**HOUSE PRICE PREDICTION USING MULTIPLE LINEAR REGRESSION MODEL**

**INTRODUCTION**

This analysis is meant to understand the influence of significant variables on the price of a house and how well these variables can be used to predict the price of the house

**UNDERSTANDING THE DATA STRUCTURE**

This dataset is gotten from Kaggle. This dataset contains the price of the house as well as other significant variables. Other significant variables can be seen below:

1. Square feet
2. Bedroom
3. Bathroom
4. Year Built
5. Neighbourhood
6. Price

By understanding the data structure, correlation analysis was done to determine the relationship between these variables and the price of the house. This analysis will provide insights into the factors that may influence the prices of the house as well as the dynamics of the house market.

**TOOLS**

Tools used for this analysis was Microsoft Excel Together with the Data Analysis add-in in Excel.

**DATA CLEANING AND PREPARATION**

* The dataset was imported as a csv format and converted to an excel format using *text to columns*
* The Price column value was converted from 2 decimal places to whole numbers

**DESCRIPTIVE ANALYSIS**

|  |  |
| --- | --- |
| Mean | 224827.3 |
| Standard Error | 340.5167 |
| Median | 225052 |
| Mode | 237828 |
| Standard Deviation | 76141.84 |
| Sample Variance | 5.8E+09 |
| Kurtosis | -0.40801 |
| Skewness | -0.00825 |
| Range | 528783 |
| Minimum | -36588 |
| Maximum | 492195 |
| Sum | 1.12E+10 |
| Count | 50000 |

**CORRELATION AND MULTIPLE REGRESSION ANALYSIS**

**Numerical Variables**

The dataset contains 5 numerical variables and correlation analysis was performed on them to determine the relationship between these variables and the dependent variable(Price) and how strong the relationships are. The dependent variable was purposely ommitted to focus on the relationships between the other variables and the dependent variable.

|  |  |  |
| --- | --- | --- |
| Independent Variable | Correlation | Result |
| Square feet | 0.7507 | Strong positive correlation |
| Bedroom | 0.0726 | Weak positive correlation |
| Bathroom | 0.0284 | Weak positive correlation |
| Year Built | -0.0023 | Weak negative correlation |

**Significant numerical variable after the correlation**

* Square feet
* Bedroom

**Regression analysis on significant numerical variables**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| SUMMARY OUTPUT |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| Regression Statistics |  |  |  |  |  |  |  |  |
| Multiple R | 0.754417693 |  |  |  |  |  |  |  |
| R Square | 0.569146055 |  |  |  |  |  |  |  |
| Adjusted R Square | 0.56912882 |  |  |  |  |  |  |  |
| Standard Error | 49980.09835 |  |  |  |  |  |  |  |
| Observations | 50000 |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |  |  |
|  | df | SS | MS | F | Significance F |  |  |  |
| Regression | 2 | 1.65E+14 | 8.25E+13 | 33022.32192 | 0 |  |  |  |
| Residual | 49997 | 1.25E+14 | 2498010231 |  |  |  |  |  |
| Total | 49999 | 2.90E+14 |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  | Coefficients | Standard Error | t Stat | P-value | Lower 95% | Upper 95% | Lower 95.0% | Upper 95.0% |
| Intercept | 7694.161981 | 1072.764236 | 7.172276742 | 7.48E-13 | 5591.531812 | 9796.79215 | 5591.531812 | 9796.79215 |
| Square Feet | 99.34814649 | 0.38838517 | 255.7979919 | 0 | 98.58690712 | 100.1093859 | 98.58690712 | 100.1093859 |
| Bedrooms | 5088.620915 | 200.229013 | 25.41400389 | 1.4E-141 | 4696.16976 | 5481.072069 | 4696.16976 | 5481.072069 |

The R square of **0.57** shows that 57% of change in the dependent variable was influenced by the independent variable. The *p-value* seems to agree with the hypothesis.(<0.05).

**Categorical Variable**

There is only 1 categorical variable (Neighbourhood) and it was converted to a dummy variable and regression analysis was performed on the dummy variable.

Regression analysis Result

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *Regression Statistics* | |  |  |  |  |  |  |  |
| Multiple R | 0.022147 |  |  |  |  |  |  |  |
| R Square | 0.00049 |  |  |  |  |  |  |  |
| Adjusted R Square | 0.000431 |  |  |  |  |  |  |  |
| Standard Error | 76124.69 |  |  |  |  |  |  |  |
| Observations | 50000 |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |  |  |  |
| Regression | 3 | 1.42E+11 | 4.74E+10 | 12.26744 | 5.1E-08 |  |  |  |
| Residual | 49997 | 2.9E+14 | 5.79E+09 |  |  |  |  |  |
| Total | 50000 | 2.9E+14 |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
| Intercept | 227166.2 | 590.7888 | 384.5134 | 0 | 226008.2 | 228324.1 | 95065.77 | 359266.6 |
| Rural | -3070.07 | 834.5867 | -3.67855 | 0.000235 | -4705.87 | -1434.27 | -189684 | 183543.6 |
| Suburb | -3932.01 | 834.0262 | -4.71449 | 2.43E-06 | -5566.71 | -2297.31 | -190420 | 182556.3 |
| Urban | 0 | 0 | 65535 | #NUM! | 0 | 0 | 0 | 0 |

The R square shows that only 0.05% of variance of the dependent variable(Price) can be attributed to this categorical variable.

This result is almost negligible and so we cannot rely on the result of this estimate.

Evidence of this result shows that the categorical variable has little of no influence on the variance of the dependent variable.

**Final regression analysis result**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| SUMMARY OUTPUT | |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| *Regression Statistics* | |  |  |  |  |  |  |  |
| Multiple R | 0.75072 |  |  |  |  |  |  |  |
| R Square | 0.56358 |  |  |  |  |  |  |  |
| Adjusted R Square | 0.563571 |  |  |  |  |  |  |  |
| Standard Error | 50301.39 |  |  |  |  |  |  |  |
| Observations | 50000 |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |  |  |  |
| Regression | 1 | 1.63E+14 | 1.63E+14 | 64566 | 0 |  |  |  |
| Residual | 49998 | 1.27E+14 | 2.53E+09 |  |  |  |  |  |
| Total | 49999 | 2.9E+14 |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
| Intercept | 25549.96 | 815.878 | 31.31591 | 3.3E-213 | 23950.83 | 27149.09 | 23950.83 | 27149.09 |
| Square Feet | 99.32211 | 0.39088 | 254.0984 | 0 | 98.55598 | 100.0882 | 98.55598 | 100.0882 |

The multiple-R value of 0.75 shows that there exist a strong relationship between the variable and the house price. The R square value shows that 57% of variance in the house price is influenced by the independent variable which is squarefeet. The *p-value* strongly agree with this with a *p-value* of 0.

However, 43% of variance in the dependent variable can be attributed to other variables not in the dataset.

**Linear regression model equation**

Dependent variable(House price) = Coefficient of Intercept + (99.322\*Squarefeet)

**Testing the model using prediction**

Squarefeet: 2130

**Using the regression model equation;**

Dependent variable(House price) = Coefficient of Intercept + (99.322\*Squarefeet)

H = 25549.96 + (99.322\*2130)

H = 25549.96+ 211,555.86

H = 237,105.82

H = 237,106

Predicted price: 237,106

Actual price: 272,436

The predicted price is 35,330 lower than the actual price

**CONCLUSION**

The analysis above shows that Square feet has significant impact on the price of a house. The variable also showed positive relationship of 0.75 on the corellation analysis to the price of a house which means that the bigger the squarefeet, the higher the price of the house. Other numerical variables like Bedroom,Bathroom and Year built were not included in the final regression analysis because they lack statistical significance.

The analysis also shows that the cathegorical variable, neighbourhood does not have any influece on the price of a house. Their analysis showed weak relationship to the price of the house and their contribution to the variance of the dependent variable was negligible.

Futhermore, 47% of the remaining variance on the dependent variable could be attributed to other factors not in the dataset.

**RECOMMENDATION**

The regression model developed could only account for approximately 57% of variance in house price. 47% is dependent on other unknown variables. Other alternative modelling approaches could be considered.