**House Prediction Analysis**

**Data Summary:**

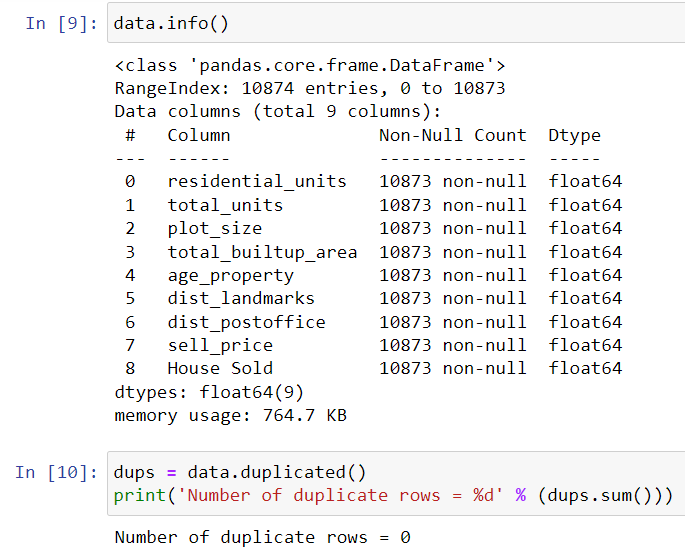
We have chosen a dataset from the MIT data directory to predict the sales of houses in New York.

The variables in the dataset are as follows:

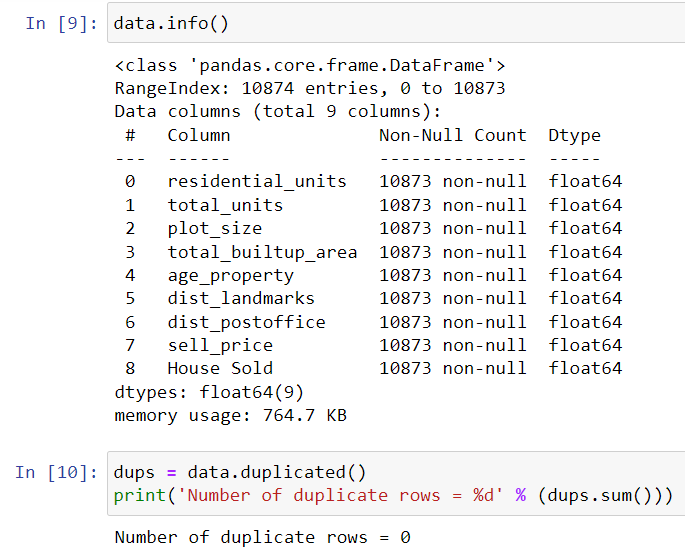
* **address:** address of the house
* **zip\_code:** postal code of the area where the house is located (numerical)
* **residential\_units:** residential unit of the area where the house is located (numerical)
* **total\_units:** No of unit houses built per area
* **plot\_size:** Plot size of the land of the House
* **total\_builtup\_area:** The total area in which the house is built is less than or equal to the plot size.
* **age\_property:** The age of the land on which the house is built.
* **dist\_landmarks:** The distance of the nearest landmark.
* **dist\_postoffice:** The distance of the nearest post office.
* **sell\_price:** The selling price of the house.
* **House\_Sold:** This is the indicator variable is whether the house is sold (1) or not (0).

**Preliminary Data Pre-processing and Cleaning:**

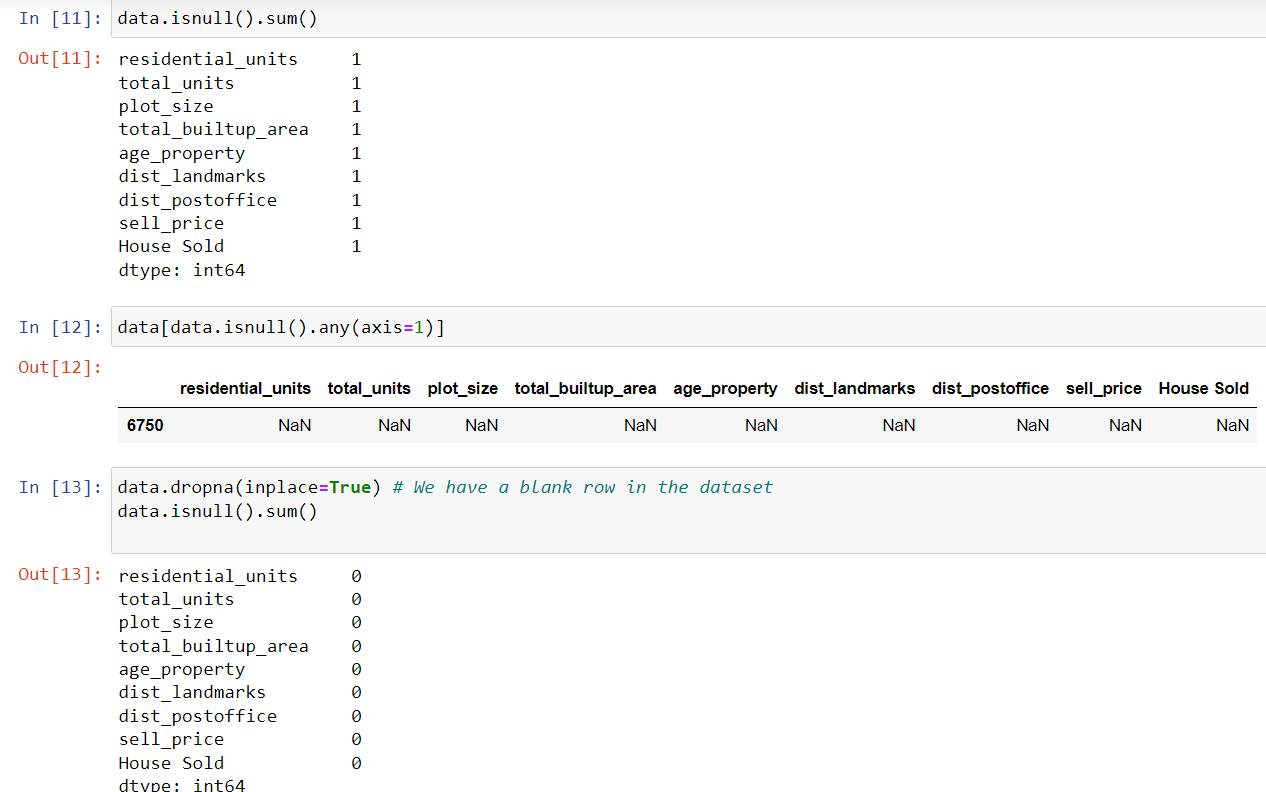
**To verify if the given dataset is practicable, we verified if the data contains any duplicate or null values as illustrated in the pictures below.**

****

**Checked:** *No Duplicate Rows in the Dataset*

****

**Checked:** *1 Null Value in the Data Set*

****

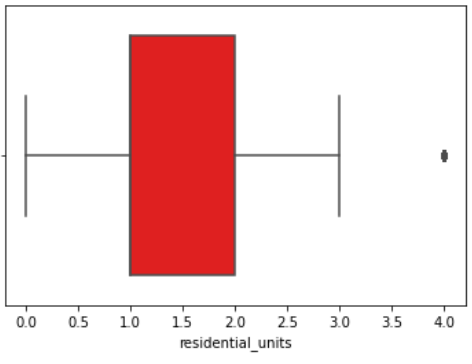
*Checking the row with null values, we get that row no. 6750 has null values in the entire row,*

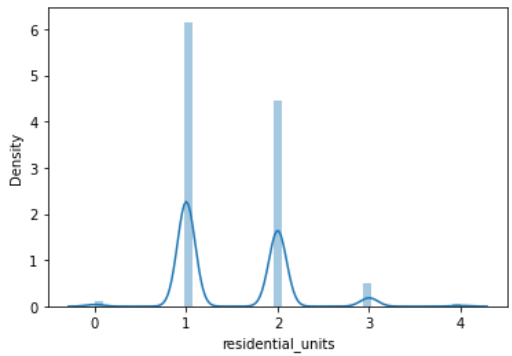
*Hence, we eliminate row no 6750.*

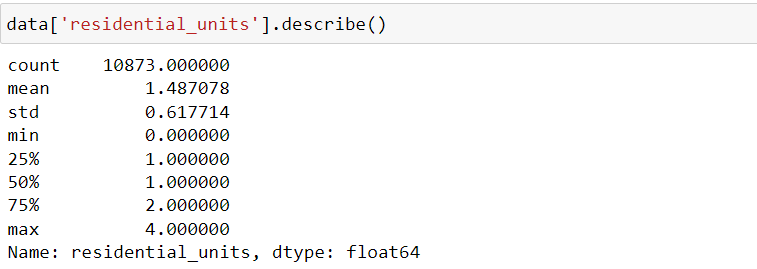
**Exploratory Data Analysis (Univariate)**

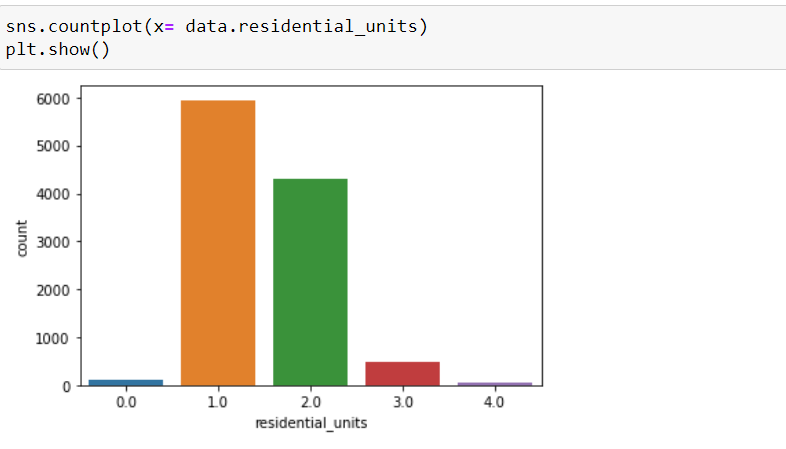
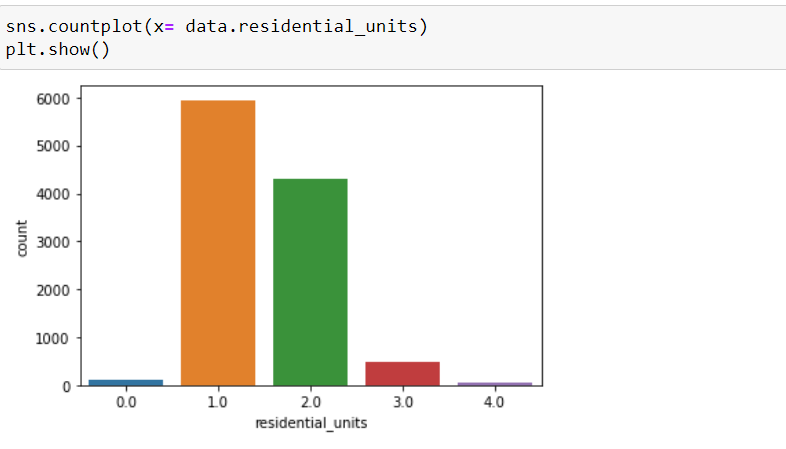
1. **Residential Units**

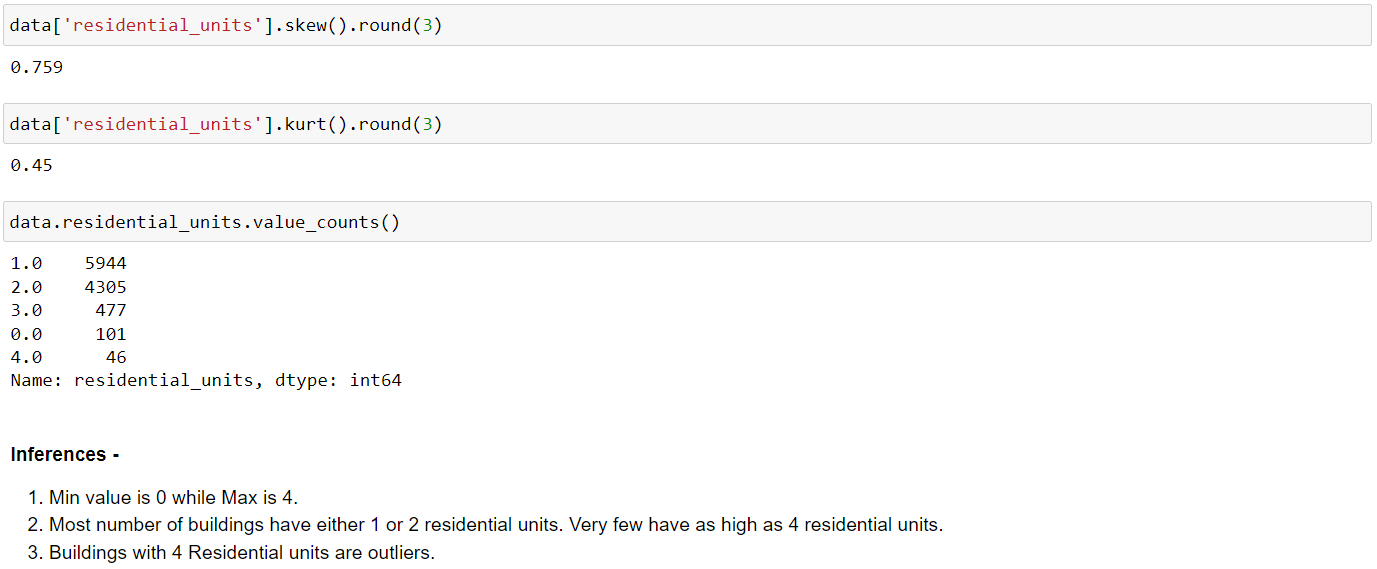


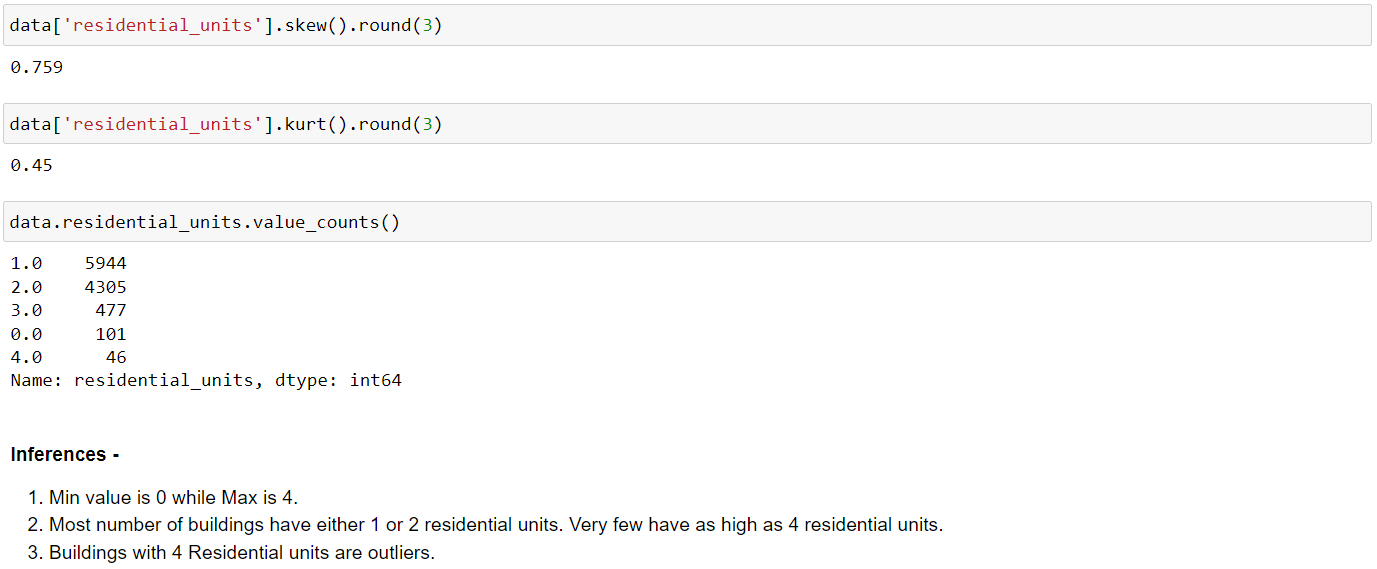








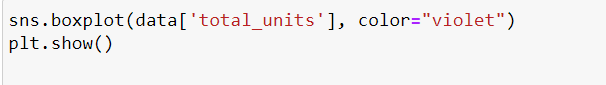
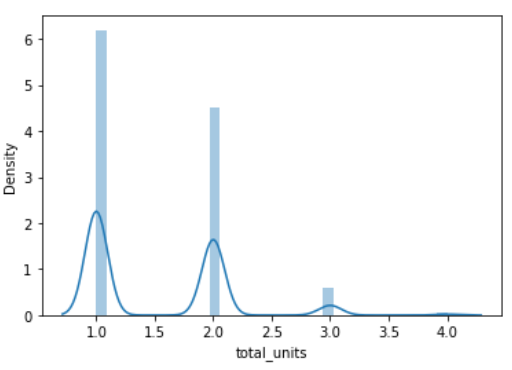


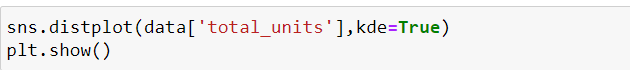


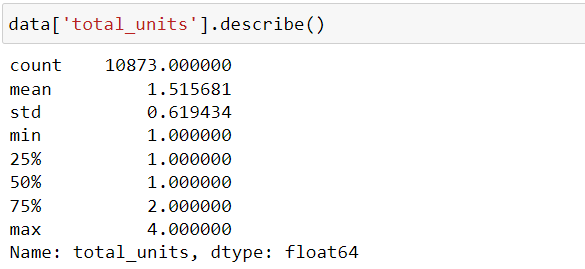
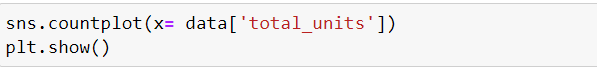
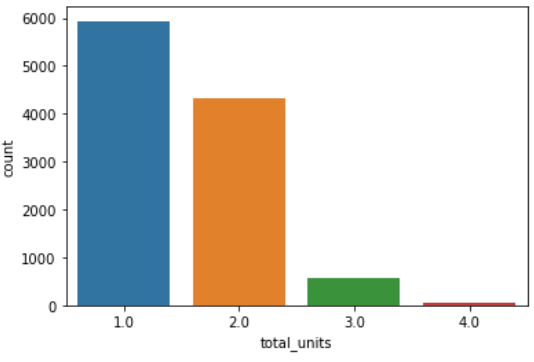
*Inferences:*

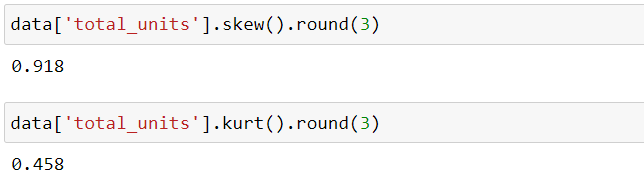
1. Minimum Value is 0, while Maximum Value is 4
2. Most buildings have either 1 or 2 residential units. Very few have as high as 4 residential units.
3. Buildings with 4 Residential Units are outliers
4. Residential Units are **Positively Skewed**, and **Leptokurtic**
5. **Total Units**



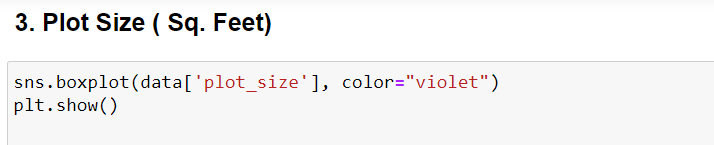
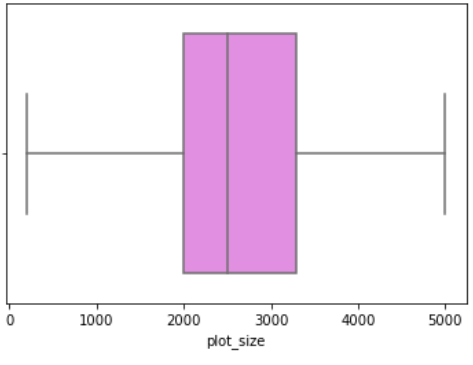


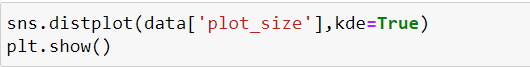
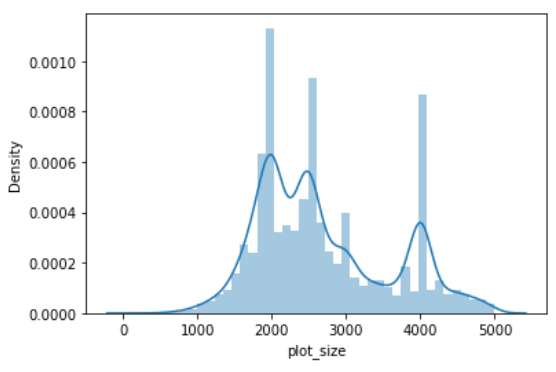
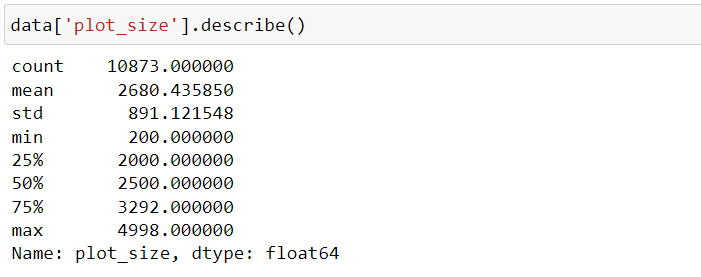


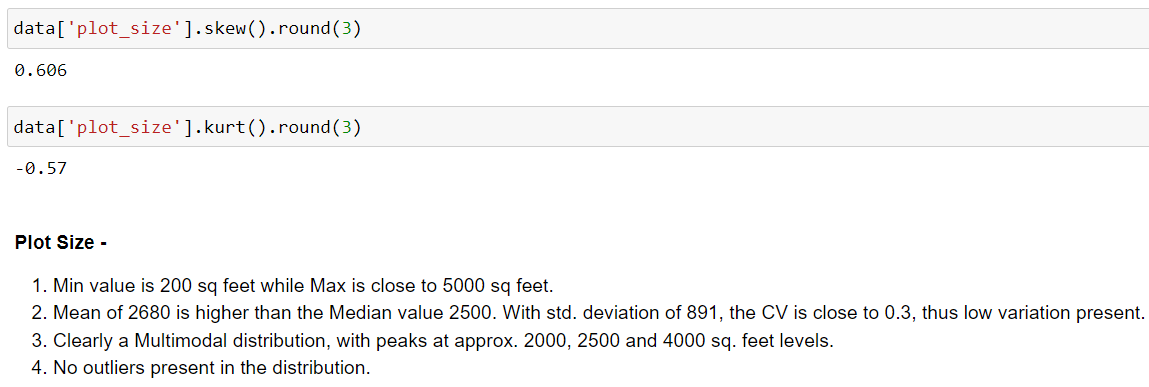


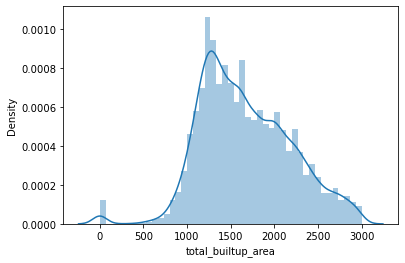
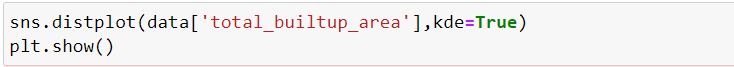
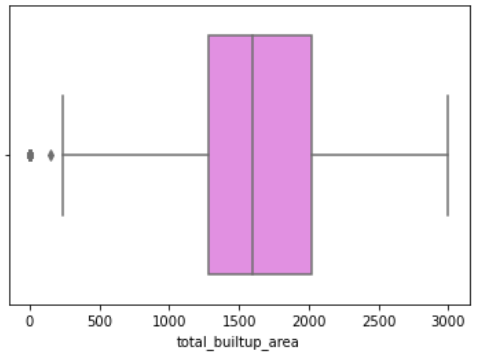
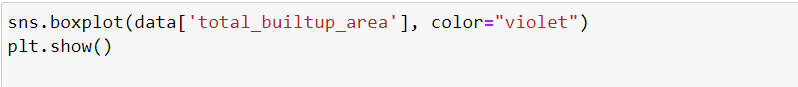


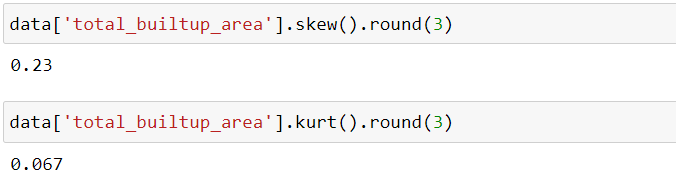
*Inferences:*

1. Minimum Value is 1 while Maximum Value is 4
2. Similar to residential units, most buildings have either 1 or 2 Total Units. Very few have as high as 4 total units.
3. Buildings with 4 Total units are outliers.
4. The Values of Total Units are **Positively Skewed** and **Leptokurtic.**
5. **Plot Size**



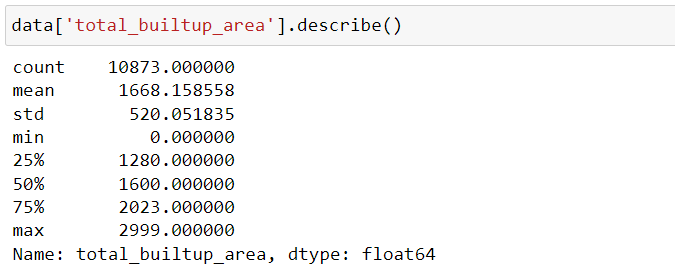
*Inferences:*

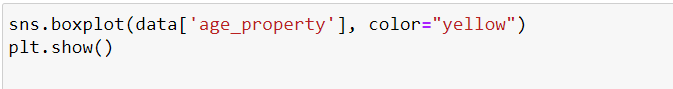
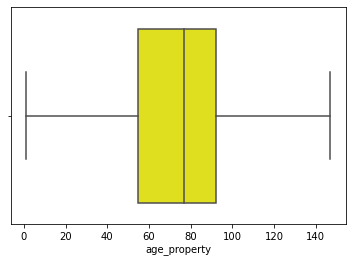
1. Min value is 200 sq feet while Max is close to 5000 sq feet.
2. Mean of 2680 is higher than the Median value of 2500. With std. deviation of 891, the CV is close to 0.3, thus low variation is present.
3. a Multimodal distribution, with peaks at approx. 2000, 2500, and 4000 sq. feet levels.
4. No outliers are present in the distribution.
5. The values of Plot Size are **Positively skewed** and **Platykurtic**
6. ******Total Builtup Area**

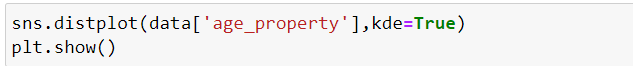


*Inferences*

1. Min value is 0 sq feet(Odd) while Max is close to 3000 sq feet.
2. Mean of 1668 is approx. equal to the Median value of 1680. With std. deviation of 520, the CV is again close to 0.3, thus low variation is present.
3. Few outliers are present in the left tail of the distribution. (Values lesser than mean)
4. The values in the total built-up area are **slightly positively skewed** and **slightly Leptokurtic**, i.e. near to Normal Distribution.

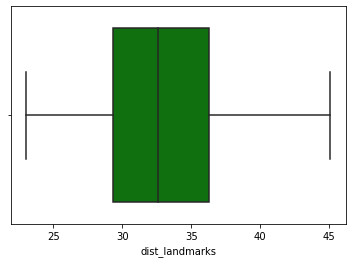
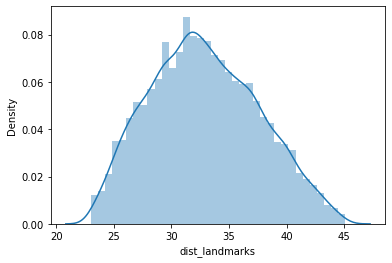
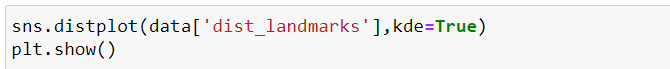
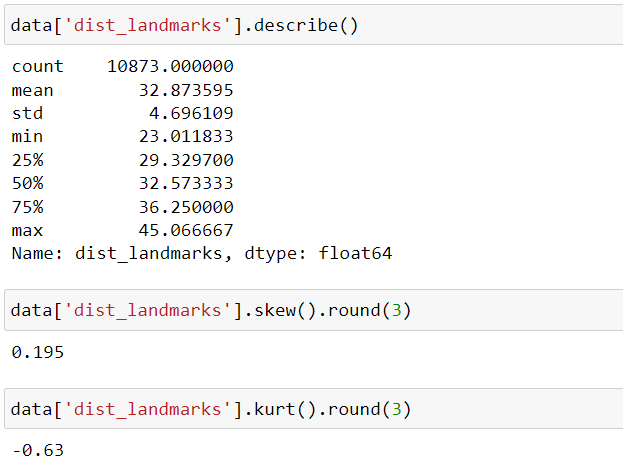


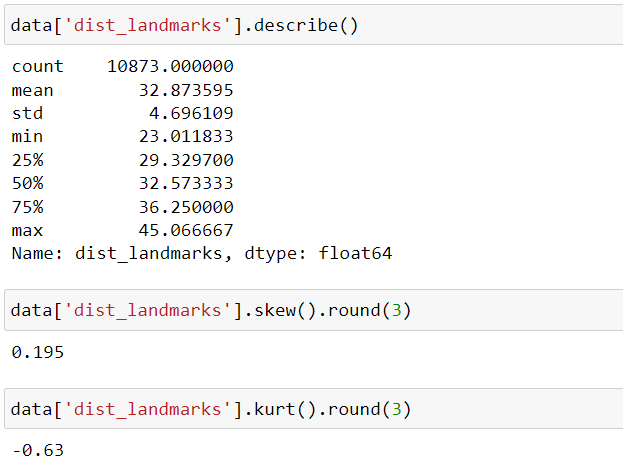
1. **Age Property**

****

****

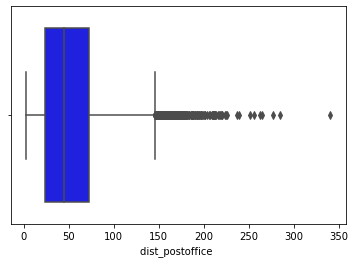
*Inferences:*

1. Min age of properties is 1 yr while the oldest buildings have an age of close to 147 yrs.
2. Mean of 71 yrs is approx. equal to the Median value of 77. With std. deviation of 30, the CV is close to 0.42, thus higher variation than plot\_size and total\_builtup\_area.
3. No outliers are present in the distribution.
4. Age can also be considered a Multimodal distribution, with three peaks at approx. 90-100 yrs, 50-75yrs and 0-25 yrs levels.
5. Values of Age Property are **Negatively Skewed** and **Platykurtic**
6. **Dist Landmarks**

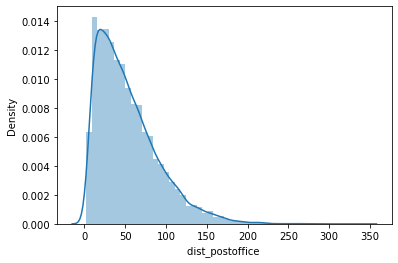
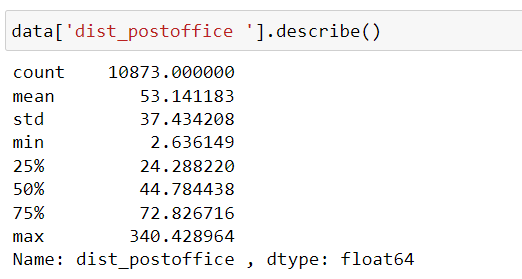
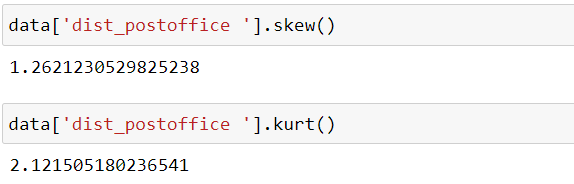
****

Inferences:

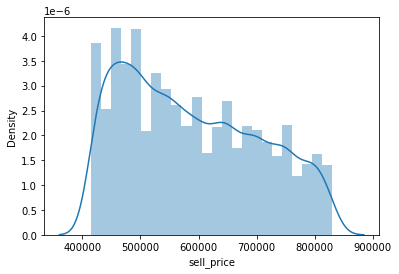
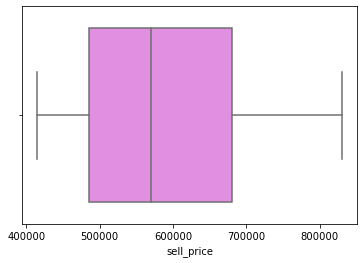
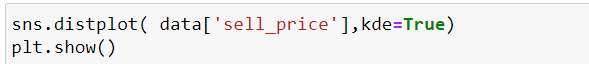
1. Min distance is 23 miles while the max distance is 45 miles.
2. Mean of 32.8 is approx. equal to the Median value of 32.5. With std. deviation of 4.6, the CV is close to 0.13, thus the lowest variation so far.
3. No outliers are present in the distribution.
4. With a fair bit of approximation, Distance to landmarks can be considered as a Symmetric distribution.
5. The values in dist\_landmarks are **positively skewed** and **platykurtic.**
6. **Dist Postoffice**

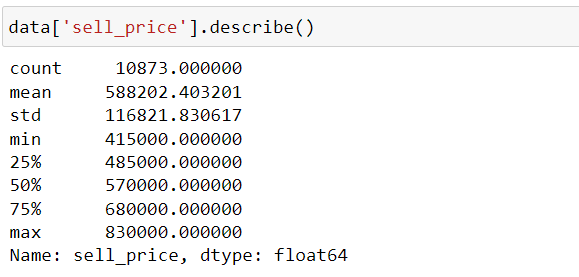
****

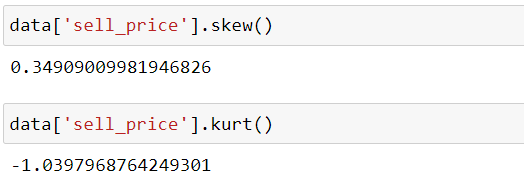
****

****

*Inferences:*

1. Min distance is 2.6 miles while the max distance is 340 miles.
2. Mean of 53.14 is greater than the Median value of 44.7. With std. deviation of 37.43, the CV is close to 0.7, thus the highest variation so far.
3. Few outliers are present in the right tail of the distribution. (Percentage outliers will be analyzed in subsequent sections)
4. The value in the dist\_postoffice is highly **positively skewed** and **leptokurtic**.
5. Positive skewness evident from the long right tail, skewness is 1.26.
6. **Sell Price (Target Variable)**

****

****

*Inferences:*

1. Min sell price is 415000 USD while the max is 830000 USD.
2. Mean of 588202 is approx. equal to the Median value of 570000. With std. deviation of 116821, the CV is close to 0.2, thus low variation is present.
3. No outliers are present in the distribution.
4. Values of Sell Price are **Positively Skewed** and **Platykurtic**.

**Coefficient of Variance**

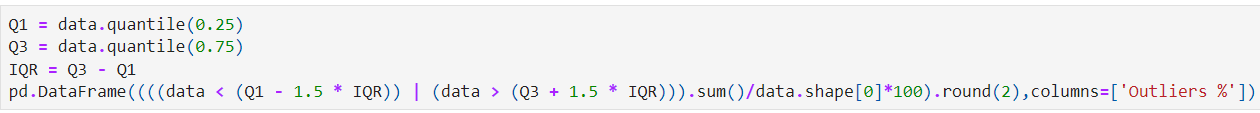
****The coefficient of variation (relative standard deviation) is a statistical measure of the dispersion of data points around the mean. A higher value indicates a greater spread. It can be used to compare variation among data series with different means.

****

|  |  |  |  |
| --- | --- | --- | --- |
|  | Std Dev | Mean | CV |
| House Sold | 0.500 | 0.510 | 0.980 |
| dist\_postoffice | 37.434 | 53.141 | 0.704 |
| age\_property | 30.558 | 71.117 | 0.430 |
| residential\_units | 0.618 | 1.487 | 0.415 |
| total\_units | 0.619 | 1.516 | 0.409 |
| plot\_size | 891.122 | 2680.436 | 0.332 |
| total\_builtup\_area | 520.052 | 1668.159 | 0.312 |
| sell\_price | 116821.831 | 588202.403 | 0.199 |
| dist\_landmarks | 4.696 | 32.874 | 0.143 |

Note:

* Dist\_Postoffice has the highest value of Coefficient of Variation at 0.704, while dist\_landmarks has the lowest CV of 0.143.

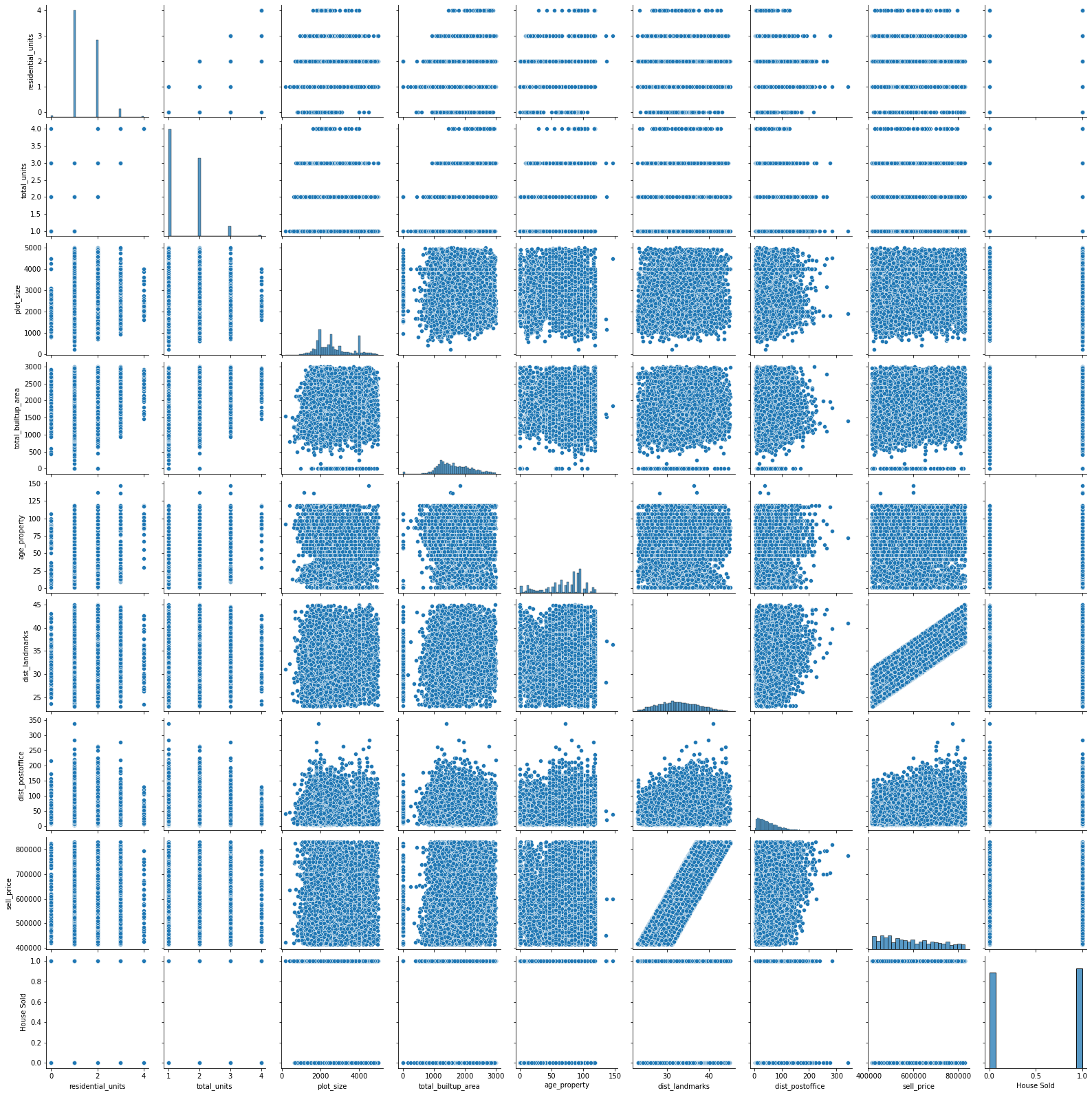
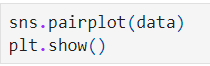
**Percentage Outliers**

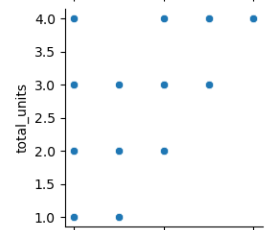
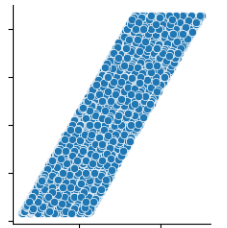
|  |  |
| --- | --- |
|  | Outliers % |
| residential\_units | 0.42 |
| total\_units | 0.52 |
| plot\_size | 0.00 |
| total\_builtup\_area | 0.83 |
| age\_property | 0.00 |
| dist\_landmarks | 0.00 |
| dist\_postoffice | 2.62 |
| sell\_price | 0.00 |
| House Sold | 0.00 |

As the percentage outliers are very low (2.62% being the highest), we will not be doing outlier treatment or removal for the current analysis. An outlier percentage of orders of 15-20% is a cause for concern and thus must be thoroughly investigated.

**Multivariate Analysis**

**Pair Plot**

A pairs plot allows us to see both distributions of single variables and relationships between two variables. Pair plots are a great method to identify trends for follow-up analysis. The histogram on the diagonal allows us to see the distribution of a single variable while the scatter plots on the upper and lower triangles show the relationship (or lack thereof) between two variables.

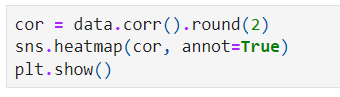
**

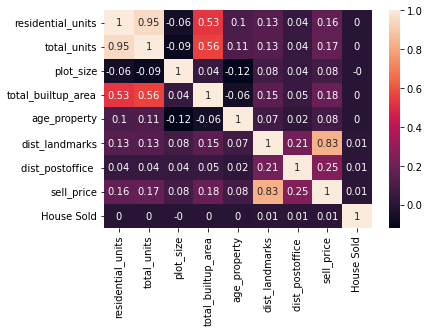
1st Plot Total Units (y) vs Residential Units(x)

2nd Plot Sell\_Price (y) vs dist\_landmark

*Inferences:*

1. In the above pair plot diagram, it is seen that the plots in the diagonal are the histogram of each variable.
2. Remaining, non-diagonal plots are pair-wise scatter plots of every pair-wise combination of variables.
3. We can observe that sell\_price (target variable) and dist\_landmark (feature variable) have a better correlation as compared to the other plots of the target variable.
4. Other than target variable plots, there is a high correlation between *residential\_units* and *total\_units,* and the remaining plots have comparatively less correlation, as the plots don’t have much relation.

**Correlation Plot**



*Inferences*

Top Positive Correlation Pairs -

* residential\_units and total\_units (+0.95)
* dist\_landmarks and sell\_price (+0.83)
* total\_builtup\_area and total\_units (Moderate correlation +0.56)
* Negative correlations are not significant, the highest being -0.12 between age\_property and plot\_size.

We have very few pairs of features with significant positive correlations. This shows there is very less multicollinearity in the data. Component.

In case, if there would have been multicollinearity in the data then one of the methods to remove it is PCA (Principle Component Analysis) which is also useful when the features are too many, and we want to reduce the feature set without losing a significant amount of information (variability).

Another method to do so is calculating the VIF for each independent (feature) variable.

**Regression Models**

1. **Linear Regression**

Simple linear regression is a statistical method that models the relationship between two continuous variables by fitting a linear equation to the observed data. The goal of simple linear regression is to determine whether there is a significant relationship between the two variables, and to predict the value of one variable based on the value of the other. The equation for simple linear regression is:

**y = β0 + β1x + ε**

where:

y is the **dependent variable** (response variable)

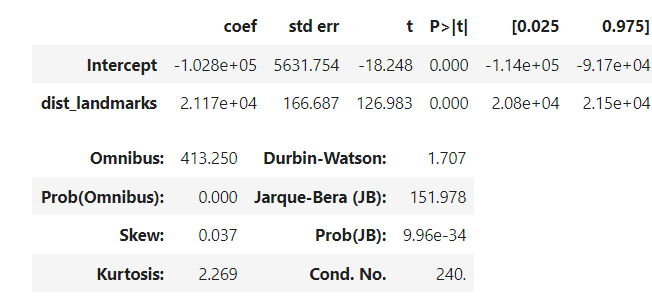
x is the **independent variable** (explanatory variable)

β0 is the **intercept** (the value of y when x is 0)

β1 is the **regression coefficient** (the change in y for a unit change in x)

ε is the **error term** (the difference between the predicted value of y and the actual value of y)

Once we have estimated the values of β0 and β1, we can use the regression equation to predict the value of y for any value of x. We can also use the regression equation to test whether there is a significant relationship between the two variables, by calculating the p-value associated with the regression coefficient β1. If the p-value is less than a predetermined significance level (usually 0.05), we can conclude that there is a significant relationship between the two variables.



Observations:

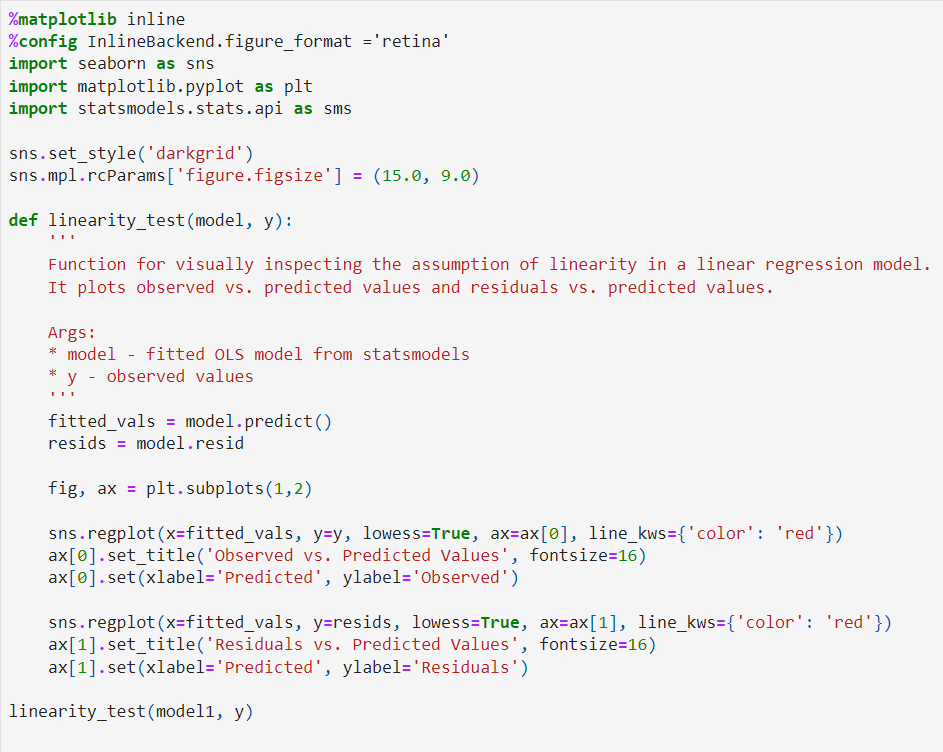
* The linear regression is fitted with the target variable being the *sell\_price* and the feature variable being the *dist\_landmark (the only feature variable with a considerably good correlation value(0.83)).*
* Standard Errors assume that the covariance matrix of the errors is correctly specified.
* The value of R2 is .705, i.e., *70.5% is the model accuracy*.
* Method used for regression is *Ordinary Least Squares.*
* *Confidence Intervals*

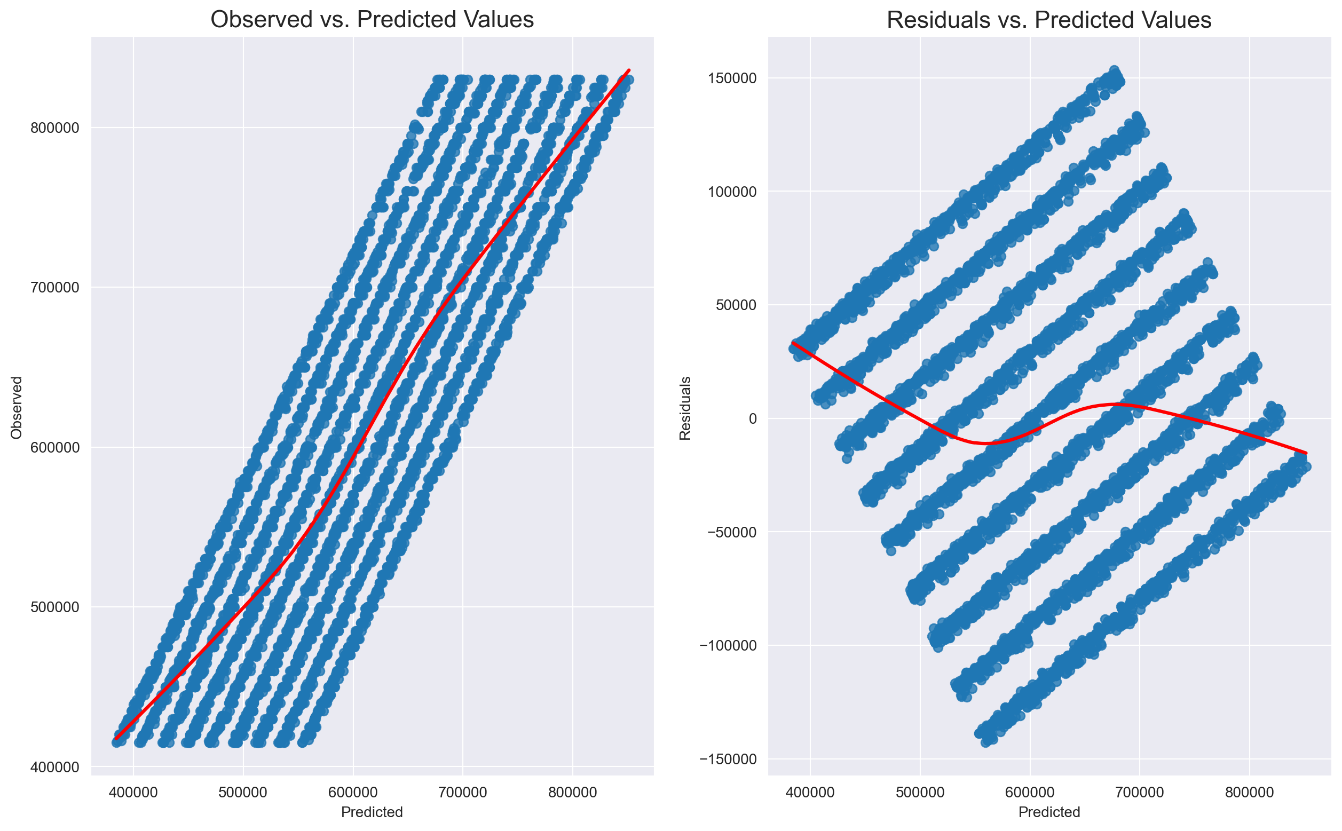
Confidence Interval for *Intercept Variable* at 5% L.O.S is **[-1.14e+05, -9.17e+04]**

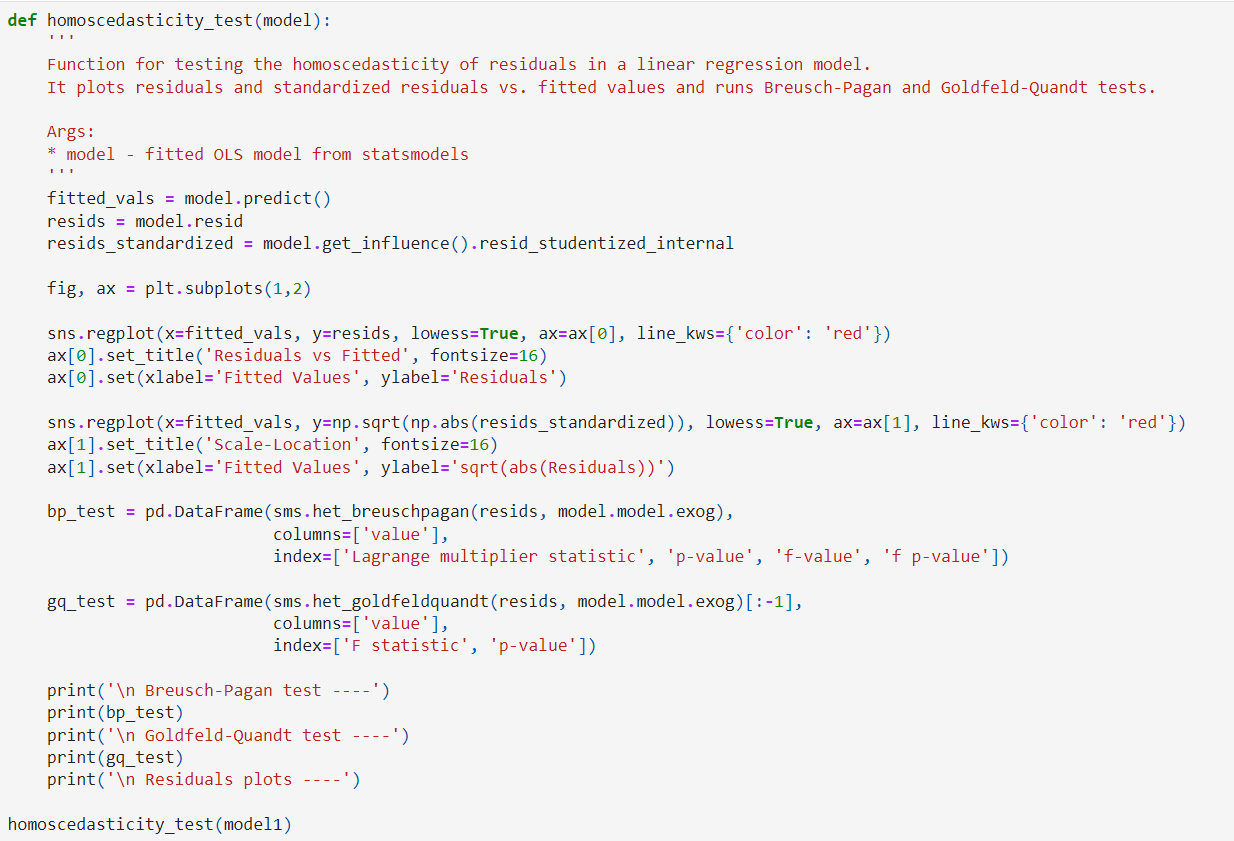
Confidence Interval for *dist\_landmark* at 5% L.O.S is **[2.08e+04, 2.15e+04]**

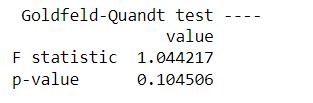
* *Durbin-Watson Test Statistic* Value, not much less than 2 (1.7) states that there is a negligible autocorrelation.
* Skewness in the model is close to zero, hence the model is close to normal fit.
* Kurtosis is 2.269, it states that it is leptokurtic.

**Checking Assumptions of the Simple Linear Regression Model**

* **Linearity**



* **Homoscedasticity**

Test Result

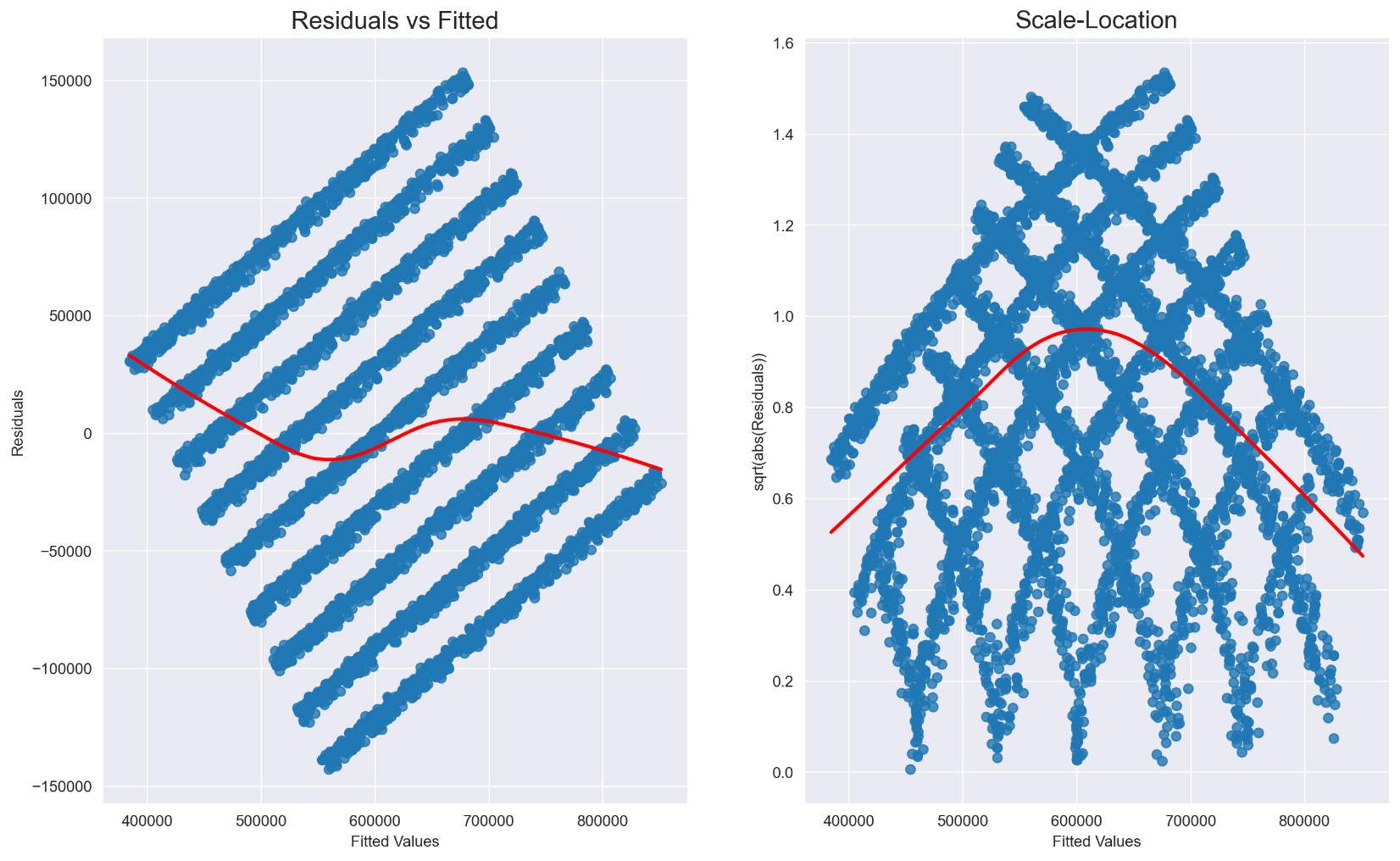
Inference:

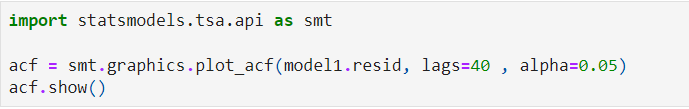
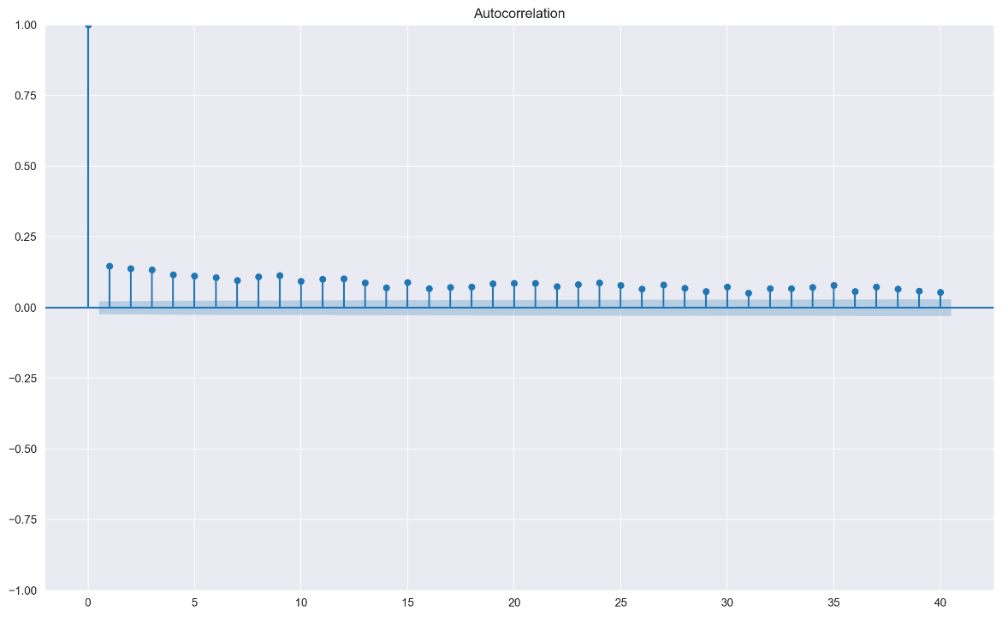
1. *Null Hypothesis: There is Homoscedasticity*

*Alternate Hypothesis: There is no Homoscedasticity*

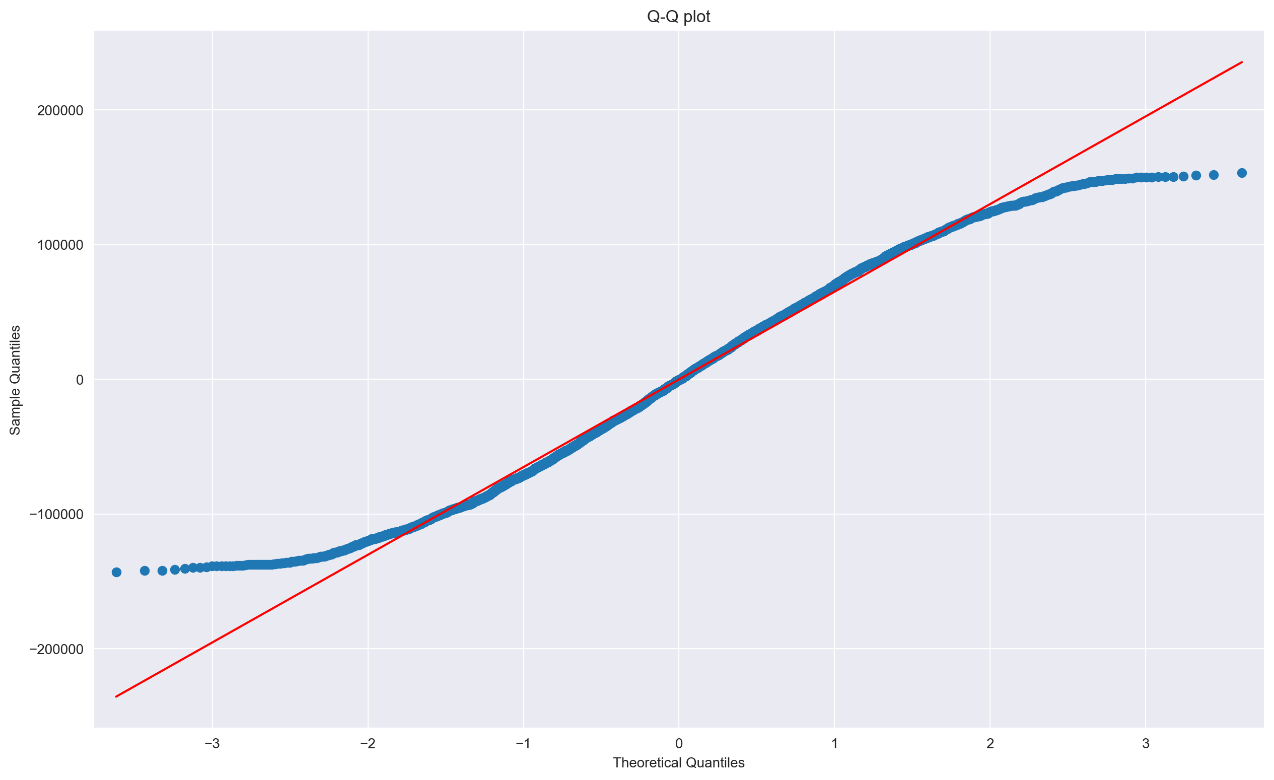
We reject the null hypothesis at a 5% Level of Significance if the p-value <0.05

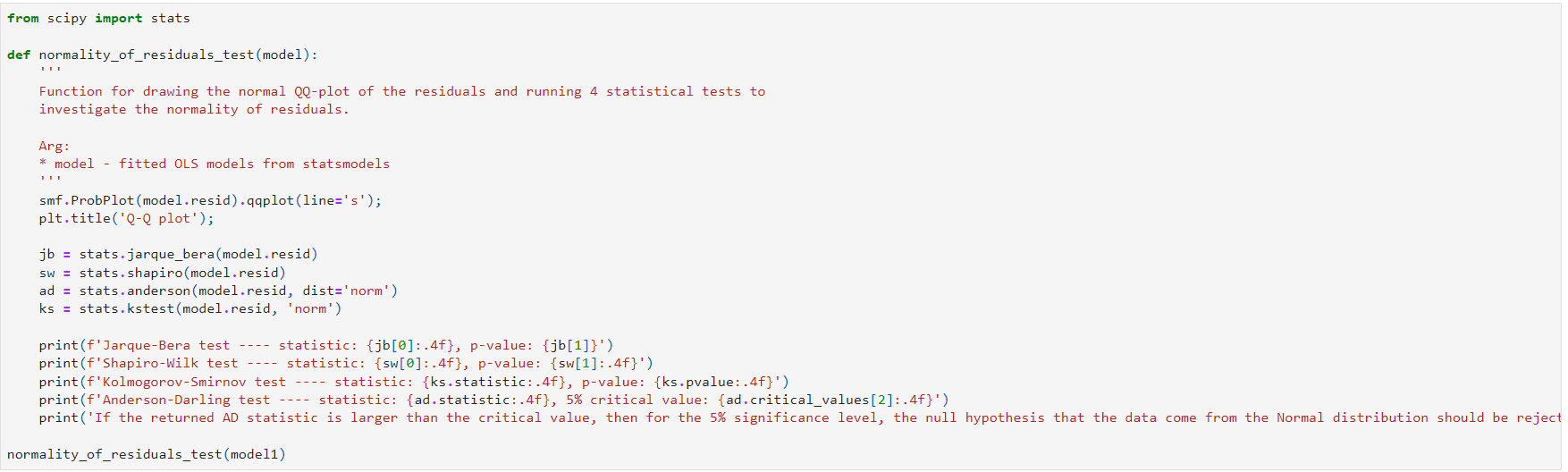
1. Since p-value > 0.05 (for *Goldfeld-Quandt Test Statistic*), hence we accept Null Hypothesis at 5% LOS, and conclude that the *data has homoscedasticity.*

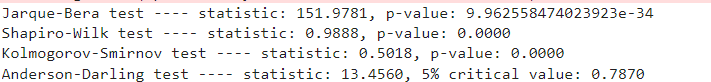


* **Autocorrelation**

There is a negligible autocorrelation in the data model

* **Normality**

****



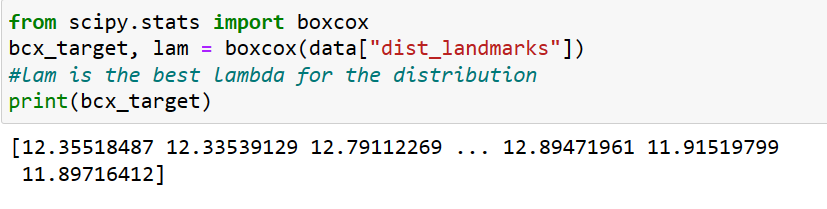
* Hypothesis:

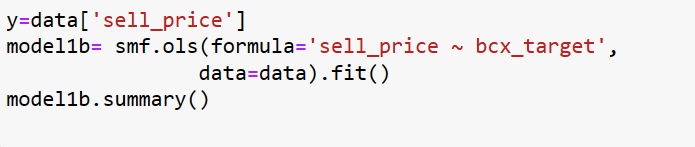
Null: Data comes from the normal distribution.

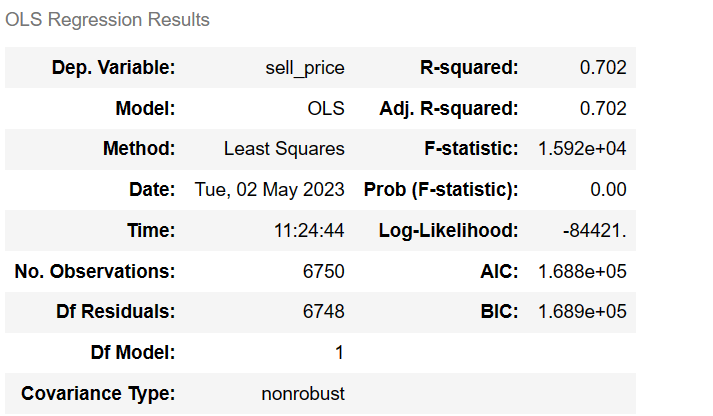
Alternate: Data doesn’t come from the normal distribution.

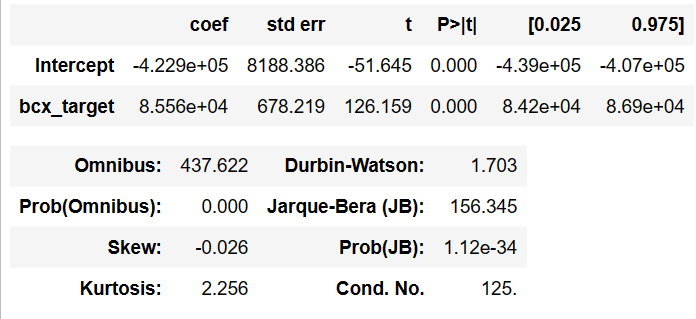
Reject the Null Hypothesis at 5% L.O.S if the critical value is less than the test statistic.

* Since all the test statistic values are greater than the critical value (p-value), hence we conclude that the Data doesn’t come from normal distribution, i.e. Data doesn’t follow the normality assumption.
* Since the model doesn’t follow normality assumption, hence we apply Box-Cox Transformation as follows:

****

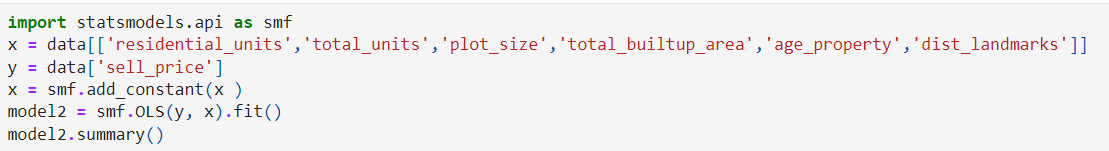
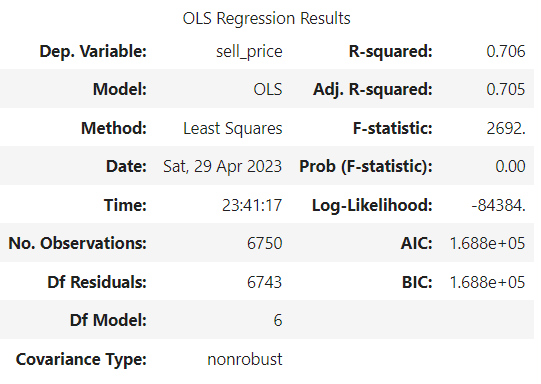
****

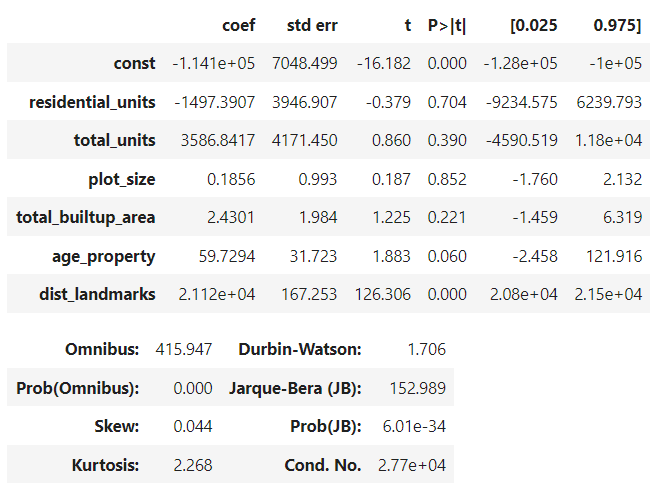
****

****

* After applying the box-cox transformation, we get the same results, with negligible difference in the values.
* Hence we can conclude that the Data satisfies Normality Assumption, with the model accuracy of 70.2%

1. **Multiple Linear Regression**

****

****

Notes:

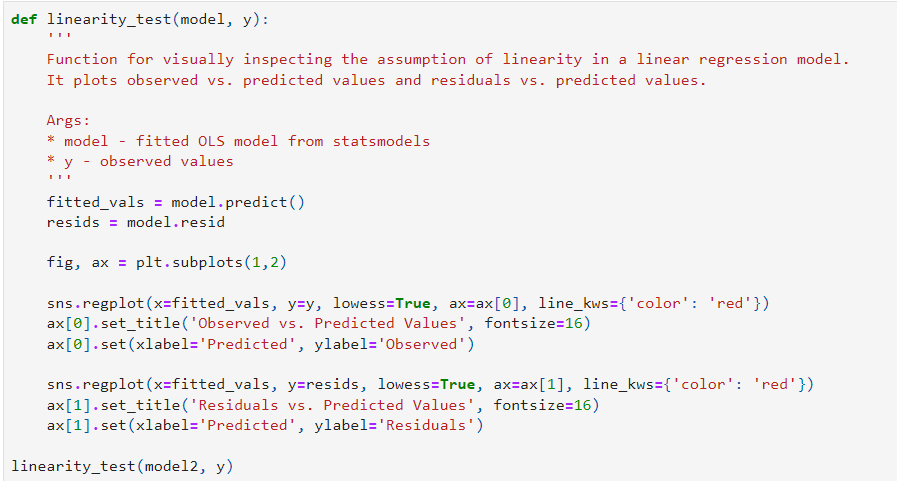
1. Standard Errors assume that the covariance matrix of the errors is correctly specified.
2. The linear regression is fitted with the target variable being the *sell\_price* and the feature variables being the other remaining variables
3. The value of R2 is .706, i.e., *70.6% is the model accuracy*.
4. Method used for regression is *Ordinary Least Squares*
5. The model is slightly positively skewed.
6. The model is Positively Kurtosis, i.e. Leptokurtic
7. *Jarque-Bera Test Statistic* is used to check for normality, larger the value, more the normality.

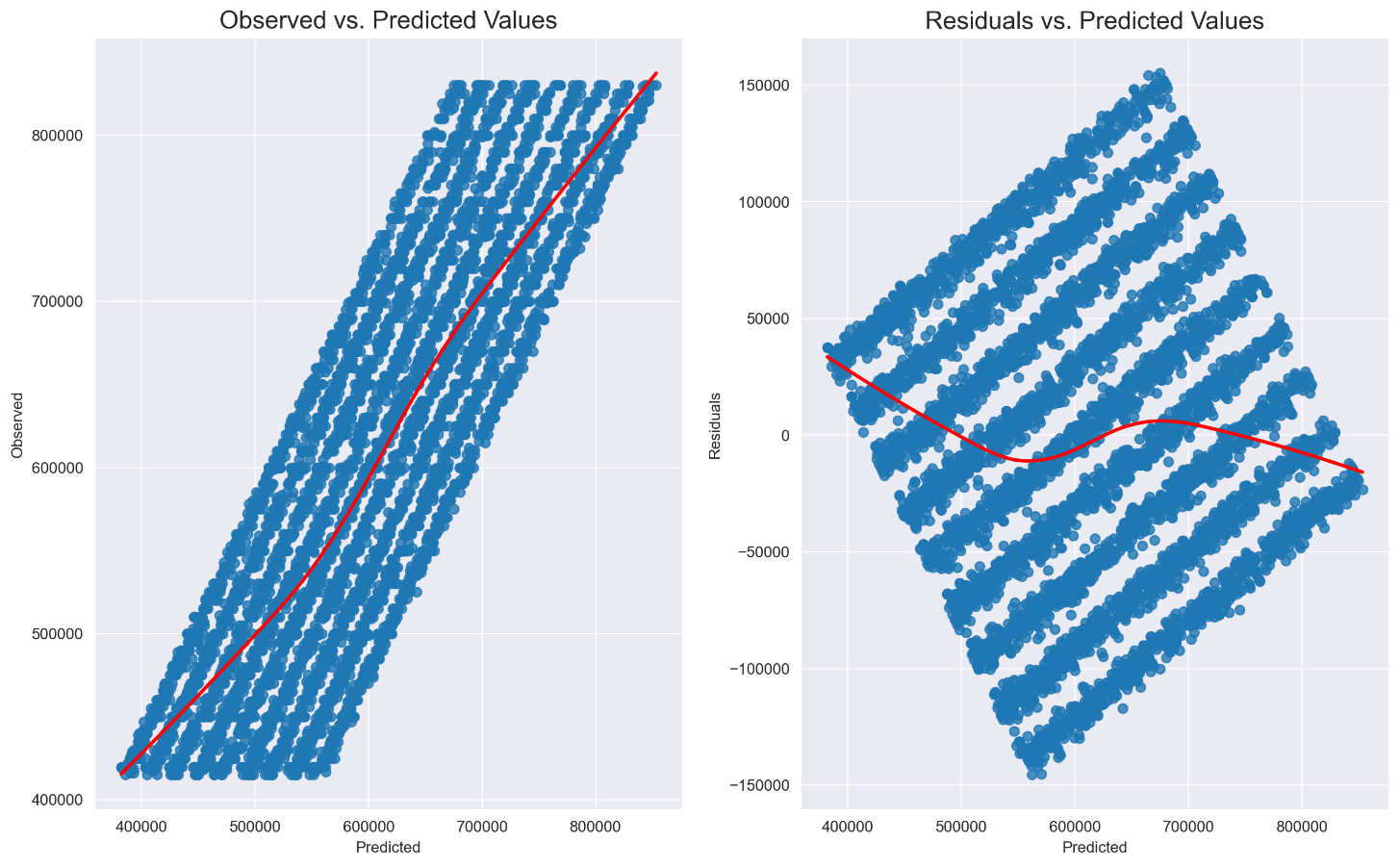
Since the *JB Test Statistic* Value is large enough, we can conclude that the model follows the normality assumption.

1. For there to be no autocorrelation in the Data, Durbin-Watson Test Statistic Value should be equal to 2.

*Durbin-Watson Test Statistic* Value, here in not much less than 2 (1.7) which states that there is a negligible autocorrelation.

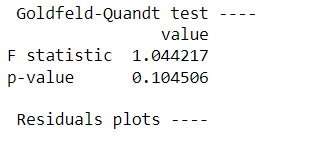
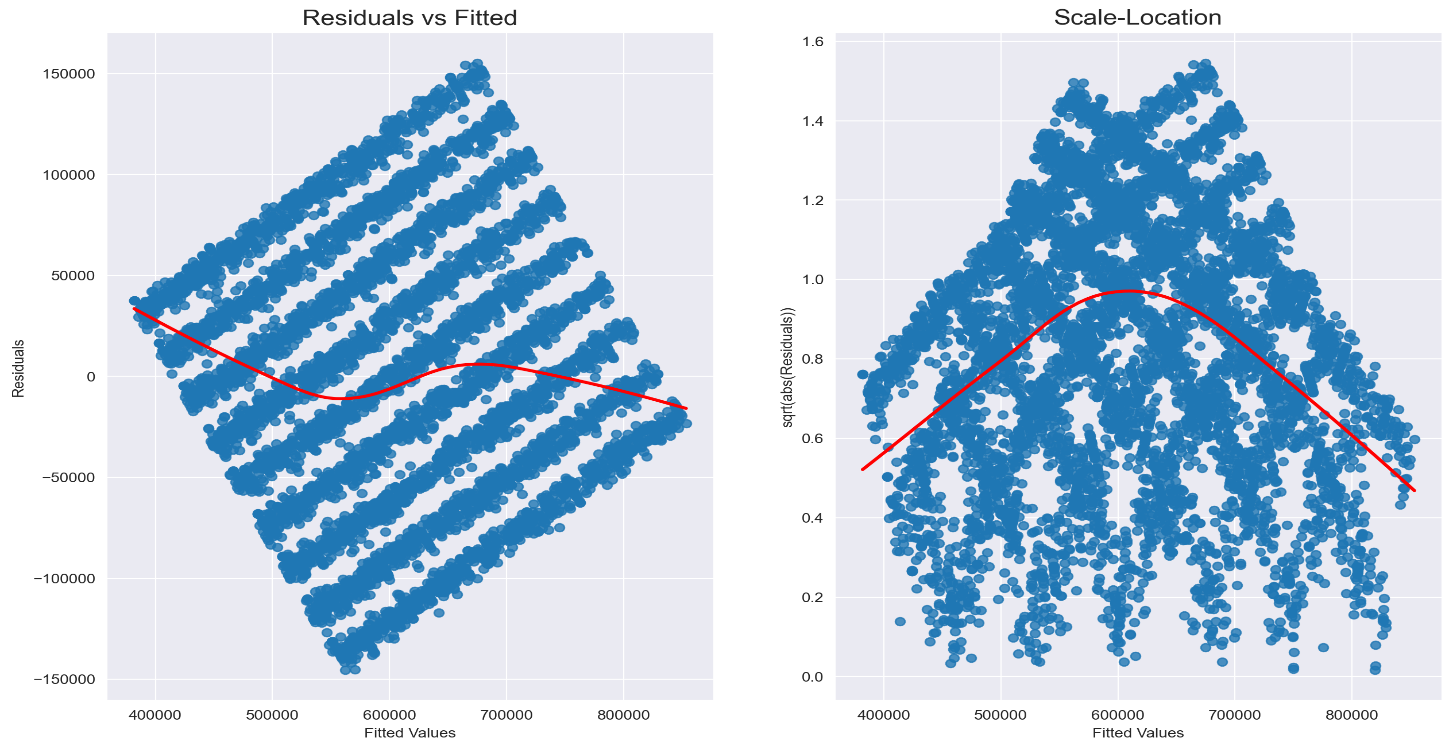
**Checking Assumptions of the Multiple Linear Regression Model**

* **Linearity**



* **Homoscedasticity**

****

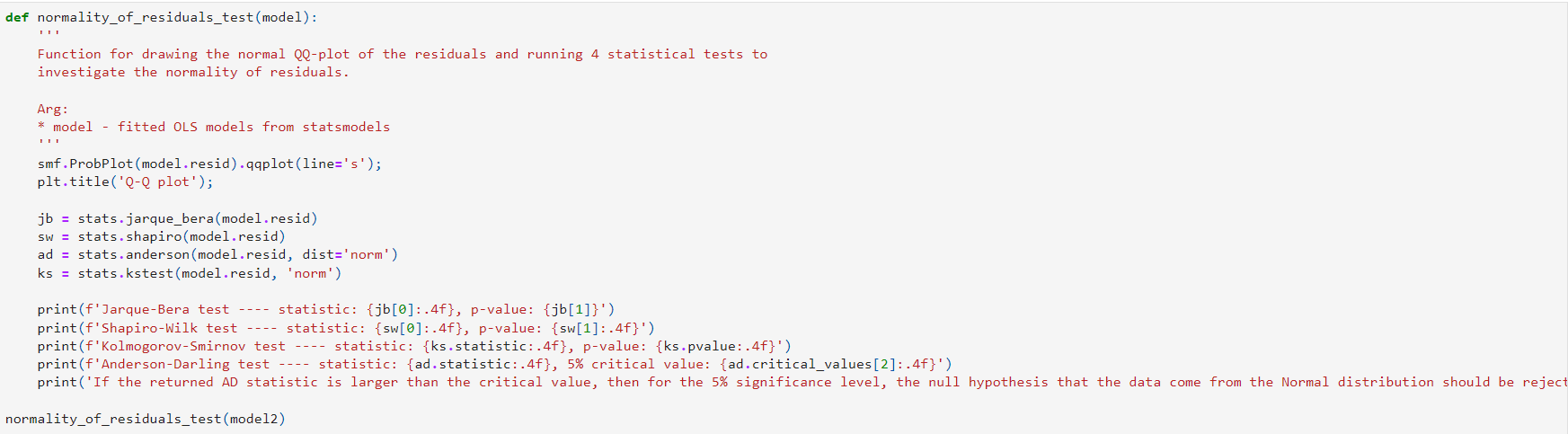
****

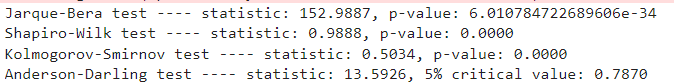
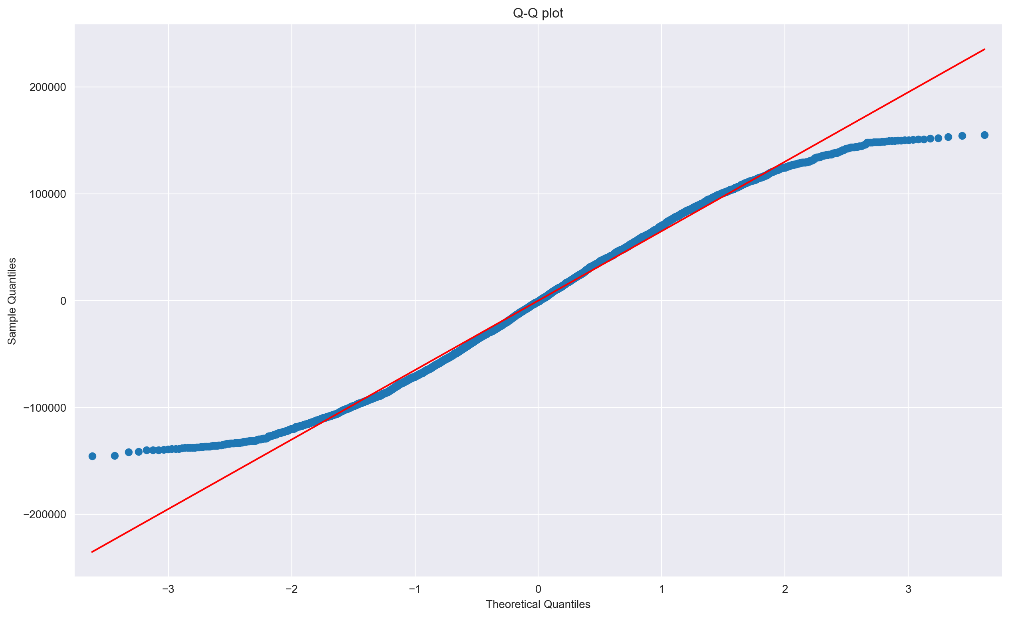
1. *Null Hypothesis: There is Homoscedasticity*

*Alternate Hypothesis: There is no Homoscedasticity*

We reject the null hypothesis at a 5% Level of Significance if the p-value <0.05

1. Since p-value > 0.05 (for *Goldfeld-Quant Test Statistic*) , hence we accept Null Hypothesis at 5% LOS, conclude that the data has homoscedasticity.

* **Normality**

****

*Inference:*

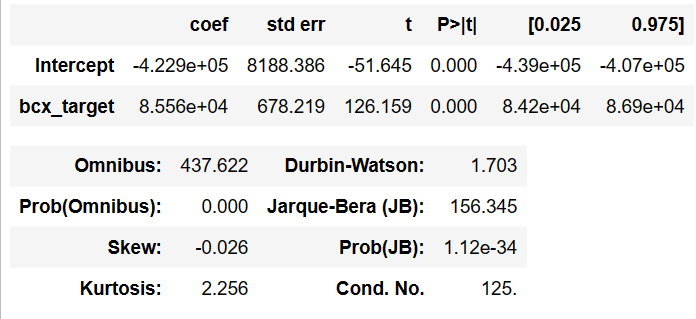
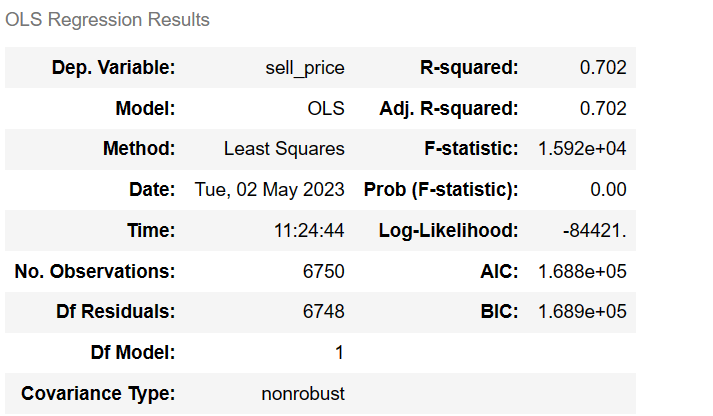
* *Hypothesis*:

Null Hypothesis: Data comes from the normal distribution.

Alternate Hypothesis: Data doesn’t come from the normal distribution.

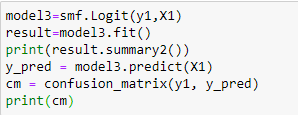
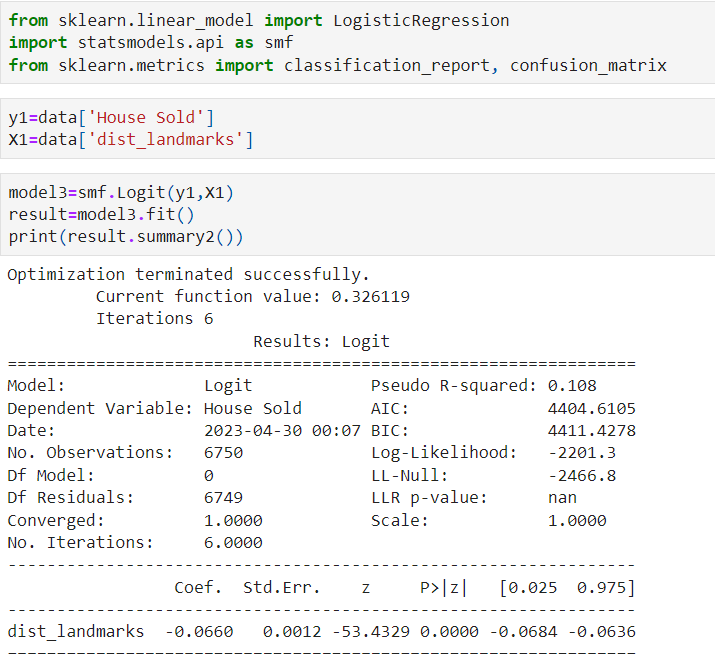
Reject the Null Hypothesis at 5% L.O.S if the critical value is less than the test statistic.

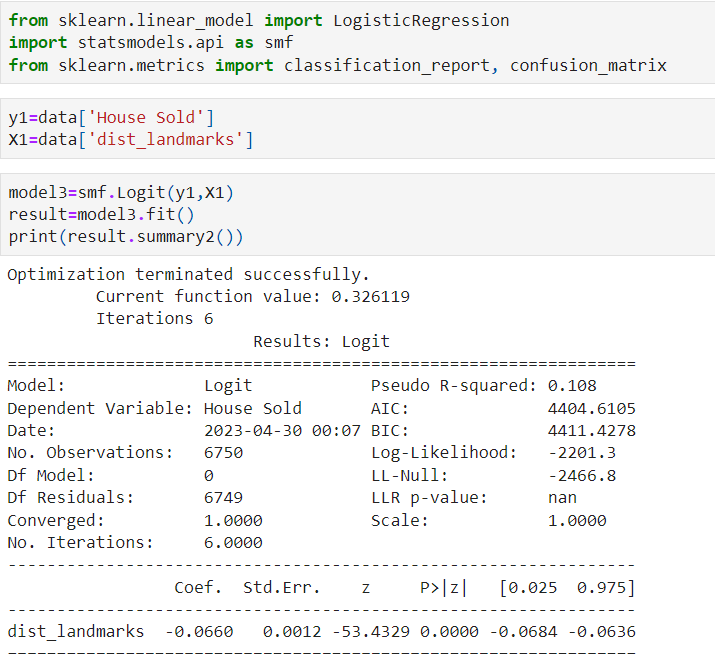
* Since all the test statistic values are greater than the critical value (p-value), hence we conclude that the Data doesn’t come from normal distribution, i.e. Data doesn’t follow the normality assumption.
* Since the model doesn’t follow normality condition hence we apply Box-Cox Transformation as follows:

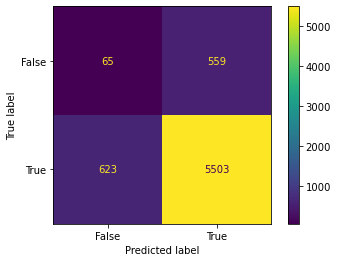
****

Post applying box-cox transformation, we get the same results, with the slight change(negligible) in the model accuracy, with the new model accuracy of 70.2%, as compared to the earlier one of 70.6%.

Hence we conclude that the data follows normality assumption with model accuracy of 70.2%

1. **Logistic Regression**

****

****

*Inference*

1. The linear regression is fitted with the target variable being the *house\_sold (categorical variable)* and the feature variable (dist\_landmark)
2. The value of the pseudo R2 is 0.108
3. *Confidence Intervals*

The confidence Interval for *Intercept Variable* at 5% L.O.S is **[-0.0684, -0.0636]**

1. No of iterations in the model are 6.
2. Critical Value (p-value) is less than the Absolute value of z statistic, hence we reject the null hypothesis, i.e. There is a significant relation between Feature and Target Variable (House\_Sold).
3. A confusion matrix is a table that is often used to describe the performance of a classification model on a set of data for which the true values are known. The matrix provides a summary of the number of correct and incorrect predictions made by the logistic regression model, with respect to each class.
4. A confusion matrix generally has four components:

* True Positives (TP): The number of correctly predicted positive instances (correctly classified as positive).
* False Positives (FP): The number of incorrectly predicted positive instances (incorrectly classified as positive).
* True Negatives (TN): The number of correctly predicted negative instances (correctly classified as negative).
* False Negatives (FN): The number of incorrectly predicted negative instances (incorrectly classified as negative).

------------------