**Logo

Description automatically generated**

**Regression Diagnostics with R**

**A blue and white logo

Description automatically generated with low confidence**

Name: Shivani V

Date: 2023/03/04

**Introduction:**

In this project, a regression analysis of the housing dataset that was downloaded from Kaggle is being done. We construct two regression models, analyze the data, and apply diagnostic tools to find and fix model flaws.

**Analysis:**

1. **Loading the Dataset:**



**Table

Description automatically generated**

The readcsv() method is used to import the dataset. The dataset has 4600 rows and 18 total columns. The remaining columns are of the character and numeric types, with one column of date type.

1. **Exploratory Data Analysis:**

With the summary() function, the dataset's descriptive statistics are obtained.

**Text

Description automatically generated**A screenshot of a computer

Description automatically generated with low confidence

1. **There are no missing values in the dataset.**
2. **Correlation matrix using cor():**

**A screenshot of a computer

Description automatically generated with medium confidence**

The above correlation matrix is obtained by using the cor() function.

1. **Plot of Correlation Matrix:**

**Chart, scatter chart, bubble chart

Description automatically generated**

From the above correlation matrix and plot, it is clear that for the price variable, sqft living has the highest correlation with a value of 0.43 and yr built has the lowest correlation with a value of 0.021.

1. **Scatter plots of variables :**

**Chart

Description automatically generated**

The variable sqft living has the highest connection with price (0.43). The two variables' scatter plot appears as follows. - We can observe from the plot above that price and sqft living have a positive association.

**Chart

Description automatically generated**

The yr\_built variable's association with price has a lowest value of 0.02 for price. The following is the scatter plot of the two variables:

**7. Regression model with 3 continuous variables:**

A Regression model is created for price with sqft\_living, bedrooms and bathrooms as the 3 continuous variables.

Text

Description automatically generated

**8. Reporting the model in equation:**

Price = 123151.7 + 274.7(sqft\_living) + 13303.3(bathrooms) – 55142.9(bedrooms)

Text

Description automatically generated

From the result, we can see the intercept value is 123151.72 which is the price for a total of 0. According to the R squared value of 0.1897, the price is not dependent of the above variables.

**9. Plotting the Regression Model:**

**A picture containing chart

Description automatically generated**

* In the above plot we see there is a random pattern. We can say there is no linearity.

Diagram

Description automatically generated

* We can see the points are all on the diagonal line, there is no normality. There are 3 outliers that are found in this plot.

Chart, scatter chart

Description automatically generated

* We can see a continuous variation and the pattern is random.

Chart

Description automatically generated

**10. Checking for Multicollinearity:**

**Text

Description automatically generated with medium confidence**

As the values are below 5, we can say there is no multicollinearity.

**11. Outliers:**

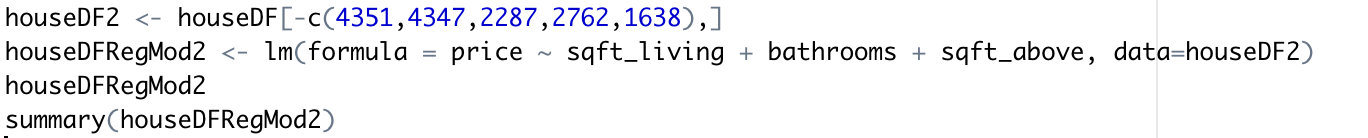
By checking for outliers using outlierTest() function we saw the below outliers. Outliers in the model may cause the model to be inaccurate.

**Text

Description automatically generated**

**12.**

* We remove the outliers and create another regression model to test it. We see no substantial difference in the model’s output.



Text, table

Description automatically generated

**13. Using Subset regression method for best model to identify the best model:**

Graphical user interface, text, application, email

Description automatically generated

In the above result, we can see the selection algorithm has identified that when only one predictor is used, sqft\_living would give the better outcome and when two predictors are used, sqft\_living and sqft\_above would give the bnetter outcomes.

14. **Conclusion:**

The regression model in step 13 which was identified by using the regsubsets() method, would give the more suitable models for this project.

**References:**

*Bommae, W. by. (n.d.). University of Virginia Library Research Data Services + Sciences. Research Data Services + Sciences. Retrieved March 4, 2023, from https://data.library.virginia.edu/diagnostic-plots/*

*Regression Analysis: Step by Step Articles, Videos, Simple Definitions. (2023, February 22). Statistics How To. https://www.statisticshowto.com/probability-and-statistics/regression-analysis/*