MIDDLE EAST TECHNICAL UNIVERSITY

STAT 467
MULTIVARIATE ANALYSIS
TERM PROJECT

"Real Estate Data"

Submitted to Prof. Dr. Barış SÜRÜCÜ

Bennur Kaya 2146181
Beste Karaçay 2146165
Merve Erşahin 2146116

TABLE OF CONTENT

1. INTRODUCTION

1.1 Abstract

The real estate company, which the data used in this project came from, has been the Turkey's most known online real estate market. Construction sector has an important place in Turkey's economy. Real estate sales for the sector are increasing in parallel. However, the purchase cost is also important for those who are willing to buy a real estate. In the acquisition of real estate, factors such as size, location, age of the house and other features of the house are taken into consideration. The aim of the article is to conduct research on factors affecting real estate values by multivariate analysis. This project was conducted to see which factors affecting the price of the house or a work place.

1.2 Data Description

Real estate data is mainly constructed with the continuous and categorical variables based on 5282 observations. The data consists of 17 variables which 8 of them are continuous and 9 are categorical.

- "**Price**" (continuous) is the value of the house (in Turkish Liras).
- "**Room**" (continuous) is the number of rooms in the house.
- "**Salon**" (continuous) is the number of salons in the house.
- "**Sqm**" (continuous) represents the square meter of the house.
- "**Age**" (continuous) is the age of the building.
- "**Bathroom**" (continuous) is the number of bathrooms in the house.
- "**Floor**" (continuous) is the floor where the house is.
- "**FloorCount**" (continuous) is the total number of floors in the building.
- "**Heating**" (categorical) is the type of heating type in the building (in MerkeziPayOlcer, Merkezi, Kombi, Kalorifer).

- "**Type**" (categorical) is the purpose of the real estate (in 0: House, 1: Work place).
- "**Fuel**" (categorical) is the fuel type used in the house (in Dogalgaz, Elektrik, Komur).
- "**Build**" (categorical) represents what the house is made of (in Betonarme, Celik, Prefabrik, Ahsap).
- "**BuildState**" (categorical) represents the condition of the house (in 0: Brand-new, 1: Under construction, 2: Second-hand).
- "**Furnished**" (categorical) represents if the house is furnished or not (in 0: Not furnished, 1: Furnished)
- "**Usage**" (categorical) shows if someone lives in the house or not (in Bos, Kiraci, Ev Sahibi)
- "**Register**" (categorical) is the type of the land registry (in KatIrtifaki, KatMulkiyeti, Arsa)
- "**City**" (categorical) is the location of the house (in 0: İstanbul, 1: Ankara and 2: İzmir)

1.3 Aim of Research

The aim in this project is to explore the dataset and find basic relationship of different features with house's price. We started with some brief exploratory analysis involving summary statistics, graphical visualization and PCA. We also wanted to see if there is a way to predict the price of the house given the information we have. We will be comparing a few different logistic regression models to find the best one. And then, end with Canonical Analysis.

2. SURVEY METHODOLOGY

Data Visualization

- **Histogram**: A histogram is a plot that lets you discover, and show, the underlying frequency distribution (shape) of a set of continuous data. This allows the inspection of the data for its underlying distribution.

- **Box plot**: Box plots are another way to visualize the distribution of data values. In this respect, box plots are comparable to histograms, but are quite different in presentation.

- **Scree plot**: This technique is used in determining Principle Components number in PCA.

- **QQ plot**: A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another.

- **Classification Tree**: A classification tree (or a decision tree) is used to predict a qualitative response rather than a quantitative one.

Statistical Tests

- ***Principal Component Analysis (PCA):*** Principal component analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components.

- ***Factor Analysis*:** is a technique that is used to reduce a large number of variables into fewer numbers of factors. This technique extracts maximum common variance from all variables and puts them into a common score.

- ***Classification*:** A classification is an ordered set of related categories used to group data according to its similarities.

- ***K-means*:** K-means clustering is a simple unsupervised learning algorithm that is used to solve clustering problems. It follows a simple procedure of classifying a given data set into a number of clusters.

- ***Shapiro – Wilk test*:** The Shapiro-Wilk test is a way to tell if a random sample comes from a normal distribution.

- ***Logistic Regression*:** Logistic regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome.

- ***Random Forest***: Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set.

- ***Canonical Analysis***: Canonical analysis is a multivariate technique which is concerned with determining the relationships between groups of variables in a data set. The data set is split into two groups X and Y, based on some common characteristics.

3. DATA ANALYSIS, FINDINGS AND DISCUSSION

Price	Room	Salon	Sqm	Age	Heating	Type	Bathroom	Floor	FloorCount	Fuel	Build	BuildState	Furnished	Usage	Register	city
325000	1	0	55	0	MerkeziPayOlcer	0	1	11	14	Dogalgaz	Betonarme	0	0	Bos	KatMulkiyeti	0
315000	1	0	58	5	MerkeziPayOlcer	0	1	3	15	Dogalgaz	Betonarme	0	0	Kiraci	KatMulkiyeti	0
181000	1	0	48	2	Merkezi	0	1	7		Dogalgaz	Betonarme	0	0	Kiraci	KatIrtifaki	1
158000	1	0	55	5	Kombi	0	1	1	4	Dogalgaz	Celik	2	0	EvSahibi	KatMulkiyeti	0
150000	1	0	75	2	Merkezi	0	1		6		Betonarme	0	0	EvSahibi	Arsa	0
150000	1	0	49	2	Merkezi	0	1				Betonarme		0			0
140000	1	0	36	3	MerkeziPayOlcer	0	1	5	12	Elektrik	Betonarme	2	1	Kiraci	KatMulkiyeti	0
6500000	1	1	2598	1	Yok	0	1			Dogalgaz	Betonarme	0	0			2

Table 1. Preview of the data

As it can be seen from above table, the data has total of 17 variables which 8 of them are continuous and 9 are categorical. As the next step, we have correlations between continuous variables.

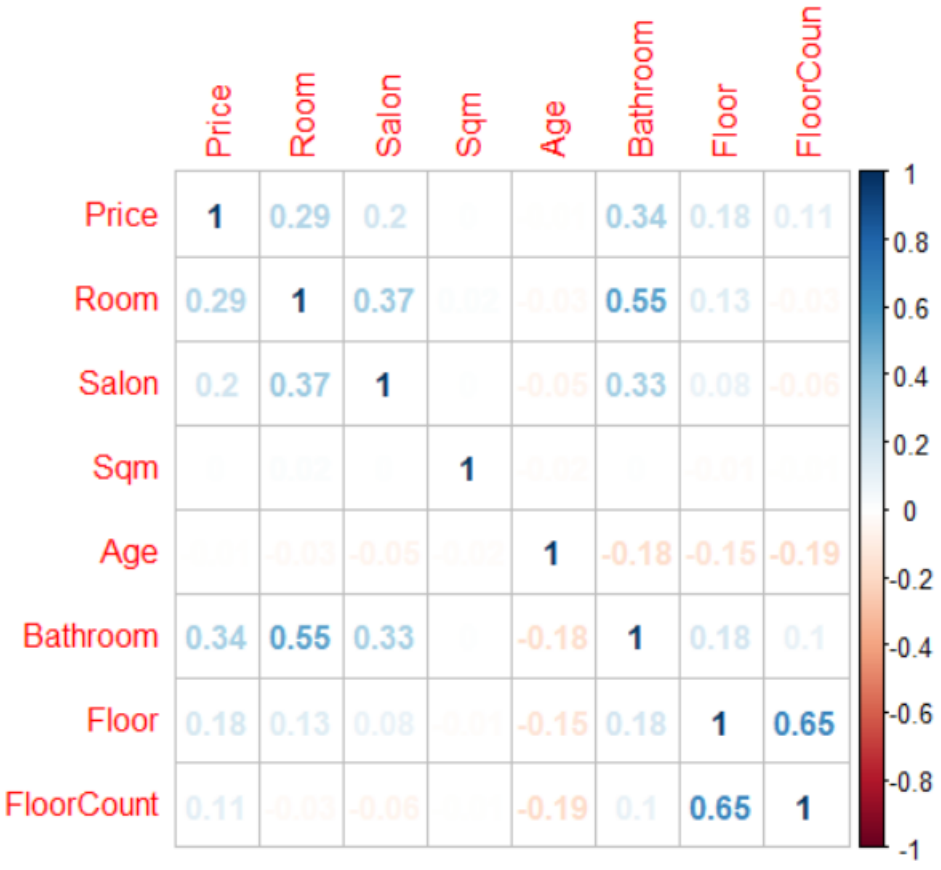


Figure 1. Correlations

The correlations are not as high as we expected. The most highly correlated variables are “Floor” and “FloorCount”. We would like to see a high correlation between

"price" and "square meter" or the "age". However, since it is a real data, this is not the case.

Also, since we only have one response variable, we tried to find another response which should be correlated with "price" variable. The logical choice was the "sqm", so we applied a test which has a null hypothesis stating that the true correlation is equal to 0. And we failed to reject H0, so in the end, we have only one response.

Pearson's product-moment correlation test

Data: price and Sqm

p-value = 0.95

alternative hypothesis: true correlation is not equal to 0

After correlations are checked, the summary table of variables can be observed. Since the range of the continuous variables are too high, especially for Price and Sqm variables, the data has been standardized.

```
     Price               Room            Salon             Sqm               Age            Bathroom
 Min.   :    1800   Min.   : 1.000   Min.   :0.000   Min.   :       1   Min.   :  0.00   Min.   : 1.000
 1st Qu.:  240000   1st Qu.: 2.000   1st Qu.:1.000   1st Qu.:      98   1st Qu.:  3.00   1st Qu.: 1.000
 Median :  360000   Median : 3.000   Median :1.000   Median :     125   Median : 10.00   Median : 1.000
 Mean   :  833475   Mean   : 2.934   Mean   :1.091   Mean   :    2543   Mean   : 13.16   Mean   : 1.543
 3rd Qu.:  635000   3rd Qu.: 3.000   3rd Qu.:1.000   3rd Qu.:     165   3rd Qu.: 20.00   3rd Qu.: 2.000
 Max.   :55000000   Max.   :17.000   Max.   :8.000   Max.   :11111111   Max.   :510.00   Max.   :21.000
                    NA's   :598      NA's   :598                        NA's   :465      NA's   :598
     Floor          FloorCount
 Min.   :-3.000   Min.   : 1.000
 1st Qu.: 0.000   1st Qu.: 4.000
 Median : 3.000   Median : 5.000
 Mean   : 3.217   Mean   : 6.599
 3rd Qu.: 4.000   3rd Qu.: 8.000
 Max.   :45.000   Max.   :45.000
 NA's   :2112     NA's   :2216
```

Figure 2. Summary of the data

```
     Price               Room             Salon              Sqm               Age             Bathroom
 Min.   :-0.4878   Min.   :-1.7800   Min.   :-3.7833   Min.   :-0.05883   Min.   :-0.8597   Min.   :-0.6699
 1st Qu.:-0.3419   1st Qu.:-0.8434   1st Qu.:-0.2315   1st Qu.:-0.03409   1st Qu.:-0.6608   1st Qu.:-0.6699
 Median :-0.2574   Median : 0.0932   Median :-0.2315   Median :-0.02490   Median :-0.1967   Median :-0.6699
 Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.0000   Mean   : 0.0000
 3rd Qu.:-0.0969   3rd Qu.: 0.0932   3rd Qu.:-0.2315   3rd Qu.:-0.01289   3rd Qu.: 0.4663   3rd Qu.: 0.6327
 Max.   :16.5888   Max.   :12.2689   Max.   :13.9753   Max.   :51.18156   Max.   :32.9528   Max.   :25.3827
     Floor          FloorCount
 Min.   :-1.6273   Min.   :-1.1479
 1st Qu.:-0.6353   1st Qu.:-0.5617
 Median :-0.1394   Median :-0.3663
 Mean   : 0.0000   Mean   : 0.0000
 3rd Qu.: 0.3566   3rd Qu.: 0.2198
 Max.   :10.2764   Max.   : 7.4491
```

Figure 3. Summary of the standardized data

After standardization, the ranges are narrower now.

Normality Test

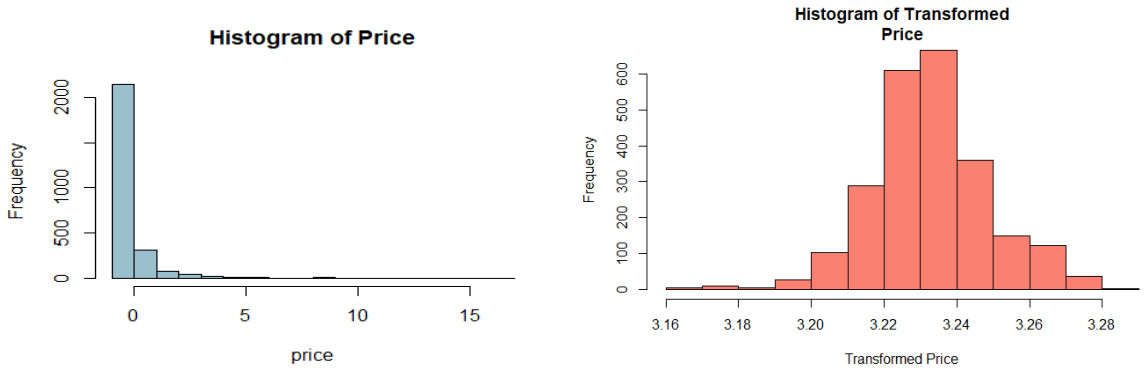


Figure 4-5. Histogram of the Price and Transformed Price

It is clear from the first histogram shown above that price has right skewed distribution. However, in order to get more accurate results, Box-Cox transformation method is selected to try to make it normal. In general, the distribution of logarithm of the price variable is checked as it is proposed to apply logarithmic transformation on the positively skewed dependent variable.

After applying Box-Cox transformation, although the transformed price seems close to normal, Shapiro-Wilk test is applied to be sure.

```
shapiro.test(T_box)

        Shapiro-Wilk normality test

data:  T_box
W = 0.98299, p-value = 1.999e-14
```

According to the output of Shapiro-Wilk test, since the p- value is less than 0.05, we reject the null hypothesis. Thus, the price variable is not normally distributed. However, we know that by applying Box-Cox transformation, the variable became closer to the normal distribution. This can be checked by using QQ-plot.

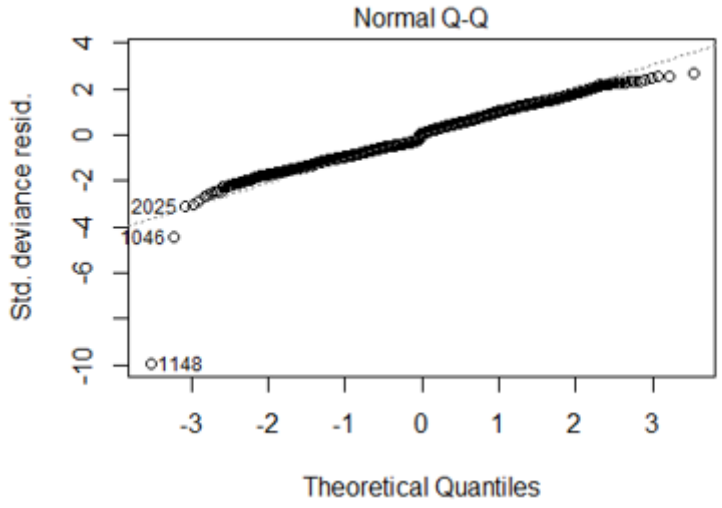


Figure 6. QQ – Plot

By looking at the QQ-plot, it can be said that the distribution is not far from normality because there is not many outliers in the data and it has only one small break point. So, to be able to continue to analysis, we can assume that the "price" variable is normally distributed.

Analysis of multiple variables in a single relationship or set of relationships is the focus of many multivariate analysis methods such as Principal Component Analysis (PCA), Factor Analysis, Clustering, etc. Before going into deep consideration on these techniques, four important assumptions and some ways that should be kept in the mind to deal with any violations against these are listed below.

Assumptions:

1-Normality

The most fundamental assumption in multivariate analysis is normality, referring to the shape of the data distribution for an individual metric variable and its correspondence to the normal distribution. As being crucial in the case of univariate analysis, the most fundamental assumption in multivariate analysis is normality. Even the multivariate normality is really difficult to obtain, these methods require at least univariate normal distribution.

- Assessing the Impact of Violating the Normality Assumption: The severity of

non-normality is based on two dimensions: the shape of the distribution and the sample
size. As we will see in the following discussion, the researcher must not only judge the extent to which the variable's distribution is non-normal, but also the sample sizes involved. What might be considered unacceptable at small sample sizes will have a negligible effect at larger sample sizes.

- Impacts Due to Sample Size: Even though it is important to understand how the distribution departs from normality in terms of shape and whether these values are large enough to warrant attention, the effects of sample size should be taken into consideration. As it is known, sample size has the effect of increasing statistical power by reducing sampling error. It results in a similar effect here, in that larger sample sizes reduce the detrimental effects of nonnormality. In small samples of 50 or fewer observations, significant departures from normality can have a substantial impact on the results. For sample sizes of 200 or more, however, these same effects may be negligible.

2-Homoscedasticity

The next assumption is related primarily to dependence relationships between variables. Homoscedasticity refers to the assumption that dependent variable(s) exhibit equal levels of variance across the range of predictor variable(s).
If this dispersion is unequal across values of the independent variable, the relationship is said to be heteroscedastic.

3-Linearity

An implicit assumption of all multivariate techniques based on correlational measures of association, including multiple regression, logistic regression, factor analysis, and structural equation modelling, is linearity. Because correlations represent only the linear association between variables, nonlinear effects will not be represented in the correlation value. This omission results in an underestimation of the actual strength of the relationship. It is always prudent to examine all relationships to identify any departures from linearity that may affect the correlation.

4-Uncorrelated Errors

Predictions in any of the dependence techniques are not perfect, and we will rarely find a situation in which they are. However, we do attempt to ensure that any prediction errors are uncorrelated with each other.

Principal Component Analysis (PCA)

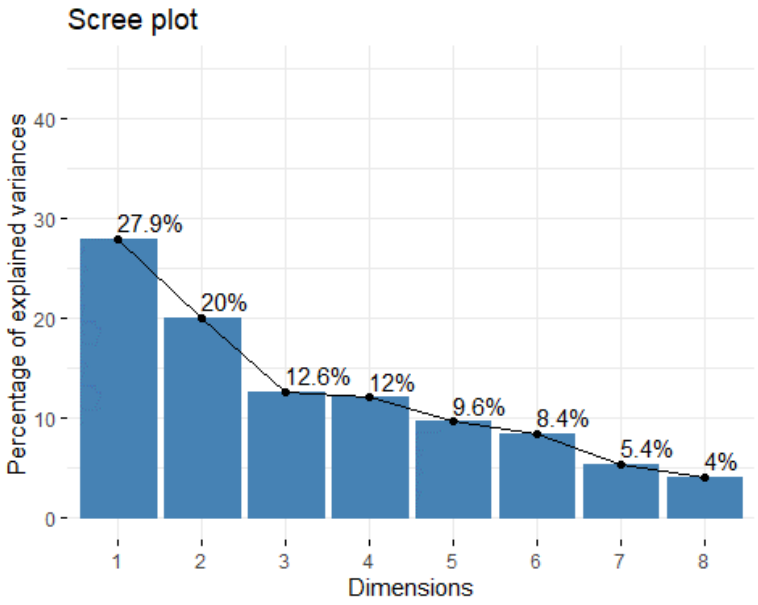
As an initial stage of the exploratory analysis with many variables, to reduce dimension and examine the relationship between all these, PCA is better to visualize the variation. The analysis was conducted on 5282 individuals, 8 continuous variables.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Standard deviation	1.4948	1.2663	1.0033	0.9816	0.87735	0.82013	0.65570	0.5657
Proportion of Variance	0.2793	0.2004	0.1258	0.1204	0.09622	0.08408	0.05374	0.0400
Cumulative Proportion	0.2793	0.4797	0.6055	0.7260	0.82218	0.90626	0.96000	1.0000

Before dividing the data into train and test sets, principal components are obtained to compare with the training data. Here, 8 principal components which correspond to the explained total variance in dataset are obtained. The “Proportion of Variance” part shows the explained variance by each individual principal component. From the first view of interpretation, there can be seen that 1st principal component has the highest explained variation; that is, PC1 explains almost quarter of the information in the dataset. Also, to make inference about the total variation explained, “Cumulative Proportion” can be checked. According to this output, it can be concluded that if one would like to explain significant amount of variance such as the value greater than 70% in the dataset, 4 or more principal components should be examined.

Here, the first 4 PC’s are enough because we reached 72%.

To visualize explained variation by each principal component, scree plot is used.



To determine the number of PCs among 8 of them, scree plot can be checked. Since it is used to examine the explained variation by each principal component, when the percentage of explained variance changes slightly with the addition of a new PC, it can be concluded as there is no need to add more PC to improve explained variance.

Figure 7. Scree Plot of Principle Components

After dividing our data into training and test sets, PCA is applied again. The results are:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Standard deviation	1.4818	1.2332	0.8958	0.8176	0.73193	0.55074	0.50207	0.01296
Proportion of Variance	0.3497	0.2422	0.1278	0.1065	0.08532	0.04831	0.04015	0.00003
Cumulative Proportion	0.3497	0.5919	0.7197	0.8262	0.91152	0.95983	0.99997	1.00000

Now, as it can be seen from above, by using 4 PCs, the total variation can be explained 82%. This can also be observed from the scree plot.

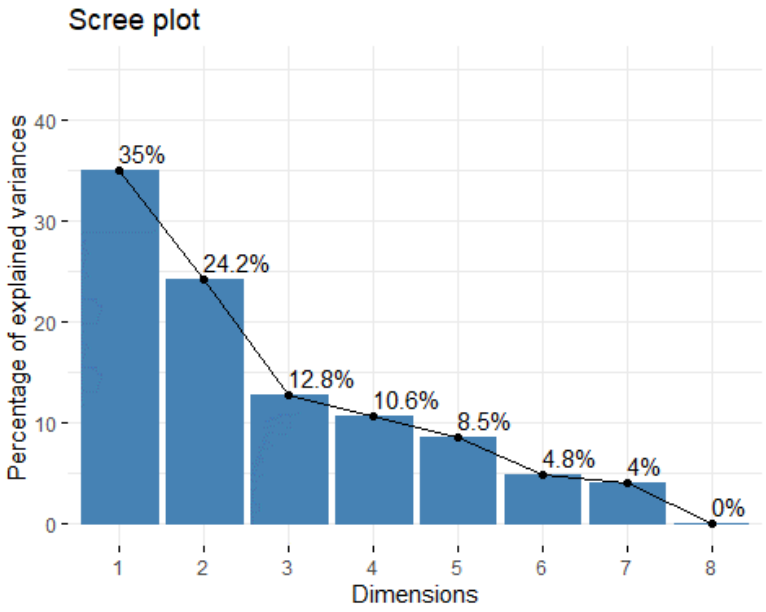


Figure 8. Scree Plot of Principle Components by using train data

After deciding how many PC should be selected, the model was conducted with them.

```
summary(model.pca)

Call:
lm(formula = na_traindata$Price ~ pc$scores[, 1] + pc$scores[,
    2] + pc$scores[, 3] + pc$scores[, 4], data = na_traindata)

Residuals:
       Min         1Q     Median         3Q        Max
-1.407e-14 -8.900e-17 -5.100e-17 -5.000e-18  6.716e-14

Coefficients:
                Estimate Std. Error   t value Pr(>|t|)
(Intercept)    3.822e-03  3.587e-17 1.065e+14   <2e-16 ***
pc$scores[, 1] 4.027e-01  2.359e-17 1.707e+16   <2e-16 ***
pc$scores[, 2] 7.749e-01  4.008e-17 1.934e+16   <2e-16 ***
pc$scores[, 3] 4.309e-01  4.388e-17 9.820e+15   <2e-16 ***
pc$scores[, 4] 2.026e-01  7.702e-17 2.630e+15   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.591e-15 on 1963 degrees of freedom
Multiple R-squared:      1,       Adjusted R-squared:    0.95
F-statistic: 1.922e+32 on 4 and 1963 DF,  p-value: < 2.2e-16
```

According to the summary of the model, all PCs and the model is significant. In addition to that, Adjusted R-squared value is 0.95. It means that 95% of the variation in price can be explained by this model.

Biplot of Variables

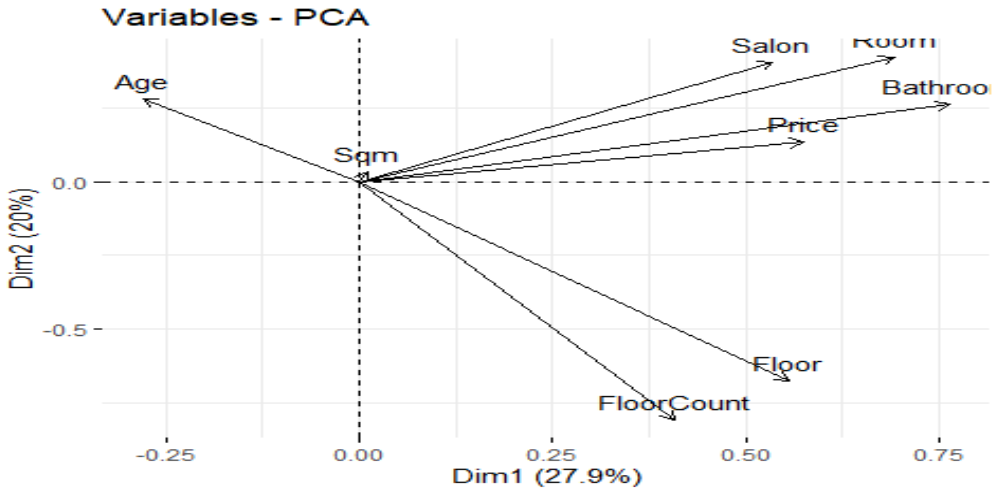


Figure 9: Biplot of Variables

In the Variables-PCA plot, the original variables are represented in the Dimension-1 (Dim1) and Dimension-2 (Dim2) space that are correspond to first two principal components. The Dim1 can be interpreted as the resultant of all the variable projected on the x-axis. The longer the projected vector means that the more important is the contribution of the variable in the dimension. From this knowledge, it can be concluded that the x-axis (Dim1) is dominated by many of the variables such as Floor, Floor Count etc. There is a high correlation between bathroom,price,sqm, and high correlation between floor and floor count, again. On the other hand, Age has different characteristics compared to other variables.

The contributon of variables on PC1 and PC2 can be verified by the following plots, seperately.

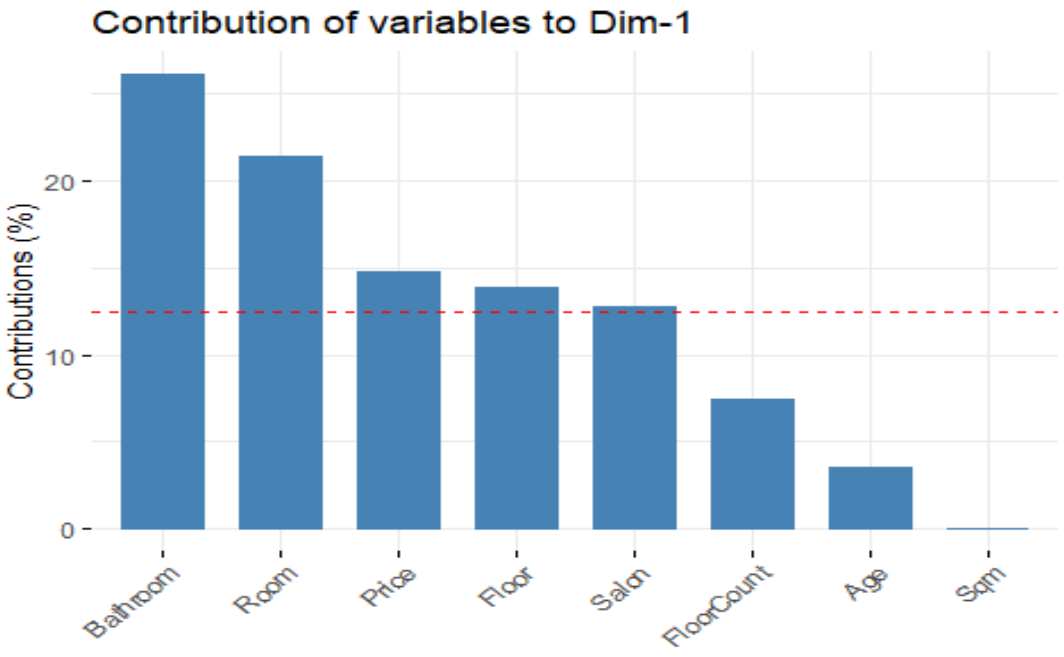


Figure10: Contribution of Variables to Dim-1

Variables that have contribution value above the red line can be specified as being explained by the first principal components significantly. So, the variables apart from Floor Count, Age and Sqm seem to be insignificantly explained by the first PC.

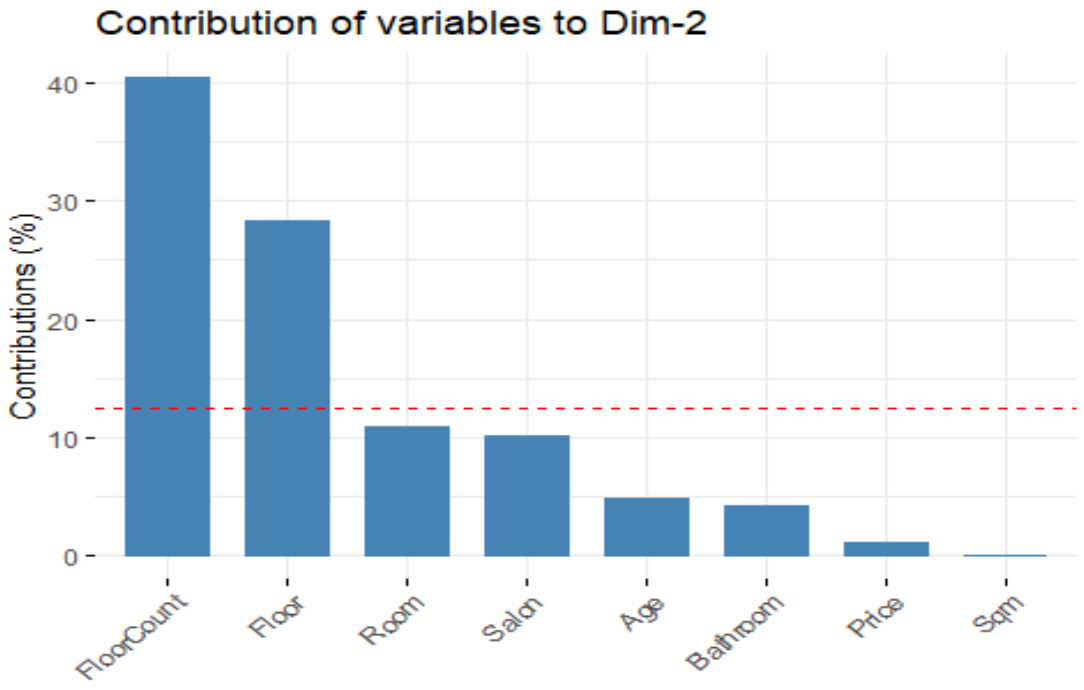


Figure11: Contribution of Variables to Dim-2

Total contribution of each variable on PC1 and PC2 is shown with the following graph.

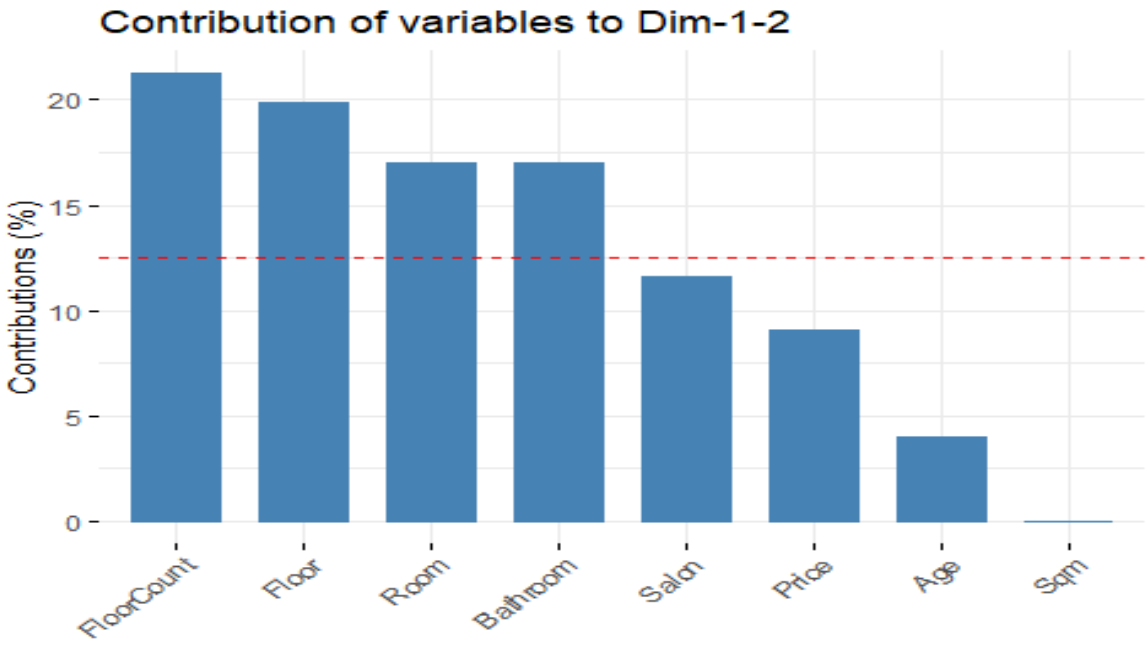


Figure12: Contribution of Variables to Dim-1-2

Salon, Price, Age and Sqm have insignificant contribution on the first two principal components.

FACTOR ANALYSIS

A crucial decision in exploratory factor analysis is how many factors to extract.

Assumptions:

- The variables used in factor analysis should be linearly related to each other. This can be checked by looking at scatterplots of pairs of variables.
- Sample size: The sample size should be large enough to yield reliable estimates of correlations among the variables.
- Normality: Statistical inference is improved if the variables are multivariate normal.

After the assumptions are checked, the normality could not be catched.. Since it did not improve after Boxcox transformation, the normality is assumed for the analysis.

At first, Maximum Likelihood Factor Analysis is entering raw data and extracting 2 factors, with varimax rotation.

```
Call:
factanal(x = st.data, factors = 2, rotation = "varimax")

Uniquenesses:
     Price       Room      Salon        Sqm        Age   Bathroom      Floor FloorCount
     0.810      0.460      0.766      1.000      0.950      0.435      0.533      0.005

Loadings:
           Factor1 Factor2
Price       0.177   0.398
Room                0.729
Salon               0.483
Sqm
Age        -0.204
Bathroom    0.214   0.720
Floor       0.678
FloorCount  0.985  -0.160

               Factor1 Factor2
SS loadings      1.557   1.483
Proportion Var   0.195   0.185
Cumulative Var   0.195   0.380

Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 92.85 on 13 degrees of freedom.
The p-value is 3.97e-14
```

Uniqueness represents the percentage of variance for the variable that is not explained by the common factors. The quantity "1 − uniqueness" is called communality. The greater the uniqueness, the more likely that it is more than just measurement error. 97% of the variance in 'Sqm' is not share with other variables in the overall factor model. Also, it does not belong to any factor. If the uniqueness is high, then the variable is not well explained by the factors. On the contrary, 'Floor Count' has low variance not accounted by other variables (0%).

Factor loadings are the weights and correlations between each variable and the factor. The higher the load the more relevant in defining the factor's dimensionality. The variable with the strongest association to the underlying latent variable. Factor 1, is 'Floor Count', with a factor loading of 0.98. It seems that most variables such as

'Price','room','saloon,'age','bathroom' and 'floor' can be explained by factor 1. Moreover, 'Price','room', saloon', 'bathroom' and 'floor count' also define factor 2.

However, sicnce the uniquness values are too high and the test of the hypothesis that 2 factors are sufficient is rejected, the number of factors was changed from 2 to 4.

```
Call:
factanal(x = st.data, factors = 4, rotation = "varimax")

Uniquenesses:
     Price       Room      Salon        Sqm        Age   Bathroom      Floor
FloorCount
      0.81       0.49       0.71       1.00       0.00       0.19       0.40
0.20

Loadings:
           Factor1 Factor2 Factor3 Factor4
Room        0.68
Bathroom    0.86
Floor               0.76
FloorCount          0.86
Age                         0.96
Price       0.41
Salon       0.45
Sqm

               Factor1 Factor2 Factor3 Factor4
SS loadings        1.6    1.36    0.98    0.25
Proportion Var     0.2    0.17    0.12    0.03
Cumulative Var     0.2    0.37    0.49    0.52

Test of the hypothesis that 4 factors are sufficient.
The chi square statistic is 1.52 on 2 degrees of freedom.
The p-value is 0.469
```

Cumulative variance showed the amount of variance explained by up to corresponding factor value. That means while factor 1, and 2 account for 37% of the total variance, factor 1, 2, 3 and 4 together account for 52% of the total variance. So, increasing the factor number method worked.

CLUSTERING

Clustering is a broad set of techniques for finding subgroups of observations within a data set.

To obtain which observations are alike, clustering is used.

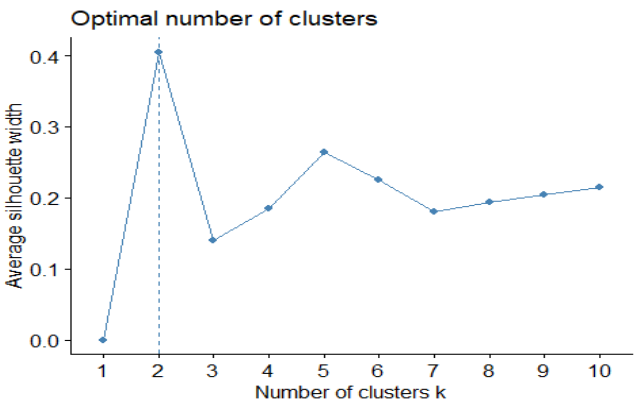
To perform a cluster analysis, generally, the data should be prepared as follows:

1. Rows are observations (individuals) and columns are variables

2. Any missing value in the data must be removed or estimated.

3. The data must be standardized (i.e., scaled) to make variables comparable.

After checking data, it is needed to find optimal number of clusters in order to apply K-means clustering. To find number of clusters, optimal number of clusters plot is created.



From the graph, it can be said that there is a high average silhouette for number of cluster K as 2. So, it is needed to take K as 2 give better results.

Figure 13: Optimum Number of Clusters Plot

After finding number of clusters, one can be found K-means clustering.

```
Cluster means:
  Price       Room      Salon          Sqm        Age    Bathroom        Floor
FloorCount
1  0.8057995  0.9975482  0.8015717  0.005013990 -0.3991322  1.0340207  0.9644446
0.7252635

2 -0.2292701 -0.2838274 -0.2280672 -0.001426605  0.1135631 -0.2942047 -0.2744086
-0.2063556
```

By looking the cluster means, it can be said that room and floor are close to each other and also price and salon are close to each other and the graph of these two groups explain K-means clustering.

```
Within cluster sum of squares by cluster:
[1] 8654.447 8611.185
 (between_SS / total_SS =  17.7 %)
```

Within cluster sum of squares by cluster (between_SS/total_SS) is expected to be small and it equals to 17.7%. To sum up, two cluster give better results. And also trying other numbers for k-means give higher probabilities so, choosing K as 2 gives better results.

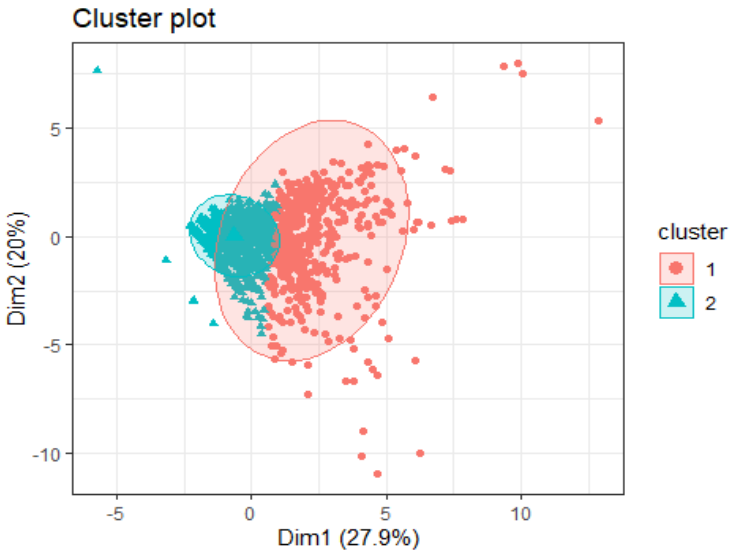


Figure 14: Cluster Plot

Each object lies within its cluster according to the cluster plot.

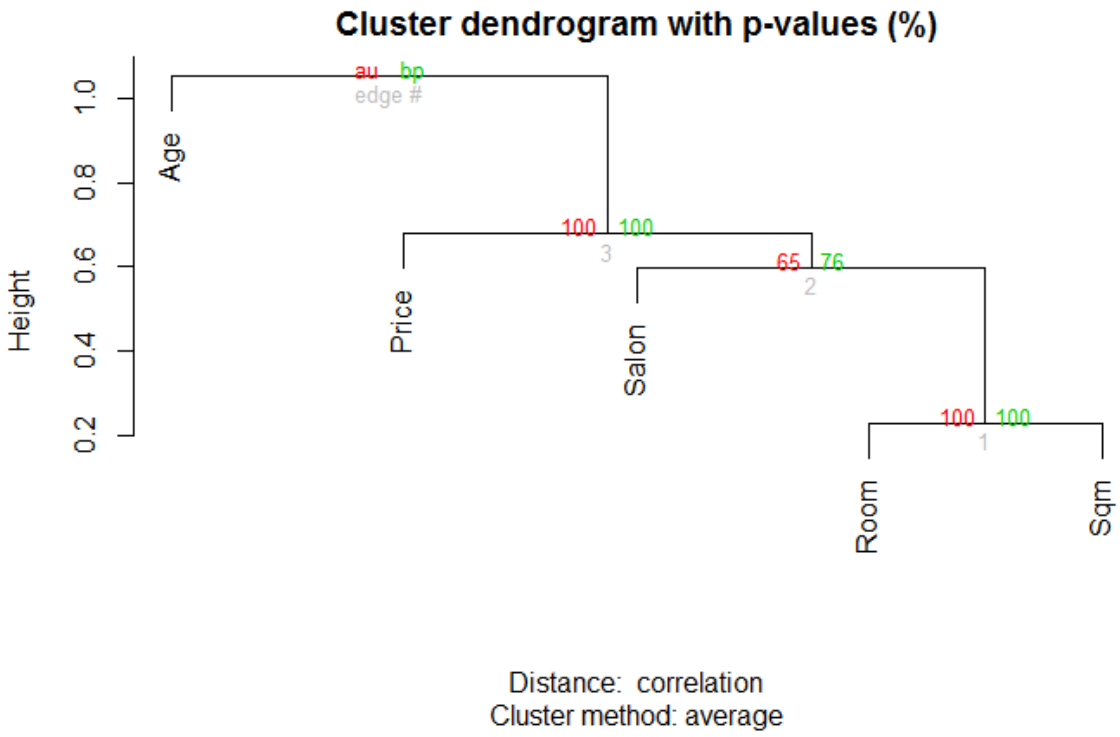


Figure 15: Dendrogram Plot

A dendrogram is a diagram that shows the hierarchical relationship between objects. In the plot above, it is seen that Room and Sqm are most similar, as the height of the link that joins them together is the smallest. The heights reflect the distance between the clusters. In this case, the dendrogram shows us that the big difference between clusters is between the cluster of Sqm versus that of all the other variables.

CLASSIFICATION

A classification is an ordered set of related categories used to group data according to its similarities in order to make classification Discriminant Analysis is applied.

Linear Discriminant Analysis

It is a multivariate classification technique that separates objects into two or more mutually exclusive groups based on measurable features of those objects. The measurable features are sometimes called predictors or independent variables, while the classification group is the response or what is being predicted. There are two discriminant analysis methods; LDA and QDA. Linear Discriminant Analysis (LDA) which assumes that the covariance of the independent variables is equal across all classes is chosen to make classification. LDA requires the number of independent variables to be less than the sample size and both assume multivariate normality among the independent variables. That is, the independent variables come from a normal distribution. After checking the assumption, LDA is applied.

```
Call:
lda(newtrain$Price1 ~ ., data = newtrain)

Prior probabilities of groups:
        0         1
0.4699793 0.5300207

Group means:
      Room    Salon      Sqm      Age HeatingKalorifer HeatingKombi
HeatingMerkezi
0 2.525698 1.014684 113.9266 12.96623       0.02496329    0.7841410
0.08076358
1 3.291667 1.145833 168.1328 11.71615       0.02864583    0.6497396
0.09765625
  HeatingMerkeziPayOlcer  HeatingYok TypeKonut Bathroom    Floor
FloorCount FuelDogalgaz FuelElektrik
0              0.1042584 0.005873715 1.0000000 1.204112 2.628488
6.167401    0.9897210  0.007342144
1              0.2187500 0.003906250 0.9973958 1.824219 4.589844
7.733073    0.9908854  0.005208333
    FuelKomur BuildBetonarme  BuildCelik BuildTasBina BuildState
Furnished UsageEvSahibi UsageKiraci
0 0.002936858      0.9941263 0.001468429  0.002936858   1.563877
0.04552129     0.3318649   0.2334802
1 0.002604167      0.9830729 0.009114583  0.005208333   1.354167
0.05729167     0.4322917   0.1640625
  RegisterKatIrtifaki RegisterKatMulkiyeti cityistanbul  cityizmir
0           0.1997063            0.7621145    0.3832599 0.08810573
1           0.1744792            0.8033854    0.6992188 0.10026042
```

The prior probabilities of groups show πi, the probability of randomly selecting an observation from class i from the total training set. For group 0 which represents the price below the 360.000TL, prior probability is approximately 47% and for group 1 which represents the price above the 360.000TL , it is 52.4%. The group means shows μi, the mean value for each of the independent variables for each class i. The coefficients of linear discriminants are the coefficients for each discriminant. The first linear discriminant (LD1) is the linear combination:

```
Coefficients of linear discriminants:
                                LD1
Room                       0.240694280
Salon                     -0.230937594
Sqm                        0.008329045
Age                        0.002908010
HeatingKalorifer          -1.249246851
HeatingKombi              -1.582856063
HeatingMerkezi            -0.979440766
HeatingMerkeziPayOlcer    -1.128986693
HeatingYok                -0.882614887
TypeKonut                 -0.254715751
Bathroom                   0.323051104
Floor                      0.033810583
FloorCount                 0.007310265
FuelDogalgaz              -0.234411311
FuelElektrik              -0.586675288
FuelKomur                 -1.657377767
BuildBetonarme            -0.649352981
BuildCelik                -0.814212125
BuildTasBina               0.025124671
BuildState                -0.142651679
Furnished                  0.236519141
UsageEvSahibi              0.407093700
UsageKiraci               -0.024972278
RegisterKatIrtifaki        0.651790560
RegisterKatMulkiyeti       0.847721581
cityistanbul               1.647863419
cityizmir                  1.266611383
```

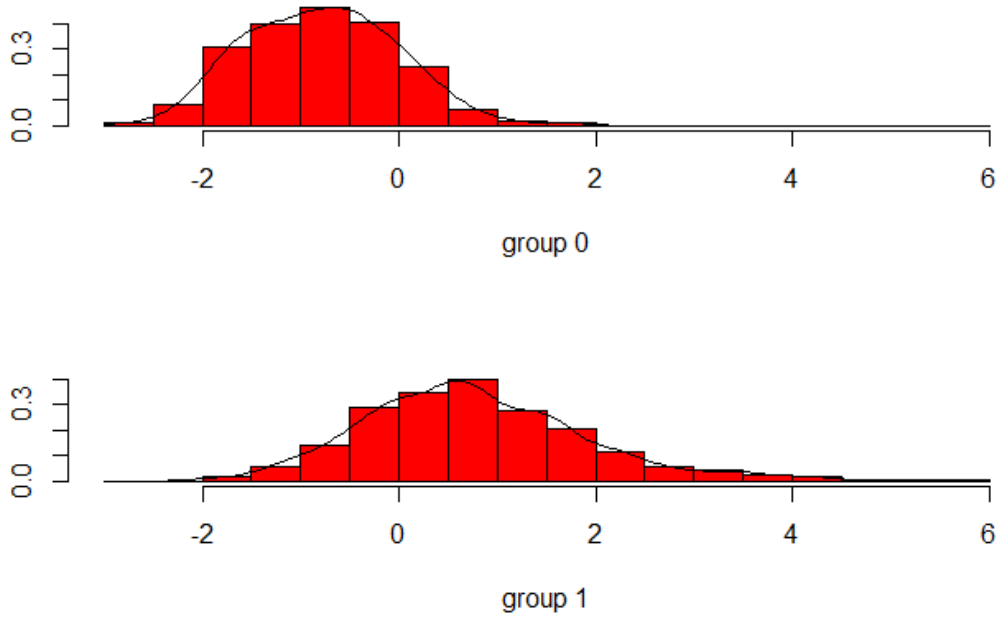


Figure 16: LDA Plot

These plots illustrate the separation between groups as well as overlapping areas that are potential for mix-ups when predicting classes. According to above graphs, when the 2 levels of the Price are combined, they overlap slightly. However it is not a huge area, so there is no series problem.

By using sampling, training data is constructed, and it is used to look whether there are misclassification variables or not.

	0	1
0	546	161
1	135	607

Misclassification Rate = 0.204

From the table, it can be said that the missclassification rate is as 0.204. It may not be low enough for the critical decisons, but there is no big problem for this data.

Logistic Regression Analysis

Logistic regression is the appropriate regression analysis to conduct when the dependent variable is binary. Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

In logistic regression our response must be categorical and take values 0 and 1. Hence we use our converted Price as categorical variable.

In order to have a Multivariate Logistic Regression, the number of responses was tried to be increased. The correlations between Price and predictor variables are examined. However, high correlation couldn't found. In order to be sure that there is no significant correlation, formal test is conducted.

```
Pearson's product-moment correlation

data:  Price and Sqm
t = -0.061077, df = 5279, p-value = 0.9513
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.02781168  0.02613165
sample estimates:
        cor
-0.0008406261
```

Significancy check for the correlations

As a result of Pearson's product-moment correlation test, it can be said that correlations are not significant. Hence, Logistic Regression model is conducted by using one response which is Price.

Firstly, logistic regression does not require a linear relationship between the dependent and independent variables. Second, the error terms (residuals) do not need to be normally distributed. Third, homoscedasticity is not required. Hence, after checking the assumption, the logistic regression model is conducted.

To conduct a logistic analysis, Price is splitted to 0 and 1. 0 represents the Price which is less than 360 thousand Turkish Liras which is the median of Price. 1 represents otherwise.

summary(fit)

```
Call:
glm(formula = Price1 ~ Room + Salon + Sqm + Age + Bathroom +
    Floor + FloorCount + as.factor(Heating) + as.factor(Type) +
    as.factor(Fuel) + as.factor(Build) + as.factor(Register) +
    as.factor(city) + as.factor(Furnished) + as.factor(BuildState) +
    as.factor(Usage), family = "binomial", data = na.omit(satilikbinary))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.4923  -0.7033   0.0449   0.6715   2.6481

Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                      6.573e+00  5.359e+02   0.012 0.990215
Room                             9.690e-01  8.675e-02  11.170  < 2e-16 ***
Salon                            4.646e-01  3.228e-01   1.439 0.150054
Sqm                             -2.376e-05  3.978e-05   2.618 0.040322 *
Age                              1.872e-02  5.998e-03   3.121 0.001802 **
Bathroom                         1.271e+00  1.299e-01   9.783  < 2e-16 ***
Floor                            7.018e-02  2.008e-02   3.495 0.000474 ***
FloorCount                       6.816e-03  1.714e-02   0.398 0.690845
as.factor(Heating)GunesEnerji    1.268e+01  5.914e+02   0.021 0.982900
as.factor(Heating)Kalorifer     -7.305e-01  8.650e-01  -0.845 0.398371
as.factor(Heating)Kombi         -1.317e+00  8.005e-01  -1.645 0.100028
as.factor(Heating)Merkezi       -4.224e-02  8.170e-01  -0.052 0.958767
as.factor(Heating)MerkeziPayOlcer -5.187e-01  8.168e-01  -0.635 0.525367
as.factor(Heating)Yok           -3.605e-01  1.001e+00  -0.360 0.718599
as.factor(Type)Konut            -1.188e+01  5.359e+02  -0.022 0.982316
as.factor(Fuel)Akaryakit        -1.499e-01  2.176e+00  -0.069 0.945080
as.factor(Fuel)Dogalgaz          3.954e-01  4.737e-01   0.835 0.403973
as.factor(Fuel)Elektrik         -8.954e-01  9.495e-01  -0.943 0.345633
as.factor(Fuel)Komur            -7.172e-01  7.699e-01  -0.932 0.351569
as.factor(Build)Ahsap            1.010e+00  1.252e+00   0.807 0.419732
as.factor(Build)Betonarme       -4.678e-02  2.307e-01  -0.203 0.839320
as.factor(Build)Celik            7.785e-01  9.240e-01   0.842 0.399511
as.factor(Build)TasBina          4.965e-01  8.514e-01   0.583 0.559808
as.factor(Register)Arsa         -1.378e+00  3.486e-01  -3.954 7.69e-05 ***
as.factor(Register)KatIrtifaki  -3.394e-01  2.070e-01  -1.640 0.101101
as.factor(Register)KatMulkiyeti  1.238e-01  1.693e-01   0.731 0.464655
as.factor(city)istanbul          2.789e+00  1.538e-01  18.133  < 2e-16 ***
as.factor(city)izmir             1.995e+00  1.917e-01  10.408  < 2e-16 ***
as.factor(Furnished)1            2.154e-01  2.531e-01   0.851 0.394749
as.factor(BuildState)1          -2.192e+00  1.258e+00  -1.743 0.081379 .
as.factor(BuildState)2          -4.358e-01  1.587e-01  -2.747 0.006016 **
as.factor(Usage)Bos             -9.741e-01  4.778e-01  -2.039 0.041476 *
as.factor(Usage)EvSahibi        -4.825e-01  4.768e-01  -1.012 0.311481
as.factor(Usage)Kiraci          -1.155e+00  4.851e-01  -2.381 0.017269 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3295.8  on 2378  degrees of freedom
Residual deviance: 2093.4  on 2345  degrees of freedom
AIC: 2161.4
```

Logistic Regression Model

The results shows that Room, Square meter, Age, Bathroom, Floor, City, Usage variable for the situations there is a tenant in the house or the house is empty, Build state variable for the house that is second hand and Register variable for the house that appears as land in the deed. Later, multicollinearity is checked.

```
                          GVIF Df GVIF^(1/(2*Df))
Room                  1.854310  1        1.361730
Salon                 1.087411  1        1.042790
Sqm                   1.003360  1        1.001678
Age                   1.704911  1        1.305723
Bathroom              1.518383  1        1.232227
Floor                 1.750143  1        1.322930
FloorCount            2.225471  1        1.491801
as.factor(Heating)    3.132422  6        1.099824
as.factor(Type)       1.000000  1        1.000000
as.factor(Fuel)       1.880171  4        1.082118
as.factor(Build)      1.191289  4        1.022121
as.factor(Register)   1.255313  3        1.038625
as.factor(city)       1.907978  2        1.175285
as.factor(Furnished)  1.050336  1        1.024859
as.factor(BuildState) 1.669945  2        1.136778
as.factor(Usage)      1.278540  3        1.041803
```

VIF values for the Logistic Regression Model

After check VIF values which all are less than 10, it is seen that there is no multicollinearity problem. That means all of the variables are uncorrelated with each other.
And then, model selection is made to select the most representative model. By looking at the AIC values, the most representative model is the model with AIC value as 1648.51

```
Step: AIC=1648.51

new_fit= glm(formula = Price1 ~ Room + Sqm + Age + Bathroom + Floor + as.factor(Heating) +
    as.factor(Register) + as.factor(city) + as.factor(BuildState) +
    as.factor(Usage), family = "binomial", data = na.omit(satilikbinary))
summary(new_fit)

 Call:
 glm(formula = Price1 ~ Room + Sqm + Age + Bathroom + Floor +
     as.factor(Heating) + as.factor(Register) + as.factor(city) +
     as.factor(BuildState) + as.factor(Usage), family = "binomial",
     data = na.omit(satilikbinary))

 Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.4286  -0.7025   0.0553   0.6734   2.5874
```

```
Coefficients:
                                  Estimate     Std. Error     z value   Pr(>|z|)
(Intercept)                     -4.787e+00   9.157e-01   -5.228   1.71e-07 ***
Room                             9.748e-01   8.525e-02   11.435    < 2e-16 ***
Sqm                             -2.374e-05   3.996e-05   -2.183    0.05237  **
Age                              1.820e-02   5.929e-03    3.069    0.00214 **
Bathroom                         1.290e+00   1.276e-01   10.113    < 2e-16 ***
Floor                            7.710e-02   1.602e-02    4.811   1.50e-06 ***
as.factor(Heating)GunesEnerji    1.076e+01   3.627e+02    0.030    0.97633
as.factor(Heating)Kalorifer     -4.500e-01   7.988e-01   -0.563    0.57318
as.factor(Heating)Kombi          -1.005e+0   7.239e-01   -1.388    0.16501
as.factor(Heating)Merkezi        2.471e-01   7.467e-01    0.331    0.74065
as.factor(Heating)MerkeziPayOlcer -2.287e-01  7.415e-01   -0.308    0.75772
as.factor(Heating)Yok           -7.418e-01   9.535e-01   -0.778    0.43655
as.factor(Register)Arsa         -1.340e+00   3.407e-01   -3.933   8.40e-05 ***
as.factor(Register)KatIrtifaki  -3.277e-01   2.005e-01   -1.634    0.10218
as.factor(Register)KatMulkiyeti  1.365e-01   1.635e-01    0.834    0.40408
as.factor(city)istanbul          2.821e+00   1.526e-01   18.489    < 2e-16 ***
as.factor(city)izmir             1.931e+00   1.889e-01   10.225    < 2e-16 ***
as.factor(BuildState)1          -1.974e+00   1.191e+00   -1.658    0.09740 .
as.factor(BuildState)2          -4.355e-01   1.574e-01   -2.766    0.00567 **
as.factor(Usage)Bos             -1.004e+00   4.709e-01   -2.133    0.03293 *
as.factor(Usage)EvSahibi        -4.965e-01   4.697e-01   -1.057    0.29055
as.factor(Usage)Kiraci          -1.189e+00   4.783e-01   -2.485    0.01295 *
---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3295.8  on 2378  degrees of freedom
## Residual deviance: 2103.7  on 2357  degrees of freedom
## AIC: 2147.7
##
## Number of Fisher Scoring iterations: 12
```

The most representative model is a model with the predictor variables as Room, Square meter, Age, Bathroom, Floor, Register, Heating, City, Build State and Usage. That means, Salon, Floor Count, Type, Fuel, Build, and Furnished are eliminated. Model adequacy checks are made for the selected model.

	Actual Values	
Predicted Values	0	1
0	928	267
1	225	959

Contingency Table of the selected model

Sensitivity	0.80
Specificity	0.78
Misclassification Rate	0.19

Model Adequacy Check values for selected model

Both sensitivity and specificity values are high as 0.80 and 0. 78 respectively. Also, the misclassification rate is 0.19 which is low.

Lastly, the normality is checked by looking at normal Q-Q Plot.

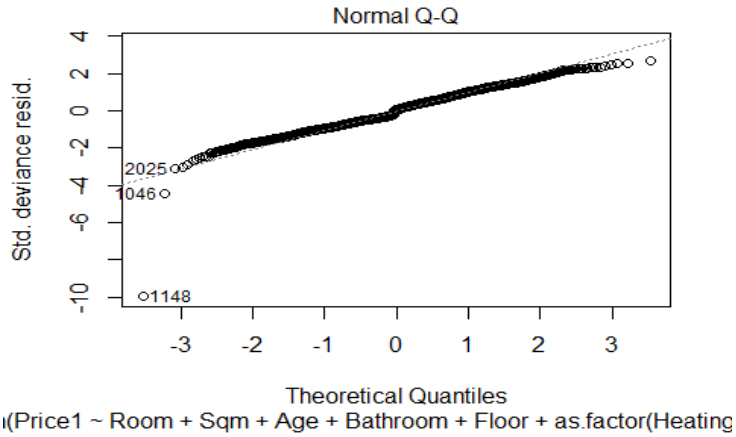


Figure 17.Normal Q-Q Plot of Standardized Deviance Residuals

Most of the Standardized Residuals values seems to follow normality line. However, there also many outliers and also there exist a break point. To be sure, formal test is conducted to see whether the normality exists or not.

```
        Shapiro-Wilk normality test

data:  new_fit$residuals
W = 0.0069267, p-value < 2.2e-16
```

Normality test for the Standardized Deviance Residuals

As a result of the test, the Standardized Deviance Residuals are not normally distributed since its p value is less than 0.05.

Decision Tree

Decision tree builds regression or classification models in the form of a tree structure. In Real Estate data, decision tree is conducted to see which variables affects price as high or low.

In this study, two different decision tree are created. By using first decision tree, the adequacy checks couldn't made. Hence, the second decision tree is created by modelling to make adequacy checks.

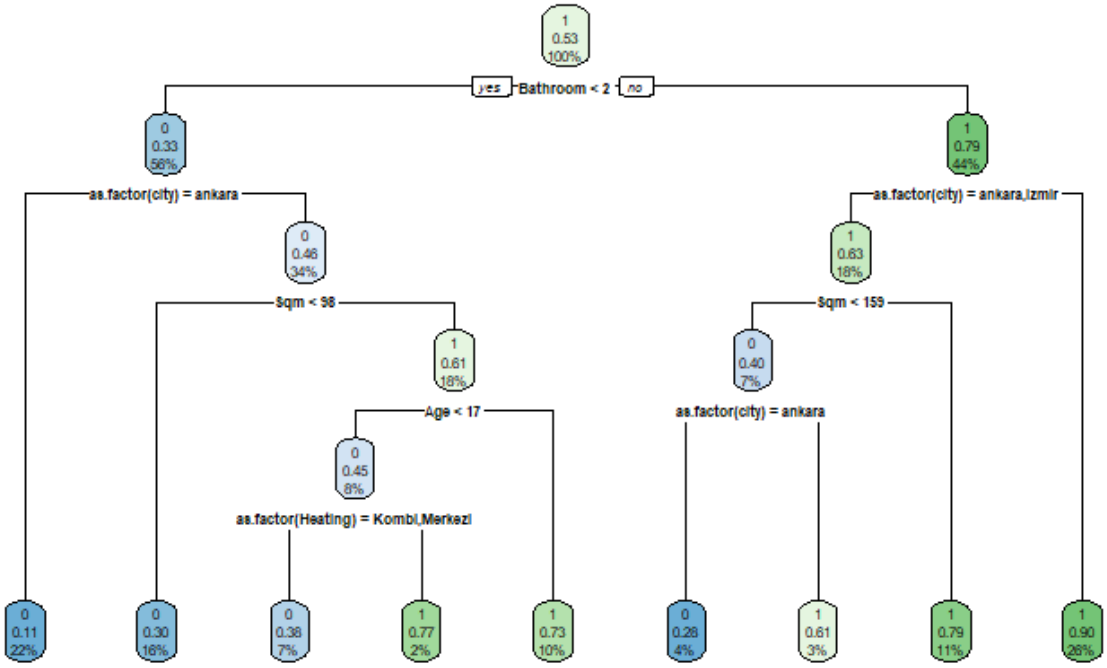


Figure 18: The first Decision Tree

The 1s and 0s represent price. 1 represents the price higher than 360 thousand Turkish Liras and 0 represents the price less than 360 thousand Turkish Liras. As it can be seen from the first decision tree, the node at the top asks whether the bathroom is smaller than 2 or not. If it is greater than 2, the root's left child node would be the next stop. That means, 44 percent of the houses have two or more bathroom. This node asks whether the city in Ankara/Izmir or Istanbul. If the city is Ankara or Izmir, the node asks if the square meter is greater than 159 square meter or not.

But if it is Istanbul, it can be directly concluded that the price of the house with 2 or more bathrooms in Istanbul has a price which is greater than 360 thousand Turkish Liras.

	attr_importance
Room	0.091777443
Sqm	0.126312856
Age	0.020980410
Bathroom	0.132083820
Floor	0.022934395
as.factor(Heating)	0.099076093
as.factor(Register)	0.033383290
as.factor(city)	0.109211789
as.factor(BuildState)	0.009300688
as.factor(Usage)	0.009695004

Importance of the variables for the first decision tree

When we look for the importance of the variables in the decision tree, we see that Bathroom, Square meter, City, Room respectively have high importance for the first decision tree.

The second decision tree is created to analyse by using adequacy checking methods. Here, we have more simple decision tree.

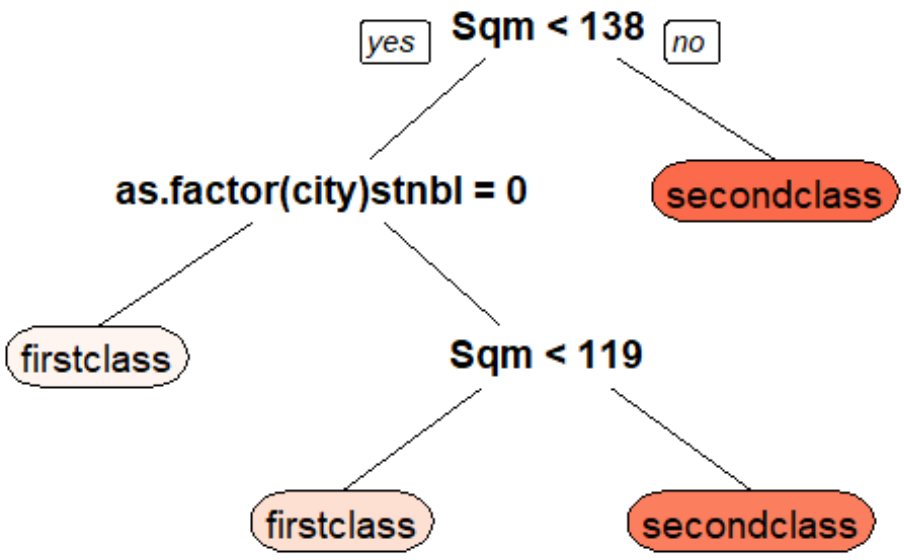


Figure 19.The second Decision Tree

The node at the top shows the square meter of the house. This node asks whether the square meter is less than 138 or not. If it is greater than 138, the root's left child node would be the next stop. That means, the house with greater than 138 square meter's price is more than 360 thousand Turkish Liras. If it is not, another node asks if the house is in İstanbul and then square meter again.
Later, model adequacy checks is made for different cut of points and similar sensitivity and specificity values are obtained.

The sensitivity and specificity values are 0.77 and 0.81 respectively. The accuracy of decision tree is 0.77. All seems high. However, it can be said that sensitivity can be improved.

Sensitivity	0.73
Specificity	0.80
Accuracy	0.77

Later, the importance of the variables in the second decision tree model is checked.

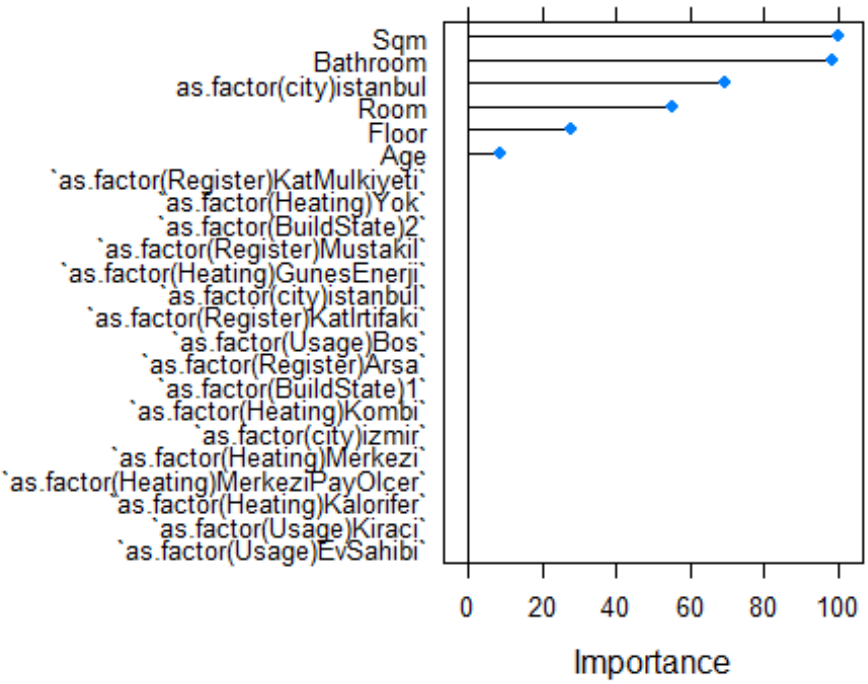


Figure 20.The importance of the values for second decision tree

By looking at the importance of the variables in the second decision tree model it can be said that Square meter, bathroom, and the city to be İstanbul or not, room, floor and age have high importance.
Hence, the importance of the variables looks quite similar for the both decision trees.

Random Forest

Random Forests are similar to a famous Ensemble technique called Bagging but have a different tweak in it. In Random Forests the idea is to decorrelate the several trees which are generated by the different bootstrapped samples from training Data.

```
randomForest(formula = Price ~ ., data = traindata)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 2

          Mean of squared residuals: 0.5865896
                    % Var explained: 40.71
```

Output of the randomForest code

By default, the number of decision trees in the forest is 500 and the number of features used as potential candidates for each split is 2.

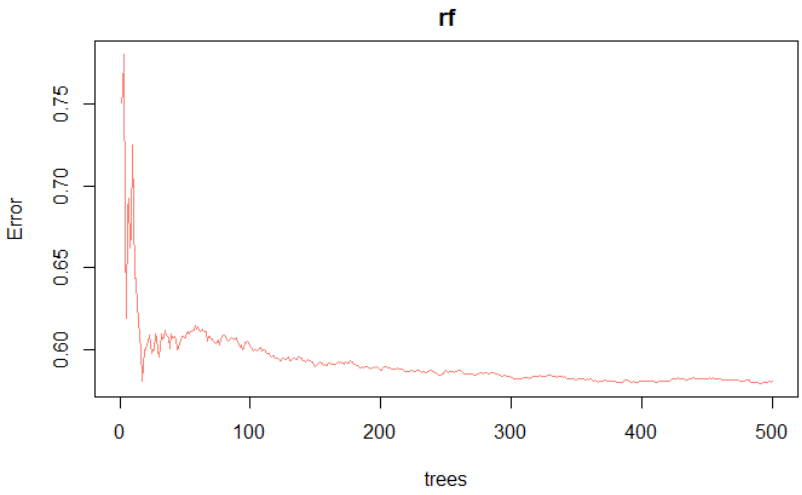


Figure 21.The Error and the Number of Trees

This plot shows the Error and the Number of Trees. We can easily notice that how the error is dropping as we keep on adding more and more trees and average them.
When we look importance, Age and Square Meter has the highest impurity.

	%IncMSE	IncNodePurity
Room	12.535280	2.169917e+14
Salon	6.963149	8.186115e+13
Sqm	23.646632	8.099063e+14
Age	16.162815	6.318861e+14
Heating	14.253394	2.786887e+14
Type	2.853135	1.167915e+12
Bathroom	8.089223	2.475165e+14
Floor	5.723951	2.187128e+14
FloorCount	6.685609	1.986574e+14
Fuel	-1.174553	3.388491e+12
Build	2.984127	7.930339e+13
BuildState	7.290999	3.439491e+13
Furnished	2.183533	1.502581e+13
Usage	6.139416	9.332873e+13
Register	6.491013	7.264461e+13
city	26.865184	3.049217e+14

Impurity of variables

Now, after creating pre-trained model and it can predicted the values for the test data. We then compare the predicted value with the actual values in the test data and analyse the accuracy of the model. To make this comparison more illustrative, it will be shown in the forms of table and plot the price and the age value.

Figure 22. Comparison of predicted values and actual values

The figure displays that predicted prices (blue scatters) coincide well with the real ones (red scatters). But to estimate our model more precisely, we will look at Mean absolute error (MAE), Mean squared error (MSE), and R-squared scores.
Later, high error values (MAE and MSE) are obtained. To improve the predictive power of the model, tune the hyper parameters of the algorithm is necessary.

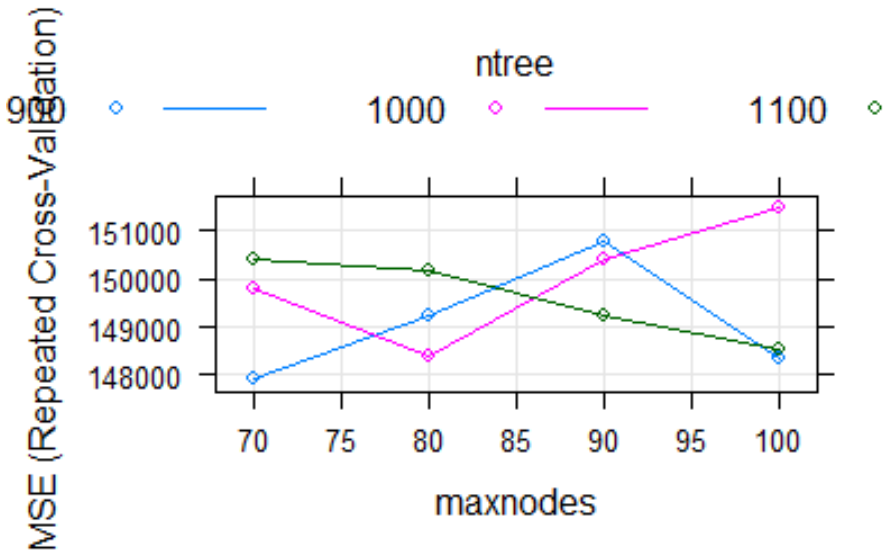


Figure 23. Impact of tuned parameters on RMSE

The plot shows how the model's performance develops with different variations of the parameters. For values maxnodes: 70 and ntree: 900, the model seems to perform best. After checking best parameters by using appropriate code, it has seen that the model performs the best for values maxnodes: 70 and ntree: 900.
Lastly, the importance of variables are defined and visualized.

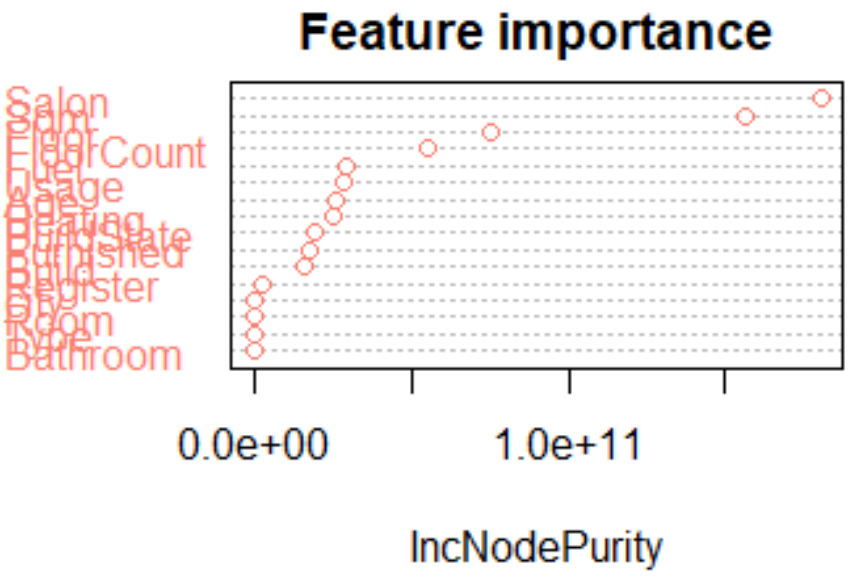


Figure24. The importance of variables

From the figure above, it can be said that Home related variables as Salon, Square meter, Usage and Furnished, and Building Related variables as Floor, Floor Count, Fuel, Age, Heating, Build State and Build contain the major part of the predictive power.

Canonical Correlation Analysis

The canonical correlation analysis is used to determine the relationships between the linear composition of the variables in a variable set and the linear composition of the variables in another set of variables. In the canonical correlation, the number of variables in the clusters is not necessarily equal.

To conduct this analysis, data set should be separated into two groups. Here, the first group consists price, age, floor, floor count and type as Building Related Variables. The second group consists room, salon, square meter, bathroom and furnish as Home Related Variables. First, correlation with in groups is checked.

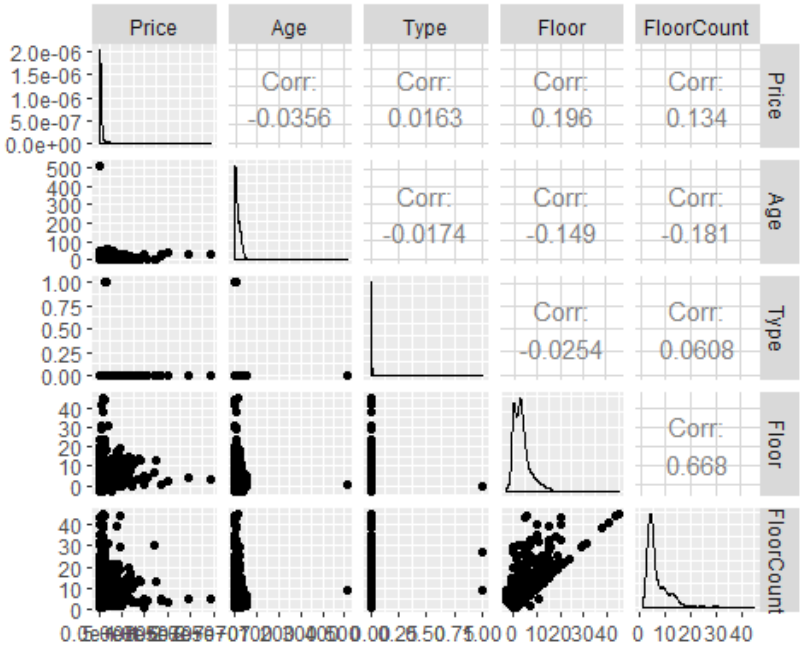


Figure 25.The correlation with in Group 1

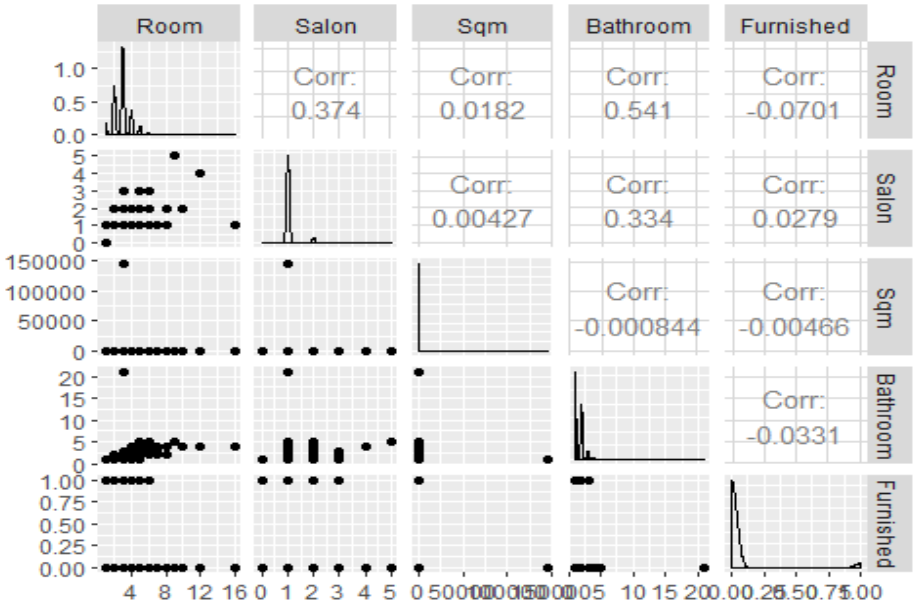


*Figure 26.*The correlation with in Group 2

In first group, floor and floor count has high correlation between them. In second group, there is no such kind of strong correlation.

After looking for the correlation between groups, correlation between groups is checked. However, it is the same thing with checking correlation for data directly.

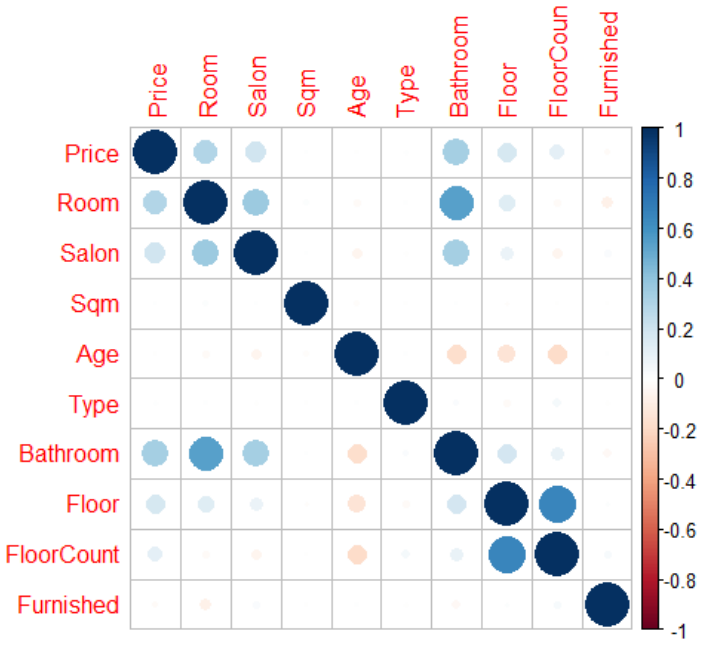


Figure 27. The correlation between variables of both groups in dataset

After checking the correlations, canonical correlations are checked. The number of possible canonical variates is equal to the number of variables in the smaller set. In this case, there exist 5 variables for both groups.

	Canonical correlation values					
Canonical values	0.428	0.203	0.061	0.021	0.007	*correlation*

The first two canonical correlations are much higher than the others. Also, it is customary to report the largest correlation as canonical correlation.
Then, the raw canonical correlation coefficients is checked. The raw canonical coefficients are interpreted in a manner analogous to interpreting regression coefficients.

xcoef

	[,1]	[,2]	[,3]	[,4]	[,5]
Price	-6.091756e-07	-2.562232e-08	-1.609751e-07	3.216507e-07	2.751665e-07
Age	1.880633e-02	-3.736573e-02	-5.082555e-02	1.818648e-03	7.381196e-03
Type	-2.884505e+00	1.495312e+00	-6.440900e+00	1.551729e+01	-3.019498e+01
Floor	-1.241106e-01	-1.371211e-01	9.062461e-03	-2.275404e-01	-1.482702e-01
FloorCount	7.948979e-02	1.887646e-01	-1.337900e-01	3.171916e-02	7.708980e-02

ycoef

	[,1]	[,2]	[,3]	[,4]	[,5]
Room	-3.078622e-01	-7.804542e-01	-7.205670e-01	0.2487020084	0.0922926981
Salon	-8.619852e-01	-1.557004e+00	3.090913e+00	-1.3209763008	0.0996171703
Sqm	5.490886e-07	2.238987e-05	8.821306e-05	0.0001986147	0.0002562049
Bathroom	-8.306434e-01	1.293552e+00	8.653320e-02	-0.0347186236	-0.0889280579
Furnished	7.385747e-02	2.994902e-01	-1.605518e+00	-3.1526701853	2.9905815853

Raw canonical correlation coefficients

For the variable room, a one unit increase in the number of room leads to a 0.3 decrease in the first canonical variate of the second set when all of the other variables are held constant. Also, for the variable floor, a one unit increase in floor leads to a 0.12 decrease in the first canonical variate of the first set when all of the other variables are held constant. Lastly, if the house has furnishings in it, it leads to a 0.3 increase in the second canonical variate of the second set when all of the other variables are held constant.

Next, the loadings of the variables on the canonical dimensions (variates) are computed. These loadings are correlations between variables and the canonical variates.

```
corr.X.xscores
                [,1]         [,2]       [,3]       [,4]        [,5]
Price      -0.87481521  0.008308423 -0.2785714  0.2753945  0.28493014
Age         0.32040039 -0.668097798 -0.6480094  0.1149728  0.13361887
Type       -0.06349717  0.127500737 -0.2203606  0.4905151 -0.83099209
Floor      -0.43748776  0.170894583 -0.3535848 -0.7636797 -0.26677805
FloorCount -0.09316159  0.714152334 -0.5775032 -0.3833085 -0.02953013
```

```
corr.Y.xscores
                [,1]         [,2]         [,3]          [,4]          [,5
]
Room      -0.333151401 -0.095076503 -0.023482057  0.0035472928  0.000276953
7
Salon     -0.252247547 -0.086181623  0.037806794 -0.0063541487  0.000428551
2
Sqm       -0.002141813  0.009793708  0.015617356  0.0124052500  0.005088547
5
Bathroom  -0.390942088  0.082043728 -0.002785672  0.0002767269 -0.000187597
3
Furnished  0.023062227  0.015791882 -0.016729867 -0.0148325102  0.004280585
1
```

```
corr.X.yscores
                [,1]         [,2]        [,3]         [,4]          [,5]
Price      -0.37436301  0.001691444 -0.01717806  0.005727471  0.0019075274
Age         0.13711016 -0.136012582 -0.03995939  0.002391127  0.0008945409
Type       -0.02717259  0.025956835 -0.01358850  0.010201404 -0.0055632590
Floor      -0.18721581  0.034791035 -0.02180375 -0.015882499 -0.0017860043
FloorCount -0.03986699  0.145388450 -0.03561164 -0.007971793 -0.0001976960
```

```
corr.Y.yscores
                [,1]        [,2]       [,3]        [,4]        [,5]
Room      -0.778511506 -0.46701857 -0.3808014  0.17056482  0.04136898
Salon     -0.589454575 -0.42332666  0.6131014 -0.30552714  0.06401332
Sqm       -0.005005011  0.04810698  0.2532620  0.59648282  0.76008374
Bathroom  -0.913557357  0.40300120 -0.0451744  0.01330589 -0.02802169
Furnished  0.053892042  0.07757019 -0.2713032 -0.71319301  0.63939723
```

The loadings of the variables on the canonical dimensions

The above correlations are between observed variables and canonical variables which are known as the canonical loadings. These canonical variates are actually a type of latent variable.

In general, the number of canonical dimensions is equal to the number of variables in the smaller set; however, the number of significant dimensions may be even smaller.

Canonical dimensions, also known as canonical variates, are latent variables that are analogous to factors obtained in factor analysis. Here, whether the canonical dimensions are significant or not is checked by formal tests. These tests are Wilks Lambda, Hotelling-Lawley Trace, Pillai's-Bartlett Trace test and Roy's Largest Root.

```
> p.asym(rho, n, p, q, tstat = "Wilks")
Wilks' Lambda, using F-approximation (Rao's F):
              stat     approx df1      df2      p.value
1 to 5:  0.7796671 24.3968248  25 8801.946 0.000000e+00
2 to 5:  0.9544537  6.9585830  16 7241.107 3.330669e-16
3 to 5:  0.9957219  1.1304751   9 5770.542 3.367822e-01
4 to 5:  0.9995227  0.2831574   4 4744.000 8.890446e-01
5 to 5:  0.9999552  0.1063609   1 2373.000 7.443546e-01
```

```
> p.asym(rho, n, p, q, tstat = "Hotelling")
 Hotelling-Lawley Trace, using F-approximation:
                  stat     approx df1   df2      p.value
1 to 5:  2.717132e-01 25.7301540  25 11837 0.000000e+00
2 to 5:  4.753218e-02  7.0389213  16 11847 2.220446e-16
3 to 5:  4.294603e-03  1.1315803   9 11857 3.358859e-01
4 to 5:  4.775374e-04  0.2833468   4 11867 8.889328e-01
5 to 5:  4.482128e-05  0.1064685   1 11877 7.442069e-01
```

```
> p.asym(rho, n, p, q, tstat = "Pillai")
 Pillai-Bartlett Trace, using F-approximation:
                  stat     approx df1   df2      p.value
1 to 5:  2.288528e-01 22.7646611  25 11865 0.000000e+00
2 to 5:  4.572547e-02  6.8500181  16 11875 6.661338e-16
3 to 5:  4.279900e-03  1.1313375   9 11885 3.360576e-01
4 to 5:  4.773482e-04  0.2839299   4 11895 8.885576e-01
5 to 5:  4.481927e-05  0.1067156   1 11905 7.439205e-01
```

```
> p.asym(rho, n, p, q, tstat = "Roy")
 Roy's Largest Root, using F-approximation:
              stat   approx df1  df2 p.value
1 to 1:  0.1831274 106.3963   5 2373       0
```

Test results for checking the significancy of canonical dimensions

As shown in the table above, the first test of the canonical dimensions tests whether all five dimensions are significant with F = 24.39. The next test tests whether dimensions 2 and 5 combined are significant with F = 6.96. The third test tests whether dimension 3 and 5 combined are significant or not. They are not significant. The dimension 4 and 5 are not significant as well. Finally, the last test tests whether dimension 5, by itself,

is significant or not. It is not significant. Therefore dimensions 1 and 2 must each be significant while dimension three, four and five are not.
When the variables in the model have very different standard deviations, the standardized coefficients allow for easier comparisons among the variables. Next, the standardized canonical coefficients are computed.

```
            [,1]        [,2]       [,3]       [,4]       [,5]
[1,] -0.81820122 -0.03441408 -0.2162102  0.4320183  0.3695840
[2,]  0.28879361 -0.57379526 -0.7804869  0.0279275  0.1133471
[3,] -0.08361761  0.04334693 -0.1867123  0.4498236 -0.8753085
[4,] -0.51501911 -0.56900850  0.0376063 -0.9442196 -0.6152738
[5,]  0.41785990  0.99229289 -0.7033037  0.1667405  0.4052437
```

```
             [,1]        [,2]        [,3]        [,4]        [,5]
[1,] -0.332898986 -0.84392431 -0.77916680  0.26892759  0.09979836
[2,] -0.248250704 -0.44841534  0.89017934 -0.38043958  0.02868962
[3,]  0.001631169  0.06651323  0.26205309  0.59002148  0.76110385
[4,] -0.649821329  1.01195979  0.06769586 -0.02716075 -0.06956938
[5,]  0.015974618  0.06477669 -0.34725729 -0.68189048  0.64683236
```

The standardized coefficients

The standardized canonical coefficients are interpreted in a manner analogous to interpreting standardized regression coefficients. For example, consider the variable room, a one standard deviation increase in number of room leads to a -0.33 standard deviation decrease in the score on the first canonical variate for group2 when the other variables in the model are held constant.

There is a lot of variation in the write-ups of canonical correlation analyses. The write-up below is fairly minimal, including only the tests of dimensionality and the standardized coefficients.

Dimension	Canonical Corr.	F	df1	df2	p
1	0.428	24.396824	25	8801.946	0.000000e+00
2	0.203	6.9585830	16	7241.107	3.330669e-16
3	0.061	1.1304751	9	5770.542	3.367822e-01
4	0.020	0.2831574	4	4744.000	8.890446e-01
5	0.006	0.1063609	1	2373.000	7.443546e-01

Tests of Canonical Dimensions

Dimension 1 had a canonical correlation of 0.428 between the sets of variables, while for dimension 2 the canonical correlation was much lower at 0.203. Also, dimension 3 had a canonical correlation of 0.061, dimension 4 had a canonical correlation of 0.02 and lastly dimension 5 had a canonical correlation of 0.006 between the sets of variables.

	Dimensions	
	1	2
Building Related Variables		
Price	-0.81820122	-0.03441408
Age	0.28879361	-0.57379526
Type	-0.08361761	0.04334693
Floor	-0.51501911	-0.56900850
Floor Count	0.41785990	0.99229289
Home Related Variables		
Room	-0.33289898	-0.84392431
Salon	-0.24825070	-0.44841534
Sqm	0.00163116	0.06651323
Bathroom	-0.64982132	1.01195979
Furnished	0.01597461	0.06477669

Standardized Canonical Coefficients

This table presents the standardized canonical coefficients for the first two dimensions across both sets of variables. For the Building Related variables, the first canonical

dimension is most strongly influenced by price (-.81) and for the second dimension floor count (0.99) and age (-0.57). For the Home related variables, the first dimension was comprised of bathroom (-0.64), room (-0.33) and salon (-0.24). For the second dimension bathroom, room and salon were also the dominating variables respectively.

4. CONCLUSION

In our data set, there exist 17 variables which 8 of them are continuous and 9 are categorical. The aim in this project was to explore the dataset and find basic relationship of different features with house's price. First , we standardized our variables to make the variables' scales same. According to cross validation, we separated the data into two parts; train data(0.8) and test data(0.2). Later, the analyze is started with some brief exploratory analysis involving summary statistics, graphical visualization and PCA. However, before these, normality has checked for the response. After summary statistics and graphical visualization, Principal Component Analysis is made. According to Principal Component Analysis, 4 PCs are sufficient since they explain 82% of the total variation. Similarly, in Factor Analysis, 4 factors became sufficient after changing the number of factors from 2 to 4. In clustering part, it is seen that the data is clustered in 2 main clusters and Age of the building has different characteristics compared to other ones. In LDA, the analysis is conducted based on the levels of Price values and the misclassification rate is obtained 0.2 which is not low enough.

After conducting Linear Discriminant Analysis, we conducted a logistic regression model to describe data and to explain the relationship between Price variable and the all other variables as the predictors. Since there is no significant correlation as a result of formal test, Price is the only response in the model. Price is splitted into 0 and 1 from the cut-off point which is median of the Price (360 Thousand Turkish Liras). The results show that Room, Square meter, Age, Bathroom, Floor, City, Usage variable for the situations there is a tenant in the house or the house is empty, Build state variable for the house that is second hand and lastly Register variable for the house that appears

as land in the deed. After seeing there is no multicollinearity problem, the most representative model is selected with the predictor variables as Room, Square meter, Age, Bathroom, Floor, Register, City, Build State and Usage by looking at AIC values. Later, the normality is checked with normality test. As a result of this test, Standardized deviance Residuals are not normally distributed. After Logistic Regression analysis, 2 different decision trees are conducted and analysed by model adequacy checks. Later, Random Forest is implemented. The data has become more comprehensible. The importance check for variables was repeated.

Lastly, Canonical correlation analysis is used to identify and measure the associations among two sets of variables. The first group consists price, age, floor, floor count and type as Building Related Variables. The second group consists room, salon, square meter, bathroom and furnish as home related variables. It is customary to report the largest correlation which is 0.42.

To conclude, it can be said that Room, Square meter, age, bathroom, floor, Heating, Register, city, Build State and Usage variables have significant effect on Price from Logistic Regression. Also, by looking at the decision trees, it can be said that Bathroom, Square meter, City and Room has the highest importance and they play the most important role in our decision-making. For further analysis, variance modelling can be modelled in order to get rid of heteroscedasticity. Also, normality assumption can be provided. Finally, more variables can be added to explain response better.

5. REFERENCES

- PCA Analysis in R. (n.d.). Retrieved from https://www.datacamp.com/community/tutorials/pca-analysis-r.

- Decision Tree in R with Example. (n.d.). Retrieved from https://www.guru99.com/r-decision-trees.html

- HOME. (n.d.). Retrieved from https://stats.idre.ucla.edu/r/dae/canonical-correlation-analysis/.

- How to perform a Logistic Regression in R. (2015, September 13). Retrieved from https://www.r-bloggers.com/how-to-perform-alogistic-regression-in-r/.

- Random Forest in R, Maklin, C., (2019, Jul). Retrieved from https://towardsdatascience.com/random-forest-in-r-f66adf80ec9