**HOUSE PRICE PREDICTION**

**Introduction**

The real estate home pricing is a system of determining the price of a house by taking into account several factors such as square footage, age of the house, number of bedrooms, number of bathrooms, number of floors, make of the house, its location, etc. All these variables help one person determine the price when selling or buying a home.

The rising cost of homes is a result of industrialization, urbanization, and government building construction. Real estate is a critical driver of economic growth as it is a critical indicator of strong or weak market conditions.

**Data Preparation**

A clean dataset is required to obtain an accurate result. This dataset contains no missing values, which is a good thing. Upon uploading an excel file into IBM SPSS, the measure of variables such as beds and bathtubs is automatically scaled to the nominal value of the characteristic, however the values of these characteristics are discrete integer values. Another alteration we made to the dataset was to add a decimal. Following the download of the data, SPSS converts Price's decimal characteristics to 15 decimal focuses.

Table

Description automatically generated

1.Total sample size :

Effective variables are:

|  |  |  |
| --- | --- | --- |
| Slno | Scale | Nominal |
| 1. | Date | bedrooms |
| 2 | Price | waterfront |
| 3 | bathrooms | view |
| 4 | sqft\_living | condition |
| 5 | sqft\_lot | city |
| 6 | floors |  |
| 7 | sqft\_above |  |
| 8 | sqft\_basement |  |
| 9 | yr\_built |  |
| 10 | yr\_renovated |  |
|  |  |  |

**Descriptive Statistics**

Table

Description automatically generated

The descriptive statistics shows the statistics like mean, variance, standard deviation for the variables.

**Correlation**

Graphical user interface, application, table, Excel

Description automatically generated

Correlation is drawn against all the variables. Basically, its value lies between -1 and +1. +1 indicates the strongest positive correlation that mean there is a direct relationship between the variables. and -1 represents the strongly negative correlation which means the variables are indirectly proportional. If the modulus of pearson’s coefficient between the variables are lies in between the 0.8-1.0, they are very strong correlation, 0.6-0.79 indicates strong correlation, 0.4-0.59 indicates moderate correlation, 0.2-0.39 represents weak positive and 0-0.19 indicates very weak correlation.

As price is a targeted variable, correlation of price is drawn against all the other variables. Here, price is more strongly correlated with sqft\_living with a correlation coefficient of 0.430. So, we are performing simple regression with price as dependent variable and sqft\_living as independent variable.

**ANOVA:** Bedrooms, waterfront, view, condition and city are compared with price.

Test(I): PricevsBedroom

Chart, histogram

Description automatically generated

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Descriptives** | | | | | | | | |
| price | | | | | | | | |
|  | N | Mean | Std. Deviation | Std. Error | 95% Confidence Interval for Mean | | Minimum | Maximum |
| Lower Bound | Upper Bound |
| 0 | 2 | 1195324.0000 | 141879.56143 | 100324.00000 | -79413.2840 | 2470061.2840 | 1095000.00 | 1295648.00 |
| 1 | 38 | 274076.3158 | 123506.87186 | 20035.46027 | 233480.6172 | 314672.0144 | .00 | 540000.00 |
| 2 | 566 | 391621.9183 | 195119.61170 | 8201.48901 | 375512.7869 | 407731.0497 | .00 | 1695000.00 |
| 3 | 2032 | 488613.0221 | 690508.79070 | 15318.18676 | 458572.0251 | 518654.0190 | .00 | 26590000.00 |
| 4 | 1531 | 635119.3638 | 380893.46168 | 9734.55095 | 616024.8894 | 654213.8383 | .00 | 4489000.00 |
| 5 | 353 | 770185.9887 | 664666.30165 | 35376.61767 | 700609.8668 | 839762.1106 | .00 | 7062500.00 |
| 6 | 61 | 817362.8415 | 591322.87881 | 75711.13644 | 665918.0202 | 968807.6628 | .00 | 3100000.00 |
| 7 | 14 | 1049428.5714 | 852215.36771 | 227764.13755 | 557374.0676 | 1541483.0753 | 280000.00 | 3200000.00 |
| 8 | 2 | 1155000.0000 | 1152584.05333 | 815000.00000 | -9200556.8600 | 11510556.8600 | 340000.00 | 1970000.00 |
| 9 | 1 | 599999.0000 | . | . | . | . | 599999.00 | 599999.00 |
| Total | 4600 | 551962.9885 | 563834.70255 | 8313.28915 | 535664.9519 | 568261.0251 | .00 | 26590000.00 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **ANOVA** | | | | | |
| price | | | | | |
|  | Sum of Squares | df | Mean Square | F | Sig. |
| Between Groups | 62356559332612.650 | 9 | 6928506592512.517 | 22.720 | .000 |
| Within Groups | 1399709561359138.800 | 4590 | 304947616853.843 |  |  |
| Total | 1462066120691751.500 | 4599 |  |  |  |
|  |  |  |  |  |  |

H0: There is no significant difference between the price and different types of bedrooms

H1 : There is significant difference between the price and different types of bedrooms

The anova results suggest that the price of the groups differ significantly is 0.000 (p<0.05) which is less than p value and hence we can conclde that we can reject the null hypothesis.

Test (II) : Price vs Waterfront

Chart, waterfall chart

Description automatically generated

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Descriptives** | | | | | | | | | |
| price | | | | | | | | | |
|  | N | Mean | Std. Deviation | Std. Error | 95% Confidence Interval for Mean | | Minimum | Maximum |
| Lower Bound | Upper Bound |
| 0 | 4567 | 545462.2831 | 547781.08632 | 8105.71868 | 529571.1540 | 561353.4122 | .00 | 26590000.00 |
| 1 | 33 | 1451621.2121 | 1425993.68331 | 248233.63780 | 945985.8383 | 1957256.5859 | .00 | 7062500.00 |
| Total | 4600 | 551962.9885 | 563834.70255 | 8313.28915 | 535664.9519 | 568261.0251 | .00 | 26590000.00 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Test of Homogeneity of Variances** | | | | | |
|  | | Levene Statistic | df1 | df2 | Sig. |
| price | Based on Mean | 74.551 | 1 | 4598 | .000 |
| Based on Median | 54.414 | 1 | 4598 | .000 |
| Based on Median and with adjusted df | 54.414 | 1 | 4031.958 | .000 |
| Based on trimmed mean | 63.344 | 1 | 4598 | .000 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **ANOVA** | | | | | |
| price | | | | | |
|  | Sum of Squares | df | Mean Square | F | Sig. |
| Between Groups | 26902699969673.773 | 1 | 26902699969673.773 | 86.191 | .000 |
| Within Groups | 1435163420722076.200 | 4598 | 312127755702.931 |  |  |
| Total | 1462066120691750.000 | 4599 |  |  |  |

H0: There is no significant difference between the price and different types of waterfront

H1 : There is significant difference between the price and different types of waterfront

The anova results suggest that the price of the groups differ significantly is 0.000 (p<0.05) which is less than p value and hence we can conclde that we can reject the null hypothesis

Test (III) : Price Vs View

Chart, bar chart

Description automatically generated

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Descriptives** | | | | | | | | |
| price | | | | | | | | |
|  | N | Mean | Std. Deviation | Std. Error | 95% Confidence Interval for Mean | | Minimum | Maximum |
| Lower Bound | Upper Bound |
| 0 | 4140 | 510684.9071 | 534350.08609 | 8304.73374 | 494403.1668 | 526966.6473 | .00 | 26590000.00 |
| 1 | 69 | 867010.9669 | 506748.96112 | 61005.40598 | 745276.5746 | 988745.3592 | 180785.71 | 2400000.00 |
| 2 | 205 | 808935.2878 | 702005.66162 | 49030.20219 | 712264.3557 | 905606.2199 | .00 | 7062500.00 |
| 3 | 116 | 998207.5172 | 536378.38385 | 49801.48269 | 899560.3648 | 1096854.6697 | .00 | 2475000.00 |
| 4 | 70 | 1190666.7000 | 851943.11130 | 101826.67802 | 987528.0356 | 1393805.3644 | .00 | 4668000.00 |
| Total | 4600 | 551962.9885 | 563834.70255 | 8313.28915 | 535664.9519 | 568261.0251 | .00 | 26590000.00 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Test of Homogeneity of Variances** | | | | | |
|  | | Levene Statistic | df1 | df2 | Sig. |
| price | Based on Mean | 24.839 | 4 | 4595 | .000 |
| Based on Median | 22.176 | 4 | 4595 | .000 |
| Based on Median and with adjusted df | 22.176 | 4 | 4480.263 | .000 |
| Based on trimmed mean | 23.347 | 4 | 4595 | .000 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **ANOVA** | | | | | |
| price | | | | | |
|  | Sum of Squares | df | Mean Square | F | Sig. |
| Between Groups | 79095335331419.380 | 4 | 19773833832854.844 | 65.700 | .000 |
| Within Groups | 1382970785360333.500 | 4595 | 300972967434.240 |  |  |
| Total | 1462066120691753.000 | 4599 |  |  |  |

H0: There is no significant difference between the price and different types of views

H1 : There is significant difference between the price and different types of views

The anova results suggest that the price of the groups differ significantly is 0.000 (p<0.05) which is less than p value and hence we can conclude that we can reject the null hypothesis

**Test (IV): Price Vs Condition**

Chart, bar chart

Description automatically generated

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Descriptives** | | | | | | | | | |
| price vs condition | | | | | | | | | |
|  | N | Mean | Std. Deviation | Std. Error | 95% Confidence Interval for Mean | | Minimum | Maximum |
| Lower Bound | Upper Bound |
| 1 | 6 | 306633.3333 | 186665.92262 | 76206.04380 | 110739.4614 | 502527.2053 | 7800.00 | 550000.00 |
| 2 | 32 | 324373.7500 | 426231.61728 | 75347.81673 | 170700.8646 | 478046.6354 | .00 | 2555000.00 |
| 3 | 2875 | 550111.5164 | 434311.75636 | 8099.95602 | 534229.2056 | 565993.8271 | .00 | 12899000.00 |
| 4 | 1252 | 533647.2861 | 813713.87287 | 22996.91372 | 488530.5129 | 578764.0593 | .00 | 26590000.00 |
| 5 | 435 | 637041.3223 | 429941.51559 | 20614.11785 | 596525.4061 | 677557.2384 | .00 | 3200000.00 |
| Total | 4600 | 551962.9885 | 563834.70255 | 8313.28915 | 535664.9519 | 568261.0251 | .00 | 26590000.00 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **ANOVA** | | | | | |
| price vs condition | | | | | |
|  | Sum of Squares | df | Mean Square | F | Sig. |
| Between Groups | 5597147333690.162 | 4 | 1399286833422.541 | 4.415 | .001 |
| Within Groups | 1456468973358061.200 | 4595 | 316968220534.943 |  |  |
| Total | 1462066120691751.500 | 4599 |  |  |  |

H0: There is no significant difference between the price and different types of condition

H1 : There is significant difference between the price and different types of condition

The anova results suggest that the price of the groups differ significantly is 0.000 (p<0.05) which is less than p value and hence we can conclude that we can reject the null hypothesis.

Test (V) : Price vs City

Chart, bar chart

Description automatically generated

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **ANOVA** | | | | | |
| price | | | | | |
|  | Sum of Squares | df | Mean Square | F | Sig. |
| Between Groups | 156823060063042.340 | 43 | 3647047908442.845 | 12.730 | .000 |
| Within Groups | 1305243060628708.800 | 4556 | 286488819277.592 |  |  |
| Total | 1462066120691751.000 | 4599 |  |  |  |

H0: There is no significant difference between the price and different types of city

H1 : There is significant difference between the price and different types of city

The anova results suggest that the price of the groups differ significantly is 0.000 (p<0.05) which is less than p value and hence we can conclude that we can reject the null hypothesis

**Simple Linear Regression**

Table, timeline

Description automatically generated with medium confidence

**Price = 12954.242 + 251.950\*sqft\_living**

Price = 12954.242 + 251.950\*sqft living is the regression equation derived from the basic linear regression result. When the value of sqft living is 251.950, the value of price is 12954.242, suggesting that the price will increase by $251.950 for every increase in the value of sqft living. The significant value of the coefficient of sqft living is <0.001, which is nearly 0. As a result, sqft living is a crucial factor.

The overall significance F-test analyzes whether or not the model fits the data better. (n.d., Wikipedia) The F value for statistics is 1045.468, which is extremely high, and the F-significance test result is < 0.001 (0.05), which is significant.

The R-squared value of the model is a measure of how well it matches the data. The R-square value of the model is 0.185. ( Wikipedia, n.d.). It explains 18.5 percent of the price variation described by square feet of living space. 18.5 percent of the data was fit by the regression model.

The Standard Error Estimate for this study is 508991.42327. The Standard Error of Estimate is defined as the variation of data points around the regression line.

Chart, line chart

Description automatically generated

The histogram states that the residuals are normally distributed. So, the variable sqft\_living is significant.

Chart, scatter chart

Description automatically generated

**The residuals show no identifiable pattern. All of the data points are dispersed. As a result, there are no issues with the data. We used all of the independent variables in a multiple linear regression to generate reliable results.**

**Multiple Linear Regression:**

Table

Description automatically generated

Table

Description automatically generated

Chart, line chart

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

Stepwise (Forward):

Table

Description automatically generated

Table

Description automatically generated

Table

Description automatically generated

Price = 4843303.400 + 249.898\*(sqft\_living) - 2540.488\*(yr\_built) + 40150.586(floors)-0.558(sqft\_lot) + 1726.559(city) + 38105.787(bathrooms)

Chart, line chart

Description automatically generated

Chart, scatter chart

Description automatically generated

By comparing the scatter plot to the fitted values, it appears that the data points are randomly distributed and do not show any patterns. A negative and a positive value is present. Therefore, the model is linear, and the residuals (histogram) are normal. The best model when all the analyses are compared is stepwise regression because it contains significant variable coefficients, a higher R-square value, and a lower standard error estimate.

**Conclusion:**

Multiple stepwise forward regression is the most suitable model for all significant variables and because it has a significance value of< 0.05, it is significant. According to the adjusted R-square, the regression model fits 21.2% of the data.