

DATA ANALYTICS USING R- HOUSE PRICES PROJECT OVERVIEW

PRESENTED BY: DIMITRIOS NANOS




PROJECT GOALS

- Data cleaning (Task 1)
 - Visualizations and Descriptive Statistics (Task 2)
 - Data mining (Task 3)
-

DESCRIPTION

Our data are a random sample of 117 houses from an estate agency in USA. They refer to sales of houses from different cities. The time period of this sales is from 15 February to 30 April in 1993. Our variables are:

- **PRICE:** sale price of the house in USD
 - **SQFT:** square feet of the house
 - **AGE:** age of the house
 - **FEATS:** number of electrical devices or extras
 - **NE:** if the house is in the northeast side the value of NE is 1, else 0
 - **COR:** if the house is in a corner the value of COR is 1, else 0
 - **TAX:** house tax per year in USD
- 

DATA CLEANING (TASK 1)

PRICE ▼	SQFT ▼	AGE ▼	FEATS ▼	NE ▼	COR ▼	TAX ▼
2050	2650	13	7	1	0	1639
2080	2600	*	4	1	0	1088
2150	2664	6	5	1	0	1193
2150	2921	3	6	1	0	1635
1999	2580	4	4	1	0	1732
1900	2580	4	4	1	0	1534
1800	2774	2	4	1	0	1765
1560	1920	1	5	1	0	1161
1450	2150	*	4	1	0	*
1449	1710	1	3	1	0	1010
1375	1837	4	5	1	0	1191
1270	1880	8	6	1	0	930
1250	2150	15	3	1	0	984
1235	1894	14	5	1	0	1112
1170	1928	18	8	1	0	600
1180	1830	*	3	1	0	733
1155	1767	16	4	1	0	794
1110	1630	15	3	1	1	867
1139	1680	17	4	1	1	750

That's a look of our 20 first observations from our data, before the cleaning process. First of all, we changed missing values from * to *NA*. After that, we changed the values of NE and COR by matching 1 to YES and 0 to NO. In the next step, we deleted from the sample the lines with at least 2 *NA* values. Afterward, we changed the value of SQFT variable according to this mathematical relationship $1m^2 = 10.764ft^2$. In addition, we renamed SQFT variable to SQM. In our final step, we used 2 linear regression models to predict *NA* values. We used as a subset of our data all the complete lines of the 7 variables. We noticed that we had to change 41 NAs in AGE variable and 2 NAs in TAX variable. The two models had in common that their independent variables were the 6 other variables of the subset. Their difference was that in the first model the depended variable was AGE and in the second was TAX. Finally, we replaced missing values from AGE and TAX with the fitted values of our models.

PRICE ▾	SQM ▾	AGE ▾	FEATS ▾	NE ▾	COR ▾	TAX ▾
2050	246.19	13	7	YES	NO	1639
2080	241.55	28	4	YES	NO	1088
2150	247.49	6	5	YES	NO	1193
2150	271.37	3	6	YES	NO	1635
1999	239.69	4	4	YES	NO	1732
1900	239.69	4	4	YES	NO	1534
1800	257.71	2	4	YES	NO	1765
1560	178.37	1	5	YES	NO	1161
1449	158.86	1	3	YES	NO	1010
1375	170.66	4	5	YES	NO	1191
1270	174.66	8	6	YES	NO	930
1250	199.74	15	3	YES	NO	984
1235	175.96	14	5	YES	NO	1112
1170	179.12	18	8	YES	NO	600
1180	170.01	29	3	YES	NO	733
1155	164.16	16	4	YES	NO	794
1110	151.43	15	3	YES	YES	867
1139	156.08	17	4	YES	YES	750
995	160.26	18	3	YES	NO	923

That's how our first 20 rows of data looks like after the cleaning process.

VISUALIZATIONS AND DESCRIPTIVE STATISTICS(TASK 2)

At first we will provide you some statistics about our arithmetic variables:

	mean	median	sd	min	max
PRICE	1076.11	975	383,01	540	2150
SQM	154.78	145.39	49,01	77.76	348.38
AGE	17.85	17	11.67	1	53
FEATS	3.53	4	1.4	0	8
TAX	793.43	731	306.11	223	1765

And about our categorical:

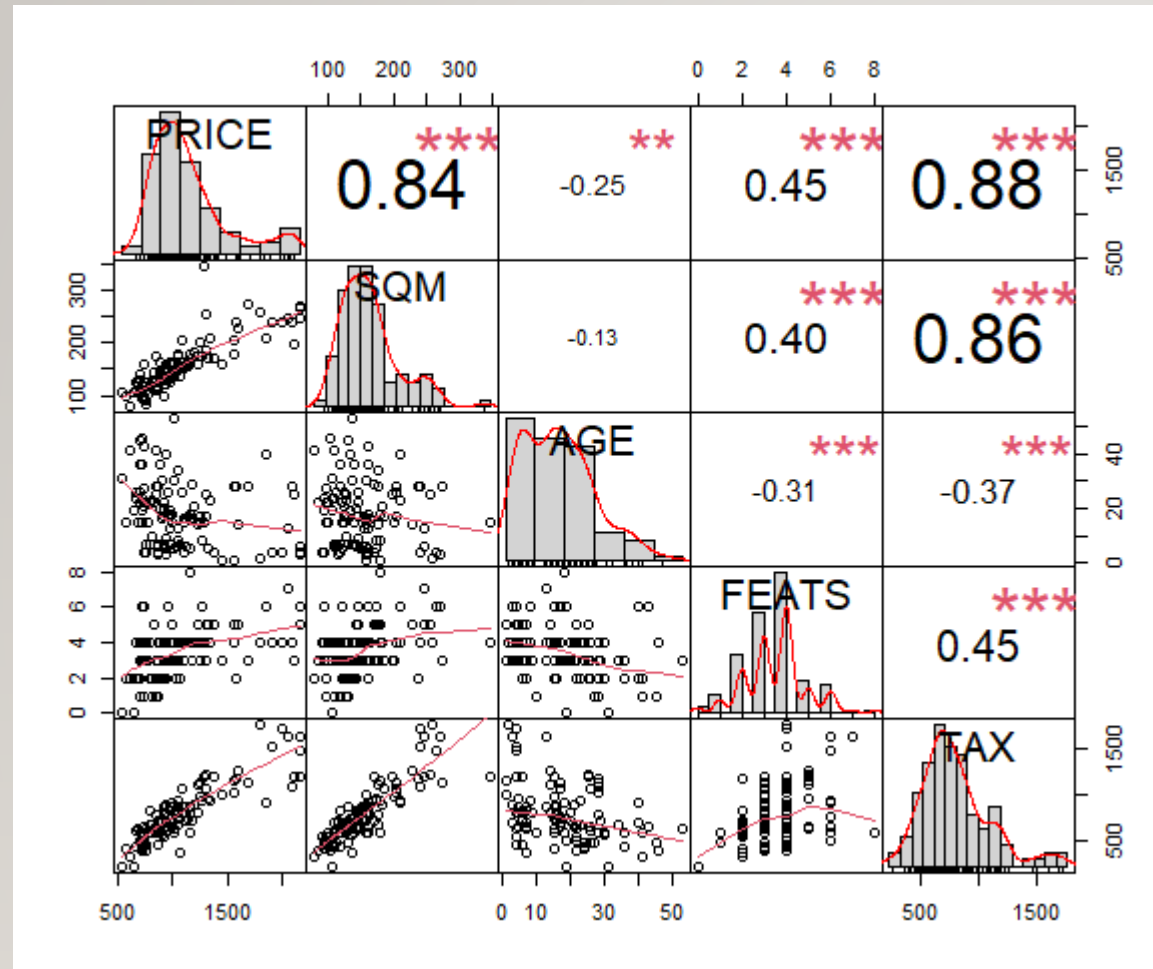
	NO	YES
NE	37	72
COR	87	22

And a probability table about FEATS:

FEATS	0	1	2	3	4	5	6	7	8
-----	0.0183	0.0459	0.1468	0.2569	0.3578	0.0826	0.0734	0.0092	0.0092

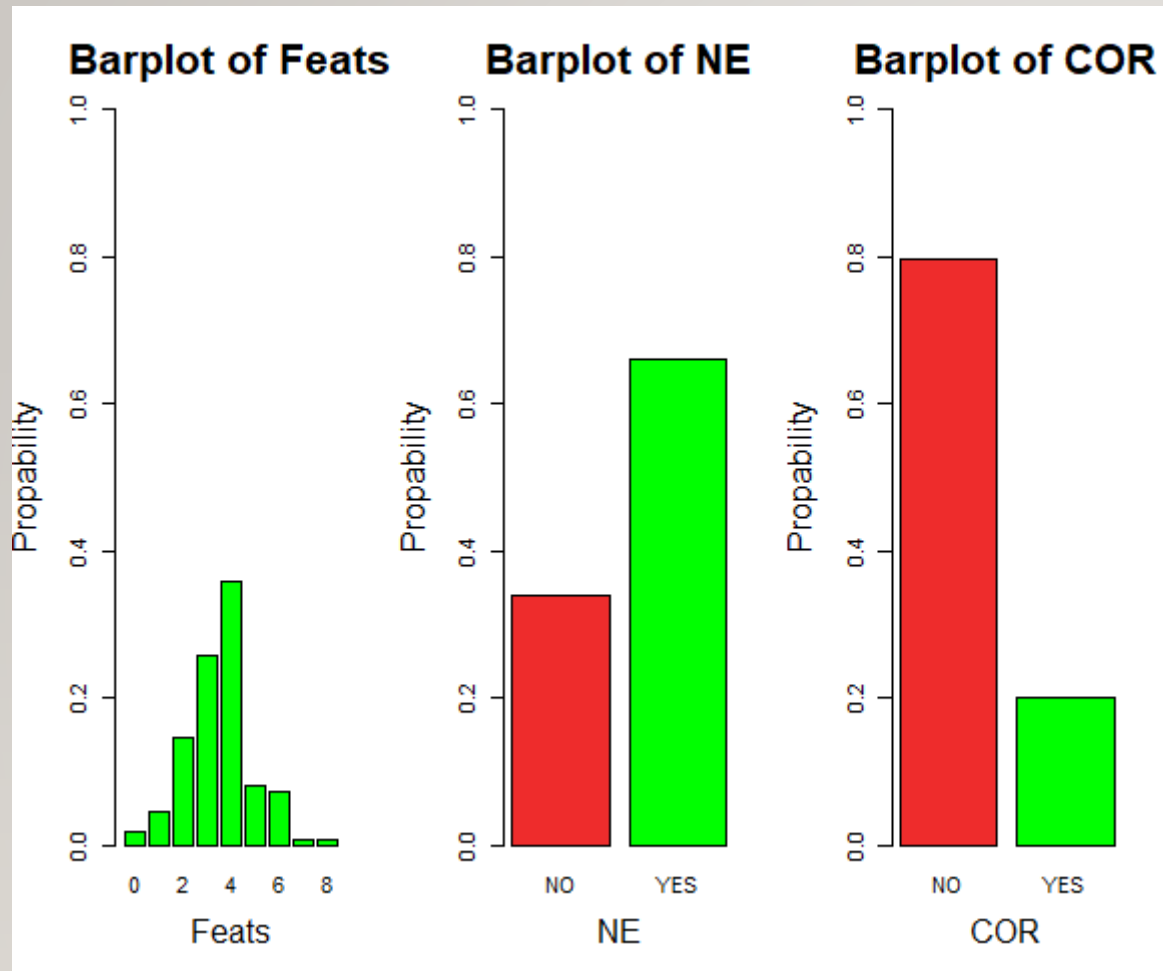


VISUALIZATIONS AND DESCRIPTIVE STATISTICS(TASK 2)



This is a scatterplot of all arithmetic variables of the sample. At first, we observe that PRICE~SQM, PRICE~TAX, SQM~TAX, PRICE~FEATS, FEATS~SQM, have a strong positive linear correlation. In parallel, AGE~FEATS, AGE~TAX have a strong negative linear correlation. Finally, we notice the highest correlation between PRICE, TAX and SQM.

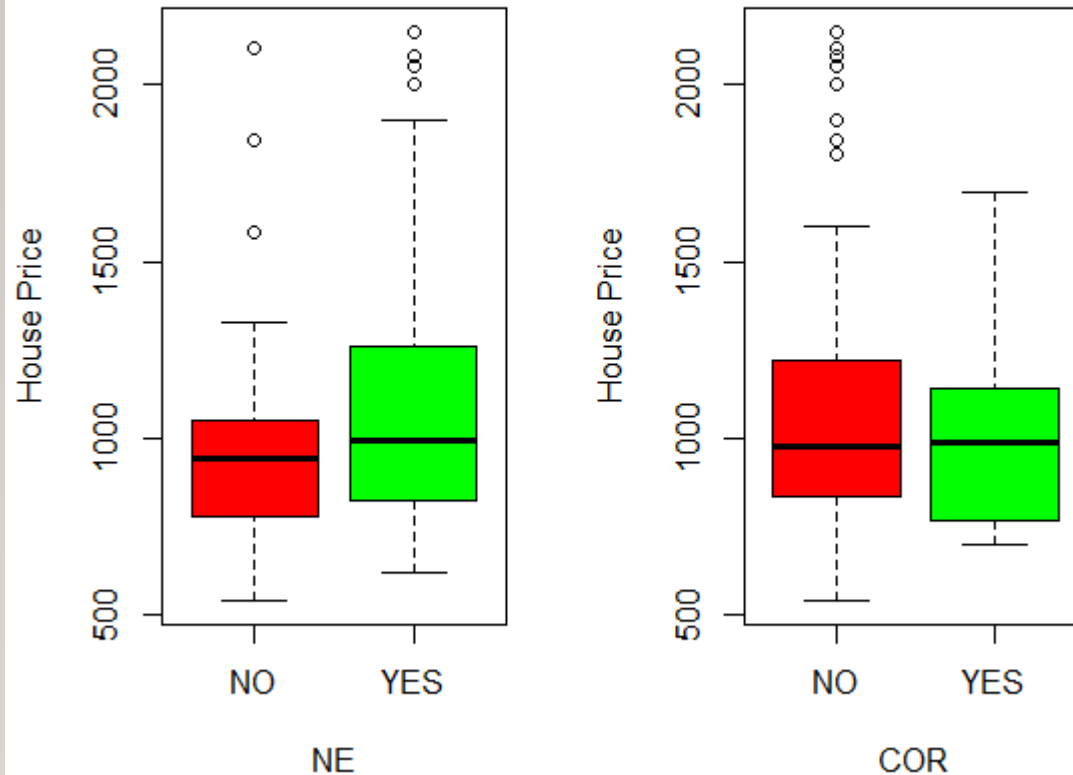
VISUALIZATIONS AND DESCRIPTIVE STATISTICS(TASK 2)



In the left we see some bar plots of the discrete variable FEATS and the two categorical variables NE and COR. We notice that the most houses are in the NE side of town, they are not cornered and they have 4 feats.

VISUALIZATIONS AND DESCRIPTIVE STATISTICS(TASK 2)

Boxplots of PRICE for the categorical variables



In the left, we got two boxplots of PRICE for the variables NE and COR. We notice that the median of PRICE through the 2 levels of NE has no big difference. Similarly, for the COR variable. As a conclusion for the box plots, we observe that the prices for the NE houses are higher, and the prices for the COR variable are in the same level approximately.

VISUALIZATIONS AND DESCRIPTIVE STATISTICS(TASK 2)

In the next step, we will take a look in means of PRICE, for the different levels of our categorical variables. Then we will run some hypothesis tests.

NE	YES	NO
Mean of PRICE for different levels	1119.111	992.4324

Through Shapiro tests for PRICE, based on the levels of NE, we observed that there is no normality, so we ran a Wilcoxon test for the 2 means and we decided that there is no statistical difference between the levels of NE for PRICE. Now, copying the last procedure for the variable COR, gives us the following results:

COR	YES	NO
Mean of PRICE for the different levels	1000.318	1095.276

Finally, after the Wilcoxon test, we are sure that there is no statistical difference between the levels of COR for PRICE.

VISUALIZATIONS AND DESCRIPTIVE STATISTICS(TASK 2)

Next, we calculate the skewness of our 5 arithmetic variables and we notice that PRICE, SQM, and TAX have it greater than 1. After 3 Shapiro tests for these 3 variables we are sure that they don't follow the normal distribution. So, we decided to use the logarithmic transformation. After this step, we ran Shapiro tests and checked again skewness. The results were a lot of better.

	SKEWNESS	SHAPIRO TEST p-value
PRICE	1.34222	7.206e-09
SQM	1.157332	3.258e-06
TAX	0.6441834	2.107e-05
LOG(PRICE)	0.6603108	0.0005961
LOG(SQM)	0.374416	0.1968
LOG(TAX)	-0.2493093	0.1333

Because of this results, we decided to transform our variables PRICE, SQM and TAX to logPRICE, logSQM and logTAX

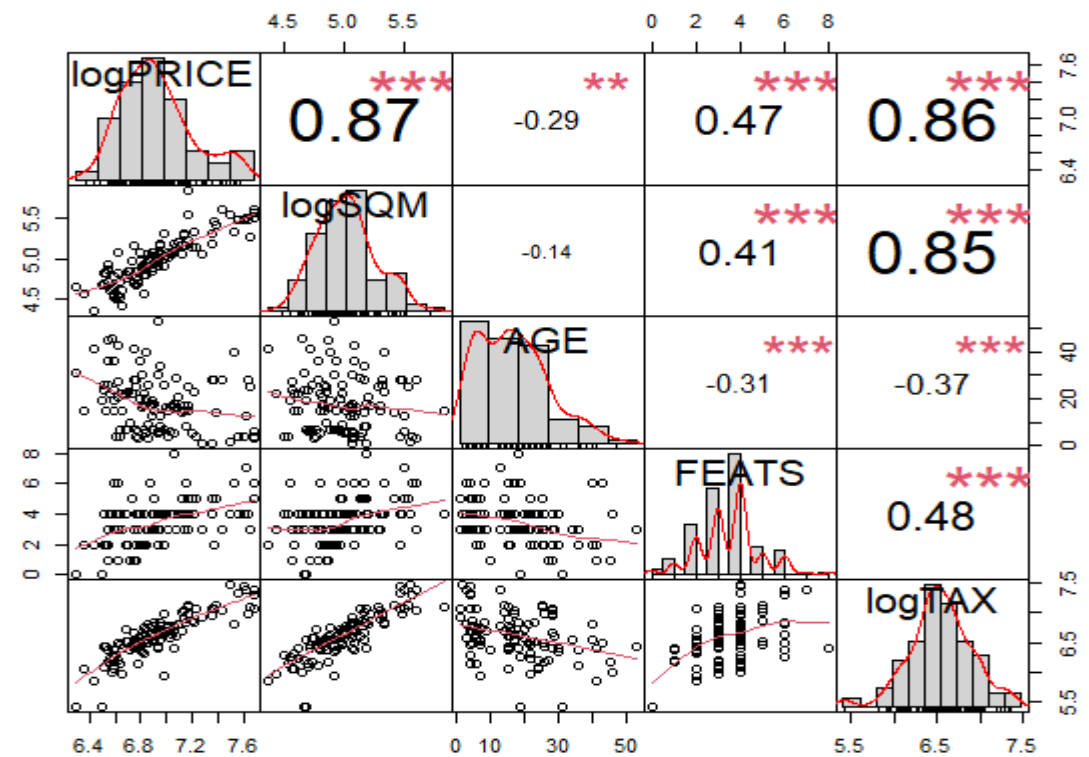
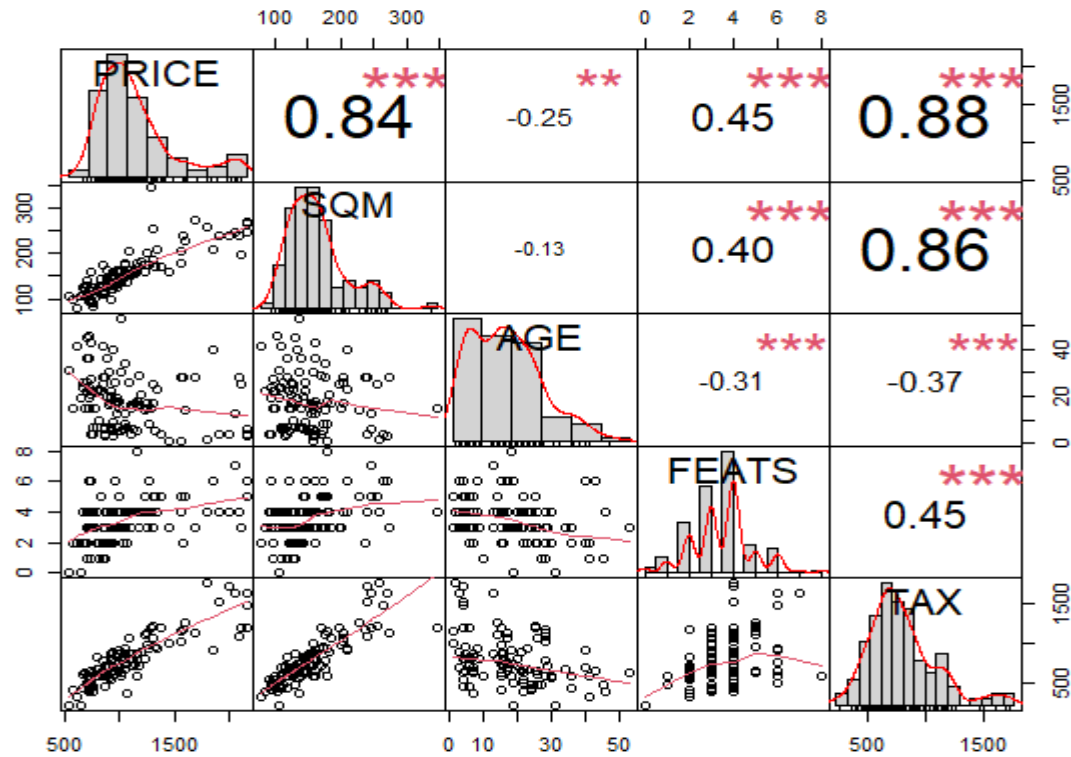
VISUALIZATIONS AND DESCRIPTIVE STATISTICS(TASK 2)

logPRICE	logSQM	AGE	FEATS	NE	COR	logTAX
7,6256E+14	5,5061E+13	13	7	YES	NO	7,40184E+14
7,64012E+14	5,48708E+14	28	4	YES	NO	6,9921E+14
7,67322E+14	5,51137E+14	6	5	YES	NO	7,08423E+14
7,67322E+14	5,60348E+14	3	6	YES	NO	7,3994E+14
7,6004E+13	5,47935E+13	4	4	YES	NO	7,45703E+14
7,54961E+14	5,47935E+13	4	4	YES	NO	7,33563E+13
7,49554E+14	5,55183E+14	2	4	YES	NO	7,47591E+13
7,35244E+14	5,18386E+14	1	5	YES	NO	7,05704E+14
7,27863E+14	5,06802E+14	1	3	YES	NO	6,91771E+13
7,22621E+14	5,13967E+14	4	5	YES	NO	7,08255E+13
7,14677E+14	5,16284E+14	8	6	YES	NO	6,83518E+13
7,1309E+14	5,29702E+14	15	3	YES	NO	6,89163E+14
7,11883E+14	5,17026E+14	14	5	YES	NO	7,01392E+14
7,06476E+13	5,18806E+13	18	8	YES	NO	6,39693E+14
7,07327E+14	5,13586E+14	29	3	YES	NO	6,59715E+14
7,05186E+14	5,10084E+13	16	4	YES	NO	6,67708E+14
7,01212E+14	5,02012E+14	15	3	YES	YES	6,76504E+14
7,03791E+14	5,05037E+14	17	4	YES	YES	6,62007E+14
6,90274E+14	5,0768E+14	18	3	YES	NO	6,82763E+14

This is how our data look like after log transformation.



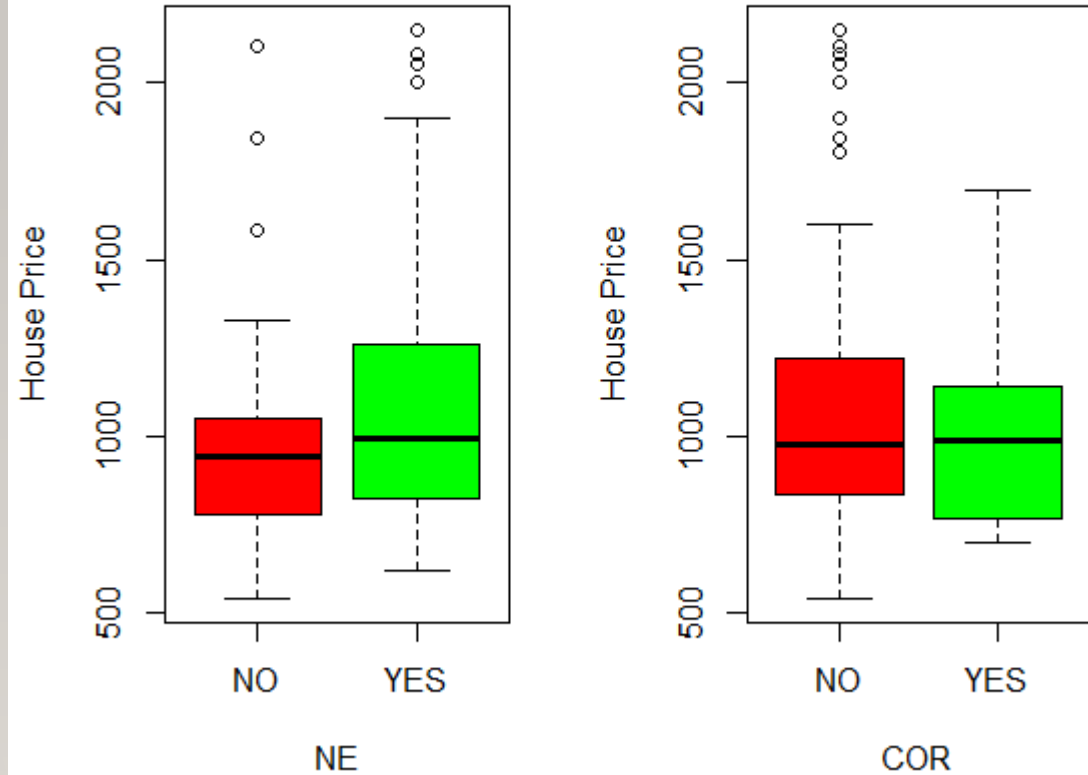
VISUALIZATIONS AND DESCRIPTIVE STATISTICS(TASK 2)



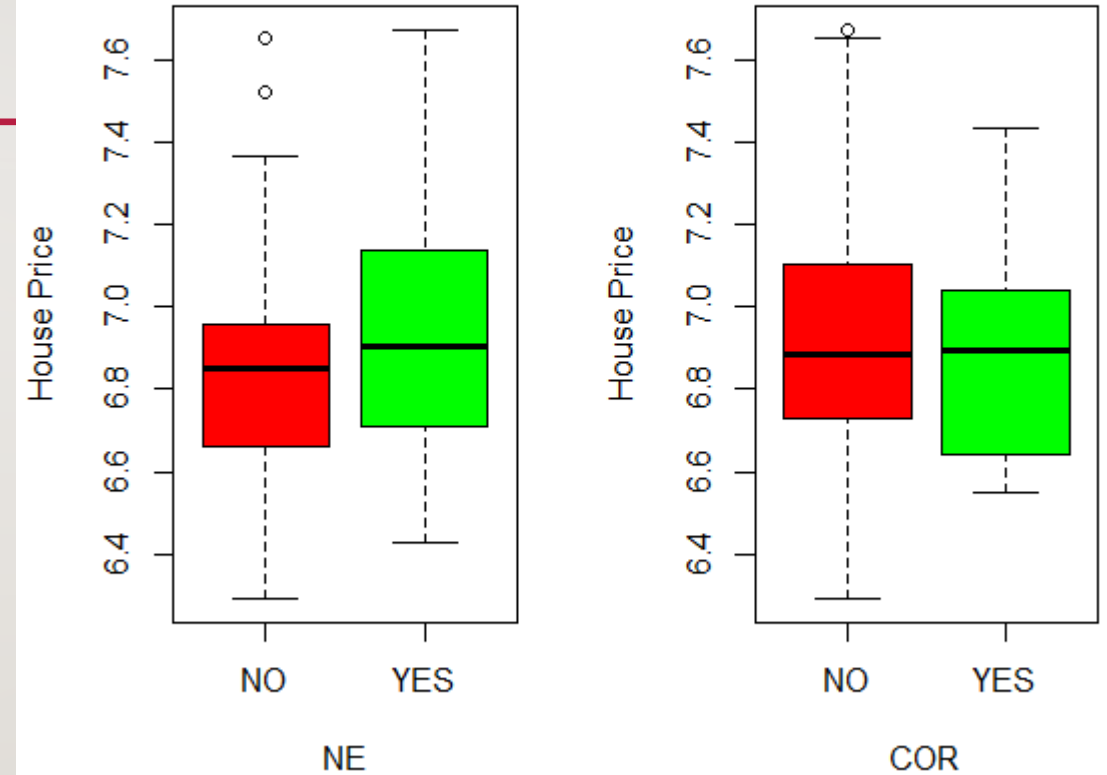
In these two charts, we observe only a little difference between the correlation of the variables.

VISUALIZATIONS AND DESCRIPTIVE STATISTICS(TASK 2)

Boxplots of PRICE for the categorical variables



Boxplots of logPRICE for the categorical variables



From this comparison, we notice that after the transformation, we significantly reduced the outliers.

DATA MINING(TASK 3)

In this section we are going to dig in our data by using some ways of statistical analysis. At first we will run a multiple linear model using logPRICE as depended variable and all the others independent. Let's take a look at the coefficients:

	Estimate	FEW WORDS ABOUT COEFFICIENTS:
(Intercept)	2.084990	- Intercept gives us the expected value of logPRICE if all the other variables are set to zero.
logSQM	0.565750	- logSQM gives us the expected value of change in logUSD, for logPRICE, if logSQM go up by 1 unit(1 logsqm) and all the other independent variables stay the same.
AGE	-0.001488	- AGE gives us the expected value of change in logUSD, for logPRICE, if AGE go up by 1 unit(1 year) and all the other independent variables stay the same.
FEATS	0.012241	- FEATS gives us the expected value of change in logUSD, for logPRICE, if FEATS go up by 1 unit(1 feat) and all the other independent variables stay the same.
NEYES	0.003296	- NEYES gives us the expected value of change in logUSD, for logPRICE, if a house is in the NE side of the town and all the other independent variables stay the same.
CORYES	-0.058426	- CORYES gives us the expected value of change in logUSD, for logPRICE, if a house is in corner and all the other independent variables stay the same.
logTAX	0.304076	

Residual standard error: 0.139 on 102 degrees of freedom

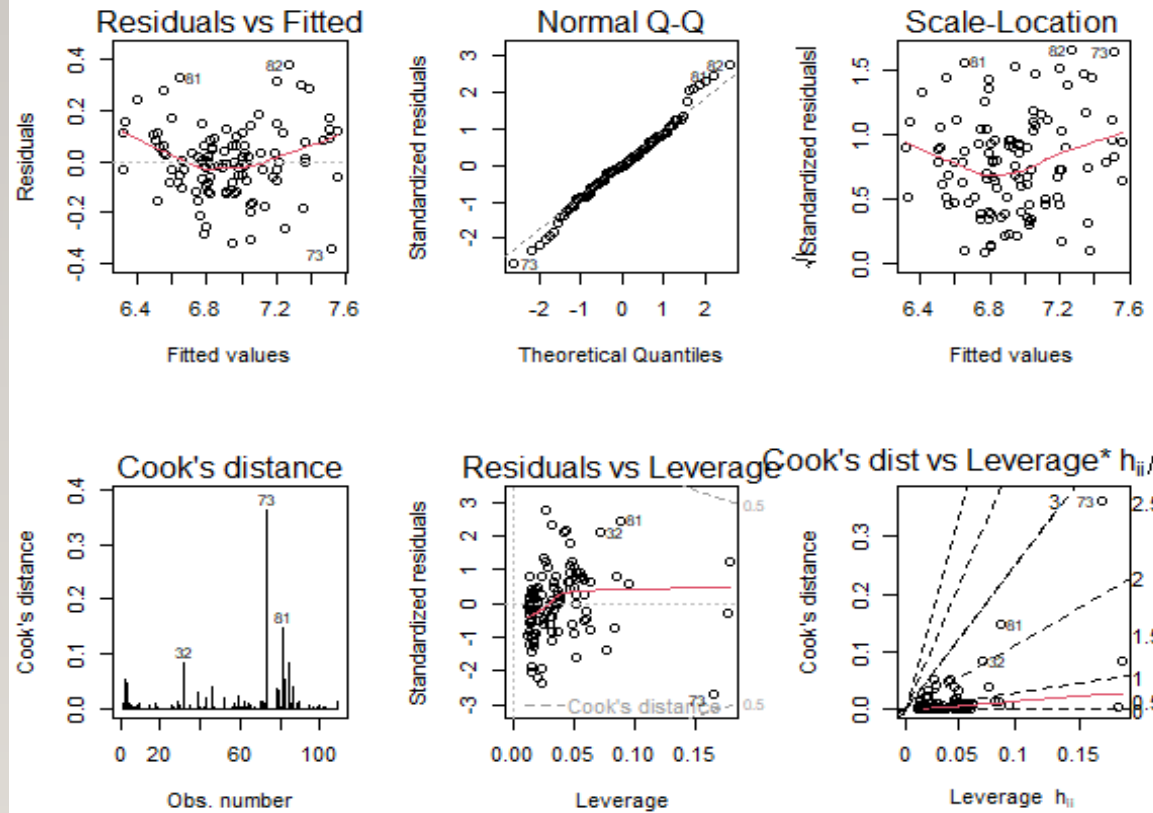
Multiple R-squared: 0.8188, Adjusted R-squared: 0.8081

F-statistic: 76.82 on 6 and 102 DF, p-value: < 2.2e-16



DATA MINING(TASK 3)

Now we will use stepwise algorithm to find the model that performs best. After running the appropriate codes in Rstudio, we ended up that the best model for logPRICE has as independent variables logSQM, COR, and logTAX. In the next steps, we will give you some diagnostic plots and assumption tests for the model.



	p.value
Shapiro	0.275
Bartlett	0.073
DW	0.046
Tuckey	0.003

After those steps, we see that we can trust the model for a confidence level equal to 99,9%.

DATA MINING(TASK 3)

In this part, we insert a new categorical variable named catFEATS, which has 3 levels. Level 1 made up of 0-3 feats and has the label of Low, level 2 made up of 4-7 feats and has the label of Moderate, and level 3 made up of 8 feats and has the label of High (In our dataset, minimum feats are 0 and maximum are 8). Then we run an ANOVA model with depended variable logPRICE and independent variable catFEATS. This gives us a p-value equal to 0.000881, which means that for the different levels of catFEATS our mean of logPRICE has statistical differences. In conclusion, catFEATS has an impact to logPRICE.



DATA MINING(TASK 3)

In this next to last paragraph, we will try to find the best logistic regression model for NE variable and calculate a propability. After using stepwise procedure, we have as outcome that the best model for NE as depended variable, has independent variables logSQM, AGE, FEATS and logTAX. Finally, with the help of this model, we calculated the propability of a house to be in the northeast side with PRICE=1200, SQM=180, AGE=15, FEATS=5, COR=NO, and TAX=1000 (we used log for PRICE, SQM and TAX). Our result was approximately 87%



DATA MINING(TASK 3)

This is our last paragraph of analysis. We choose to run a decision tree for our COR variable. Through our decision tree we can predict that if a house has PRICE=1000, SQM=150, AGE=17, FEATS=4, NE=YES, TAX=800 it is not in the corner. (we used $\log(1000)$ and $\log(800)$)

Our model has a percentage approximately 85% of right estimated values and looks like that:



$\log \text{TAX} < 6.17169$ 