**Introduction**

Zillow’s Zestimate is a home evaluation real-estate based in U.S.A. The company has been in operation for the last 11 years. Since then, the company has been giving the consumers insights related to the market prices of houses in the market. Therefore, the company has become one of the largest and most trusted marketplaces for real estate information in the U.S.A. As part of the student learning, data from the company was used to understand the application of multiplication regression model in the real-world context.

**Project Goal**

The goal of the project is to apply multiple regression model to accurately predict the price of house based on the following predictor variables:

1. Lot area/size in square feet
2. Overall quality (overall material and finish of the house)
3. Year Built (original construction date)
4. Year remod Add (remodel date)
5. BsmtFinSF1 (finished square feet)
6. FullBath-Full bathrooms
7. HalfBath-Half baths
8. BedroomAbvGr: Number of Bedrooms above the ground
9. TotRmsAbvGrd: Number of rooms above the ground
10. Fireplaces: Number of fireplaces
11. GarageArea: Size of garage in square feet
12. YrSold: Year sold

**Data Exploration and Analysis**

Before the analysis, the data was explored for linearity assumptions using scatter plot as shown in figure 1 below. This revealed that there is a linear relationship between independent and dependent variables.

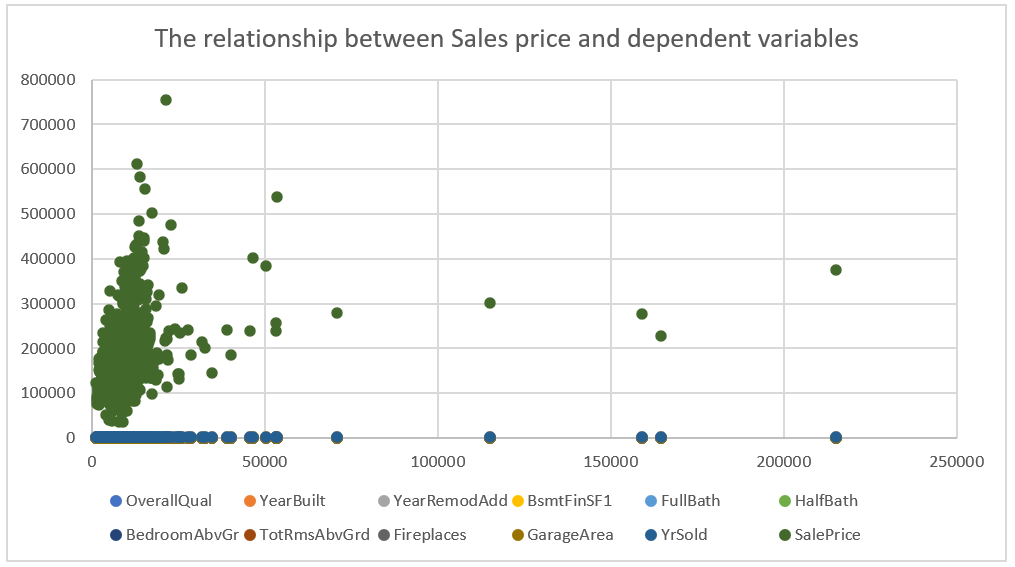


Fig 1: The relationship between Sales price and dependent variables

Secondly, multiple linear regression assumes that there is no multicollinearity among the independent variables. This was explored using the correlation matrix which revealed a correlation co-efficient of less than .80 as shown in table 2 below. To build an accurate regression model, it is recommended that the magnitude of the correlation among all independent variables should be less than .80

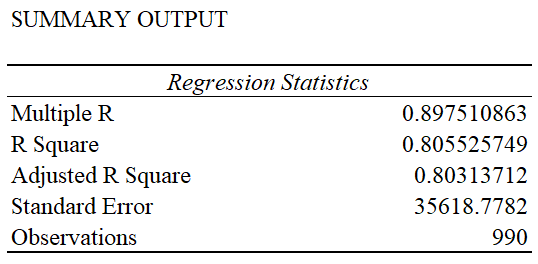
Table

Description automatically generated

Table 1: Correlation matrix of the independent variables

**Regression Analysis Findings**

R square revealed that 80.5% of the variability in Sales prices of the houses can be explained by the set of the independent variables in the study as shown in table 2 below.



*Table 2: Regression analysis Summary Output*

The ANOVA table revealed significance F statistics that is lower than 0.5 significance level hence the test is significant.

Table

Description automatically generated

*Table 3: ANOVA Table*

Table 4 below shows the individual contributions of the independent variables on the house price. The table revealed that all the independent variables stated in the table except FullBath-Full bathrooms, HalfBath-Half baths and Year Sold have a significant contribution on sales price since their p values are less than 0.05.

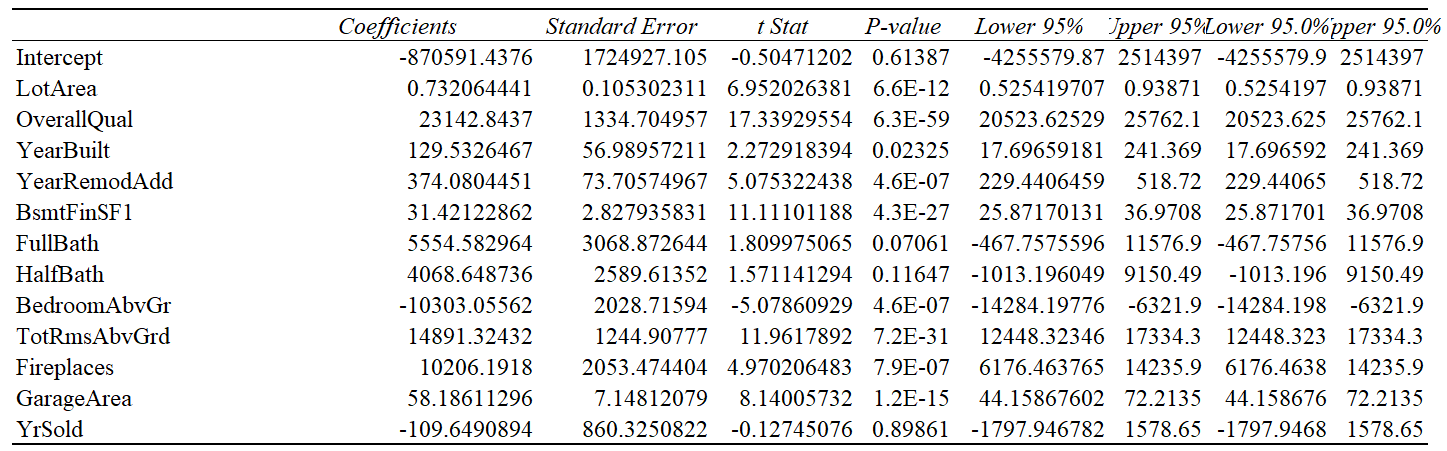


Table 4: co-efficient for the independent variables.

**Predicting the house sales price**

To predict the house sales price, only the independent variables with significant effect on the house sales price were be used for a regression model as shown below.

Table

Description automatically generated

Table 5: Regression statistics for significant independent variables.

 ,

Whereby

 is the predicted or expected value of the dependent variable

X1 through Xp are p distinct independent or predictor variables

b0 is the Y intercept

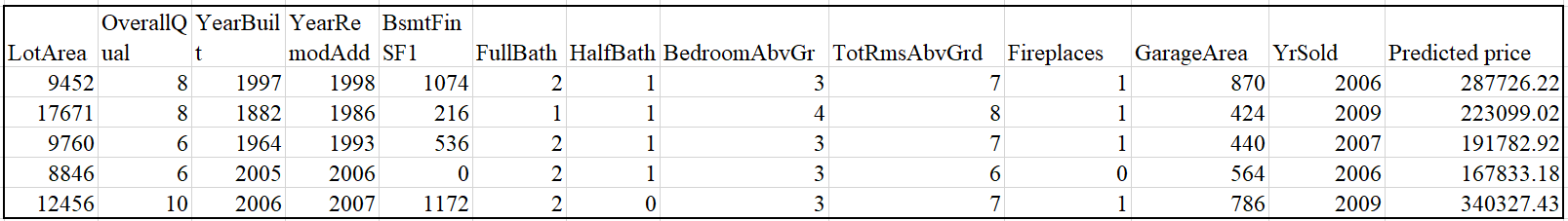


Table 6: Predicted house price