Data Analytics eLearning Academy
Final Project

Author: Sidiropoulou Nikoleta

Agency: College Link

23/10/2021

Summary

The following paper consists of three main parts. In the first section, the introduction of the data file in the R programming language, as well as their preparation and appropriate cleaning (Data Cleaning) are presented. In the next section, descriptive statistics and the visualization of the data through hypothesis tests, as well as the presentation and explanation of tables and diagrams (Descriptive statistics and Visualization) are applied. In the last section, several regression models and algorithm classification are conducted to effectively derive new useful information from the data (Data Mining) and drive the overall conclusions of the analysis.

Introduction

Data Analysis is a widespread and rising discipline of our time. With the use and continuous development of information systems and programming languages, it is a field of study and data management of statistics. One of the best known programming languages used internationally to solve statistical problems and analyzes is R.

The statistical analysis to be applied to this report is based on data from a random sample of 117 cases from the house sales files of a large brokerage company in the US from 15 February to 30 April 1993. The available variables with the measurement units consist of PRICE (selling price in thousands of USD), SQFT (square feet of residence), AGE (age of residence in years), FEATS (number of additional features from 11), NE (North East, binary variable indicating whether the dwelling is in the north-east side of the city), COR (binary variable indicating whether residence is angular) and finally TAX (annual taxation in USD).

In the following pages is presented the analysis in three main parts. Initially, the data for analysis are cleaned and processed. Descriptive statistics and visualization of the individual variables are then applied with appropriate diagrams for understanding statistical relationships. Finally, data mining techniques are carried out through statistical models for deriving the final data conclusions.

First Part: Data Cleaning

Initially, the .txt format data file (House Prices) is imported into the R programming language, saved with the name (Housedata.txt), and the import format is checked.

The columns containing missing values represented by ("*") are then identified to be replaced with NA values. This procedure is done by converting these columns, AGE and TAX, from character vectors to numeric variables, leading to the creation in this way of the requested NA. The type of variables in the sample is then checked with PRICE, SQFT, FEATS, NE and COR being integer variables, the others retaining the final type of the previous procedure.

In the next step, binary numerical variables, NE and COR, are converted into factors by substituting for each variable the 0 and 1 values with "No" and "Yes" given in the requested format.

Then, these rows containing at least two NA values are subtracted from the data. By storing them in a new vector, del, after the application of proper logical operations, the value vector is subtracted from the Housedata and the result is stored in the initial sample from the outset.

The following is the conversion of the variable SQFT (square feet of residence) to SQM (square meters of residence). With their equivalence, $1m^2=10.746ft^2$, the value vector of SQFT is divided by 10.764, and the new values are stored in the existing variable in data. It is also renamed from SQFT to SQM.

The last step in this process is to replace all missing values in the sample with the new ones, rounded to the values that are expected. The latter are to be predicted by using an applied linear regression model each time for variables containing NA, AGE and TAX. First of all, it is necessary to define the sample data of Housedata rows that do not contain any NA and then store them as subfull for use in regression models. The two corresponding ANCOVA linear models are then applied. The lmfullA has a response variable, AGE, and explanatory variables all the rest of subfull and lmfullT has a response variable, TAX, and explanatory all other variables respectively.

To estimate the expected values, the values of the individual variable corresponding to the observations containing NA for the first and the second case in a similar way shall be stored

in data frame form in the corresponding newdataA and newdataT. Then the expected values of AGE and TAX are predicted, by initially applying newdataA to lmfullA, rounding the output and storing it in the predA vector. Finally, this vector is replaced at NA records in Housedata. By a similar process, the expected values are estimated and stored via lmfullT in the predT to be replaced at the TAX records in the Housedata.

Second Part: Descriptive Statistics & Visualization

This section starts with the provision of some descriptive measures. Namely, the average value (mean), median, standard deviation (sd), the lower value (minimum), and the higher value (maximum), for all numeric variables PRICE, SQM, AGE, FEATS and TAX.

The first variable is PRICE, with the following measures for its distribution: mean= 1076.11, sd= 383.01, median= 975, min= 540 and max= 2150. The mean is a measure of central tendency and is more "sensitive" to its calculation in extreme distributional observations. However, the median by definition is the observation of the distribution found in 50% of the ordered values of its range. The fact that the mean is higher than the median indicates the existence of a positively asymmetric distribution. There is a concentration of values around the former, which due to the extreme observations is drawn to them in the range. The standard deviation is the square root of the variance of the distribution and gives the spread of observations around the mean. As for the minimum and maximum value, this is a fairly large range of values concerning the prices of the sampled houses.

For the second variable SQM, the measures are given: mean= 154.78, sd= 49.01, median= 145.39, min= 77.76, max= 348.38. With the mean value slightly higher than the median, it points to a potentially slightly positive asymmetric distribution for SQM. The range is remarkably smaller than that of the previous variable, which also depends on its unit of measurement. This interval represents the spread of observations in the distribution of the square meters of each dwelling i.e. the standard deviation.

Regarding AGE, its descriptive measures are: mean= 18.1, sd= 11.54, median= 17, min= 1, max= 53. Similar to before the suspicion of another slightly positive asymmetric distribution arises. The fairly small value of standard deviation is proportional to the range of values of the variable and arises as explained above.

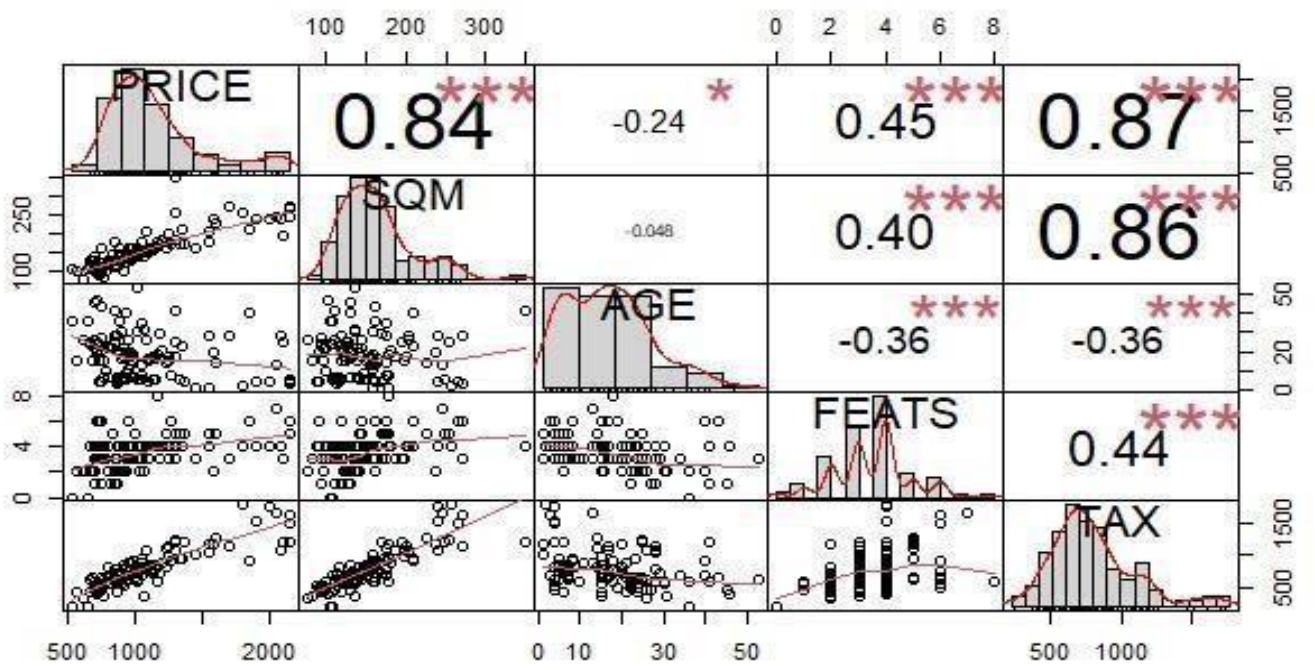
Then, the measures of FEATS variable are: mean= 3.53, sd= 1.4, median= 4, min= 0, max= 8. The mean is slightly lower than the median of the distribution. However, because it is a discrete variable with a small range of values, comparing the results might give a misleading conclusion. As regards the sample, it informs the existence of a small number of dwellings with few or no additional features as indicated for the variable. An expected small value is also observed for standard deviation.

Finally, for TAX variable, are provided the following descriptive measures: mean=793.75, sd=305.83, median=731, min=223, max=1765. Another case of a remarkable positively asymmetric distribution is observed. There is also a fairly large range of values and quite large spread of observations around the mean as easily deduced from the standard deviation value. The distribution is quite similar to that of PRICE, which is derived from the study of statistical correlation in the table below.

Then, is calculated the number of observations for each level of the factors. More specifically for NE, is provided a frequency table with 37 versus 72 observations for the value "No" and "Yes" respectively. Also, the COR frequency table, returns 87 observations for "No" and just 22 for "Yes". For the latter the difference in the frequencies of occurrence of the values is much larger than the first.

For the discrete variable FEATS, a relative frequency table is calculated, that is the range of the frequencies of occurrence of each observation divided by the length of the observations as a whole. The latter returns the frequency of each observation as a part of the total population. The fairly low frequency for the higher values is concluded mainly from value 7 onwards.

Then, is illustrated a correlation matrix for all numeric variables of Housedata. The graph requested is below:



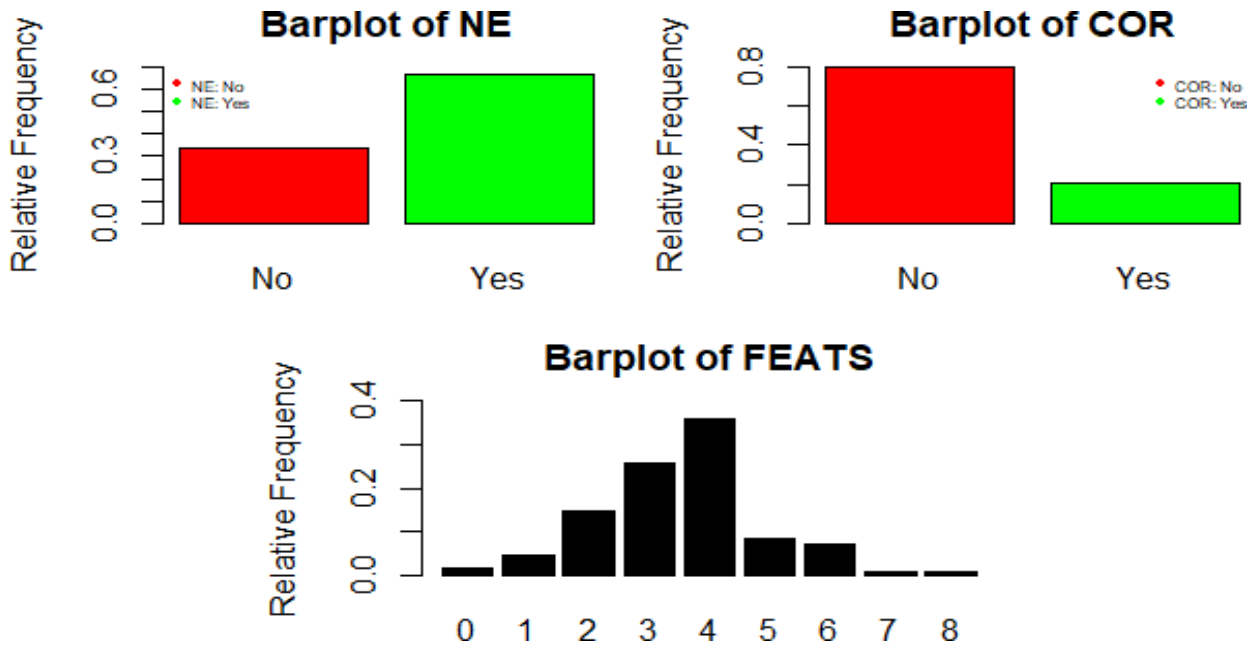
Picture 1 Correlation Matrix of Housedata

Initially, the method followed is the calculation of the correlation coefficient Pearson which gives the intensity of the linear relationship of numeric variables in the graph. Its value is given in the top right part and corresponds each time to the variables on the vertical and horizontal axes. The factor signs indicate the, on average, either positive or negative correlation, while the asterisks give information about the significance level that the null hypothesis for the statistical equality of parameters with 0 would be rejected in any case. A strong positive linear correlation can be identified for the PRICE and SQM variable pair, as well as for the PRICE with TAX and SQM with TAX pairs, with coefficients higher than 0.8. It also follows for the other variables however with lower coefficients indicating a weaker statistical relationship, as for the negative linear relationship of AGE with FEATS, similarly for AGE and TAX.

The main diagonal shows the histograms initially of the continuous numeric variables but also of the discrete variable FEATS. The non-parametric estimation of the density represented by the red line captures the height of the distribution curve. Also, the asymmetries of the above mentioned distributions, on average, are evident. The visualization of these relationships is also given by the scatter plots at the lower left part of the graph. For each graph based on the concentration of observations around the trend line it is feasible to get information on the degree of correlation and of course the type of it illustrated by the

slope of the line. For a pair of variables with positive linear correlation, large values one variable mean, on average, large values for the other, whereas for the negative correlation large values for one variable would imply small expected values for the other.

In the next step, the bar plots are given for the factors, COR and NE, and the discrete numeric variable FEATS:

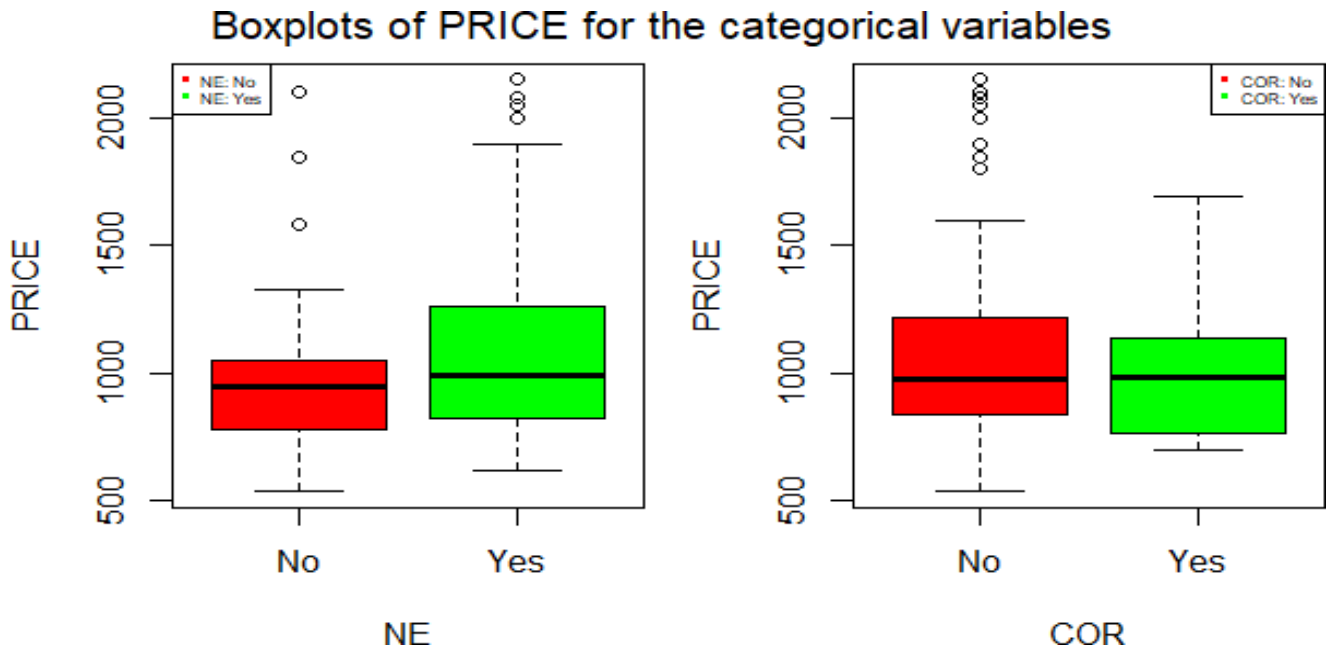


Picture 2 Bar plots of NE, COR & FEATS

The above graphs represents the bar plot of relative frequency of NE variable on the left and the corresponding bar plot of relative frequency of COR variable on the right.

For the first one, the relative frequency of "Yes" value is obviously higher than that of "No" value. This explains that most of the dwellings belong to the north-eastern geographical position compared to the rest of the whole. Continuing with the graph for the relative frequencies of COR, the appearance of "No" value is almost four times higher than that of "Yes" value. In other words, non-angular dwellings make up almost one quarter of the total dwellings in the sample. In the lower graph, the bar plot of the discrete variable is illustrated. The mode corresponds to 4 and is close enough to the mean of the distribution. Therefore, a concentration of values around the mean and a possibly asymmetric distribution are observed.

The box plots of PRICE variable, by category for the different levels of the categorical variable NE, then the COR, are represented below:



Picture 3 Box plots of PRICE for the categorical variables

The distributions of PRICE variable for "No" value for both the NE and COR categorical variables are given in red, while the corresponding distributions for "Yes" value are given in green.

Starting from the left graph are the boxes for PRICE distributions for the two above mentioned NE levels. Initially, it is quite obvious that the distribution of "Yes" is slightly higher than that of "No", in other words the north-eastern position shows, on average, a positive effect on the price of housing.

The spread of each distribution is also distinguished, in particular through its position of the median which makes up the 50% of the middle observations relative to the size of the boxes. Their median positions are defined by the black horizontal line within the boxes and are close enough for the PRICE distributions, so they move in a similar range of values. As for their upper and lower whiskers, the distribution of "Yes" value seems to be positive asymmetric, while both have some positive outliers points.

PRICE distributions for the different levels of COR are also close enough, as are their median levels. This also explains the possible non-statistical effect of the categorical variable on the numeric variable but in any case it changes according to the size of each distribution, with positive outliers for the "Yes" value.

The next step in the analysis is to find the mean values of the variable PRICE for the different levels of the variable NE and respectively COR. As a result, the values 992.43 and 1119.11 for the first, similarly 1095.28 and 1000.32 for the second respectively are given at the "No" and "Yes" levels. However, in order to check whether a statistical significant difference exists between the means, a Shapiro test for PRICE distribution initially applied as a condition for the further implementation of the former test. With the alternative hypothesis being that the population is not normally distributed there are strong indications for rejection of the null hypothesis for normality at significance level a=1%.. As a result, the non-parametric Wilcoxon test is applied to check the statistical equality of the medians. The results for the distributions of PRICE in the levels of both categorical variables, NE and COR, result in the non-rejection of the null hypothesis referring to the statistical equality of individual medians at significance level a=10%, versus the alternative hypothesis claiming the statistical difference of the medians.

The variables with high positive asymmetry are then checked and selected to apply a normality test to them. The variables are PRICE SQM and TAX with skewness>1. Then, by using a Shapiro test, the initial conclusion for all three variables is confirmed, as the null hypothesis referring to the existence of normality is rejected at significance level a=1%.

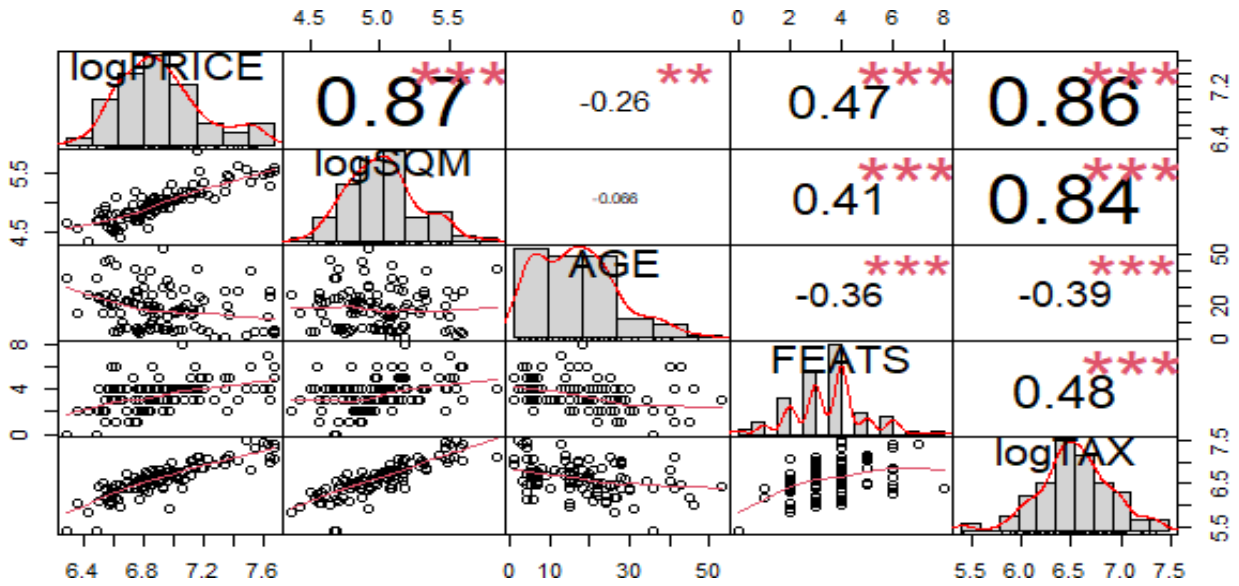
Then their logarithm is calculated and the Shapiro test is reapplied for each of the distributions. First of all, from the test of the PRICE distribution, an extremely small p-value is returned, allowing the null hypothesis to be rejected at significance level a=1%. For further verification of this, an additional test is also applied, the Lillie test, where the returned result does not allow the rejection of null hypothesis at significance level a=1%, which refers to a normal distributed variable versus the alternative hypothesis claimed to the contrary.

The interpretation of these slightly different results is due to the way each test statistic is calculated. In particular, Shapiro test is more 'sensitive' to extreme observations in distribution, as opposed to Lillie test, which does not follow this method of calculation.

For variables SQM and TAX, the results of applied Shapiro test do not allow us to reject the existence of normality in distributions at significance level a=10%. Because of these results it can be claimed that the problem of the high asymmetrical distributions of the numeric variables was tackled effectively so now they approach the normal distribution. With the new transformed values after calculating the logarithms for PRICE, SQM and TAX, the replacement is applied to the previous column-by-column of each variable in Housedata. The

new corresponding names are then given for each column.

Below is the new correlation matrix for the transformed numeric variables of Housedata:



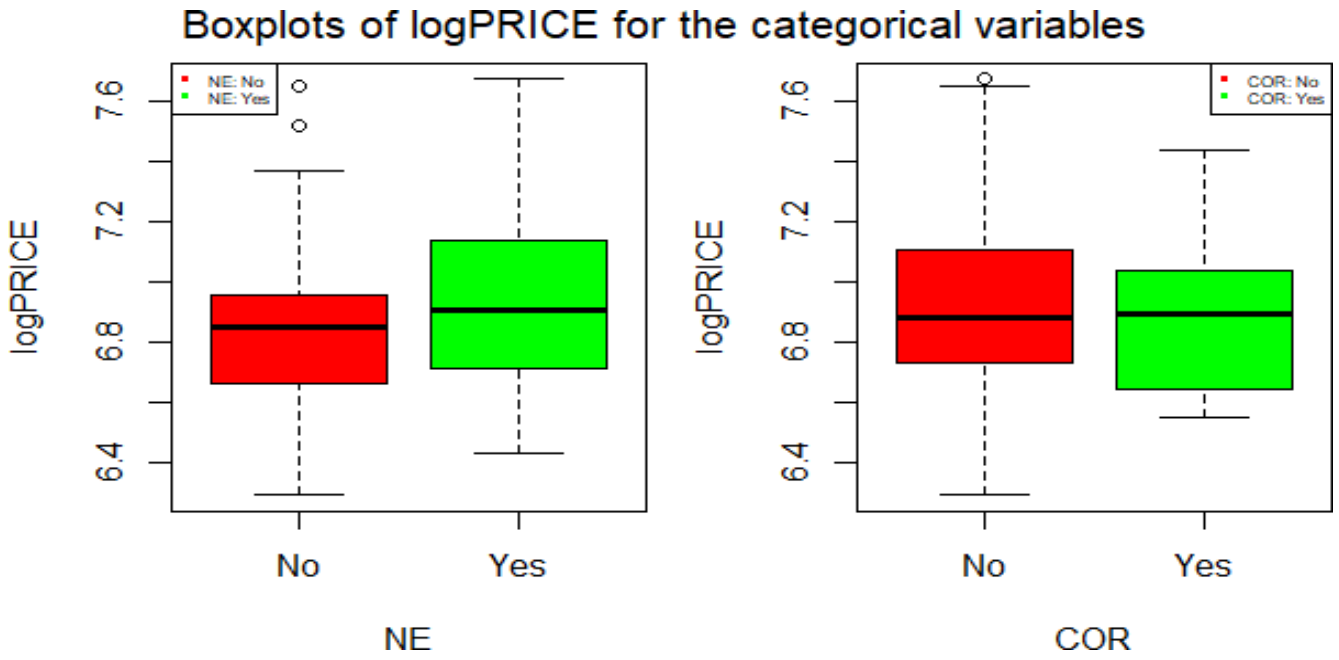
Picture 4 New Correlation Matrix of Housedata

From the graph above, the distributions of numeric variables can be easily distinguished in more detail after the transformation and the concentration of observations in each histogram separately. With the latter, the assumptions we made earlier about logPRICE, logSQM and logTAX are visually apparent.

There are also stronger correlations, on average, for the variables per pair, as interpreted by Pearson's correlation coefficients, which are in absolute terms higher than the corresponding values previously. The asterisks lead, auxiliary, to the same conclusion.

Finally, stronger linear correlations are observed, on average, mainly for the pairs of observations with the corresponding higher Pearson's correlation coefficients, such as the logPRICE with logSQM, or with logTAX, following all the others. Also, stronger linearity may be observed for pairs of variables with negative correlation, such as AGE with logTAX.

To interpret the last graph of the second part of the analysis, the two box plots are given concerning the distributions of logPRICE samples, by category for the different levels of the categorical variables NE and COR respectively:



Picture 5 Box plots of logPRICE for the categorical variables

By comparing the graph with the previous one, it is observed that for the values "No" in both categorical variables, NE and COR, the upper and lower whiskers i.e. right and left of the distributions are higher than the previous. This means that the data are spread in a wider range and more symmetrically. The positions of the medians move in a similar range between the levels of each variable.

As for the distributions per graph they appear to move in the same range again, in other words, logPRICE does not seem to be significantly affected by the levels of the categorical variables.

Third Part: Data Mining

In the third and last part, the first step is to apply an ANCOVA model, with response variable the logPRICE and explanatory variables all the rest of Housedata. The interpretations of the (bi) parameters of the model follow in detail.

Due to the existence of both numerical and categorical variables in our model, the (intercept) is initially interpreted, as the reference level for NE and COR variables. In other words, is the average expected logarithm of the price value of a dwelling, if it is not on the north-east side of the city and is not angular.

Regarding the interpretations of the numeric variables the reference level is the expected value of logPRICE, for values equal to 0 in the first, which is not always interpreted practically as is in the case with the data being considered. An example of this is a dwelling with a zero SQL value that cannot actually take place.

The interpretation of the estimated parameter for each numeric variable respectively is equal to the expected value of the dependent variable in its unit of measurement. Due to the appearance of logarithms in variables it is important to interpret them correctly, using inverse transformation, but also the type of correlation between the explanatory and response variablethat can be converted to exponential.

As regards the parameters of the explanatory categorical variables, if reference was made to an observation with value “No” for COR, by subtracting the value of the estimated parameter, for the level of COR, from the intercept value, would derive the expected difference of logPRICE for the observation with these characteristics. The interpretation of logPRICE with a change in the level of NE is similarly given.

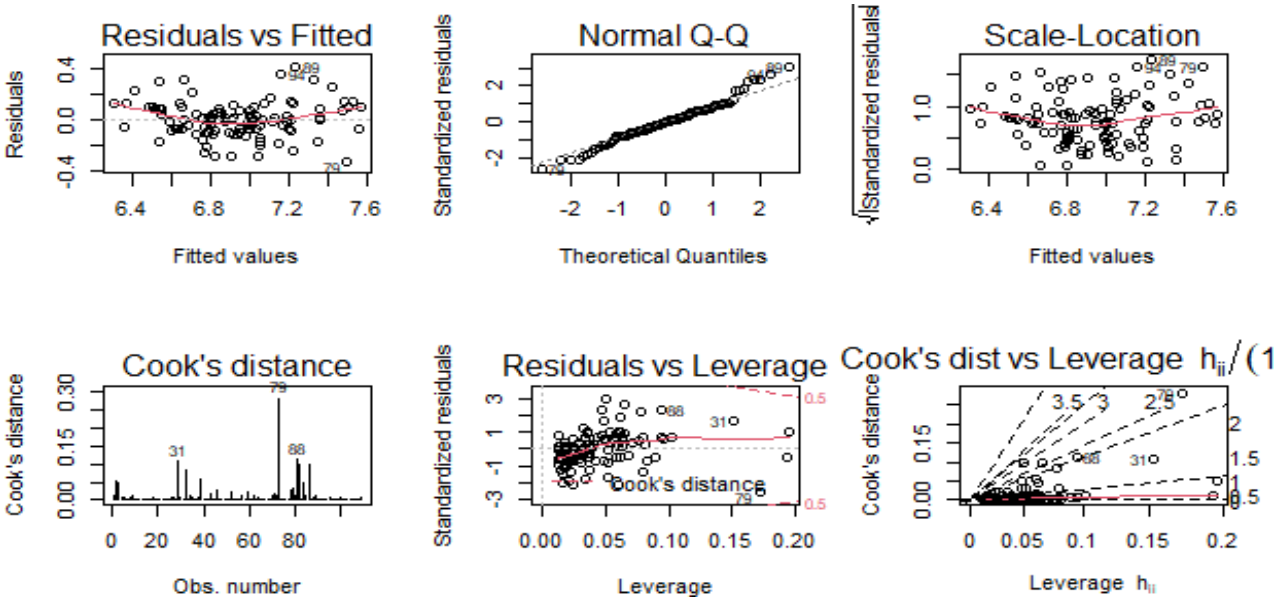
An algorithmic stepwise test procedure is then applied to select the final model, mfinal, with a response variable, logPRICE. The variables that explain our model are logSQM, AGE, CORYes and finally logTAX.

As regards the intercept of the model, it is the reference level of the unique categorical variable COR, so it refers to a dwelling with characteristics the value “No” for COR when the parameter of CORYes as well as the other numeric variables are equal to 0 and gives the average expected value of logPRICE.

It is worth mentioning that some of the estimated parameters, such as the logSQM, AGE

and logTAX, appear to have a positive statistical correlation with the response variable of the model, while the others have a negative sign in their estimate. However, special attention to further analysis of these statistical relationships would be required in case of existence of multicollinearity between the variables. The majority of these are also statistically significant variables, which is expected by observing the given p-values of their estimates. Finally, a high enough value is given for R^2adj=0.81 which identifies the percentage of variance that is explained by the model, weighted by the number of explanatory variables i.e. the inputs of model.

The hypotheses of the model are then tested, accompanied by the subsequent diagnostic graphs:



Picture 6 Diagnostic Plots of mfinal model

The assumptions checked are the normality, the homoscedasticity and the independence of the estimated residuals as well as the linearity in the statistical correlation of the response with the explanatory variables. Finally the assumption of the non-existence of multicollinearity is also made i.e. the independence between the explanatory variables.

From the Residuals vs Fitted plot, is observed a "cloud" of points which does not seem to indicate the appearance of a structure in the residuals. However, higher variance of residuals towards the lower and upper ends of the distribution shall be discerned. Some nonlinearity may be indicated in the relationship of the response and explanatory variable. These assumptions

will be checked further, by applying a number of hypothesis tests. Furthermore, ways to tackle the possible violation of linearity are given below.

Regarding the check of the assumption of normality, the Normal Q-Q Plot, helped to determine if the residuals of the regression model are normally distributed, according to the theoretical and sample quantiles. In the graph there are no great deviations except for some observations in the tails of distribution which do not concern us.

In the Scale-Location plot is checked the assumption of the existence of heteroscedasticity in the sample, defined as the non equal variance among the residuals in our regression model. An increase of variance may be occurred at the tails of the distribution, which will be checked more precisely below.

Finally, for the remaining three graphs, the Cook's distance indicator is a measurement of the influence that an outlier has on the slope of the regression line. It refers to an observation which is far away from the regression line and has a significantly high value for its explanatory variable (influence point).Some examples of them were observed to take values as 31, 79 and 88. The latter are also (Leverage) points, i.e. points whose independent variable value is distant from that of the other observations. From the graphs does not occurred a great violation of the assumptions examined.

To establish the validity of the above, a number of diagnostic hypotheses tests are also applied. Initially, a Shapiro test for the distribution of the residuals is applied, where according to the given p-value and the alternative hypothesis as used in previous tests, it is not possible to reject the null hypothesis for existence of normality at significance level a=10%.

It is then checked whether there is a violation of homoscedasticity. Based on strong indications occurred from the results of Bartlett test the null hypothesis does not rejected at significance level a=10% for equal variances of the sample of residuals across the groups of the expected values, against the null hypothesis which it claims that the variance of at least one of these three groups is statistically different from the others.

A check on the independence of the residuals is then carried out. It is initially applied the Durbin-Watson test with null hypothesis that the data are not being auto-correlated against the alternative hypothesis which claims that the autocorrelation is not equal to zero. The p-value returned is less than 0.05 which means that the null hypothesis is not rejected at significance level a=10%. It can be further investigated with some tests such as Runs test which gives strong

indications to not reject the null hypothesis that the sequence of residuals is a random one i.e. is not existed a structure in their distribution at significance level a=10%.

Last hypothesis checked is the linearity through the residualPlots test. Due to the returned result of Tukey test, the null hypothesis for non-existence of statistically significant secondary term which was not taken into account is rejected and converges in the alternative hypothesis which claiming its existence, at significance level a=1%. The small p-value returned, derived from small values of individual variables for the same test. One way to tackle this is to add the secondary term to the corresponding variable of the model, as mentioned above, and applying a new algorithmic stepwise test. The new returned model is the mfinal1 with statistically significant explanatory variables according to the given p-values by majority, and a slightly higher value for the R^2adj about 0.82 interpreted as explained above.

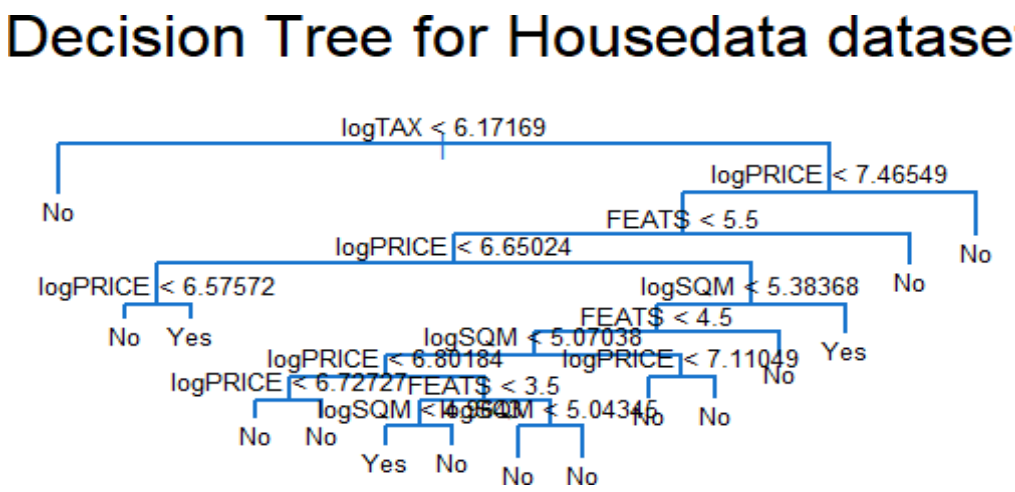
The new categorical variable catFEATS is then created, which orders the values of FEATS to the three levels “Low”,”Moderate” and “High”, based on the number of additional characteristics of each dwelling. An ANOVA model for logPRICE is then applied being attributable to the different levels of catFEATS. In other words, the statistical significance i.e. the impact that the categorical has on the response variable, is being checked. With null hypothesis the absence of statistical difference in mean values, according to the returned p-value of the model there are strong indications to reject the first, at a significance level a=1% converging in the alternative one claiming that at least one of the mean values is statistically different from the others.

With the addition of catFEATS to Housedata, the selection of an optimal logistic regression model, finalN, is performed through an algorithmic stepwise test procedure. This model has as a dependent variable the NE and independent all other from data except the variable previously introduced.

Then with the use of finalN it is estimated the probability that a house is located on the northeast side with the other values of variables specified, SQM=180, AGE=15, FEATS=5 and TAX=1000, using the logarithms where required in the model. By applying these new data and raising the finalN to the exponent, the probability of success is calculated, which is approximately equal to 0.88.

Our last step is the application and visualization of a decision tree for the COR variable, as well as the provision of the correctly predicted cases rate. By using the homonym package, the

necessary conversions of the types of variables to binaries are initially made and the decision tree for the variable fitCOR is applied for the data without the use of catFEATS with the variables used by the algorithm being logTAX, logPRICE, FEATS and logSQM. Then the COR confusion matrix is provided where the correct and incorrect predictions correspondingly for the two levels are given, “Yes” and “No”. By dividing the sum of the values of the main diagonal into the total number of observations is returned the ratio of the correctly predicted cases i.e. the classifier accuracy.



Picture 7 Decision Tree for Housedata dataset

The interpretation of the graph above starting with the root node and continuing with the following nodes, each time is making the logical decision according to the given value of the individual variable. If the logical value returned is TRUE, then the left direction is followed otherwise the right one is chosen, until reach the terminal root and the estimated final value of COR predicted.

Given some values for the variables contained in the decision tree, it is asked to estimate whether a dwelling is angular or not. By using logarithms where required for PRICE=1000, SQM=150, FEATS=4 and TAX=800, as previously done, and applying the procedure explained above, it is concluded that a dwelling of the above features is not angular.

Conclusion

Moving from the first part to the second one, the appropriate graphs are given followed by the tests on the statistical correlations of variables. Positive asymmetric distributions are observed for almost each numeric variable of the data that reasonably leads to the conclusion of existed outlier observations. The estimations of these relationships shall verified by statistical tests. The influence of categorical variables on the numeric dependent variable is further checked, concluding the non-statistical effect on the latter. In the final analysis, various linear regression models are applied with a fairly good assessment of their adaption as well as the great statistical effect of the ordering on the response numeric variable is concluded. Finally, some useful algorithmic procedures are applied and presented related to the classification of the data.