**ADTA 5130 DATA ANALYTICS 1**

Project Report

PREDICTING THE PRICE OF THE HOUSES

NISHITH REDDY MANNURU 11467417

Contents

[INDUSTRY ANALYSIS 3](#_Toc70619321)

[INTRODUCTION AND BUSINESS OBJECTIVE 4](#_Toc70619322)

[DATA DICTIONARY 4](#_Toc70619323)

[HYPOTHESIS GENERATION WITH RESPECT TO PROBLEM STATEMENT 5](#_Toc70619324)

[IMPORTING THE DATASET INTO SPSS 6](#_Toc70619325)

[DATA VIEW 7](#_Toc70619326)

[VARIABLE VIEW 7](#_Toc70619327)

[DESCRIPTIVE STATISTICS 8](#_Toc70619328)

[CHECKING THE MISSING VALUES IN THE DATASET 9](#_Toc70619329)

[EXPLORATORY DATA ANALYSIS (EDA) 9](#_Toc70619330)

[CHECKING FOR OUTLIERS 13](#_Toc70619331)

[CORRELATION 16](#_Toc70619332)

[CHECKING FOR MULTI COLLINEARITY 17](#_Toc70619333)

[REGRESSION ANALYSIS -1(trimming the mean to remove outliers) 18](#_Toc70619334)

[REGRESSION ANALYSIS – 2 (RUNNING THE MODEL WITH OUTLIER CONSIDERED) 23](#_Toc70619335)

[WHICH MODEL TO USE 28](#_Toc70619336)

[RESUDIAL ANALYSIS 28](#_Toc70619337)

[RESIDUAL STATISTICS 28](#_Toc70619338)

[CHECKING THE NORMALLY DISTRIBUTED ASSUMPTION 29](#_Toc70619339)

[CHECKING FOR CONSTANT VARIANCE (HOMOSCEDASTICITY) 30](#_Toc70619340)

[LOG TRANSFORMING MY TARGET VARIABLE AND RE-RUNNING THE REGRESSION ANALYSIS 31](#_Toc70619341)

[REGRESSION ANALYSIS WITH THE TRANSFORMED VARIABLE 31](#_Toc70619342)

[RESIDUAL ANALYSIS WITH THE TRANSFORMED VARIABLE 36](#_Toc70619343)

[CHECKING FOR NORMAL DISTRIBUTION 37](#_Toc70619344)

[CHECKING FOR CONSTANT VARIANCE(HOMOSCEDASTICITY) 39](#_Toc70619345)

[CONCLUSION TO THE PROBLEM STATEMENT 40](#_Toc70619346)

# INDUSTRY ANALYSIS

Me Nishith and my partner Charunya, I are an expert at DFW land financial backer club. Our work is to show our customers a couple of the best properties around the DFW area, in light of their spending which may intrigue them, and furthermore where they can see themselves living for quite a while cheerfully later on.

At whatever point a land financial backer, approaches us with the possibility of purchasing a property, the primary thing I ask our customer is how much their spending plan for purchasing a house, and in which region they would like to claim a property, and furthermore not many of the fundamental inquiries, for example, the number of rooms they might want, the number of floors of a house they are searching for, the number of washrooms they might want and furthermore how huge of a house they need (square feet). Also, in light of the appropriate responses of the customer, I will show them the houses.

Our association is connected to over 100+ houses. How might I know precisely which houses to show our customer? I run a relapse investigation thinking about a couple of highlights of the house, like number of rooms, restrooms, number of stories in a house, where the house is situated, through which the cost of the property can be anticipated or the other way around.

Thus, when our customer discloses to us their spending plan for the house, I consequently suggest them few properties, referencing the highlights of the house also. Or then again if the customer advises us in the event that they are searching for a house with a specific number of rooms, washrooms, region, stories, and so on, I give them a gauge on how much a house can cost dependent on their prerequisites.

# INTRODUCTION AND BUSINESS OBJECTIVE

The main aim of this model is to build a predictive model which can predict the price of the houses based on the parameters such as, number of bedrooms, number of bathrooms, number of stories, the square feet of the house, the age of the house and also the vacancy of the house.

This can really help the real estate agents to predict the prices of the houses based on their clients’ requirements or even they can give an estimate to their clients on the number of bedrooms, bathrooms, sqft of the house and many more based on their budget to buy a house.

# DATA DICTIONARY

This data set contains a set of **551 records** under **7 attributes**.

The attributes being as follows:

|  |  |
| --- | --- |
| **Attribute Name** | **Description** |
| PRICE | The price at which the house is being sold. |
| SQFT | The total number of square feet of the house. |
| BEDS | The number of bedrooms in a particular house. |
| BATH | The number of bathrooms in a particular house. |
| AGE | Describes how old the house is. |
| STORIES | Tells us the number of floors or stories the house contains. |
| VACANT | Tells us if that particular house is vacant or occupied.  “0” represents it is vacant.  “1” represents it is occupied. |

# HYPOTHESIS GENERATION WITH RESPECT TO PROBLEM STATEMENT

**NULL HYPOTHESIS (Ho)**:

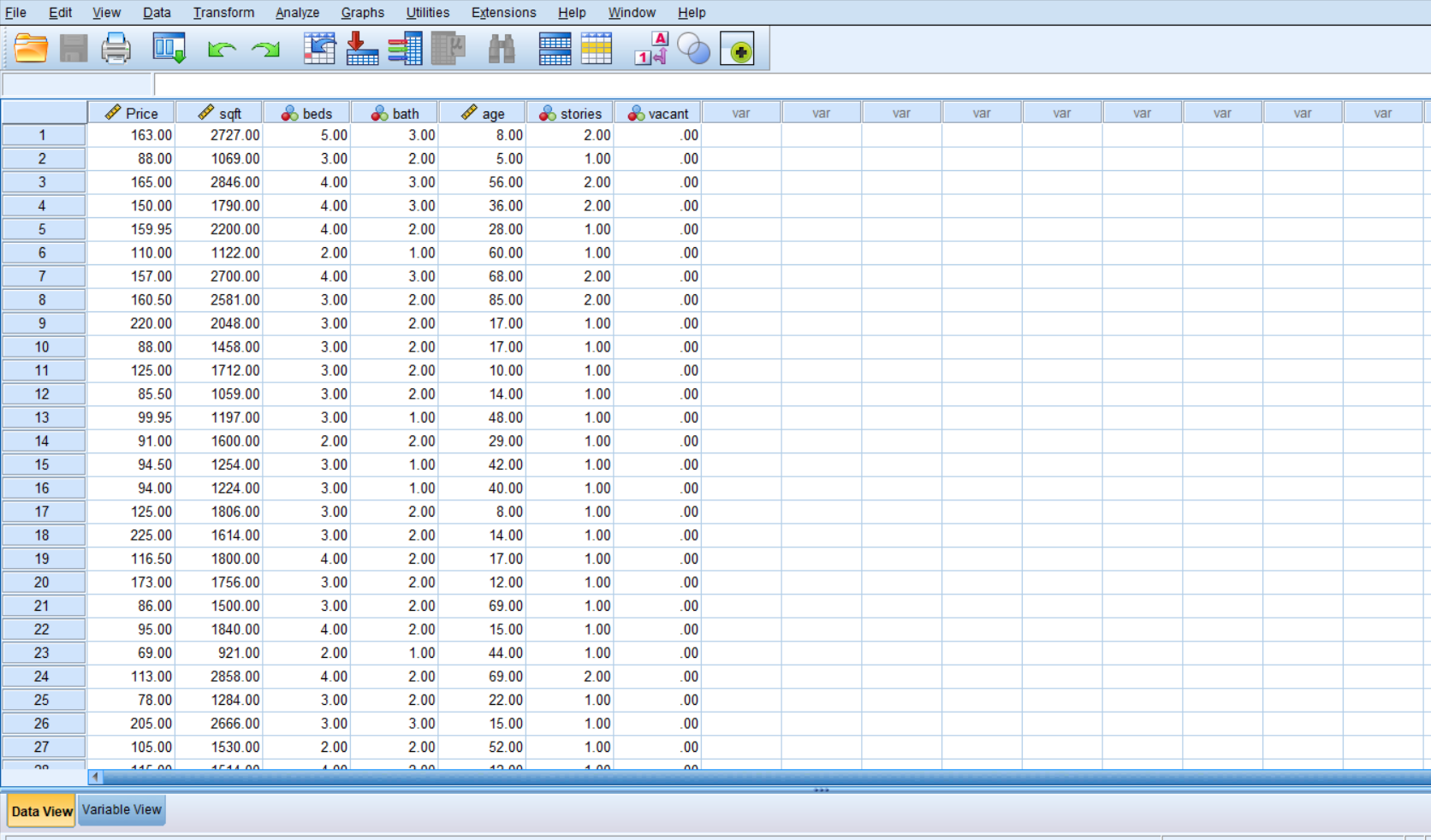
The price of the houses cannot be estimated by the parameters such as number of bedrooms, number of bathrooms, the square feet of the house, number of stories in a house and also on how old the house is.

**ALTERNATE HYPOTHESIS (Ha):**

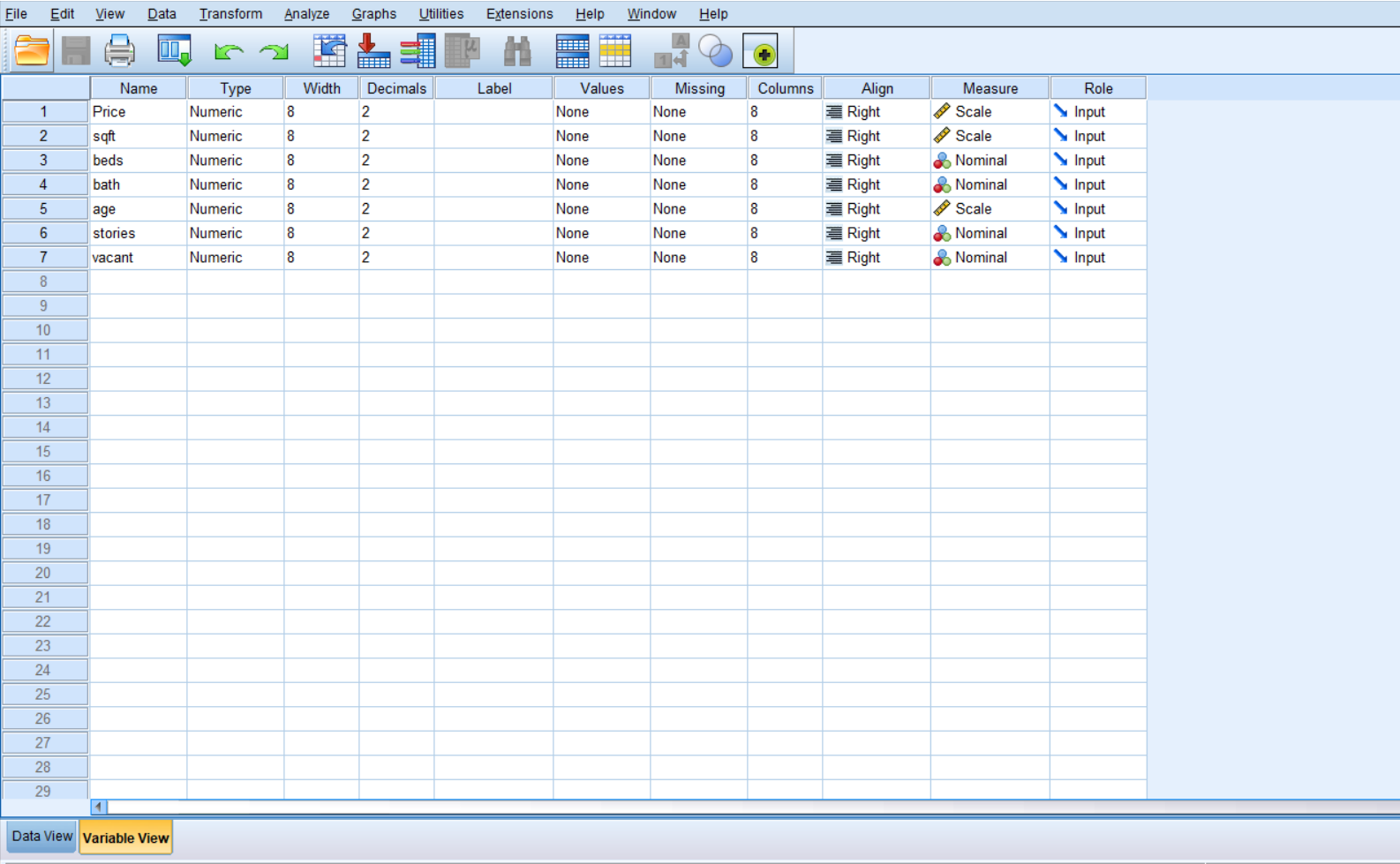
The price of the houses can be estimated by the parameters such as number of bedrooms, number of bathrooms, the square feet of the house, number of stories in a house and also on how old the house is.

# IMPORTING THE DATASET INTO SPSS

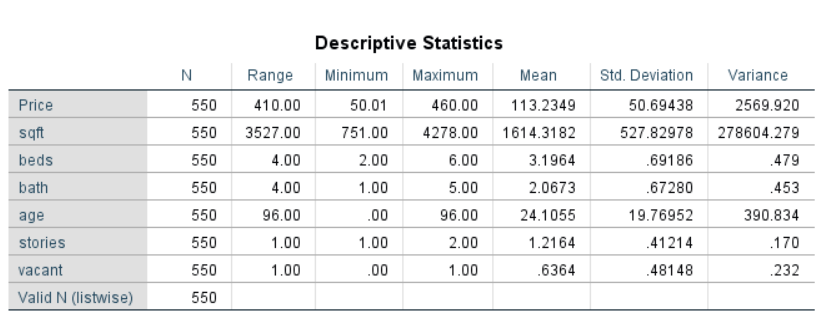
## DATA VIEW



## VARIABLE VIEW



# DESCRIPTIVE STATISTICS



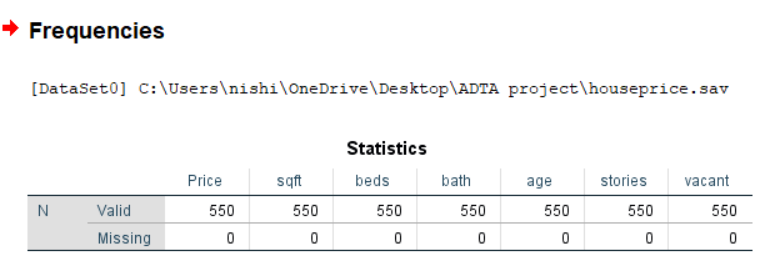
# 

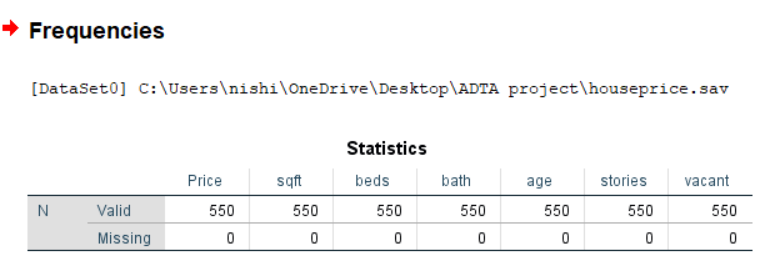
# CHECKING THE MISSING VALUES IN THE DATASET

Cleaning is one of the important steps before I run any mode. It is a step of data preparation. The cleaner data I have, the more accurate the model can predict the data. There are many ways to clean the data like making use of mean, median or mode if the missing variable is quantitative. Also, I can predict the missing values sometimes using the existing data points.

All in all, one must understand the domain of the business case and then accordingly perform the missing value imputation.

But luckily, this dataset does not contain any missing values, and is perfect to be used for analysis.



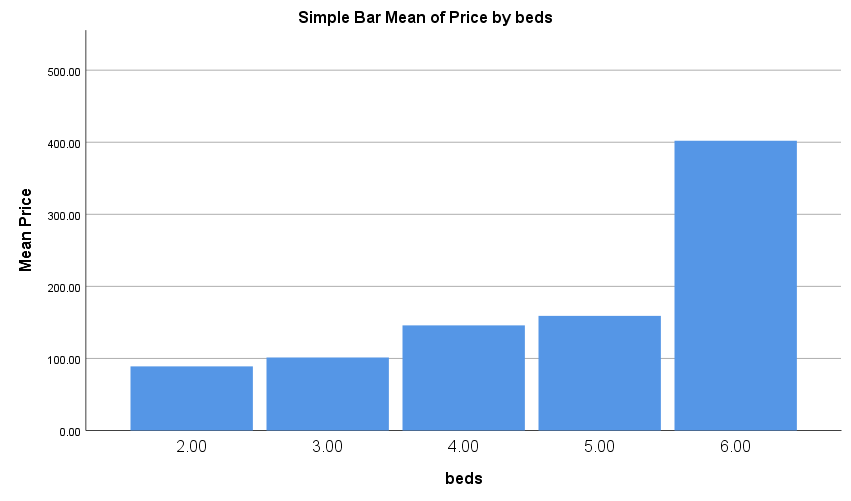


# EXPLORATORY DATA ANALYSIS (EDA)

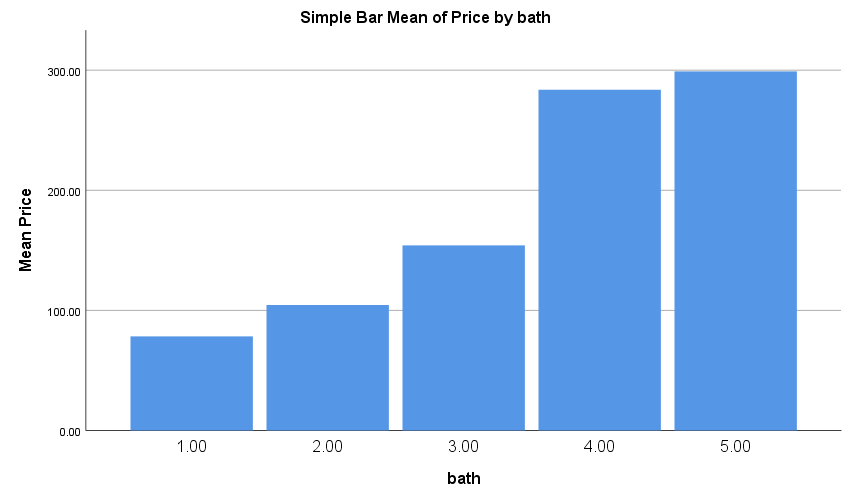
Now, I are conducting exploratory data analysis to figure out in what ways the data sets are similar or different. I used bar graphs to analyze them better.

Generally speaking, as the number of rooms in a room increases, then even the price of the property also increases.

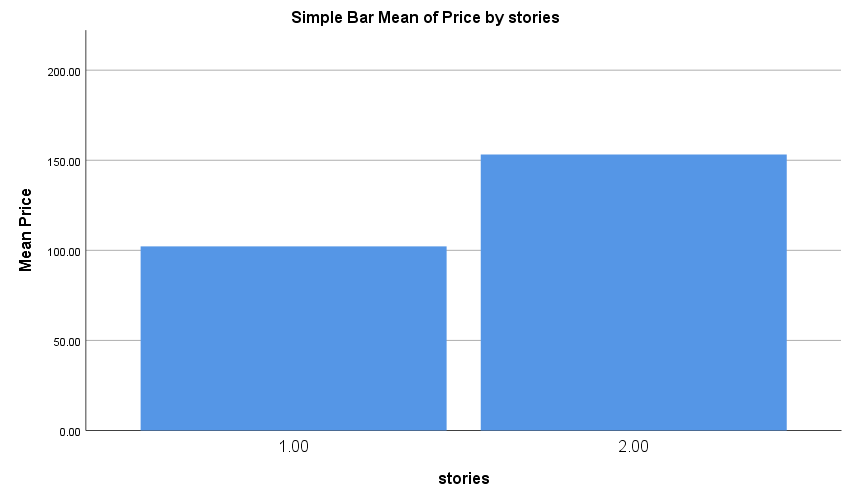
Attached below are some bar graphs which represents the prices of the property based on number of bedrooms, number of bathrooms, and also number of floors.



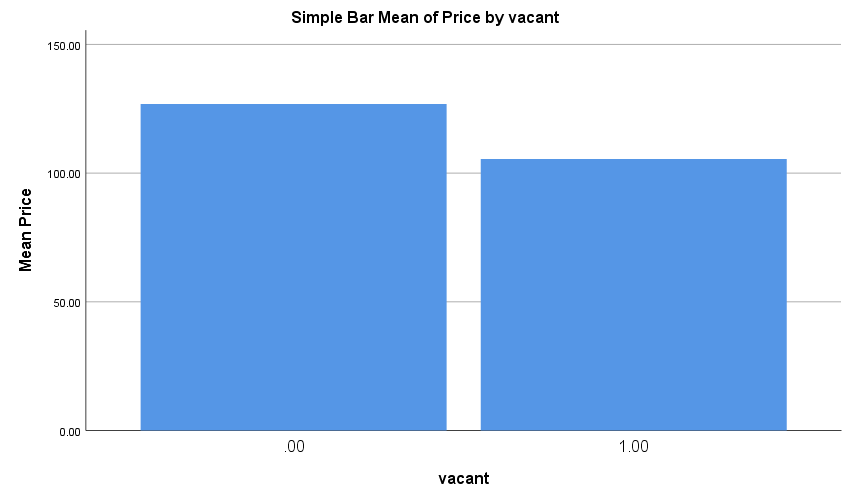
* From this graph, I could observe, as the number of bedrooms in a house increase, the price of that property also increases.



* It is the same even in this case, more the number of bathrooms, more is the price of the property.

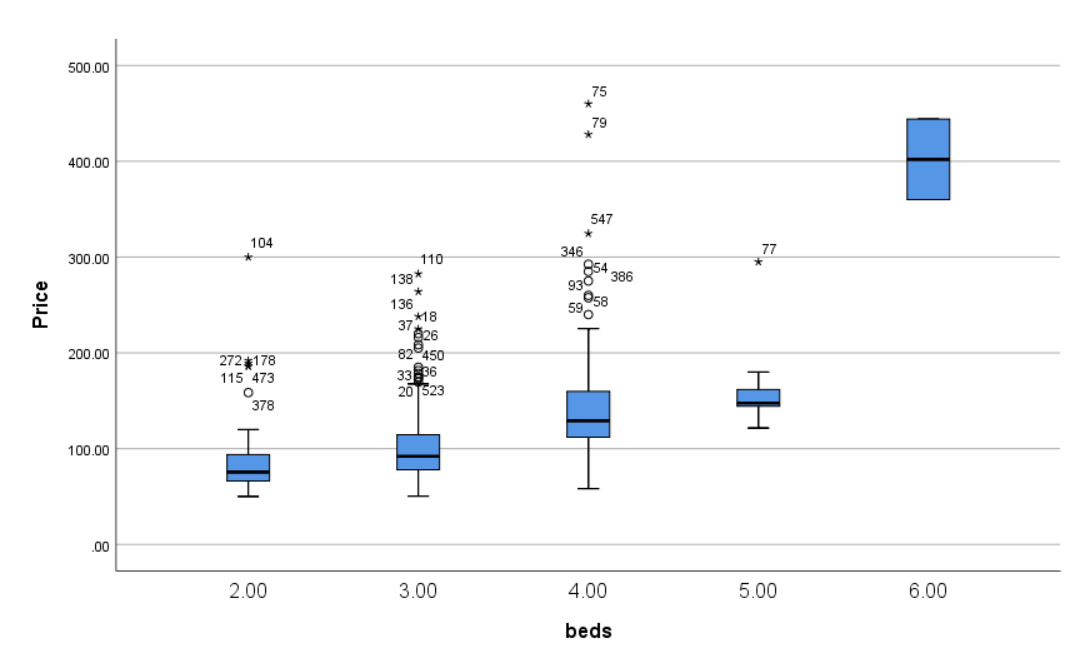


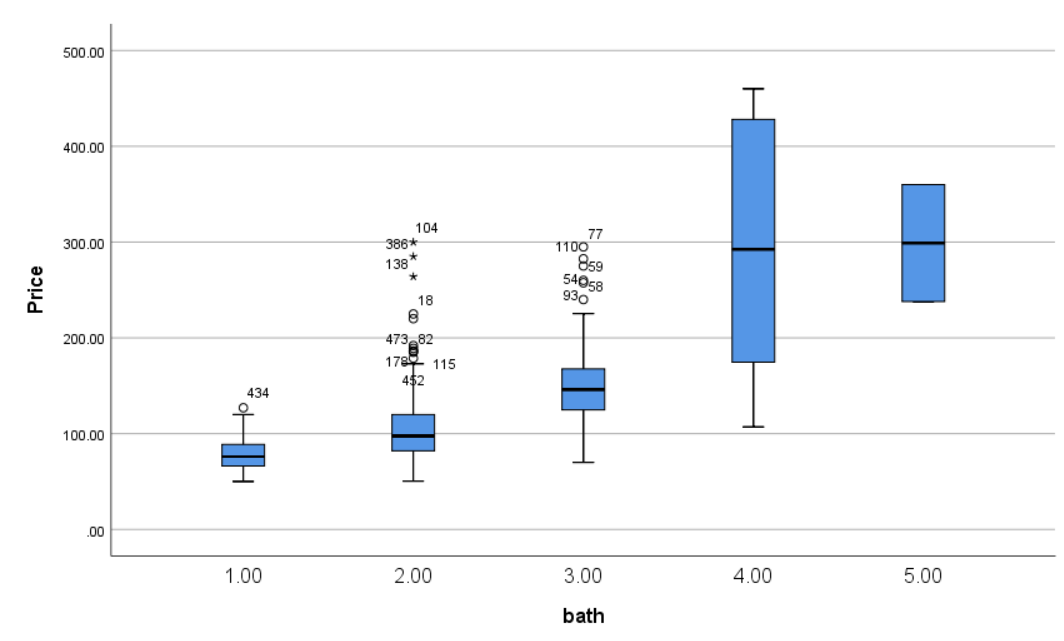
* The price of the property also depends on the number of stories it has. As the number of stories increases, so does the price of the property.

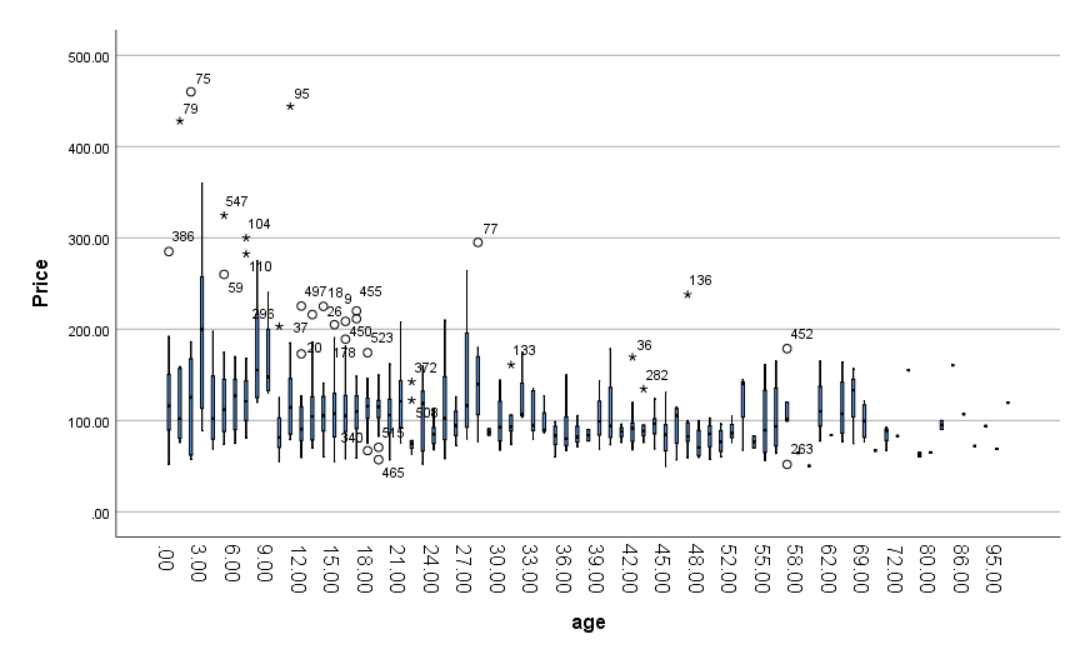


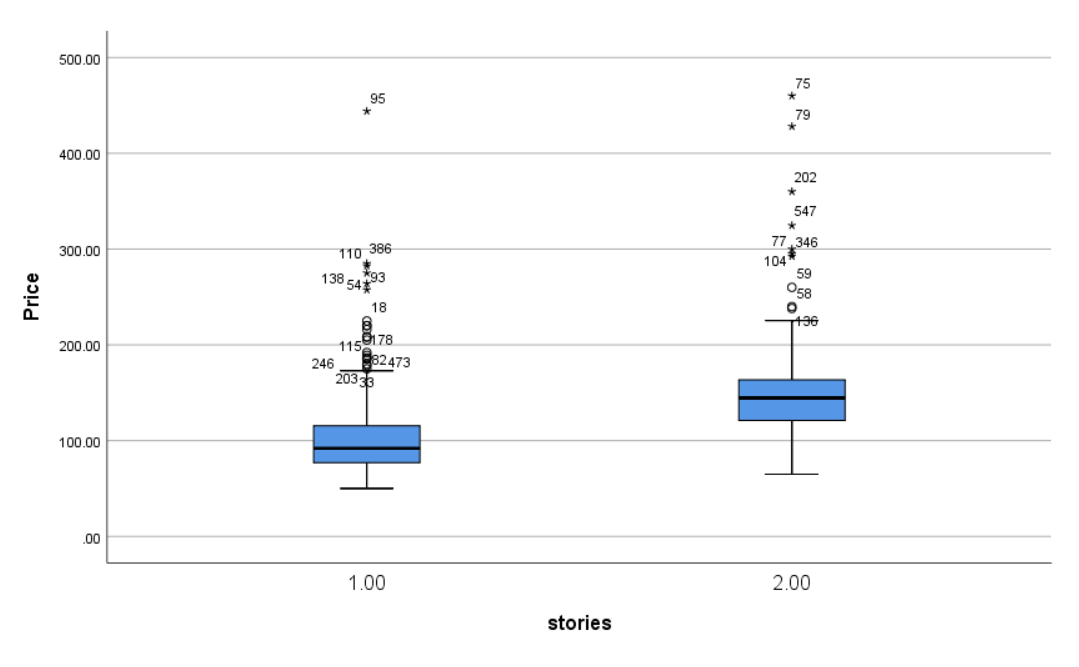
* “0” represents the property being vacant.
* “1” represents that they are occupied.
* I could notice that the vacant properties are actually costlier when compared to the occupied properties.

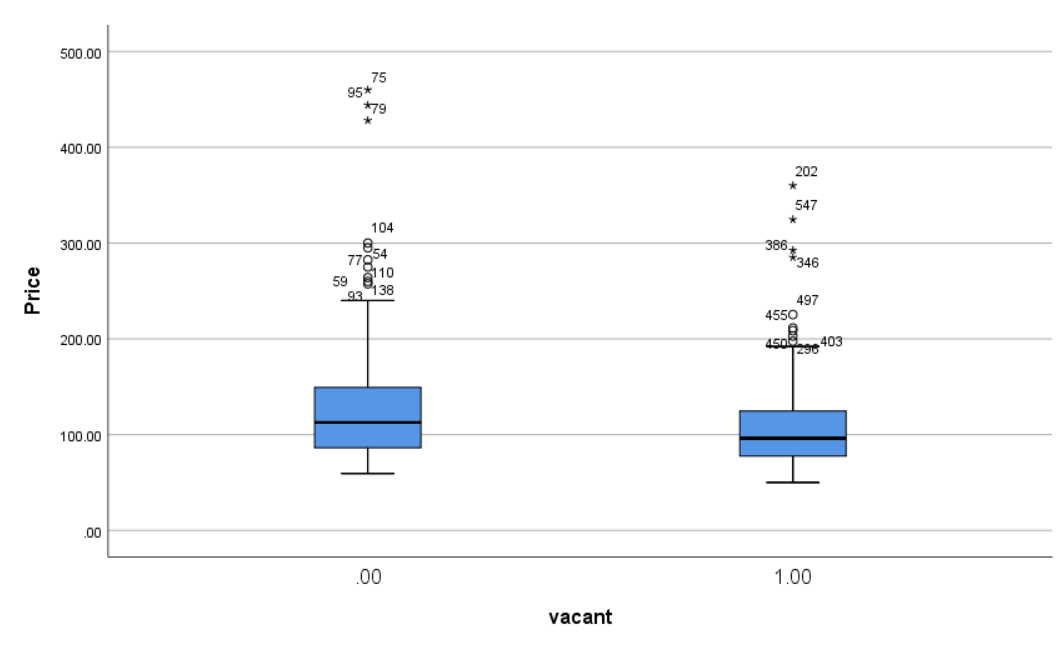
# CHECKING FOR OUTLIERS











* As you can observe there are a lot of outliers being observed in our data.
* Therefore, in this case I will be running two regression model.

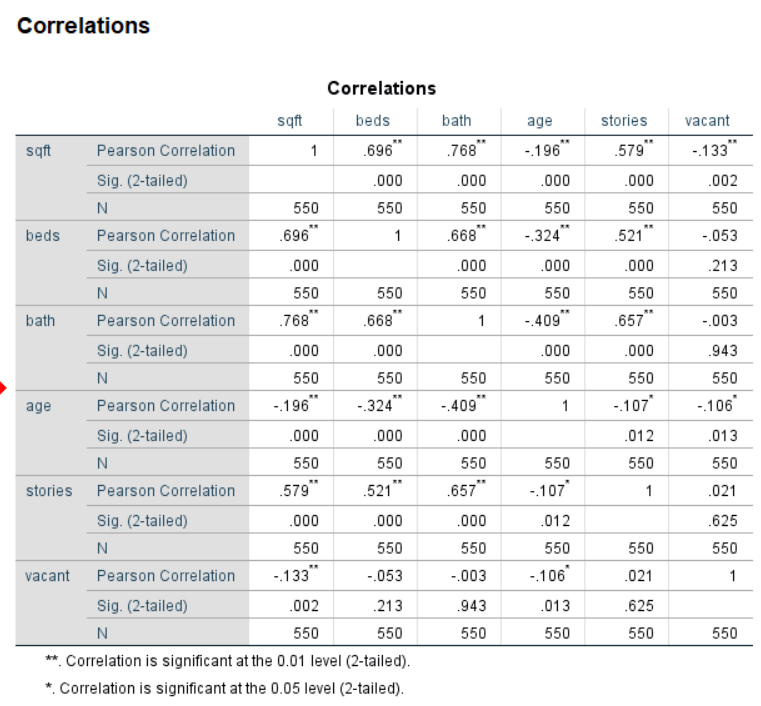
1. By removing the outliers.
2. By considering the outliers.

* And the best model to forecast the price of the house, will be based on the two regression models.

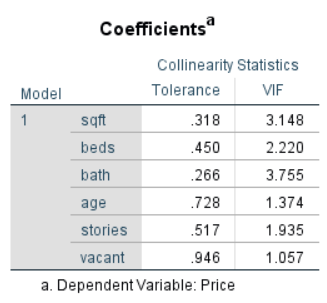
# CORRELATION

I would like to see if there are any variables which are highly correlated with each other.

Now, I are going to use the correlation to measure the strength of linear relationship betIen the quantitative variables.



# CHECKING FOR MULTI COLLINEARITY



* If for any of the independent variable, if the tolerance statistics value is less than 0.1 or if the VIF statistic value is greater than 10, then I can say multi collinearity exits betIen the variables.
* But, in this case the tolerance statistic is greater than 0.1 and the VIF statistic is less than 10 for all the independent variables.
* Therefore, I can say that there is no multi correlation betIen the variables.
* Therefore, all the independent variables can be used in our regression model.

# REGRESSION ANALYSIS -1(trimming the mean to remove outliers)

In this regression analysis, I are trimming of the mean to eliminate the outliers. I are using the Z score to verify the outliers and any observations with the Z score value more than 3 standard deviations away from the mean is being trimmed off.

I are using the stepwise regression to know exactly which independent variables could be significant to produce an analysis on the dependent variable.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Statistics** | | | | | | | | | |
|  | | Price | sqft | beds | bath | age | stories | vacant | log\_price |
| N | Valid | 528 | 528 | 528 | 528 | 528 | 528 | 528 | 528 |
| Missing | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

After eliminating the outliers, you can notice that the number of observations in the dataset has reduced to 528.

|  |  |  |  |
| --- | --- | --- | --- |
| **Descriptive Statistics** | | | |
|  | Mean | Std. Deviation | N |
| Price | 108.3158 | 38.09790 | 528 |
| sqft | 1570.6610 | 463.40526 | 528 |
| beds | 3.1723 | .65387 | 528 |
| bath | 2.0265 | .62160 | 528 |
| age | 23.5606 | 18.28896 | 528 |
| stories | 1.1932 | .39517 | 528 |
| vacant | .6439 | .47929 | 528 |

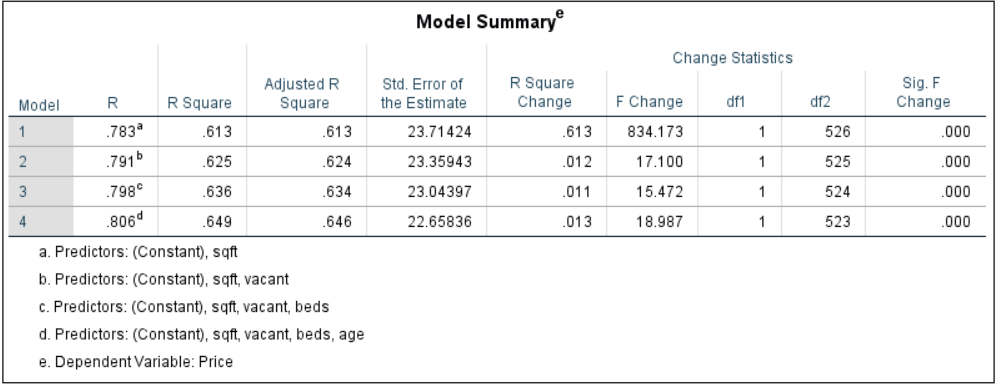
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | | | | | | |
| **Correlations** | | Price | sqft | beds | bath | age | stories | vacant |
| Pearson Correlation | Price | 1.000 | .783 | .450 | .594 | -.231 | .409 | -.201 |
| sqft | .783 | 1.000 | .676 | .737 | -.211 | .559 | -.117 |
| beds | .450 | .676 | 1.000 | .638 | -.322 | .532 | -.058 |
| bath | .594 | .737 | .638 | 1.000 | -.455 | .651 | .013 |
| age | -.231 | -.211 | -.322 | -.455 | 1.000 | -.166 | -.109 |
| stories | .409 | .559 | .532 | .651 | -.166 | 1.000 | .033 |
| vacant | -.201 | -.117 | -.058 | .013 | -.109 | .033 | 1.000 |
| Sig. (1-tailed) | Price | . | .000 | .000 | .000 | .000 | .000 | .000 |
| sqft | .000 | . | .000 | .000 | .000 | .000 | .004 |
| beds | .000 | .000 | . | .000 | .000 | .000 | .091 |
| bath | .000 | .000 | .000 | . | .000 | .000 | .386 |
| age | .000 | .000 | .000 | .000 | . | .000 | .006 |
| stories | .000 | .000 | .000 | .000 | .000 | . | .223 |
| vacant | .000 | .004 | .091 | .386 | .006 | .223 | . |
| N | Price | 528 | 528 | 528 | 528 | 528 | 528 | 528 |
| sqft | 528 | 528 | 528 | 528 | 528 | 528 | 528 |
| beds | 528 | 528 | 528 | 528 | 528 | 528 | 528 |
| bath | 528 | 528 | 528 | 528 | 528 | 528 | 528 |
| age | 528 | 528 | 528 | 528 | 528 | 528 | 528 |
| stories | 528 | 528 | 528 | 528 | 528 | 528 | 528 |
| vacant | 528 | 528 | 528 | 528 | 528 | 528 | 528 |

From the correlation table, you can notice that no two variables are correlated.



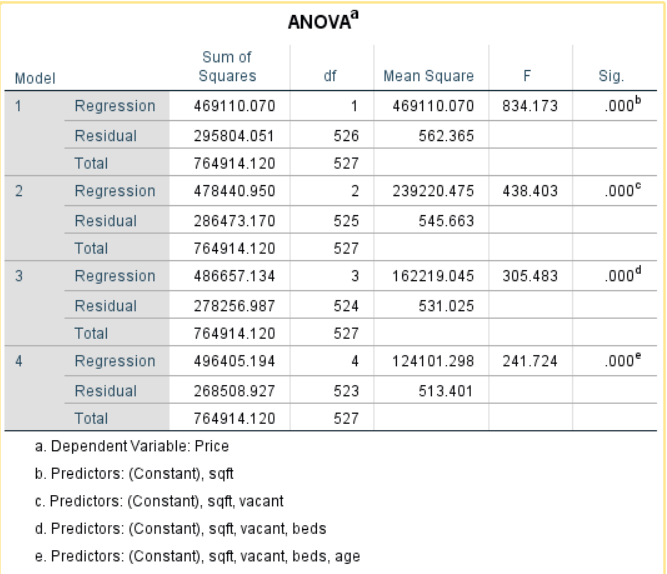


* Therefore, you can notice that sqft, vacant, beds and age alone are enough to significantly predict the price of the property.



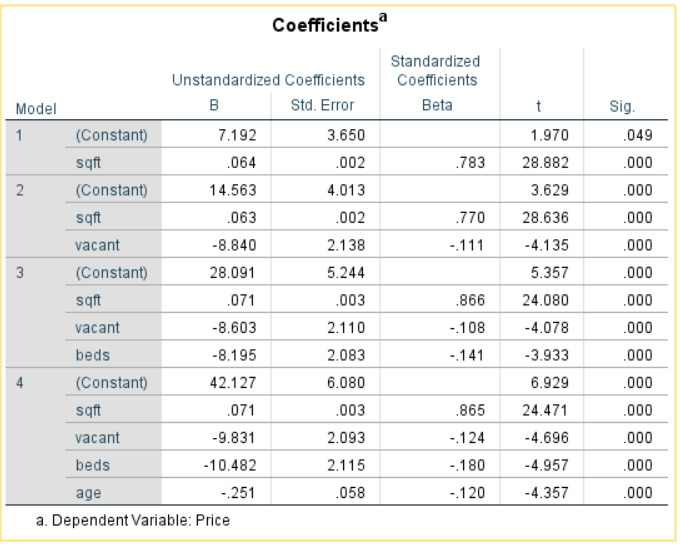


* Additionally, you can notice that the R-Square value to be 0.649 i.e., 64%. Which is not bad.
* Therefore, I can say that the variance explained in this model is 64%.





* Explained variation = 496405.194.
* Unexplained variation = 268508.927.
* Total variation = 764914.120.





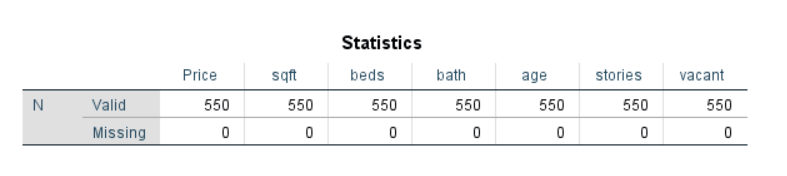
* You can notice that the p-values (sig.) for all my independent variables is less than 0.05.
* Therefore, I can say that all my independent variables are significant.
* Therefore, my regression equation is

**Price of the house = 0.071(sqft) – 9.831(vacant) -10.482(beds) -0.251(age).**

# REGRESSION ANALYSIS – 2 (RUNNING THE MODEL WITH OUTLIER CONSIDERED)

In this analysis I will be running a regression model with all the outlier included in the model.

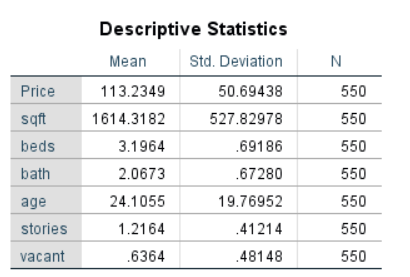
By running this regression model with all outliers, I want to check if the fit of the model is affected or not.

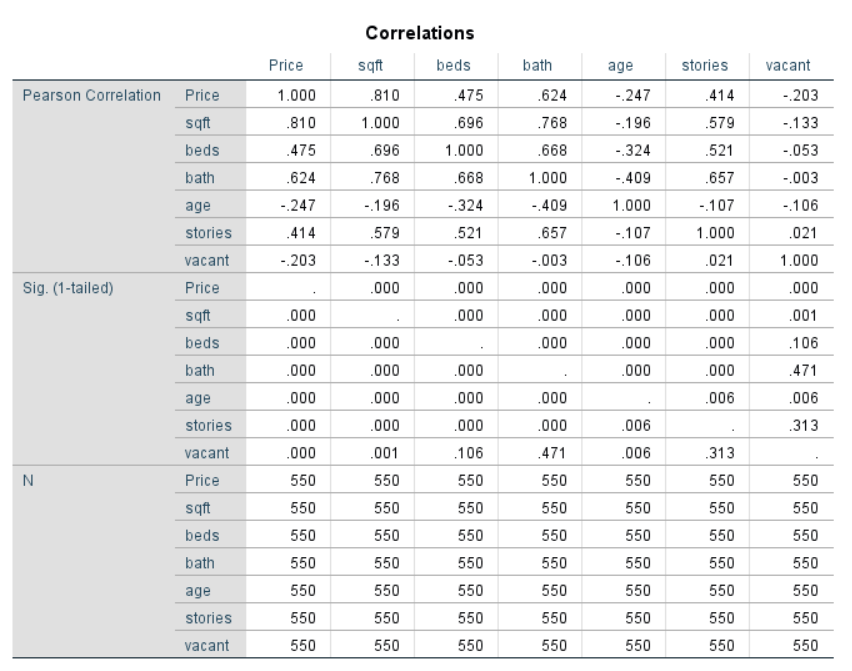


In this case, you can see that I are dealing with all 550 observations in the model.

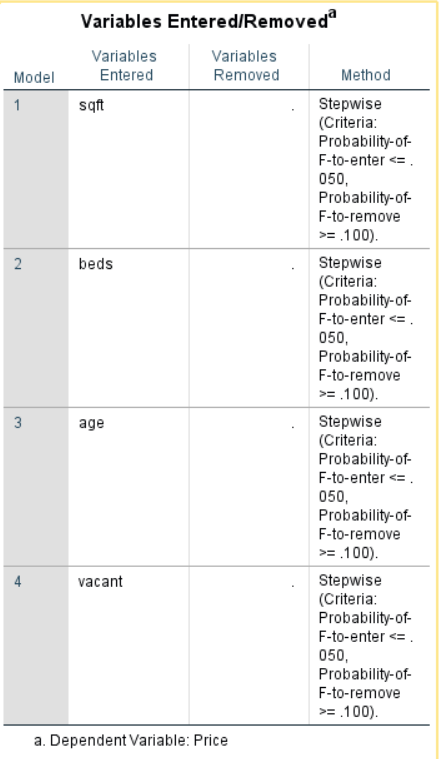
And also, no observations are missing.

**RUNNING A STEPWISE REGRESSION ANALYSIS**



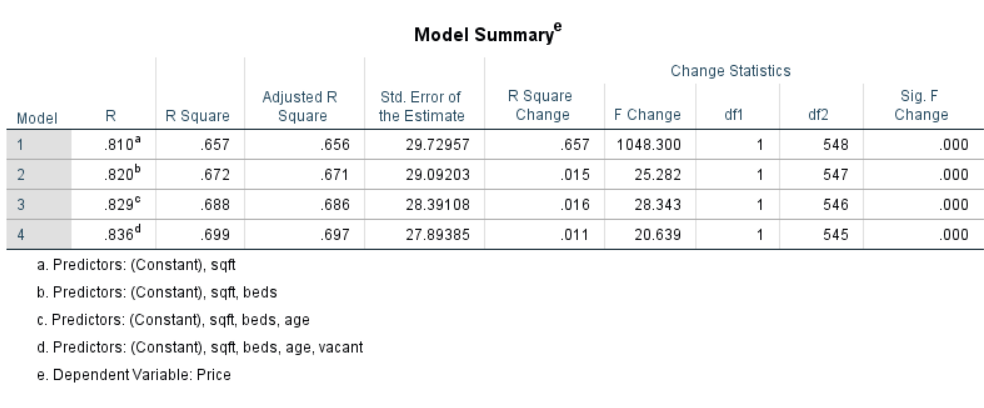


* From the Pearson correlation table, I can observe that no two variables are highly correlated.
* Therefore, I can use all the variables in our regression analysis.



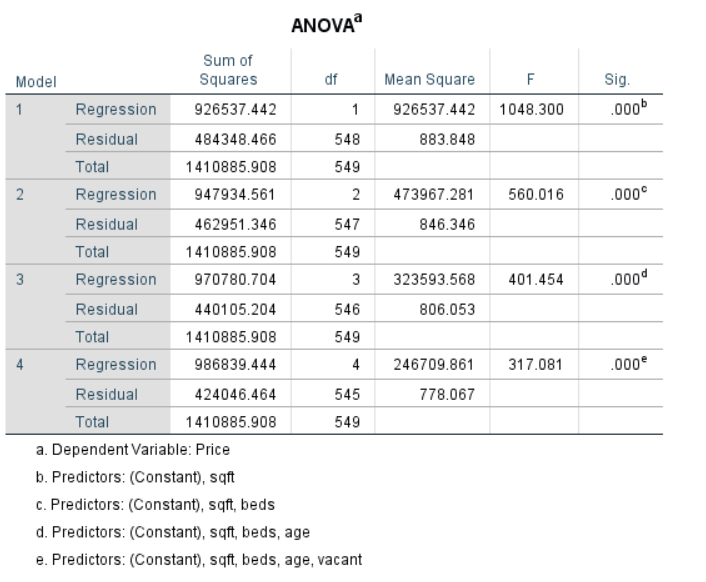


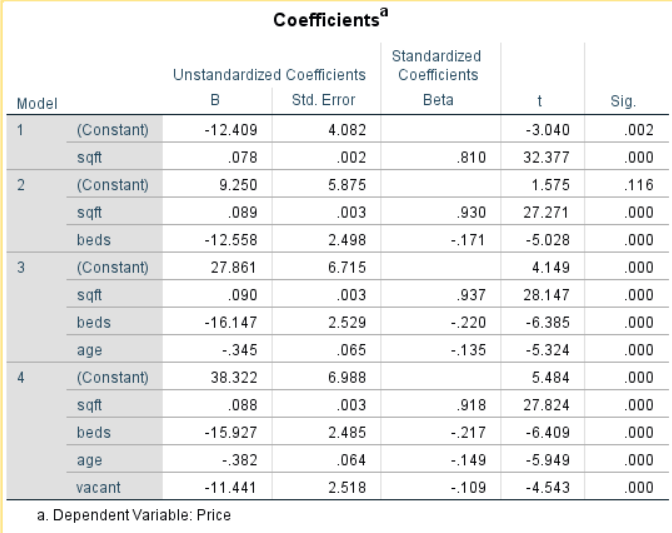
* Therefore, you can notice that sqft, vacant, beds and age alone are enough to significantly predict the price of the property.





* The R-square has significantly increased to 0.699 i.e., 69%, due to the considerations of the outliers.
* Therefore, I could say that the variability explained in this model 69%, which is much better than the variability explained by the model 1.







* You can notice that the p-values (sig.) for all my independent variables is less than 0.05.
* Therefore, I can say that all my independent variables are significant.
* Therefore, my regression equation is

**Price of the house = 0.088(sqft) – 15.927(beds) – 0.382(age) – 11.441(vacant).**

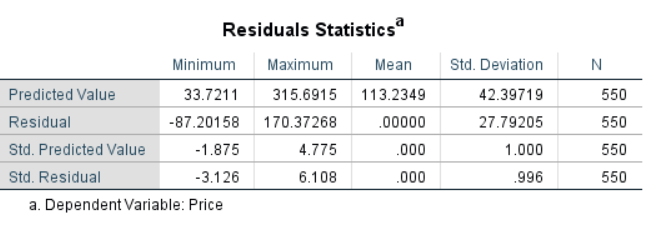
# WHICH MODEL TO USE

Since the R-square value of regression analysis 2 is higher when compared to the regression analysis 1, and also since all the variables are significant, I are considering the regression analysis 2 model to make analysis on the price of the house.

# RESUDIAL ANALYSIS

* The residuals should have a mean of zero, should be normally distributed and should also follow constant variance assumption (Homoscedasticity).

## RESIDUAL STATISTICS

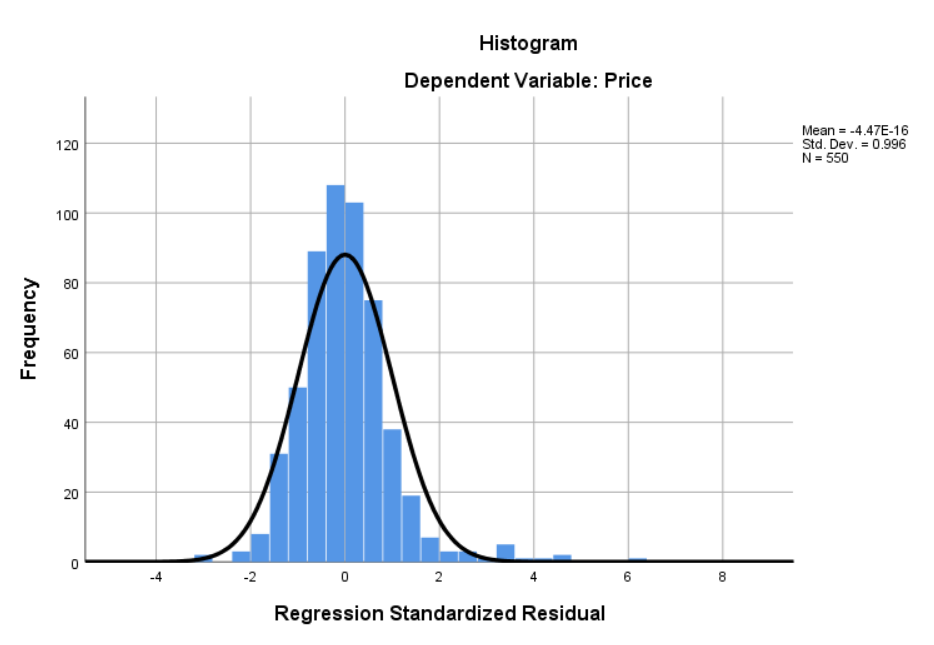


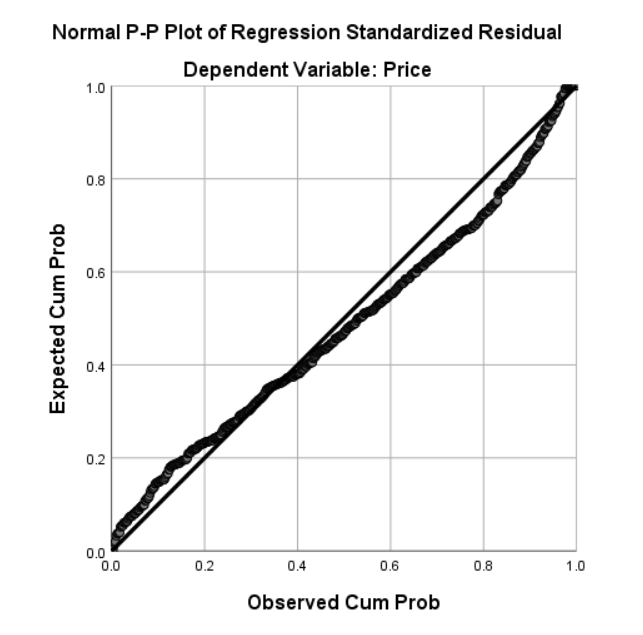


As you can see, the mean of my residual is zero.

Therefore, the condition is satisfied.

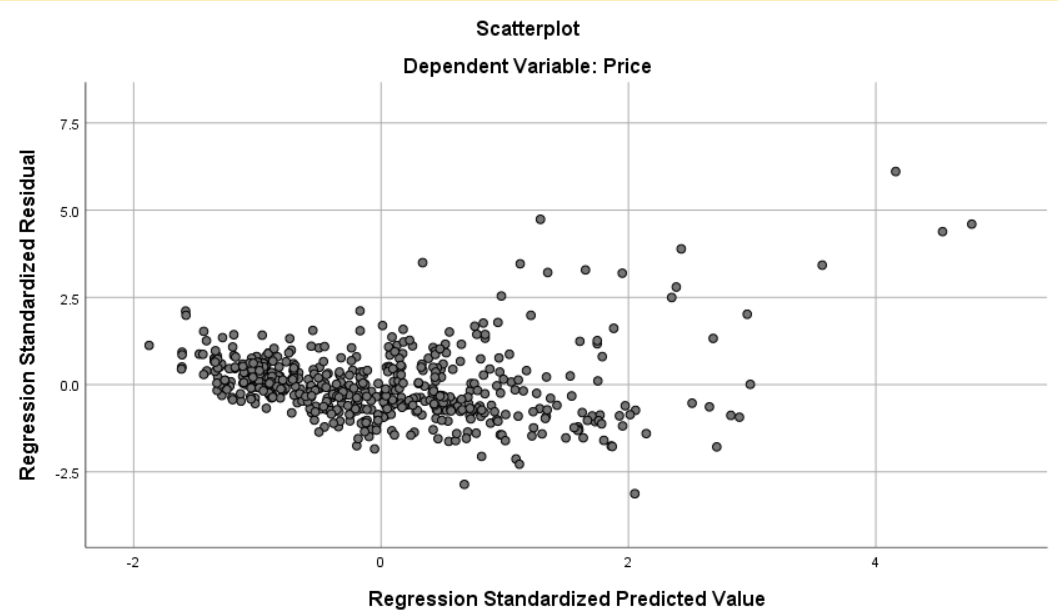
## CHECKING THE NORMALLY DISTRIBUTED ASSUMPTION





* As can be observed from the histogram and also the P-P plot, the residuals are not normally distributed.
* Therefore, it violates the assumption.

## CHECKING FOR CONSTANT VARIANCE (HOMOSCEDASTICITY)



From this scatter plot you can actually observe that the constant variance assumption is being violated.

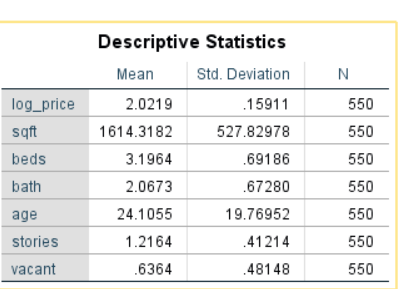
Therefore, I are going you transform our target variable (price), to make it satisfy all the residual assumptions.

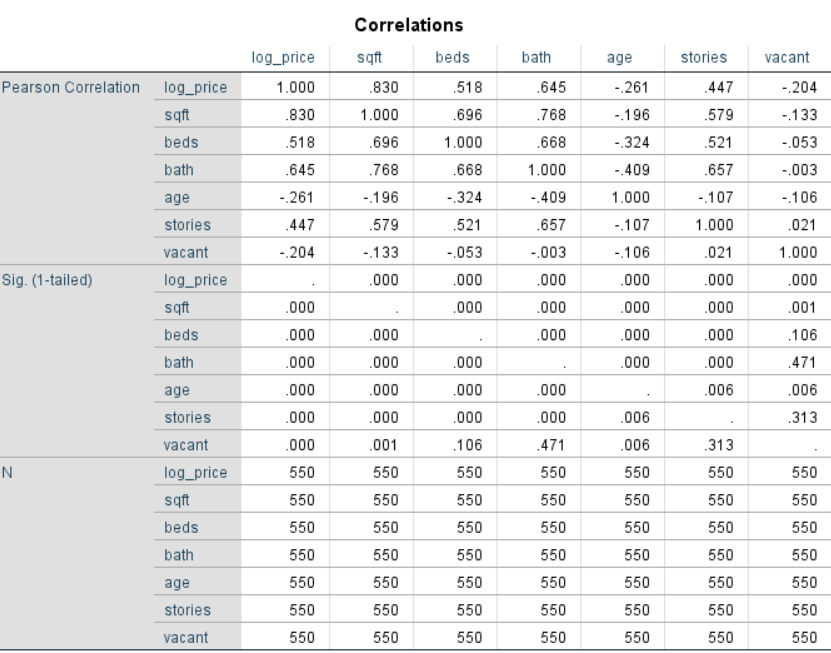
Therefore, I are going to use LOG transformation.

# LOG TRANSFORMING MY TARGET VARIABLE AND RE-RUNNING THE REGRESSION ANALYSIS

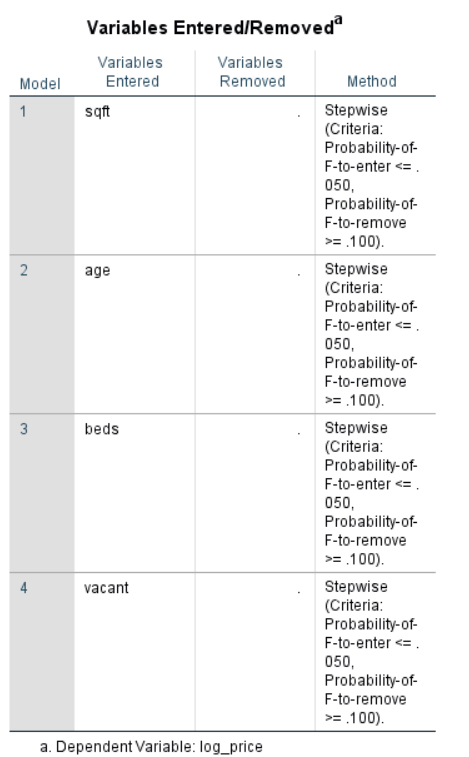
My target variable has changed to Log\_Price, after log transformation.

## REGRESSION ANALYSIS WITH THE TRANSFORMED VARIABLE



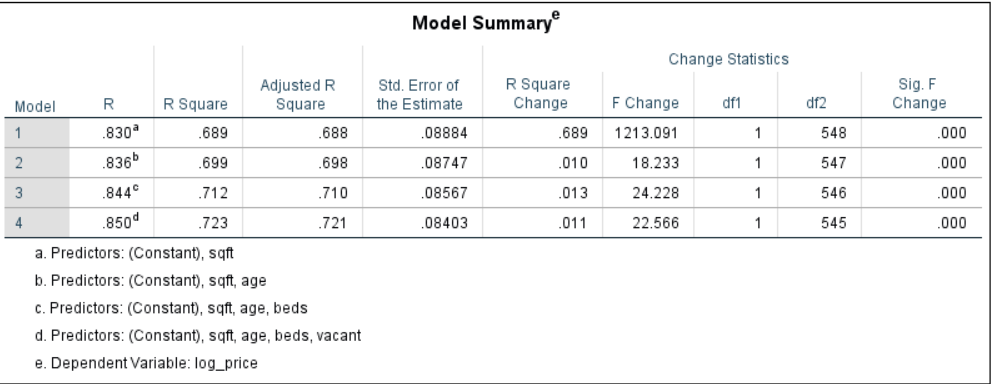


* There appears to not be any strong correlation betIen the variables.
* Therefore, all the variables can be used for the regression analysis.



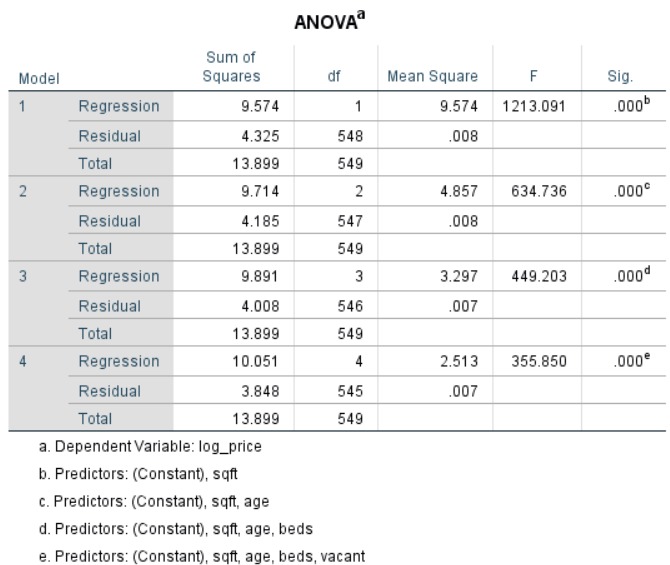


* After running the stepwise regression analysis with the transformed target variable, the variables sqft, age, beds, and vacant are said to be significant with respect to the target variable price.





* The R-square value has also increased from 0.699 to 0.723 after log transforming the target variable.
* Therefore, the amount of variance explained by this model is 0.723 i.e., 72%.

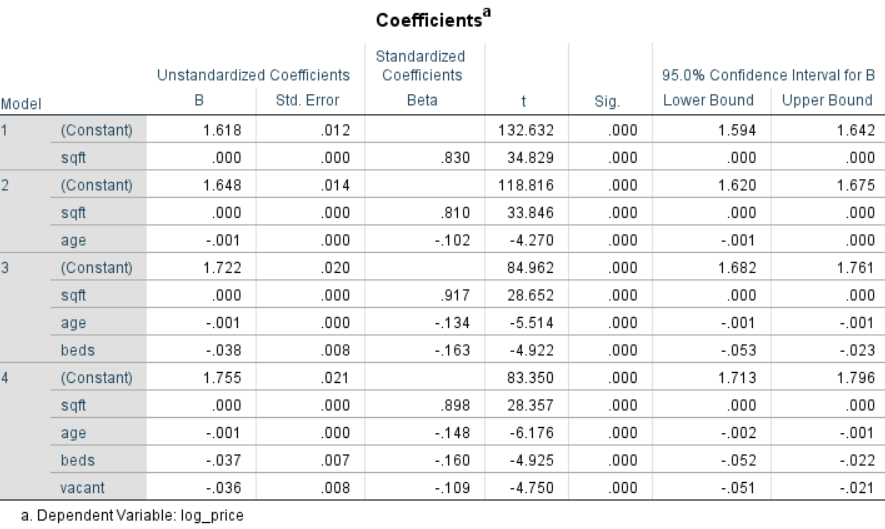




The explained variance through this regression model = 10.051

The un-explained variance for this regression model = 3.848

The total variance for the regression model = 13.899



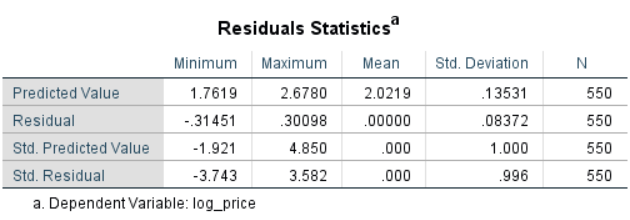


* I can even see that the P- values (Sig) for all the independent variables is less than the alp value (0.05).
* Therefore, all the independent variables i.e., sqft, age, beds, and vacant are all significant.
* Therefore, the regression model is.

**Price = sqft – 0.001(age) – 0.037(beds) – 0.036(vacant).**

# RESIDUAL ANALYSIS WITH THE TRANSFORMED VARIABLE

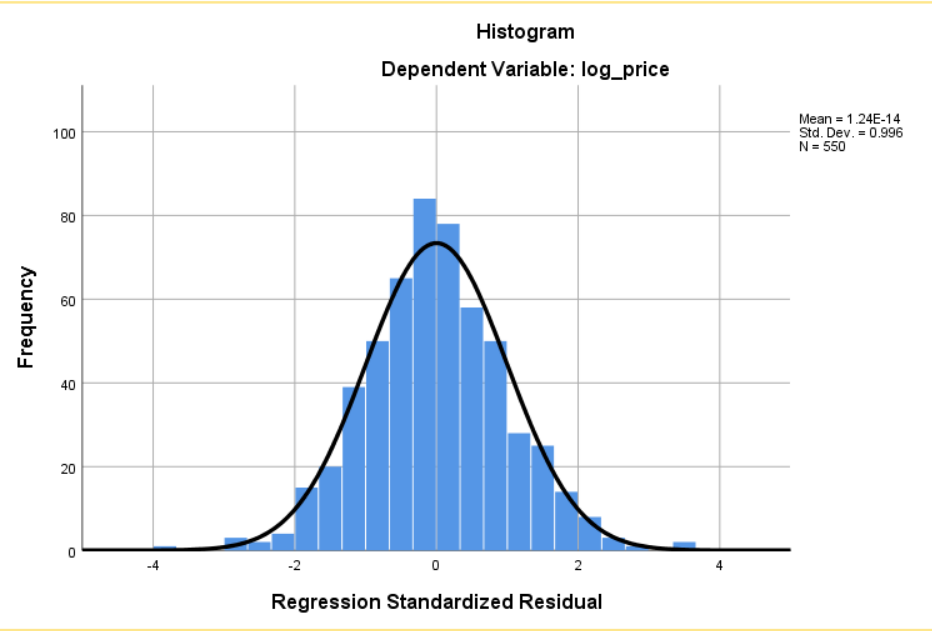
As I have mentioned before, the residuals should be normally distributed, with a mean of zero and should follow constant variance (Homoscedasticity).

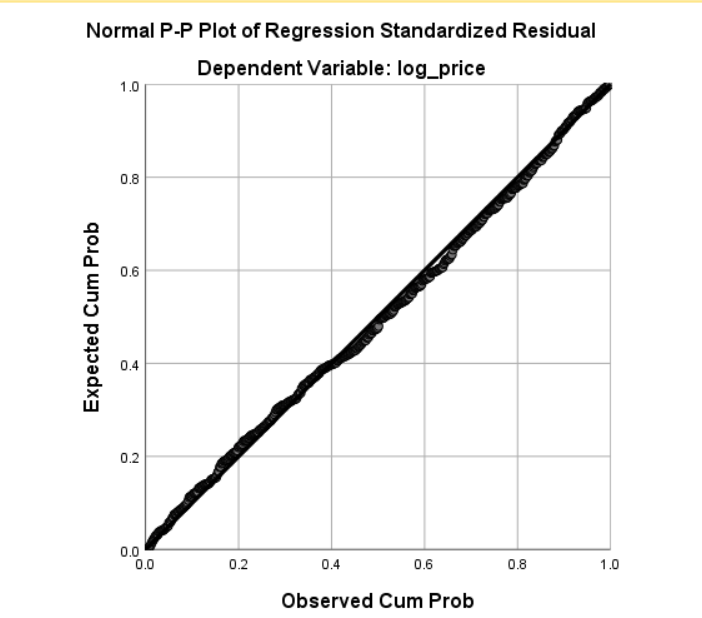




* The mean of the residuals is zero. Therefore, the condition is satisfied.

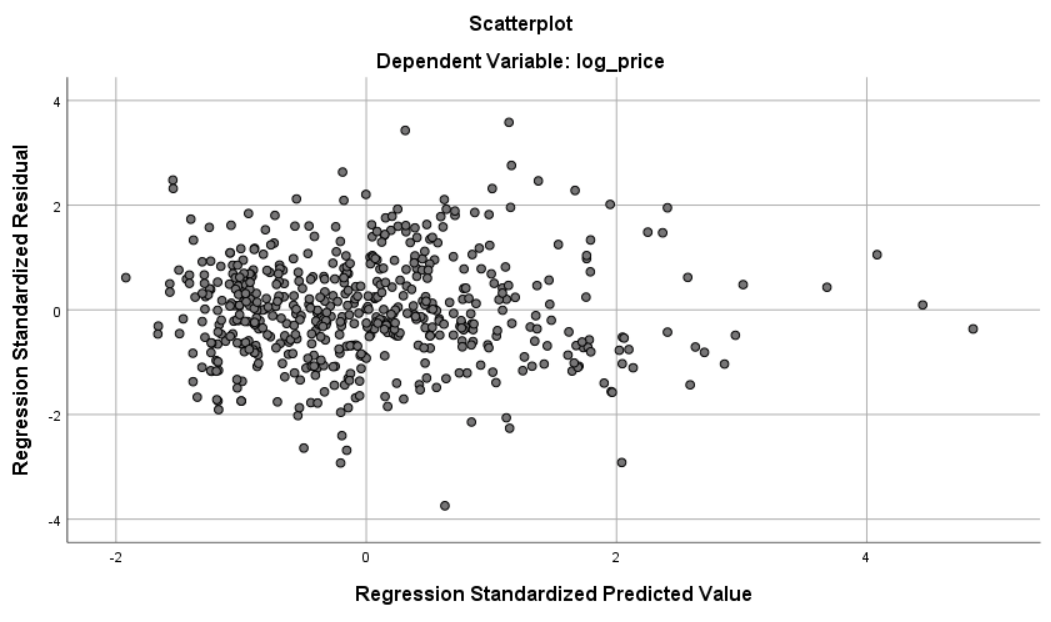
## CHECKING FOR NORMAL DISTRIBUTION





* As observed from the histogram and the normal P-P plot, you can observe that, it is normally distributed.
* Therefore, the normal distribution assumption is satisfied.

## CHECKING FOR CONSTANT VARIANCE(HOMOSCEDASTICITY)



* Therefore, from the regression standardized residual vs regression standardized predicted value scatter plot, it is clear that the data is satisfying the constant variance assumption.
* Therefore, the homoscedasticity condition is satisfied.
* Therefore, all the residual assumptions are satisfied.

# CONCLUSION TO THE PROBLEM STATEMENT

**Table

Description automatically generated**



Since the p-value(sig.) is less than alpha (0.05), I reject the null hypothesis.

Therefore, I can conclude saying that the price of the houses can be estimated by the parameters such as number of bedrooms, number of bathrooms, the square feet of the house, number of stories in a house and also on how old the house is.