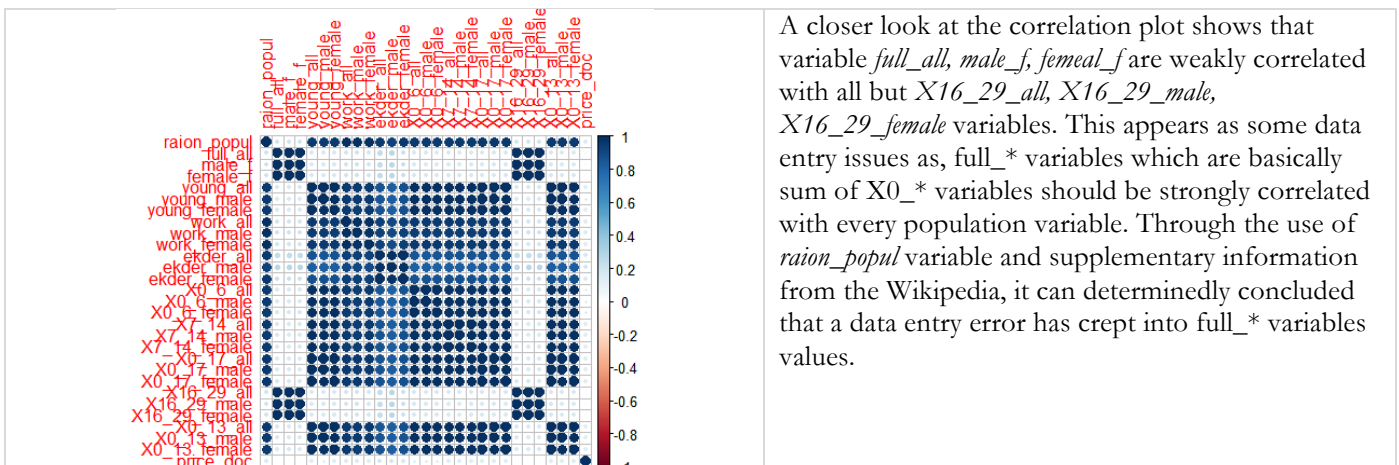
Material	Count
1	14,197
2	2,993
3	1
4	1,344
5	1,561
6	803



### 2.2.4 Volumetric Analysis of Variables Governing Population

This section covers volumetric analysis of variables that describe population parameter.

A correlation plot between the parameter shows that most of variables describing the population characteristics are highly correlated.





admin_area <fctr>	sub_area <fctr>	raion_popul <int>	full_all <int>
Central	Basmannoe	108171	28179
Central	Presnenskoe	123280	57999
Central	Taganskoe	116742	123280
Central	Hamovniki	102726	75377
Central	Krasnosel'skoe	47245	55590
Central	Tverskoe	75377	116742
Central	Jakimanka	26578	102726
Central	Zamoskvorech'e	55590	108171
Central	Meshhanskoe	57999	47245
Central	Arbat	28179	741887

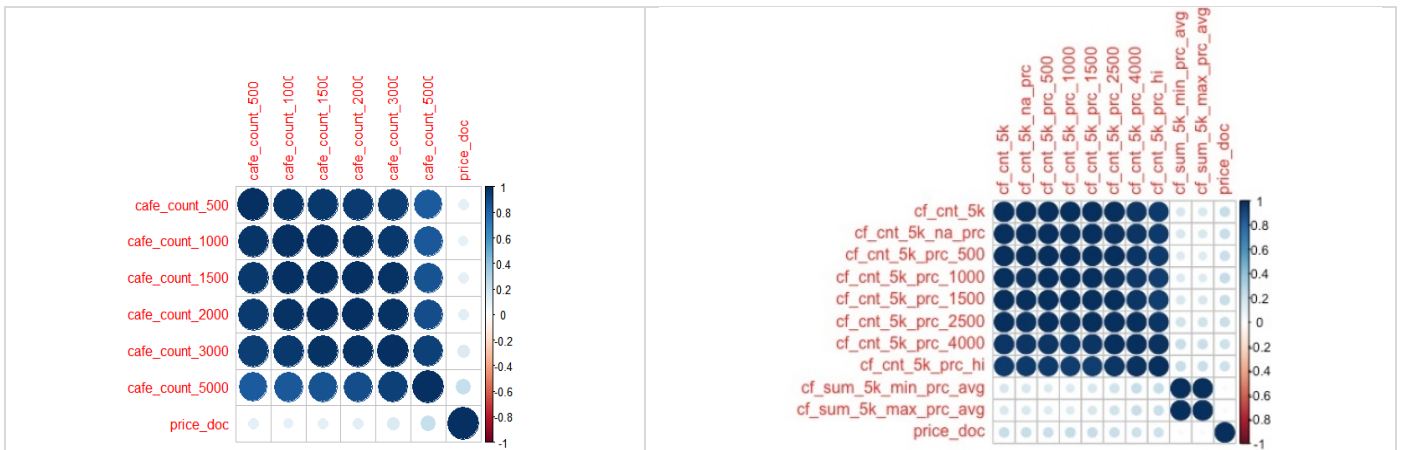
This is a sub part of the table formed by joining admin area and sub area. It can be generally seen that *raion\_popul* and *full\_all* are very distant.

### 2.2.5 Volumetric Analysis of Variables Governing Amenities

This section covers volumetric analysis of variable describing the amenities near the property.

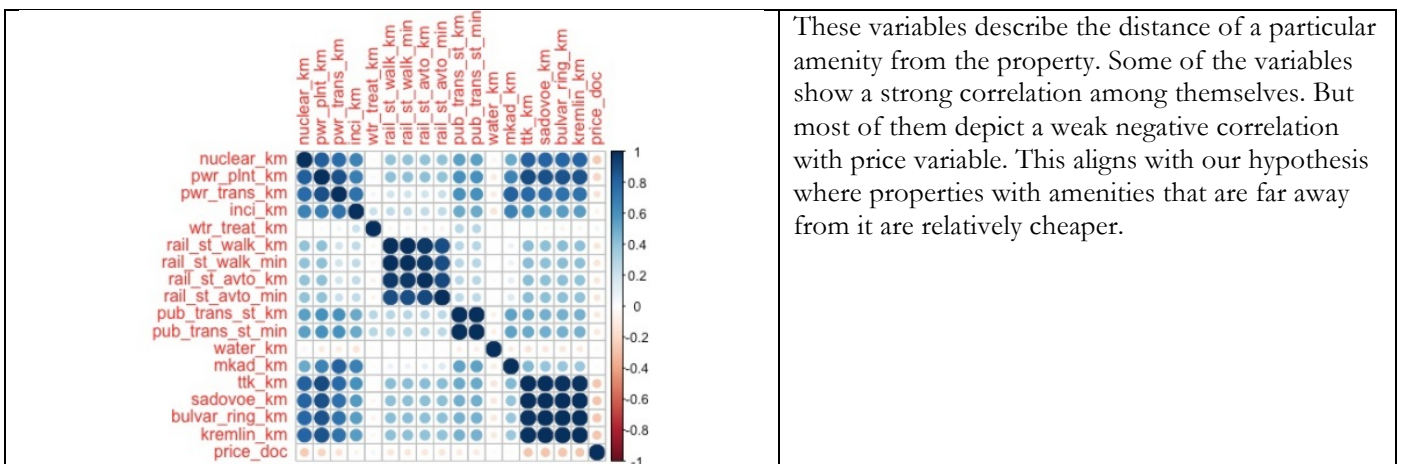
#### 1. Variables related to café in the neighbourhood

	cafe_count_500	cafe_count_1000	cafe_count_1500	cafe_count_2000	cafe_count_3000	cafe_count_5000
Min.	0.0	0.0	0.0	0.0	0.0	0.0
1 <sup>st</sup> Qu.	0.0	1.0	2.0	3.0	6.0	20.0
Median	1.0	4.0	10.0	18.0	41.0	108.0
Mean	3.872	15.41	32.46	55.03	110.9	265.5
3 <sup>rd</sup> Qu.	3.00	11.0	23.00	37.00	78.0	222.0
Max	120.0	449.0	784.0	1115.0	1815.0	2645.0



From the correlation plot, it can be concluded all the cafe count and cafe average price variables are strongly correlated to all other cafe count and cafe average price variables. These variables also show a weak positive correlation with the price of the property.

#### 2. Variables related to some miscellaneous amenities

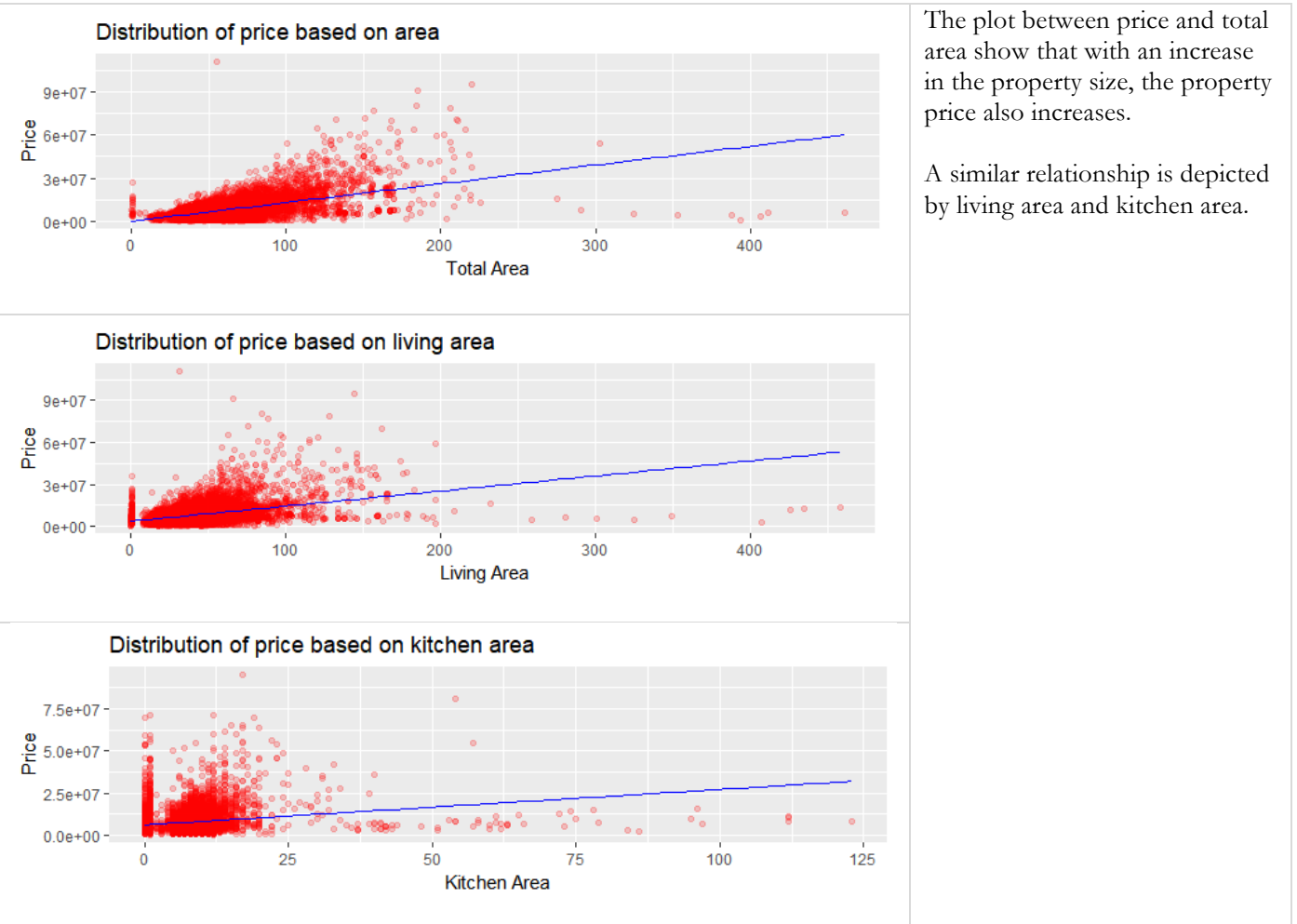


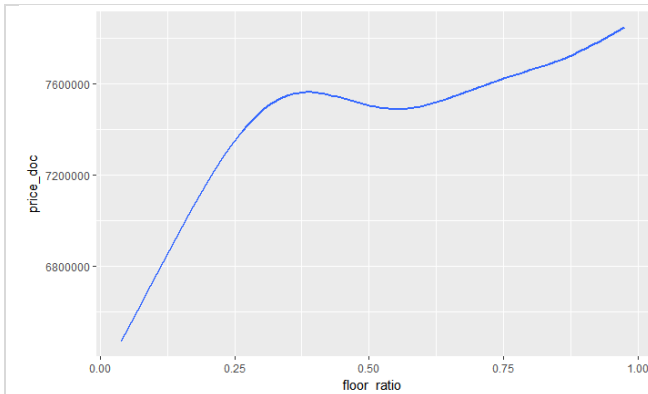
These variables describe the distance of a particular amenity from the property. Some of the variables show a strong correlation among themselves. But most of them depict a weak negative correlation with price variable. This aligns with our hypothesis where properties with amenities that are far away from it are relatively cheaper.

## 2.3 EXPLORE DATA

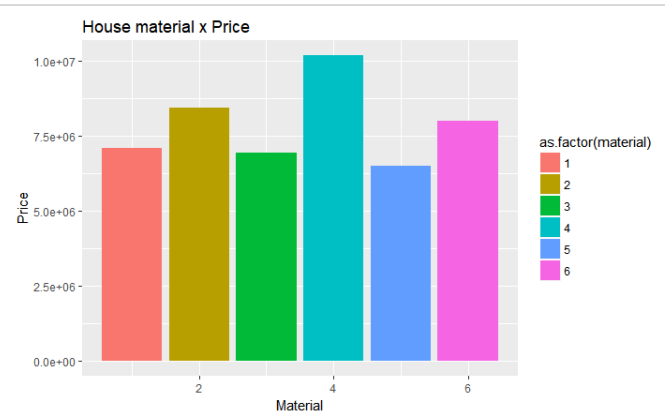
Data exploration is carried out with an aim to establish a relationship between price and other variables and find trends that affect the sale price of the property. The result from this exploration will provide empirical proof to the hypotheses. The results will also help in building more refined and better predictive models.

### 2.3.1 Variables describing house characteristics

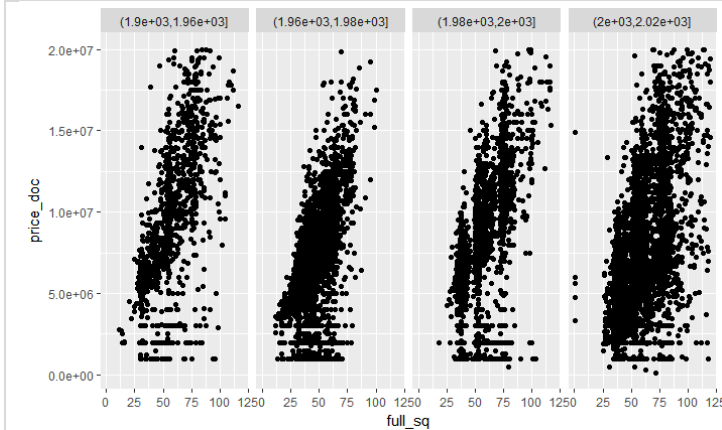




The plot above between property price and the ratio of the floor to the maximum floor of the building is used to relatively compare different properties with dissimilar elevation above the ground level. Price of the property increases with the increase in the value of floor ratio. The plot shows that higher floor properties in relative to the building's maximum floor are higher in price.

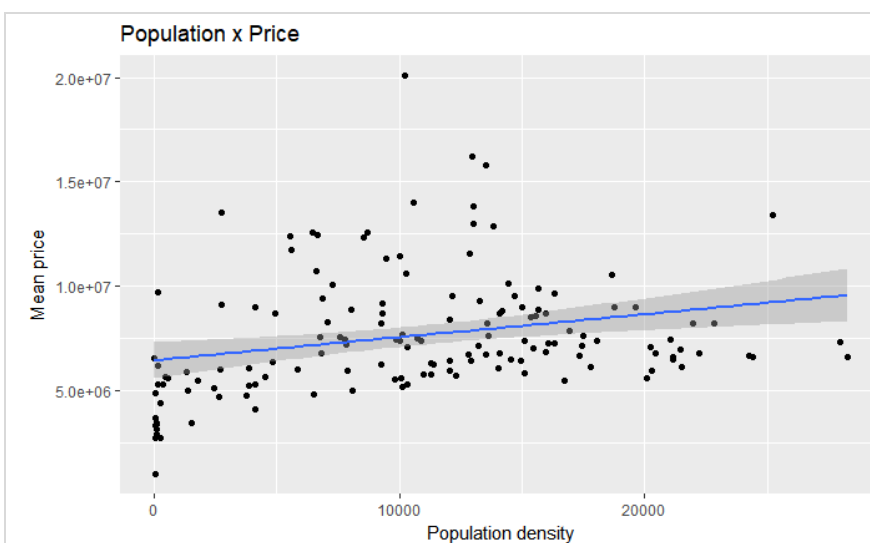


The bar plot above between mean property price and the property material shows that property built with material 4 are generally more expensive, followed with houses built with material 2 and 6.



The plot on the left between property price and property size, grouped by the built year shows that for each period property price increases with the property size.

### 2.3.2 Variables describing area population



The line plot on the left between the mean property price and population density indicates that price of properties increases with population density, though the trend is not very strong.

### 2.3.3 Variables describing amenities

The primary dataset have large number of variables describing the locality of the property. The correlation plots for all related variables are given in Appendix B.

Variable Category	Appendix B	Exploration Summary
Cafe	Plot 1 to 6	Most of these variables depict high positive correlation with each other and weak positive correlation price.
Religious Sanctuaries	Plot 7 to 9	The property price has a weak positive correlation with count variable and weak negative relationship with distance.
Industrial	Plot 10	The plot does not depict any significant relationship.
Market	Plot 11	Weak positive relationship with count and negative relationship with distance.
Leisure, Theater, Museum, Exhibition, Culture	Plot 12	Property prices are higher for properties which are closer to these amenities.
Healthcare, Hospice, Cemetery	Plot 13	Property prices are higher for properties which are closer to these amenities.
Green Zone	Plot 14	Insignificant relationship with the price for most variables but negative correlation with park distance.
Sports Facilities	Plot 16	Most count variables are positively correlated with price and negatively correlated with distance.
Office	Plot 17	The further the property is from the workplace, the lower the price of the property.
Shopping and Retail	Plot 18	Count variables are positively correlated with price, distance variables are negatively correlated.
Public Transport	Plot 21	Properties that are far from public transports are lower in price.
Education	Plot 23	Count variables have week positively correlation with price, distance variables are negatively correlated.

## 2.4 VERIFY DATA QUALITY

### 2.4.1 Inconsistent raw population data

As discussed in section 2.2.4, population variables describing total count of individuals in the municipality (full\_\*) have been erroneously entered during data entry. These variables need to be corrected or removed to prohibit them from inducing error in the predictive models.

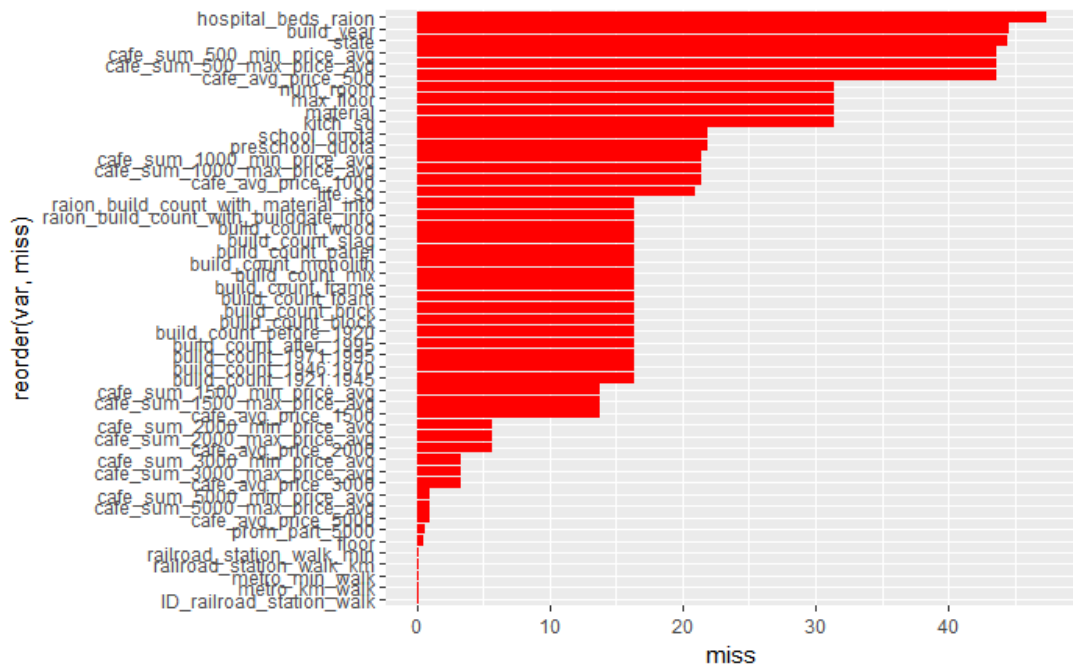
### 2.4.2 Frequent data entry errors

During the volumetric analysis and data mining, multiple data entry errors were encountered. Unusually high or low values were the most frequent data entry errors. Some errors were a violation of natural constraints like living area greater than the total area of the property. These errors need to be suitably dealt with to increase the data quality and reliability.

### 2.4.3 Investigating missing values

The primary dataset contains many missing values. The following barplot shows the percentage of missing values for variables. Some of the variables have more 40% missing values, which is quite high.

The missing values need to be dealt with before building predictive models. Approaches include dropping the columns with high missing values percentage or dropping records if most of the attributes are missing. Otherwise, mean, median, mode, the correlation among variables and linear regression model were used to impute missing values. The plot below shows the variable and the percentage of missing values.



## 2.5 EXPLORATORY DATA ANALYSIS SUMMARY

The investigation was guided by three hypotheses as mentioned in section 2.2.

### Validation of Hypothesis

- Through the exploration analysis, we can conclude that there is a definite relationship between the house features and property price. The observations support the hypothesis that in general properties that are larger in size, number of rooms, fine state and a higher floor ratio are higher in price.
- Population density also has an impact on property price, although weak, on the property prices. Generally, properties near areas with higher population density are higher in price. This supports the hypothesis on the relationship between the population and property price.
- Most of the amenities have a weak positive correlation with the price and their distance is negatively correlated with the property price. This supports our hypothesis that properties which are nearer to amenities are more expensive.

### Insights Summary

- The average price of the property is related to the sub-area it is located at.
- Properties with price more than 60,000,000 are probably data errors and outliers.
- Properties that are larger in size are likely to be sold at higher price. Values above 1000 and less than 5 are most likely outliers.
- Properties with larger living areas are higher in price. Some of the records have a living area more than the total area which is not possible.
- Properties with condition 4 are sold at highest prices.
- Properties with 1 to 3 rooms are most common.
- Generally, properties on a higher floor (relative to the max floor) are more expensive.
- Properties with material 4 have a relatively higher price. Properties with material 1, which is the most common and likely for the middle market, are not too expensive and not too cheap.
- Property prices increase as population density increases. However, raw population data is inconsistent and some variables contain errors.
- The number of cafe around the house has quite a weak impact on the house price. The average of prices of the cafe also does not impact prices much.
- The further away the amenities from the property, the lower the price of the property.

### 3. DATA PREPARATION

The Data Preparation stage is further divided into six substages below.

#### 3.1 SELECT DATA

For the selection of data, the entire raw data of housing prices was taken into consideration. No attributes or observations were removed from the raw dataset.

#### 3.2 CLEAN DATA

The focus at this stage was to identify outliers and missing values in the dataset and apply the appropriate treatment to such cases to improve the quality of the dataset.

##### Outliers

From the data exploration, the following outliers were identified and removed from the dataset

Outliers Removed	Details
Total area above 300 sq	Total area above 300 sqm are likely to be outliers as there are only 14 such records ranging from 350 sq. to 5326 sq. All the rest of values out of 30471 records for the total area are less than 300 sq.
Total area lesser than 5 sq.	The total area below 5 sqm is likely to be outliers as there are only 26 such records, some of which have value 0.
A living area larger than total area:	There are 37 records where living area is larger than the total area of the house. Such records may have occurred while data entering due to human error.
Living area lesser than 5 sq.:	A living area smaller than 5 sq. Meters is quite small for a living room. It may be a description of some other room.
Kitchen area larger than total area	There are 12 records where kitchen area is larger than the total area of the house. Such records may have occurred while data entering due to human error.
Number of rooms more than 9	The price of houses with more than 9 rooms is even lesser than the price of houses with 4 rooms, which seems quite unlikely.
Max floor higher than 90	According to Wikipedia article on the list of tallest buildings in Moscow ( <a href="https://en.wikipedia.org/wiki/List_of_tallest_buildings_in_Moscow">https://en.wikipedia.org/wiki/List_of_tallest_buildings_in_Moscow</a> ), there are only 2 buildings with a number of floors greater than 90. The records in the data set for buildings having higher than 90 floors doesn't match with the record from Wikipedia.
The floor is higher than max floor	It is not possible to have a flat a floor which exceeds the maximum floor value of the building itself. These records have some data entry error.
Data error	Build year with value 20052009. The entry can be 2005 or 2009.
Build year older than 1900 and beyond 2015:	There are 530 records with value 0 for build year. Build year values below 1900 and above 2015 seems as an outlier as they account for less than 5 percent of the total dataset. Most of these records have lots of missing values for other attributes or values that make no logical sense.
House with a state other than 1 to 4	Nearly all the observations have property state value of 1 to 4. All other values are removed as this attribute seems to be a factor and other random values make no sense.
House with no rooms	There are 14 records with zero number of rooms. There can't be a house with no rooms. This is likely a data entry error.
Some of the suspicious values on the price_doc column	It looks like some observations are repeated multiple times, which resulted in the 3 distinctive peaks in the plot of the density distribution of price_doc.
Properties with price more than 60,000,000	There are 16 records of houses with a price greater than 60,000,000.
A number of population shown in the full_all attribute is incorrect	Attributes related to the full_all attribute are from the same source; hence they need to be removed from the dataset.



### Missing Values

Based on the figure in section 2.4.2, following missing values were identified and appropriate actions were taken to treat them. In some cases, missing values records were removed from the dataset and in other cases, they were imputed.

Missing Values (% missing)	Details and Treatment
Records with the number missing values that is less than 2% of total observations	Records with missing values in 7 of the attributes (floor, prom_part_5000, railroad_station_walk_min, railroad_station_walk_km, metro_min_walk, metro_km_walk and ID_railroad_station_walk) are removed from the dataset.
Living area (22.2%) & kitchen area (31.2%)	For both these attributes, the missing values were imputed with a mean ratio with the total area. These missing values were not imputed with direct average because it can exceed the total area of the house for some cases, which is not possible.
Number of rooms (31.2%)	The missing values were imputed with ceil value of an average number of rooms.
Max floor (31.2%)	The missing values are imputed with a rounded value of the mean ratio of the floor and max_floor.
Preschool (23.3%) and school (23.3%)	<p>It was observed that preschool_quota is highly correlated with children_preschool and school_quota with children_school. The missing values are imputed using linear regression model on the correlated attributes.</p> <p><b>Correlation Education Related Variables using Pearson</b></p>
Number of hospital beds for the district (48.3%)	This attribute has too many missing values and it is also not correlated with any of other health-related attributes. Hence, this attribute was dropped from the data set.
Cafe average price	All the attributes related to cafe average prices having missing values were imputed with zero. This was done based on the assumption that the values were not recorded because of the absence of any cafes near the house. There were 18 such attributes.
Attributes related to build year, material, and state:	Attributes such as state, build_year, material, raion_build_count_with_material_info, etc. having missing values can't be imputed using any of the criteria used with previous attributes. There were 19 such attributes and all of those were removed from the dataset.

### 3.3 CONSTRUCT DATA

Transaction timestamp attribute might help when building models to learn trends. The timestamp attribute was in Date format. It will be more helpful if we model based on month or year rather than a specific date. Hence, the timestamp attribute was split into three new attributes trans\_year, trans\_month and trans\_day. The timestamp attribute was removed from the data set after the construction of above-derived attributes.

### 3.4 INTEGRATE DATA

The sub\_area (municipality where the property is located at) attribute has 146 unique categorical values. This dataset was integrated with administrative area dataset to achieve higher granularity. The administrative area data set clusters

different sub-area into 12 regions such as Northern, North-Western, Southern, etc. Thus, the new attribute `admin_area` (administrative area) has 12 categorical values which can be handled more efficiently as compared to 146 values. The `sub_area` attribute was removed from the dataset.

### 3.4.1 Feature Selection

In feature selection, the main objective is to select relevant features (attributes) that will be used as input for different modelling algorithms. After the data preparation stages, the clean dataset contains 21,064 records and 261 attributes. The number of attributes is still a large number. It is a good practice to reduce the number of attributes if there is any scope for reduction that doesn't hamper the accuracy parameters for model execution. Two different strategies, one using random forest another using correlation were used to perform feature selection. This resulted in the formation of two different prepared datasets, based on each feature selection strategy.

### 3.4.2 Feature Selection using Random Forest Only

Random Forest was used to obtaining the importance of attributes in the dataset. The top 30 attributes based on importance score were selected using this feature selection strategy. The selected attributes and their importance score are listed in Table 1 below. The rest of the attributes are removed from the dataset.

### 3.4.3 Feature Selection using Correlation and Random Forest

In this approach, first the number of attributes was reduced using the correlation criteria and then random forest was applied on the reduced set of attributes.

The idea is to first remove those attributes that are redundant in the data set. This was achieved by using the correlation score for different attributes in the same category. If the correlation between attributes in that category is higher than or equal to 0.7, they are termed as highly correlated and one attribute is selected from the highly correlated attributes while the rest are removed from the dataset. The following is the list of the categories of attributes where the correlation between the attributes in that category is checked. The correlation plots of the attributes for each category are given in Appendix B.

- Cafe in 500 meters zone
- Cafe in 1000 meters zone
- Cafe in 1500 meters zone
- Cafe in 2000 meters zone
- Cafe in 3000 meters zone
- Cafe in 5000 meters zone
- Big Church
- Small Church & Synagogue
- Mosque
- Industrial
- Market
- Leisure, Theatre, Museum, Exhibition, Culture
- Healthcare, Hospice, Cemetery
- Green Zone
- Other Nature: Water and Park
- Sports Facilities
- Office
- Shopping and retail
- Utilities
- Transport: Railroad
- Public Transport: Metro, Bus and Public Transport
- Roads & City Center
- Education
- Population

After reduction, the number of attributes has reduced from 261 to 109. Random Forest was applied to obtain the attribute importance on the reduced set of attributes. The top 30 attributes based on importance score were selected using this feature selection strategy. The selected attributes and their importance score are listed in Table 1 below.

**Table 1: Top 30 features selected from (1) Random Forest only and (2) Correlation and Random Forest**

Random Forest Only			Correlation and Random Forest	
	Attributes	Importance Score (x 10 <sup>15</sup> )	Attributes	Importance Score (x 10 <sup>15</sup> )
1	full_sq	66.65	full_sq	65.60
2	life_sq	29.56	life_sq	34.14
3	num_room	13.51	zd_vokzaly_avto_km	13.25
4	cafe_count_2000	11.20	kitch_sq	13.09
5	kitch_sq	10.55	num_room	12.60
6	sport_count_3000	9.00	sport_count_5000	6.80
7	cafe_count_3000	7.74	swim_pool km	6.43



8	cafe_count_5000_price_2500	4.71	cafe_count_3000_price_500	6.28
9	cafe_count_5000_price_high	4.00	basketball_km	5.58
10	cafe_count_3000_price_1500	3.71	cafe_count_5000_price_500	5.52
11	cafe_count_3000_price_2500	2.97	admin_area	5.46
12	max_floor	2.46	kremlin_km	5.41
13	floor	1.78	trc_sqm_5000	3.17
14	zd_vokzaly_avto_km	1.76	max_floor	2.82
15	sport_count_2000	1.74	church_count_5000	2.41
16	cafe_count_3000_price_1000	1.70	cafe_sum_5000_max_price_avg	2.27
17	admin_area	1.66	exhibition_km	2.16
18	ttk_km	1.45	floor	2.10
19	cafe_count_2000_price_2500	1.35	ekder_male	2.03
20	cafe_count_5000_price_1500	1.33	mosque_km	1.97
21	trans_month	1.23	industrial_km	1.85
22	sadovoe_km	1.17	catering_km	1.82
23	cafe_count_2000_price_1500	1.14	workplaces_km	1.79
24	kindergarten_km	1.11	railroad_km	1.79
25	office_sqm_5000	1.04	cemetery_km	1.77
26	trc_count_3000	1.02	big_road1_km	1.75
27	swim_pool_km	1.00	preschool_quota	1.71
28	industrial_km	0.99	kindergarten_km	1.70
29	catering_km	0.96	additional_education_km	1.69
30	prom_part_3000	0.94	public_healthcare_km	1.65

Comparing the 2 version of features selection, the set selected by random forest only has a high concentration of café related features among the top 30. Predictive models are also preliminary fitted on this set of data, however the performance was less satisfactory as compared to the set where the features were selected using both correlation and random forest method. Therefore, the set of features on the right of Table 1, that were selected by both correlation analysis and random forest were selected to be used for predictive modelling in the next section.

### 3.5 FORMAT DATA

Formatting of data is required since different modelling algorithms require the data to be presented in specific order or specific types for attributes. For the given clean data set, following transformation activities were performed to make the data set suitable for specific modelling algorithms.

- Transformation of Categorical Attributes to Binary Attributes: For modelling algorithms such as linear regression, if raw transformation of categorical values in numerals is done, it makes them ordinal. The better way to approach this situation is to use one-hot encoding for these categorical attributes. That is why, all categorical attributes were transformed into binary attributes.
- Applying log or square root transformation: When using modelling algorithms such as linear regression, it is important to check the skewness of different attributes of the data set. All the skewed numerical attributes in the data set are log transformed.

## 4. MODELLING

### 4.1 SELECT MODELING TECHNIQUE

The 4 modelling techniques were used to achieve the project objectives.

- Random Forest (“RF”)
- Linear Regression (“LM”)
- Support Vector Machine (“SVM”)
- Extreme Gradient Boosting (“XGB”)

The models were trained and evaluated using 2 datasets:

- **Dataset A:** data in their original form with no transformation on both the target and predictor variables
- **Dataset B:** log transformation on the target variable and log or square root transformations on skewed predictors

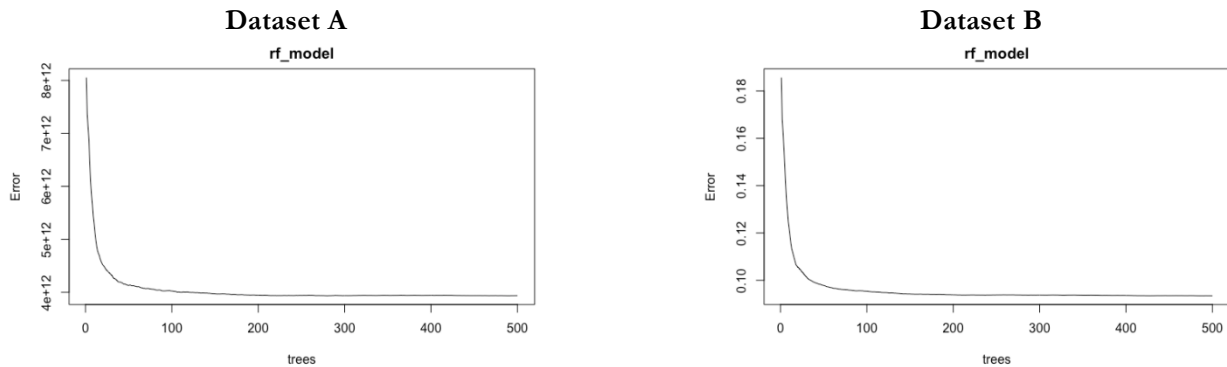
## 4.2 GENERATE TEST DESIGN

There are 21,064 observations in the both the clean datasets. As there is sufficient number of observations, the data was divided into 70:30 for training and testing set respectively.

## 4.3 BUILD MODEL

### Model 1: Random Forest

The model was built with 500 trees. The figure below showed that the model has converged.



Dataset A		Dataset B	
RMSE	R <sup>2</sup>	RMSE	R <sup>2</sup>
849,651	0.953272	909,387	0.9575287

The high R<sup>2</sup> in the training data showed some sign that the model is over fitting.

	%IncMSE	IncNodePurity
full_sq	65.308852	4.739631e+16
life_sq	36.147909	2.305878e+16
zd_vokzaly_avto_km	25.122247	1.174490e+16
kitch_sq	33.558789	1.077902e+16
num_room	19.309127	9.519171e+15
sport_count_5000	25.318293	6.118159e+15
swim_pool_km	23.782722	5.449111e+15
cafe_count_3000_price_500	27.981095	5.831316e+15
basketball_km	21.969398	5.450416e+15
cafe_count_5000_price_500	27.258666	6.306603e+15
kremlin_km	22.700190	6.295036e+15

Figure on the left showed the top important variables in the random forest model. Based on these, it can be interpreted that total area of the house, living area, distance from the station, kitchen area, and a number of the room are important features when deciding the price of the property.

### Model 2: Linear Regression

The numeric predictors are normalised before fitting the LM model.

Dataset A		Dataset B	
RMSE	R <sup>2</sup>	RMSE	R <sup>2</sup>
2,285,564	0.6101606	2,273,802	0.6325547

The RMSE is a bit high in general. The R<sup>2</sup> value is on the average when compared to other models. not that good either.

var <fctr>	Overall <dbl>
full_sq	43.8081697
cafe_count_3000_price_500	19.8308148
num_room	10.1519952
cafe_count_5000_price_500	8.5697844
mosque_km	8.5490718
exhibition_km	8.5034450
catering_km	8.4605710
south_western	7.7758832
kitch_sq	7.4783855
public_healthcare_km	7.4182021

Figure on the left shows top important variables in the LM model. According to this model, total area of the house, cheap café nearby, the number of room, and distance to the mosque are important when deciding the price of the house.

### Model 3: Support Vector Machine

The numeric predictors are normalised before fitting the SVM model.

Dataset A		Dataset B	
RMSE	R <sup>2</sup>	RMSE	R <sup>2</sup>
1,901,610	0.7329873	1,830,578	0.7521172

The RMSE is a better than the LM model. The R<sup>2</sup> value also higher in this case. There is no sign of over fitting.

full_sq	public_healthcare_km	life_sq
230.6012331	130.0856501	128.2133246
swim_pool_km	kitch_sq	kindergarten_km
117.4705638	115.9429613	110.1167198
troitsky	workplaces_km	basketball_km
104.9341234	99.0978158	91.1681005
num_room	catering_km	zd_vokzaly_avto_km
86.0353438	85.3531218	78.8765905

Figure on the left shows top important variables in our SVM model. According to this model, the total area of the house, distance to public healthcare, living area, distance to swimming pool and kitchen area are important when deciding the price of the house.

### Model 4: Extreme Gradient Boosting

A check was done to find the optimal number of trees for the XGB model before the actual model is fitted.

Dataset A		Dataset B	
RMSE	R <sup>2</sup>	RMSE	R <sup>2</sup>
1,293,262	0.8771826	1,390,407	0.8663928

The RMSE and R<sup>2</sup> value are the best among the 4 models.

Feature <chr>	Gain <dbl>	Cover <dbl>	Frequency <dbl>
full_sq	4.692065e-01	1.295774e-01	0.1744719927
zd_vokzaly_avto_km	1.786603e-01	4.901979e-02	0.0624426079
cafe_count_3000_price_500	3.099059e-02	3.130970e-02	0.0266299357
kremlin_km	3.058697e-02	1.565328e-02	0.0224977043
cafe_count_5000_price_500	2.182409e-02	2.369141e-02	0.0146923783
kitch_sq	1.854038e-02	3.256654e-02	0.0509641873
max_floor	1.767677e-02	6.711795e-02	0.0509641873
cafe_sum_5000_max_price_avg	1.625868e-02	3.770026e-02	0.0275482094
kindergarten_km	1.590784e-02	1.570766e-02	0.0275482094
big_road1_km	1.540030e-02	3.602791e-02	0.0298438935

Figure on the left shows the top important variables in the XGB model. According to this model, the total area of the house, distance to station, cheap café nearby, distance to the city are important when deciding the price of the house.

## 4.3 ASSESS MODEL

All the 4 models are tested with test data and the results are as follow:

### RMSE and R<sup>2</sup>

Model	RMSE				R <sup>2</sup>			
	Dataset A		Dataset B		Dataset A		Dataset B	
	Train	Test	Train	Test	Train	Test	Train	Test
RF	849,651	1,905,511	909,387	2,002,071	0.953272	0.716256	0.957529	0.699881
LM	2,285,564	2,198,910	2,273,802	2,182,181	0.610161	0.622090	0.632555	0.645062
SVM	1,901,610	1,961,660	1,830,578	1,942,776	0.732987	0.705677	0.752117	0.710382
XGB	1,293,262	1,966,564	1,390,407	1,999,401	0.877183	0.698355	0.866393	0.692289

Model	Results from test data
RF	From the test results, the model shows some sign of over fitting.
LM	This model is stable with both RMSE and $R^2$ quite close to each other for both train and test data. However, in terms of performance, this is not the best model among the 4.
SVM	This model is stable with both RMSE and $R^2$ quite close to each other for both train and test data. The model is performing well when compared to the rest of the models.
XGB	From the test results, the model shows some sign of over fitting.

Based on RMSE and  $R^2$  metrics, we can see that Random Forest with Dataset A is the best model. However, when comparing its performance against training and test data, we think this model is unstable. The second-best model is **SVM model with Dataset B**. It performs quite well and is stable.

#### **Plots: Actual vs Prediction and Residual vs Prediction**

The plots for all 4 models is given in Appendix C for dataset A and in Appendix D for dataset B.

From the actual vs prediction plots and residuals vs prediction plots, it can be seen that the SVM model outperforms the other models.

#### **Gain Curve comparison**

The gain curve plots are given in Appendix E for dataset A and B. From the curves, it is hard to tell which model is better. However, from the Gini score, the SVM model gives a slightly better score.

## **5. EVALUATION**

### **5.1 Evaluation Results**

The business objective is to predict the price of a property in Moscow. Using the models that are built, we are able to achieve that with minimal amount of errors. Hence, this project has met its initial business objectives.

Based on the evaluation, it can be concluded that the SVM model on Dataset B is the best model for our business needs. The model has best and most stable performance compared with the others.

### **5.2 Review Process**

Here is the summary on the processes that we have taken for this project:

1. Collect the data: property transaction data in Moscow
2. Define business objectives: to predict the property's price in Moscow.
3. We have 3 hypotheses that need to be clarified through data exploration:
  - Price of the house is affected by its physical characteristics.
  - Price of the house is affected by the population density where it is located at.
  - Price of the house is affected by availability of amenities nearby.
4. Data exploration is done based on the hypotheses that we have made earlier:
  - Physical characteristics of the house are proven as the most important factors that drive the price of a property.
  - Population of the area is found to have weak relationship to the price.
  - Number of amenities nearby is also found to have weak to no relationship to the price.
5. Through the data exploration we were taking notes of outliers, data entry errors, and missing values. All of these findings are removed or imputed during the data preparation.
6. We used random forest to extract 30 most important features from the data.
7. Using random forest alone to do feature selection gave us 30 top variables. Some of them are questionable since most of them is highly correlated. And overall, these variables made the model not interpretable.

8. To help us with correlated issue on point 7, we did elimination of the redundant features by using correlation analysis before we use random forest for feature selection.
9. Using both correlation analysis and random forest, we were able to get top 30 variables. And overall, they can describe our data better.
10. These 30 features then transformed using log or square root transformation if it is skewed.
11. We also transformed our categorical columns into binary columns.
12. We generated 2 datasets, one with transformations and one without any transformations. Both of the datasets were used to build our model.
13. We have chosen 4 modeling techniques that are proven to work well on regression problem:
  - Random Forest (RF)
  - Linear Regression (LM)
  - Support Vector Machine (SVM)
  - Extreme Gradient Boost (XGBoost)
14. We use z-score scaling to normalized our variables. This is needed for linear model and SVM only.
15. We evaluated our models using their RMSE and  $R^2$  value. And we also plot their predictions against the actual and the residuals.
16. Based on our evaluation, SVM is the best model for our business needs.

### 5.3 Determine next steps

The next steps will be deploying the model to real application. Based on our evaluation, we think we have built model that stable and good enough for real application. However, we will only be able to know that by testing the model against real application.

## APPENDIX A

No.	Column Name	Description
1	id	transaction id
2	timestamp	date of transaction
3	full_sq	total area in square meters, including loggias, balconies and other non-residential areas
4	life_sq	living area in square meters, excluding loggias, balconies and other non-residential areas
5	floor	for apartments, floor of the building
6	max_floor	number of floors in the building
7	material	wall material
8	build_year	year built
9	num_room	number of living rooms
10	kitch_sq	kitchen area
11	state	apartment condition
12	product_type	owner-occupier purchase or investment
13	sub_area	name of the district
14	area_m	Area mun. area, sqm
15	raion_popul	Number of municipality population. district
16	green_zone_part	Proportion of area of greenery in the total area
17	indust_part	Share of industrial zones in area of the total area
18	children_preschool	Number of pre-school age population
19	preschool_quota	Number of seats in pre-school institutions
20	preschool_education_centers_raion	Number of pre-school institutions
21	children_school	Population of school-age children
22	school_quota	Number of high school seats in area
23	school_education_centers_raion	Number of high school institutions
24	school_education_centers_top_20_raion	Number of high schools of the top 20 best schools in Moscow
25	hospital_beds_raion	Number of hospital beds for the district
26	healthcare_centers_raion	Number of healthcare centers in district
27	university_top_20_raion	Number of higher education institutions in the top ten ranking of the Federal rank
28	sport_objects_raion	Number of higher education institutions
29	additional_education_raion	Number of additional education organizations
30	culture_objects_top_25	Presence of the key objects of cultural heritage (significant objects for the level of the RF constituent entities, city)
31	culture_objects_top_25_raion	Number of objects of cultural heritage
32	shopping_centers_raion	Number of malls and shopping centers in district
33	office_raion	Number of malls and shopping centers in district
34	thermal_power_plant_raion	Presence of thermal power station in district
35	incineration_raion	Presence of incinerators
36	oil_chemistry_raion	Presence of dirty industries
37	radiation_raion	Presence of radioactive waste disposal
38	railroad_terminal_raion	Presence of the railroad terminal in district
39	big_market_raion	Presence of large grocery / wholesale markets
40	nuclear_reactor_raion	Presence of existing nuclear reactors
41	detention_facility_raion	Presence of detention centers, prisons
42	full_all	subarea population
43	male_f	Male population
44	female_f	Female population
45	young_all	Population younger than working age
46	young_male	Male population younger than working age
47	young_female	Female population younger than working age
48	work_all	Working-age population
49	work_male	Male working-age population

No.	Column Name	Description
50	work_female	Female working-age population
51	ekder_all	Population older than working age
52	ekder_male	Male population older than working age
53	ekder_female	Female population older than working age
54	X0_6_all	Population aged 0-6
55	X0_6_male	Male population aged 0-7
56	X0_6_female	Female population aged 0-8
57	X7_14_all	Population aged 7-14
58	X7_14_male	Male population aged 7-14
59	X7_14_female	Female population aged 7-14
60	X0_17_all	Population aged 0-17
61	X0_17_male	Male population aged 0-17
62	X0_17_female	Female population aged 0-17
63	X16_29_all	Population aged 16-19
64	X16_29_male	Male population aged 16-19
65	X16_29_female	Female population aged 16-19
66	X0_13_all	Population aged 0-13
67	X0_13_male	Male population aged 0-13
68	X0_13_female	Female population aged 0-13
69	raion_build_count_with_material_info	Number of building with material info in district
70	build_count_block	Share of block buildings
71	build_count_wood	Share of wood buildings
72	build_count_frame	Share of frame buildings
73	build_count_brick	Share of brick buildings
74	build_count_monolith	Share of monolith buildings
75	build_count_panel	Share of panel buildings
76	build_count_foam	Share of foam buildings
77	build_count_slag	Share of slag buildings
78	build_count_mix	Share of mixed buildings
79	raion_build_count_with_builddate_info	Number of building with build year info in district
80	build_count_before_1920	Share of before_1920 buildings
81	build_count_1921.1945	Share of 1921-1945 buildings
82	build_count_1946.1970	Share of 1946-1970 buildings
83	build_count_1971.1995	Share of 1971-1995 buildings
84	build_count_after_1995	Share of after_1995 buildings
85	ID_metro	Nearest metro id
86	metro_min_avto	Time to subway by car, min.
87	metro_km_avto	Distance to subway by car, km
88	metro_min_walk	Time to metro by foot
89	metro_km_walk	Distance to the metro, km
90	kindergarten_km	Distance to kindergarten
91	school_km	Distance to high school
92	park_km	Distance to park
93	green_zone_km	Distance to green zone
94	industrial_km	Distance to industrial zone
95	water_treatment_km	Distance to water treatment
96	cemetery_km	Distance to the cemetery
97	incineration_km	Distance to the incineration
98	railroad_station_walk_km	Distance to the railroad station (walk)
99	railroad_station_walk_min	Time to the railroad station (walk)
100	ID_railroad_station_walk	Nearest railroad station id (walk)
101	railroad_station_avto_km	Distance to the railroad station (avto)
102	railroad_station_avto_min	Time to the railroad station (avto)
103	ID_railroad_station_avto	Nearest railroad station id (avto)
104	public_transport_station_km	Distance to the public transport station (walk)
105	public_transport_station_min_walk	Time to the public transport station (walk)



No.	Column Name	Description
106	water_km	Distance to the water reservoir / river
107	water_1line	First line to the river (150 m)
108	mkad_km	Distance to MKAD (Moscow Circle Auto Road)
109	ttk_km	Distance to the TTC (Third Transport Ring)
110	sadovoe_km	Distance to the Garden Ring
111	bulvar_ring_km	The distance to the Boulevard Ring
112	kremlin_km	Distance to the city center (Kremlin)
113	big_road1_km	Distance to Nearest major road
114	ID_big_road1	Nearest big road id
115	big_road1_1line	First line to the road (100 m for highways, 250 m to MKAD)
116	big_road2_km	The distance to next distant major road
117	ID_big_road2	2nd nearest big road id
118	railroad_km	Distance to the railway / Moscow Central Ring / open areas Underground
119	railroad_1line	First line to the railway (100 m)
120	zd_vokzaly_avto_km	Distance to train station
121	ID_railroad_terminal	Nearest railroad terminal id
122	bus_terminal_avto_km	Distance to bus terminal (avto)
123	ID_bus_terminal	Nearest bus terminal id
124	oil_chemistry_km	Distance to dirty industries
125	nuclear_reactor_km	Distance to nuclear reactor
126	radiation_km	Distance to burial of radioactive waste
127	power_transmission_line_km	Distance to power transmission line
128	thermal_power_plant_km	Distance to thermal power plant
129	ts_km	Distance to power station
130	big_market_km	Distance to grocery / wholesale markets
131	market_shop_km	Distance to markets and department stores
132	fitness_km	Distance to fitness
133	swim_pool_km	Distance to swimming pool
134	ice_rink_km	Distance to ice palace
135	stadium_km	Distance to stadium
136	basketball_km	Distance to the basketball courts
137	hospice_morgue_km	Distance to hospice/morgue
138	detention_facility_km	Distance to detention facility
139	public_healthcare_km	Distance to public healthcare
140	university_km	Distance to universities
141	workplaces_km	Distance to workplaces
142	shopping_centers_km	Distance to shopping centers
143	office_km	Distance to business centers/ offices
144	additional_education_km	Distance to additional education
145	preschool_km	Distance to preschool education organizations
146	big_church_km	Distance to large church
147	church_synagogue_km	Distance to Christian churches and Synagogues
148	mosque_km	Distance to mosques
149	theater_km	Distance to theater
150	museum_km	Distance to museums
151	exhibition_km	Distance to exhibition
152	catering_km	Distance to catering
153	ecology	Ecological zone where the house is located
154	green_part_500	The share of green zones in 500 meters zone
155	prom_part_500	The share of industrial zones in 500 meters zone
156	office_count_500	The number of office space in 500 meters zone
157	office_sqm_500	The square of office space in 500 meters zone
158	trc_count_500	The number of shopping malls in 500 meters zone
159	trc_sqm_500	The square of shopping malls in 500 meters zone
160	cafe_count_500	The number of cafes or restaurants in 500 meters zone



No.	Column Name	Description
161	cafe_sum_500_min_price_avg	Cafes and restaurant min average bill in 500 meters zone
162	cafe_sum_500_max_price_avg	Cafes and restaurant max average bill in 500 meters zone
163	cafe_avg_price_500	Cafes and restaurant average bill in 500 meters zone
164	cafe_count_500_na_price	Cafes and restaurant bill N/A in 500 meters zone
165	cafe_count_500_price_500	Cafes and restaurant bill, average under 500 in 500 meters zone
166	cafe_count_500_price_1000	Cafes and restaurant bill, average 500-1000 in 500 meters zone
167	cafe_count_500_price_1500	Cafes and restaurant bill, average 1000-1500 in 500 meters zone
168	cafe_count_500_price_2500	Cafes and restaurant bill, average 1500-2500 in 500 meters zone
169	cafe_count_500_price_4000	Cafes and restaurant bill, average 2500-4000 in 500 meters zone
170	cafe_count_500_price_high	Cafes and restaurant bill, average over 4000 in 500 meters zone
171	big_church_count_500	The number of big churches in 500 meters zone
172	church_count_500	The number of churches in 500 meters zone
173	mosque_count_500	The number of mosques in 500 meters zone
174	leisure_count_500	The number of leisure facilities in 500 meters zone
175	sport_count_500	The number of sport facilities in 500 meters zone
176	market_count_500	The number of markets in 500 meters zone
177	green_part_1000	The share of green zones in 1000 meters zone
178	prom_part_1000	The share of industrial zones in 1000 meters zone
179	office_count_1000	The number of office space in 1000 meters zone
180	office_sqm_1000	The square of office space in 1000 meters zone
181	trc_count_1000	The number of shopping malls in 1000 meters zone
182	trc_sqm_1000	The square of shopping malls in 1000 meters zone
183	cafe_count_1000	The number of cafes or restaurants in 1000 meters zone
184	cafe_sum_1000_min_price_avg	Cafes and restaurant min average bill in 1000 meters zone
185	cafe_sum_1000_max_price_avg	Cafes and restaurant max average bill in 1000 meters zone
186	cafe_avg_price_1000	Cafes and restaurant average bill in 1000 meters zone
187	cafe_count_1000_na_price	Cafes and restaurant bill N/A in 1000 meters zone
188	cafe_count_1000_price_500	Cafes and restaurant bill, average under 500 in 1000 meters zone
189	cafe_count_1000_price_1000	Cafes and restaurant bill, average 500-1000 in 1000 meters zone
190	cafe_count_1000_price_1500	Cafes and restaurant bill, average 1000-1500 in 1000 meters zone
191	cafe_count_1000_price_2500	Cafes and restaurant bill, average 1500-2500 in 1000 meters zone
192	cafe_count_1000_price_4000	Cafes and restaurant bill, average 2500-4000 in 1000 meters zone
193	cafe_count_1000_price_high	Cafes and restaurant bill, average over 4000 in 1000 meters zone
194	big_church_count_1000	The number of big churches in 1000 meters zone
195	church_count_1000	The number of churches in 1000 meters zone
196	mosque_count_1000	The number of mosques in 1000 meters zone
197	leisure_count_1000	The number of leisure facilities in 1000 meters zone
198	sport_count_1000	The number of sport facilities in 1000 meters zone
199	market_count_1000	The number of markets in 1000 meters zone
200	green_part_1500	The share of green zones in 1500 meters zone
201	prom_part_1500	The share of industrial zones in 1500 meters zone
202	office_count_1500	The number of office space in 1500 meters zone
203	office_sqm_1500	The square of office space in 1500 meters zone
204	trc_count_1500	The number of shopping malls in 1500 meters zone
205	trc_sqm_1500	The square of shopping malls in 1500 meters zone
206	cafe_count_1500	The number of cafes or restaurants in 1500 meters zone
207	cafe_sum_1500_min_price_avg	Cafes and restaurant min average bill in 1500 meters zone
208	cafe_sum_1500_max_price_avg	Cafes and restaurant max average bill in 1500 meters zone
209	cafe_avg_price_1500	Cafes and restaurant average bill in 1500 meters zone
210	cafe_count_1500_na_price	Cafes and restaurant bill N/A in 1500 meters zone
211	cafe_count_1500_price_500	Cafes and restaurant bill, average under 500 in 1500 meters zone
212	cafe_count_1500_price_1000	Cafes and restaurant bill, average 500-1000 in 1500 meters zone
213	cafe_count_1500_price_1500	Cafes and restaurant bill, average 1000-1500 in 1500 meters zone
214	cafe_count_1500_price_2500	Cafes and restaurant bill, average 1500-2500 in 1500 meters zone
215	cafe_count_1500_price_4000	Cafes and restaurant bill, average 2500-4000 in 1500 meters zone
216	cafe_count_1500_price_high	Cafes and restaurant bill, average over 4000 in 1500 meters zone

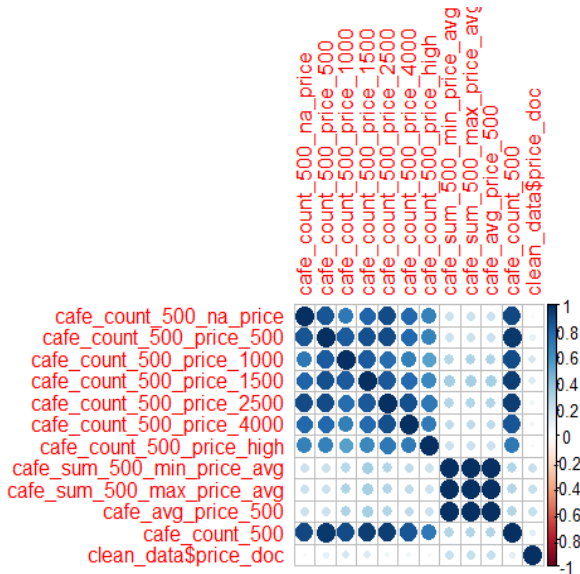
No.	Column Name	Description
217	big_church_count_1500	The number of big churches in 1500 meters zone
218	church_count_1500	The number of churches in 1500 meters zone
219	mosque_count_1500	The number of mosques in 1500 meters zone
220	leisure_count_1500	The number of leisure facilities in 1500 meters zone
221	sport_count_1500	The number of sport facilities in 1500 meters zone
222	market_count_1500	The number of markets in 1500 meters zone
223	green_part_2000	The share of green zones in 2000 meters zone
224	prom_part_2000	The share of industrial zones in 2000 meters zone
225	office_count_2000	The number of office space in 2000 meters zone
226	office_sqm_2000	The square of office space in 2000 meters zone
227	trc_count_2000	The number of shopping malls in 2000 meters zone
228	trc_sqm_2000	The square of shopping malls in 2000 meters zone
229	cafe_count_2000	The number of cafes or restaurants in 1500 meters zone
230	cafe_sum_2000_min_price_avg	Cafes and restaurant min average bill in 2000 meters zone
231	cafe_sum_2000_max_price_avg	Cafes and restaurant max average bill in 2000 meters zone
232	cafe_avg_price_2000	Cafes and restaurant average bill in 2000 meters zone
233	cafe_count_2000_na_price	Cafes and restaurant bill N/A in 2000 meters zone
234	cafe_count_2000_price_500	Cafes and restaurant bill, average under 500 in 2000 meters zone
235	cafe_count_2000_price_1000	Cafes and restaurant bill, average 500-1000 in 2000 meters zone
236	cafe_count_2000_price_1500	Cafes and restaurant bill, average 1000-1500 in 2000 meters zone
237	cafe_count_2000_price_2500	Cafes and restaurant bill, average 1500-2500 in 2000 meters zone
238	cafe_count_2000_price_4000	Cafes and restaurant bill, average 2500-4000 in 2000 meters zone
239	cafe_count_2000_price_high	Cafes and restaurant bill, average over 4000 in 2000 meters zone
240	big_church_count_2000	The number of big churches in 2000 meters zone
241	church_count_2000	The number of churches in 2000 meters zone
242	mosque_count_2000	The number of mosques in 2000 meters zone
243	leisure_count_2000	The number of leisure facilities in 2000 meters zone
244	sport_count_2000	The number of sport facilities in 2000 meters zone
245	market_count_2000	The number of markets in 2000 meters zone
246	green_part_3000	The share of green zones in 3000 meters zone
247	prom_part_3000	The share of industrial zones in 3000 meters zone
248	office_count_3000	The number of office space in 3000 meters zone
249	office_sqm_3000	The square of office space in 3000 meters zone
250	trc_count_3000	The number of shopping malls in 3000 meters zone
251	trc_sqm_3000	The square of shopping malls in 3000 meters zone
252	cafe_count_3000	The number of cafes or restaurants in 1500 meters zone
253	cafe_sum_3000_min_price_avg	Cafes and restaurant min average bill in 3000 meters zone
254	cafe_sum_3000_max_price_avg	Cafes and restaurant max average bill in 3000 meters zone
255	cafe_avg_price_3000	Cafes and restaurant average bill in 3000 meters zone
256	cafe_count_3000_na_price	Cafes and restaurant bill N/A in 3000 meters zone
257	cafe_count_3000_price_500	Cafes and restaurant bill, average under 500 in 3000 meters zone
258	cafe_count_3000_price_1000	Cafes and restaurant bill, average 500-1000 in 3000 meters zone
259	cafe_count_3000_price_1500	Cafes and restaurant bill, average 1000-1500 in 3000 meters zone
260	cafe_count_3000_price_2500	Cafes and restaurant bill, average 1500-2500 in 3000 meters zone
261	cafe_count_3000_price_4000	Cafes and restaurant bill, average 2500-4000 in 3000 meters zone
262	cafe_count_3000_price_high	Cafes and restaurant bill, average over 4000 in 3000 meters zone
263	big_church_count_3000	The number of big churches in 3000 meters zone
264	church_count_3000	The number of churches in 3000 meters zone
265	mosque_count_3000	The number of mosques in 3000 meters zone
266	leisure_count_3000	The number of leisure facilities in 3000 meters zone
267	sport_count_3000	The number of sport facilities in 3000 meters zone
268	market_count_3000	The number of markets in 3000 meters zone
269	green_part_5000	The share of green zones in 5000 meters zone
270	prom_part_5000	The share of industrial zones in 5000 meters zone
271	office_count_5000	The number of office space in 5000 meters zone
272	office_sqm_5000	The square of office space in 5000 meters zone

No.	Column Name	Description
273	trc_count_5000	The number of shopping malls in 5000 meters zone
274	trc_sqm_5000	The square of shopping malls in 5000 meters zone
275	cafe_count_5000	The number of cafes or restaurants in 1500 meters zone
276	cafe_sum_5000_min_price_avg	Cafes and restaurant min average bill in 5000 meters zone
277	cafe_sum_5000_max_price_avg	Cafes and restaurant max average bill in 5000 meters zone
278	cafe_avg_price_5000	Cafes and restaurant average bill in 5000 meters zone
279	cafe_count_5000_na_price	Cafes and restaurant bill N/A in 5000 meters zone
280	cafe_count_5000_price_500	Cafes and restaurant bill, average under 500 in 5000 meters zone
281	cafe_count_5000_price_1000	Cafes and restaurant bill, average 500-1000 in 5000 meters zone
282	cafe_count_5000_price_1500	Cafes and restaurant bill, average 1000-1500 in 5000 meters zone
283	cafe_count_5000_price_2500	Cafes and restaurant bill, average 1500-2500 in 5000 meters zone
284	cafe_count_5000_price_4000	Cafes and restaurant bill, average 2500-4000 in 5000 meters zone
285	cafe_count_5000_price_high	Cafes and restaurant bill, average over 4000 in 5000 meters zone
286	big_church_count_5000	The number of big churches in 5000 meters zone
287	church_count_5000	The number of churches in 5000 meters zone
288	mosque_count_5000	The number of mosques in 5000 meters zone
289	leisure_count_5000	The number of leisure facilities in 5000 meters zone
290	sport_count_5000	The number of sport facilities in 5000 meters zone
291	market_count_5000	The number of markets in 5000 meters zone
292	price_doc	sale price (this is the target variable)

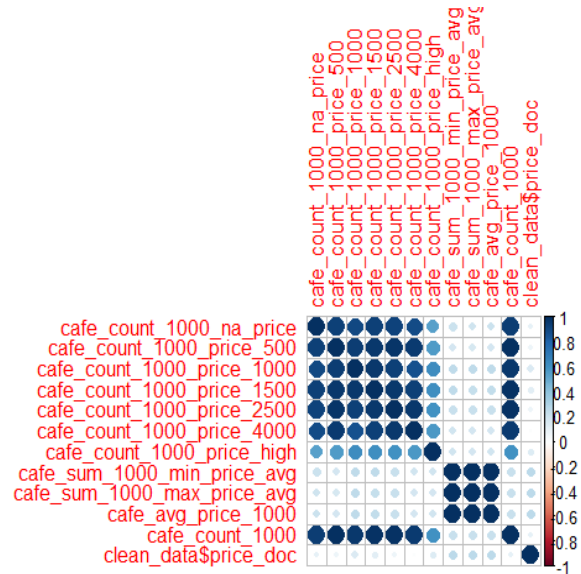
## APPENDIX B

## List of all Correlation Plots

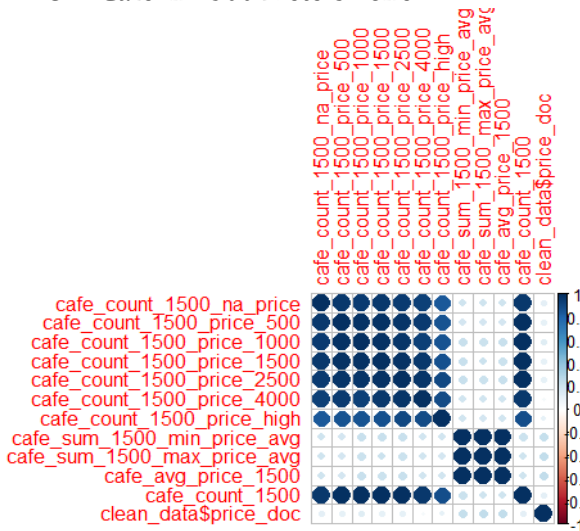
## 1. Cafe in 500 meters zone



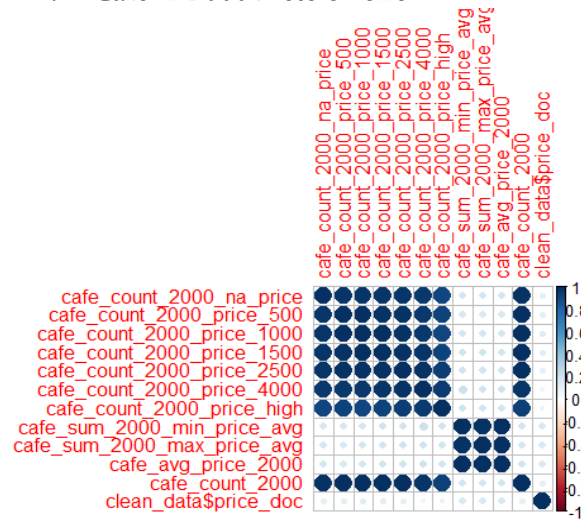
## 2. Cafe in 1000 meters zone



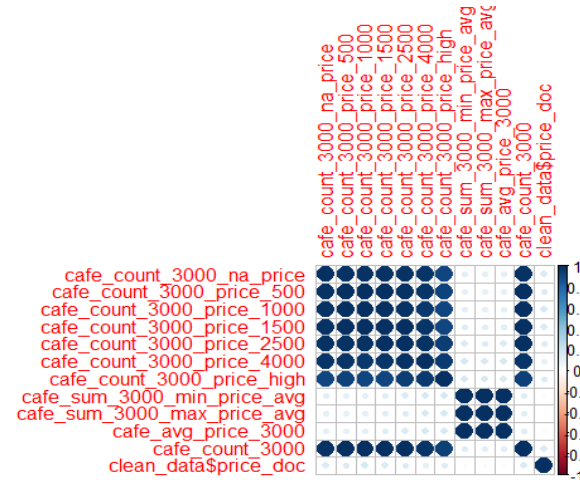
## 3. Cafe in 1500 meters zone



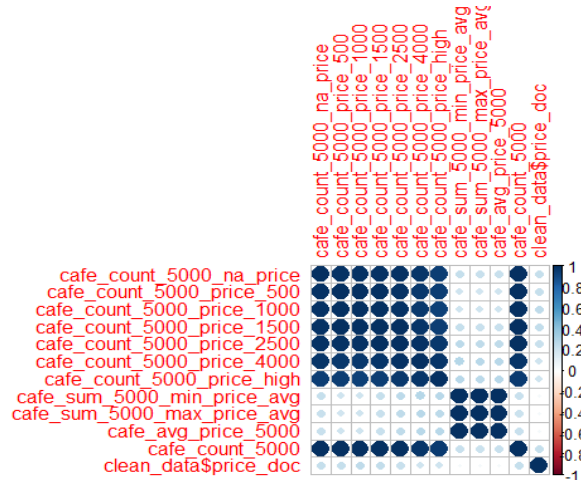
## 4. Cafe in 2000 meters zone



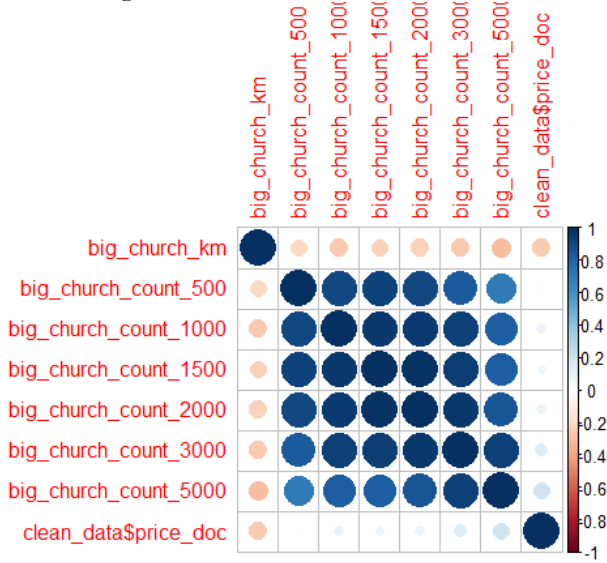
## 5. Cafe in 3000 meters zone



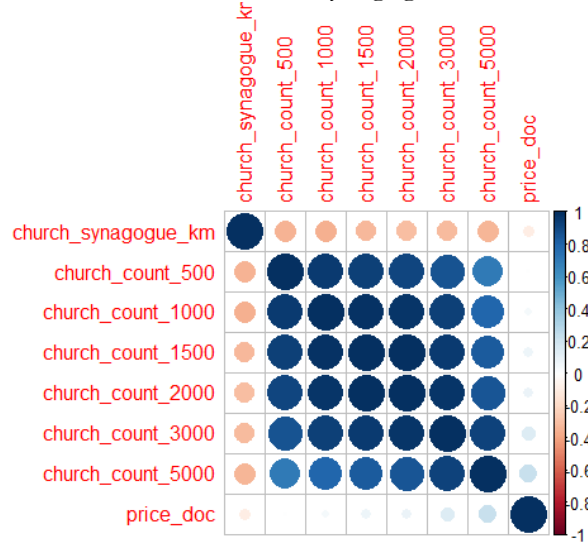
## 6. Cafe in 5000 meters zone



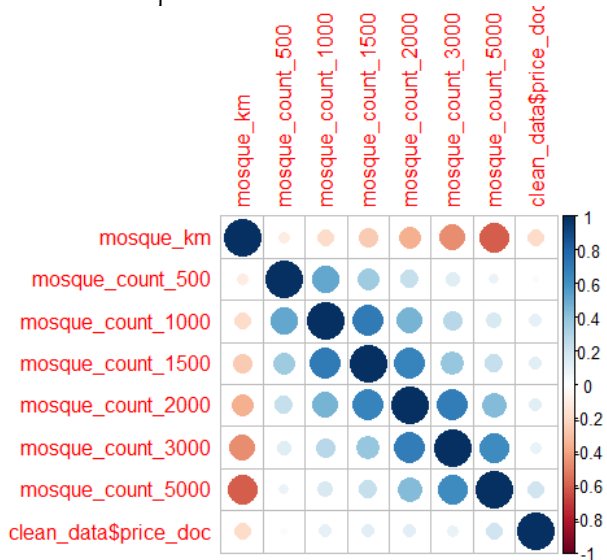
## 7. Big Church



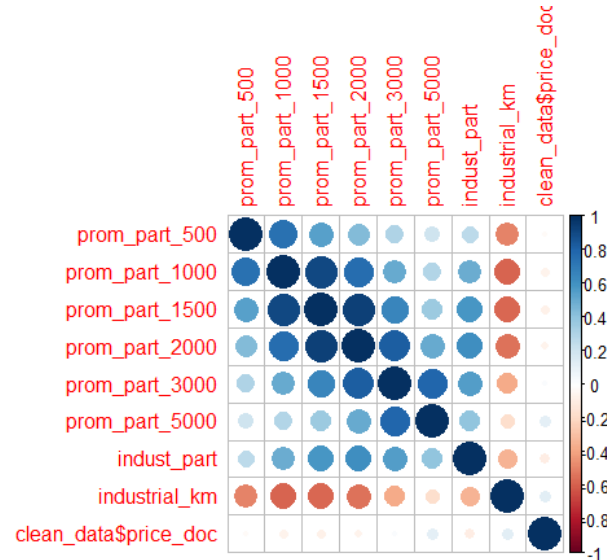
## 8. Small Church & Synagogue



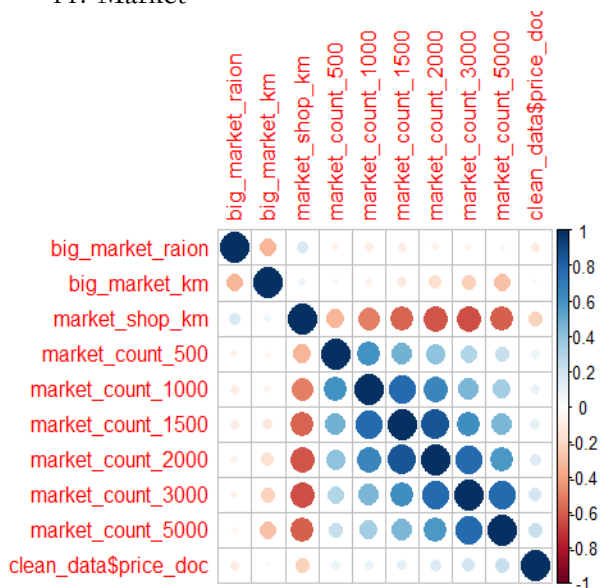
## 9. Mosque



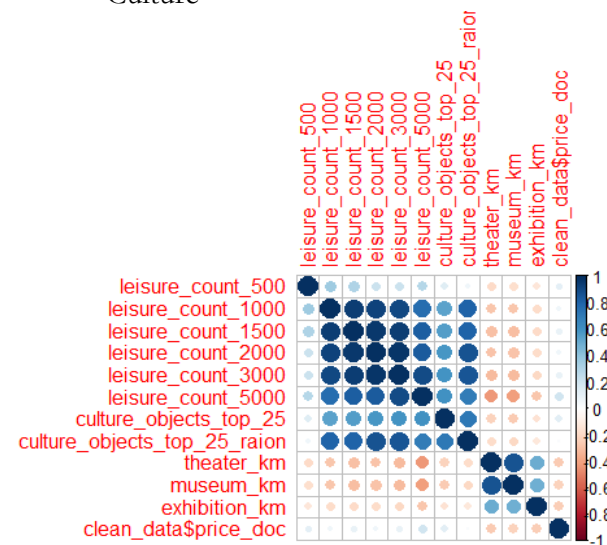
## 10. Industrial



## 11. Market



## 12. Leisure, Theater, Museum, Exhibition, Culture

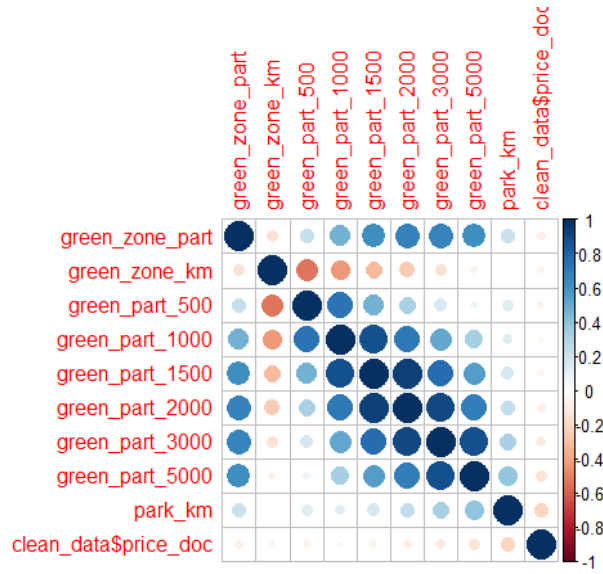




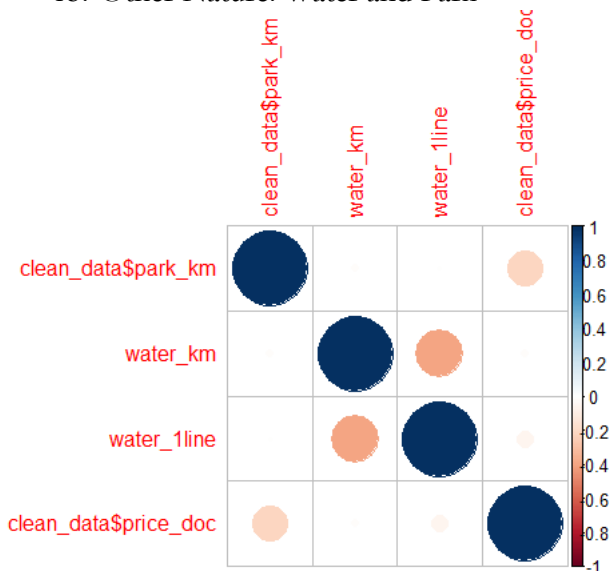
13. Healthcare, Hospice, Cemetery



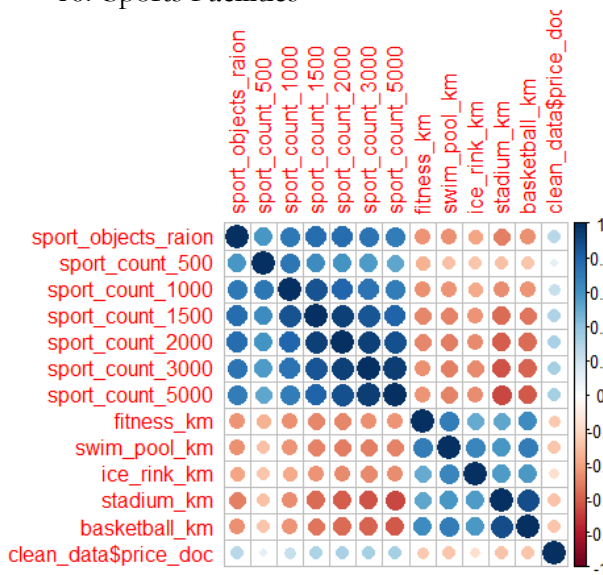
14. Green Zone



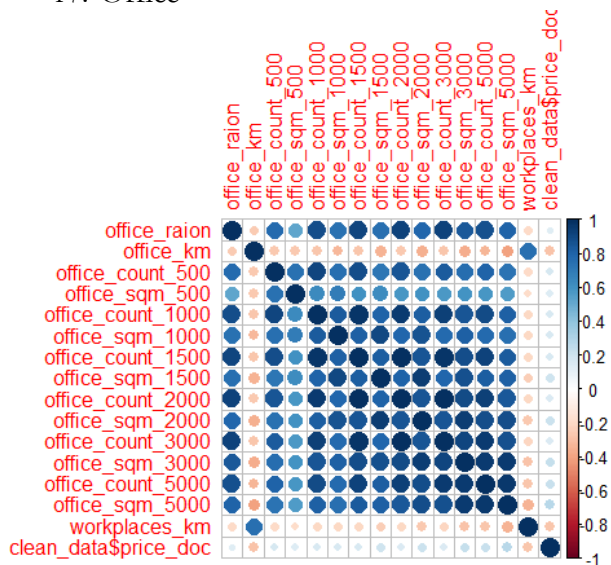
15. Other Nature: Water and Park



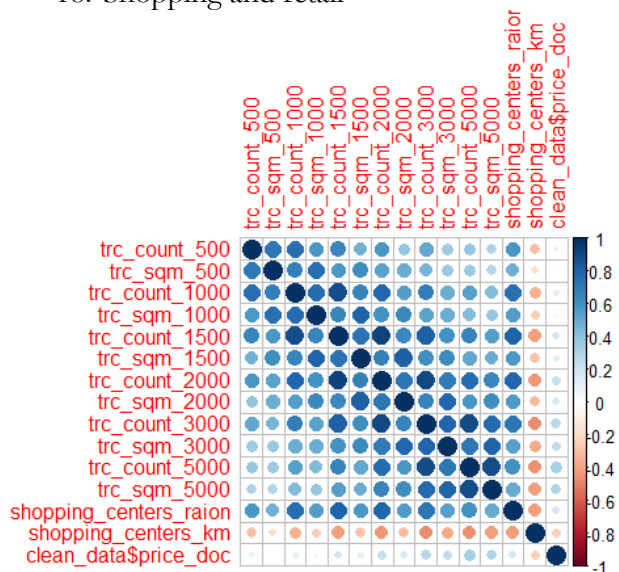
16. Sports Facilities



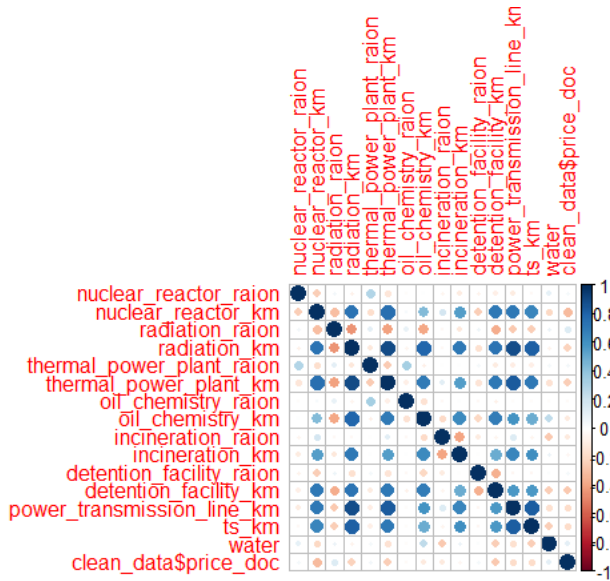
17. Office



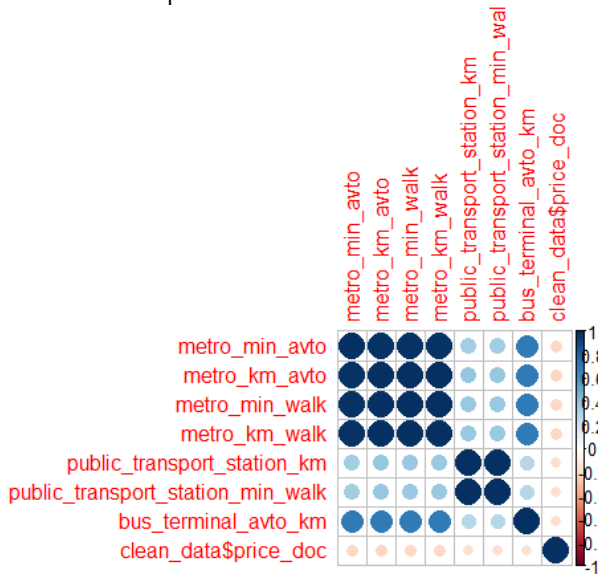
18. Shopping and retail



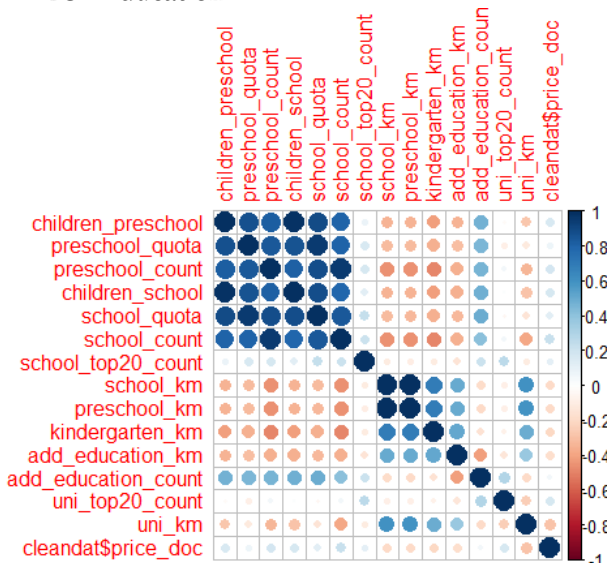
## 19. Utilities



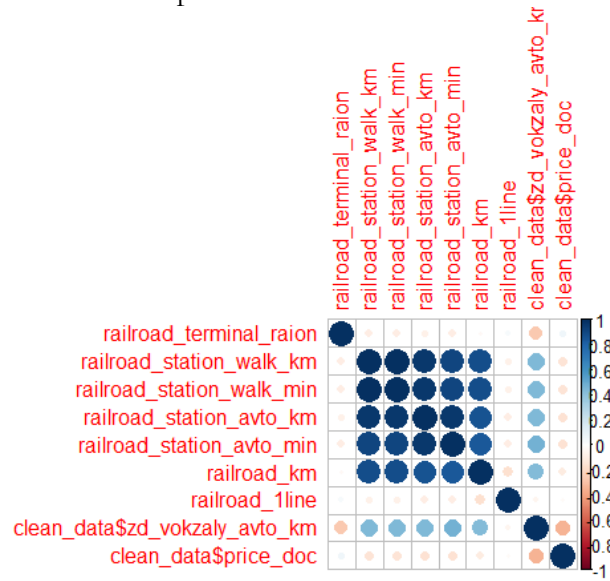
## 21. Public Transport: Metro, Bus and Public Transport



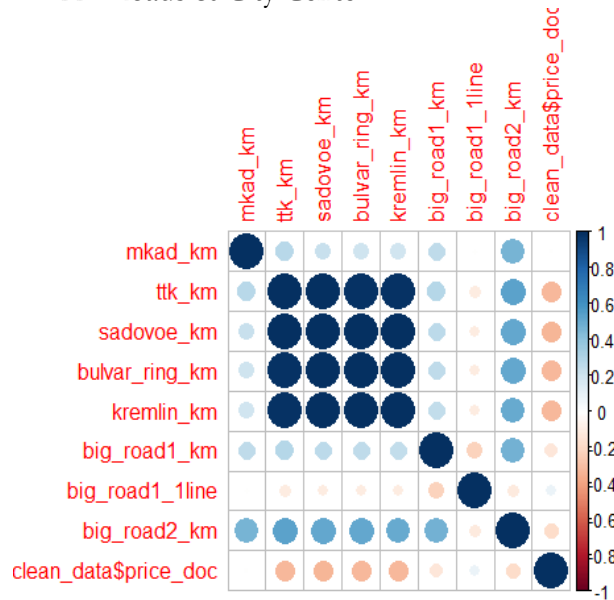
## 23. Education



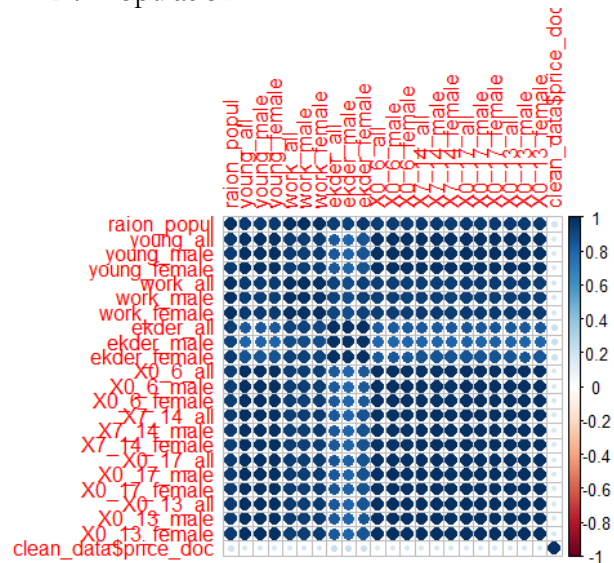
## 20. Transport: Railroad



## 22. Roads &amp; City Center



## 24. Population

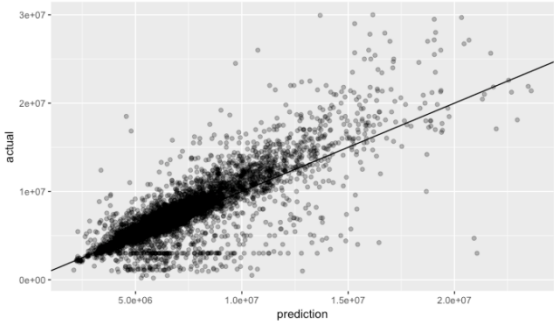
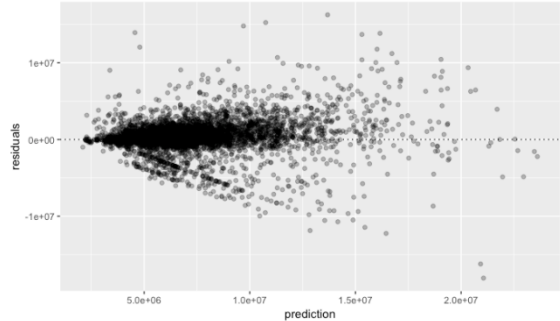
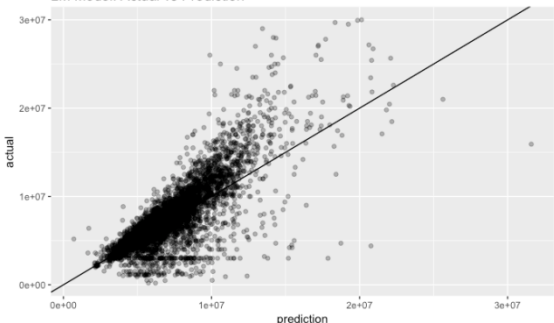
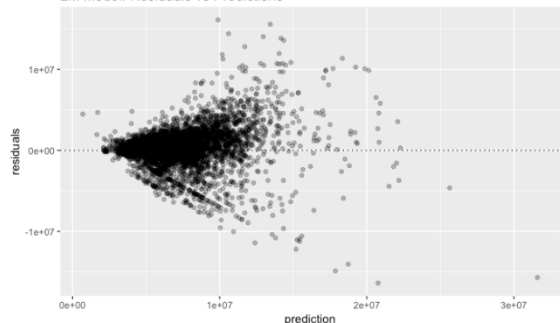
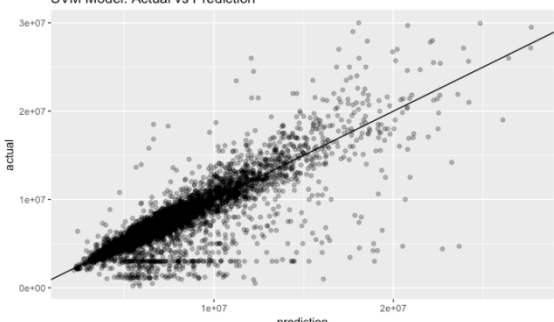
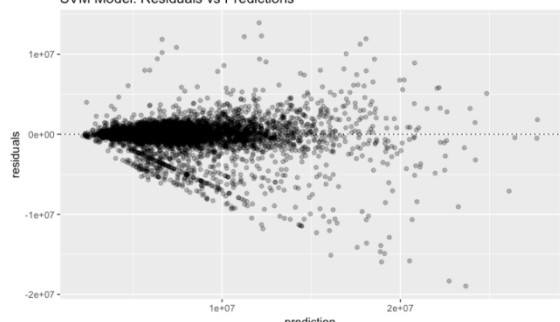
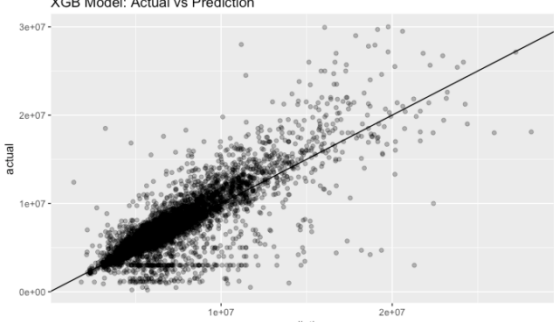
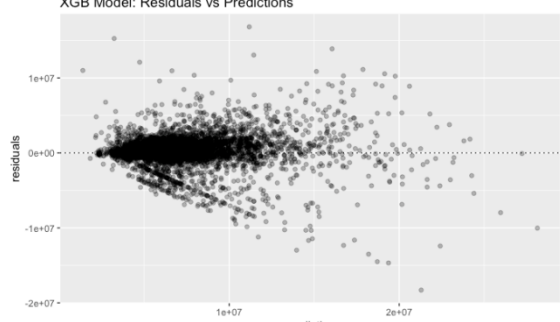


## APPENDIX C

Model	Dataset A	
	Actual vs Prediction plot comparison	Residual vs Prediction plot comparison
RF	<p>RF Model: Actual vs Prediction</p>	<p>RF Model: Residuals vs Predictions</p>
LM	<p>LM Model: Actual vs Prediction</p>	<p>LM Model: Residuals vs Predictions</p>
SVM	<p>SVM Model: Actual vs Prediction</p>	<p>SVM Model: Residuals vs Predictions</p>
XGB	<p>XGB Model: Actual vs Prediction</p>	<p>XGB Model: Residuals vs Predictions</p>

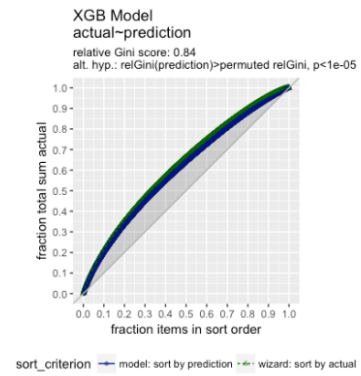
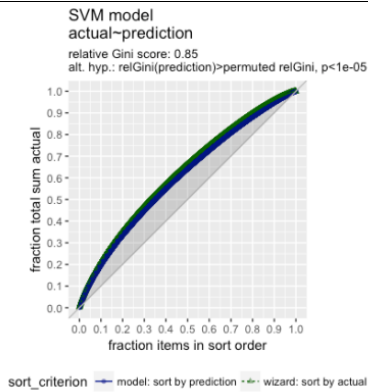
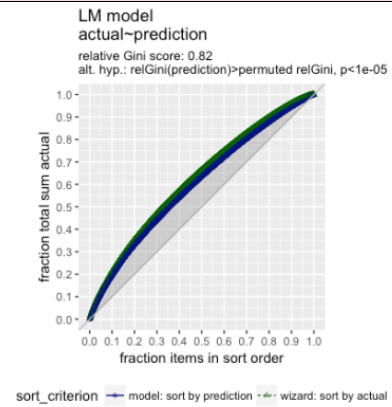
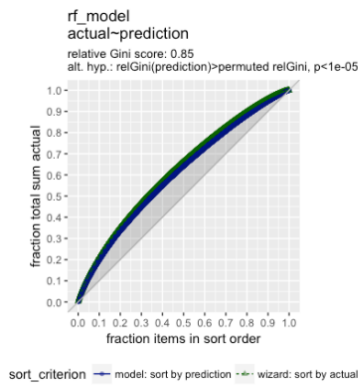


APPENDIX D

Model	Dataset B	
	Actual vs Prediction plot comparison	Residual vs Prediction plot comparison
RF	<div>RF Model: Actual vs Prediction</div> 	<div>RF Model: Residuals vs Predictions</div> 
LM	<div>LM Model: Actual vs Prediction</div> 	<div>LM Model: Residuals vs Predictions</div> 
SVM	<div>SVM Model: Actual vs Prediction</div> 	<div>SVM Model: Residuals vs Predictions</div> 
XGB	<div>XGB Model: Actual vs Prediction</div> 	<div>XGB Model: Residuals vs Predictions</div> 

## APPENDIX E

## Dataset A



## Dataset B

