Seattle, WA Housing Prices

Marjorie Blanco

DA 310 – Winnie Li

08

**Fall**



Table of Contents

[Introduction 3](#_Toc499762865)

[Data Description 3](#_Toc499762866)

[Summary Statistics 3](#_Toc499762867)

[Summary 7](#_Toc499762868)

[Histogram 7](#_Toc499762869)

[Intervals 10](#_Toc499762870)

[Z-Scores Method 11](#_Toc499762871)

[box plots (iqr method) 14](#_Toc499762872)

[OUTLIER SUMMARY 18](#_Toc499762873)

[confidence interval 18](#_Toc499762874)

[Modeling 19](#_Toc499762875)

[Modeling Techniques 19](#_Toc499762876)

[Simple Regression Modeling of Independent Variables 19](#_Toc499762877)

[Scatter Plots 19](#_Toc499762878)

[Coefficient of correlation 22](#_Toc499762879)

[Regression summary for the independent variables 23](#_Toc499762880)

[Regression: Model 1 23](#_Toc499762881)

[Summary 24](#_Toc499762882)

[Regression: Model 2 25](#_Toc499762883)

[Summary 25](#_Toc499762884)

[Regression: Model 3 26](#_Toc499762885)

[Summary 27](#_Toc499762886)

[Regression: Model 4 27](#_Toc499762887)

[Summary 28](#_Toc499762888)

[Outliers removal 29](#_Toc499762889)

[Regression: Model 5 29](#_Toc499762890)

[Summary 29](#_Toc499762891)

[Model 4 AND Model 5 COMPARISON 30](#_Toc499762892)

[Multiple Regression Modeling comparison 30](#_Toc499762893)

[Models 30](#_Toc499762894)

[Comparison Tables for Model Significance, Parameter Significance, model goodness of fit 30](#_Toc499762895)

[Best Model Selection 31](#_Toc499762896)

[Discussion for Final Model 32](#_Toc499762897)

[Conclusion 34](#_Toc499762898)

[Appendix 35](#_Toc499762899)

# Introduction

This project examines factors that help determine the prices of houses in the Seattle, WA area. Redfin published data on various factors, which might affect real estate values. The dataset is suitable for multiple linear regression modeling. The original data set includes 221 instances of 27 attributes each. For this project only assess the following explanatory variables to estimate PRICE: BEDS, BATHS, SQUARE FEET, LOT SIZE, and AGE (calculated value taken from 2018 -BUILT YEAR).

The project will evaluate the performance and predictive power of a model that has been trained and tested on data collected from homes in Seattle, WA. A model trained on this data can be used to make certain predictions about a home such as its monetary value. The model would prove to be valuable for a professional real estate agent, real estate investor, taxing authority, homeowner, home buyer, banks who could make use of such information.

# Data Description

This dataset contains information collected by the Redfin concerning housing in Seattle, Washington. It was obtained from the Redfin [website](https://www.redfin.com/city/1387/WA/Bellevue/real-estate#!v=8&sst=&region_id=1387&region_type=6&market=seattle). The original dataset is small with only 221 records.

Quantitative variables are those for which the value has numerical meaning. Categorical variables are those for which the value indicates group membership. This dataset does not contain categorical variables.

Prior to analysis data was cleaned. Rows with missing values were removed. A total of 81 rows were removed. The model will be trained with 140 records.

Table 1 Variable Classification

|  |  |  |
| --- | --- | --- |
| Variable | Classification | Clean Up |
| LIST PRICE ($) | Quantitative Ratio | No empty values |
| BEDS | Quantitative Ratio | No empty values |
| BATHS | Quantitative Ratio | 25 records had empty values |
| SQFT FEET | Quantitative Ratio | 27 records had empty values |
| LOT SIZE | Quantitative Ratio | 54 records had empty values |
| YEAR BUILT (AGE) | Quantitative Ratio | 47 records had empty values |

## Summary Statistics

Table 2 Summary Statistics for PRICE

|  |  |
| --- | --- |
| *PRICE* | |
|  |  |
| Mean | 2,047,228.71 |
| Standard Error | 152,730.11 |
| Median | 1,562,000.00 |
| Mode | 950,000.00 |
| Standard Deviation | 1,807,127.05 |
| Sample Variance | 3,265,708,170,333.58 |
| Kurtosis | 20.56 |
| Skewness | 3.60 |
| Range | 14,695,000.00 |
| Minimum | 305,000.00 |
| Maximum | 15,000,000.00 |
| Sum | 286,612,019.00 |
| Count | 140.00 |
| Confidence Level(95.0%) | 301,974.57 |
| Lower Limit | 1,745,254.13 |
| Upper Limit | 2,349,203.28 |
| Lower Quartile (Q1) | 897,250.00 |
| Upper Quartile (Q3) | 2,730,000.00 |
| Interquartile Range (IQR) | 1,832,750.00 |
| Coefficient of Variation (CV) | 88.99 |

Table 3 Summary Statistics for BEDS

|  |  |
| --- | --- |
| *BEDS* | |
|  |  |
| Mean | 4.178571 |
| Standard Error | 0.127538 |
| Median | 4 |
| Mode | 5 |
| Standard Deviation | 1.509051 |
| Sample Variance | 2.277235 |
| Kurtosis | 7.45372 |
| Skewness | 1.37334 |
| Range | 12 |
| Minimum | 1 |
| Maximum | 13 |
| Sum | 585 |
| Count | 140 |
| Confidence Level(95.0%) | 0.252165 |
| Lower Limit | 3.926406 |
| Upper Limit | 4.430737 |
| Lower Quartile (Q1) | 3 |
| Upper Quartile (Q3) | 5 |
| Interquartile Range (IQR) | 2 |
| Coefficient of Variation (CV) | 36.74886 |

Table 4 Summary Statistics for BATHS

|  |  |
| --- | --- |
| *BATHS* | |
|  |  |
| Mean | 3.316071 |
| Standard Error | 0.115911 |
| Median | 3 |
| Mode | 1.75 and 2.5 |
| Standard Deviation | 1.371475 |
| Sample Variance | 1.880945 |
| Kurtosis | 1.558553 |
| Skewness | 0.904461 |
| Range | 8.25 |
| Minimum | 1 |
| Maximum | 9.25 |
| Sum | 464.25 |
| Count | 140 |
| Confidence Level(95.0%) | 0.229176 |
| Lower Limit | 3.086895 |
| Upper Limit | 3.545248 |
| Lower Quartile (Q1) | 2.5 |
| Upper Quartile (Q3) | 4.25 |
| Interquartile Range (IQR) | 1.75 |
| Coefficient of Variation (CV) | 41.83534 |

Table 5 Summary Statistics for SQUARE FEET

|  |  |
| --- | --- |
| *SQUARE FEET* | |
|  |  |
| Mean | 3774.514 |
| Standard Error | 212.2727 |
| Median | 3375 |
| Mode | 2570 and 3770 |
| Standard Deviation | 2511.644 |
| Sample Variance | 6308357 |
| Kurtosis | 7.709784 |
| Skewness | 2.246086 |
| Range | 15165 |
| Minimum | 810 |
| Maximum | 15975 |
| Sum | 528432 |
| Count | 140 |
| Confidence Level(95.0%) | 419.7008 |
| Lower Limit | 3354.813 |
| Upper Limit | 4194.215 |
| Lower Quartile (Q1) | 2135 |
| Upper Quartile (Q3) | 4857.75 |
| Interquartile Range (IQR) | 2722.75 |
| Coefficient of Variation (CV) | 67.31379 |

Table 6 Summary Statistics for LOT SIZE

|  |  |
| --- | --- |
| *LOT SIZE* | |
|  |  |
| Mean | 44840.1 |
| Standard Error | 9665.78 |
| Median | 13281 |
| Mode | 1 |
| Standard Deviation | 114367.1 |
| Sample Variance | 1.31E+10 |
| Kurtosis | 36.23026 |
| Skewness | 5.56007 |
| Range | 976178 |
| Minimum | 1 |
| Maximum | 976179 |
| Sum | 6277614 |
| Count | 140 |
| Confidence Level(95.0%) | 19110.97 |
| Lower Limit | 25729.13 |
| Upper Limit | 63951.07 |
| Lower Quartile (Q1) | 8989.5 |
| Upper Quartile (Q3) | 31322.25 |
| Interquartile Range (IQR) | 22332.75 |
| Coefficient of Variation (CV) | 256.0985 |

Table 7 Summary Statistics for AGE

|  |  |
| --- | --- |
| *AGE* | |
|  |  |
| Mean | 31.80714286 |
| Standard Error | 1.897282538 |
| Median | 35 |
| Mode | 1 |
| Standard Deviation | 22.44894974 |
| Sample Variance | 503.9553443 |
| Kurtosis | -0.284812254 |
| Skewness | 0.306802881 |
| Range | 105 |
| Minimum | 1 |
| Maximum | 106 |
| Sum | 4453 |
| Count | 140 |
| Confidence Level(95.0%) | 3.751264772 |
| Lower Limit | 28.05587809 |
| Upper Limit | 35.55840763 |
| Lower Quartile (Q1) | 11 |
| Upper Quartile (Q3) | 49.25 |
| Interquartile Range (IQR) | 38.25 |
| Coefficient of Variation (CV) | 71.28694799 |

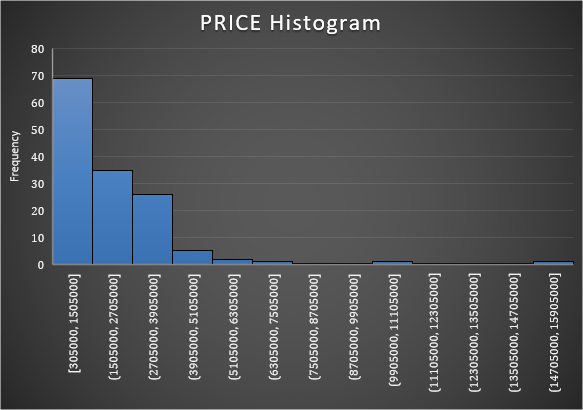
### Summary

Based on the details from the summary statistics, the data for the various dependent and independent variables do not appear normally distributed. The data is highly skewed.

## Histogram

The following are histograms of the Quantitative Ratio variables. For each histogram, the mean and median line will be displayed along with a description of the distribution.

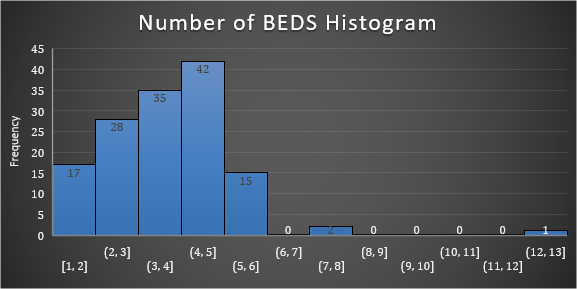
For PRICE, the mean is greater than the median. This distribution is right skewed (positively) with skewness of 3.6. The finding is consistent with the descriptive measurements for PRICE.



Median **<** Mean

Figure 1 PRICE histogram

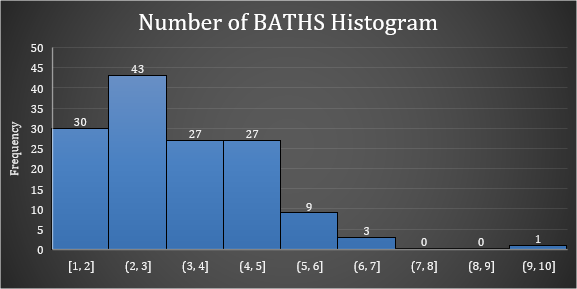
For BEDS, the mean is greater than the median. This distribution is right skewed (positively) with skewness of 1.4. The finding is consistent with the descriptive measurements for BEDS.



Median **<** Mean

Figure 2 BEDS Histogram

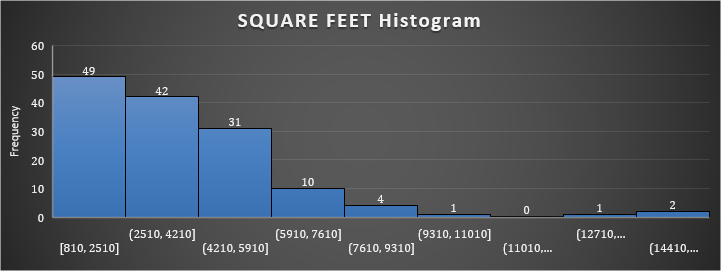
For BATHS, the mean is greater than the median. This distribution is right skewed (positively) with skewness of 0.9. The finding is consistent with the descriptive measurements for BATHS.



Median **<** Mean

Figure 3 BATHS Histogram

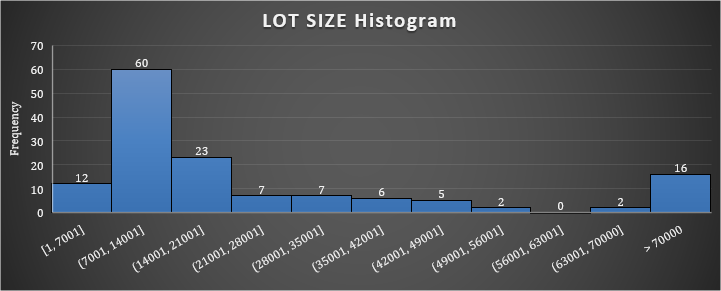
For SQUARE FEET, the mean is greater than the median. This distribution is right skewed (positively) with skewness of 2.5. The finding is consistent with the descriptive measurements for SQUARE FEET.



Median **<** Mean

Figure 4 SQUARE FEET Histogram

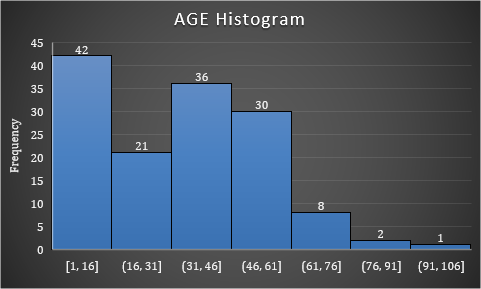
For LOT SIZE, the mean is greater than the median. This distribution is right skewed (positively). The finding is consistent with the descriptive measurements for LOT SIZE.



Median **<** Mean

Figure 5 LOT SIZE Histogram

For AGE, the mean is less than the median. This distribution is left skewed (negative) with skewness of 0.30. The finding is consistent with the descriptive measurements for AGE.



Median **>** Mean

Figure 6 AGE Histogram

## Intervals

This section provides intervals of the mean plus and minus one, two and three standard deviations for all Quantitative Ratio variables. The following are the interval to help identified normal and extreme outlier(s).

Table 8 Intervals

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | PRICE | BEDS | BATHS | SQUARE FEET | LOT SIZE | AGE |
| 𝑥̅-3𝑠 | (3,374,152.44) | -0.35 | -0.80 | (3,760.42) | (298,261.06) | -35.5 |
| 𝑥̅-2𝑠 | (1,567,025.39) | 1.16 | 0.57 | (1,248.77) | (183,894.01) | -13.1 |
| 𝑥̅-𝑠 | 240,101.66 | 2.67 | 1.94 | 1,262.87 | (69,526.95) | 9.4 |
| 𝑥̅ | 2,047,228.71 | 4.18 | 3.32 | 3,774.51 | 44,840.10 | 31.8 |
| 𝑥̅+𝑠 | 3,854,355.76 | 5.69 | 4.69 | 6,286.16 | 159,207.15 | 54.3 |
| 𝑥̅+2𝑠 | 5,661,482.80 | 7.20 | 6.06 | 8,797.80 | 273,574.21 | 76.7 |
| 𝑥̅+3𝑠 | 7,468,609.85 | 8.71 | 7.43 | 11,309.45 | 387,941.26 | 99.2 |

Table 9 Intervals Summary

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | All between 𝑥̅±2𝑠 | All between 𝑥̅±3𝑠 | Distribution | Findings consistent |
| PRICE | No | No | Right skewed with normal and extreme outliers | Yes |
| BEDS | No | No | Right skewed with normal and extreme outliers | Yes |
| BATHS | No | No | Right skewed with normal and extreme outliers | Yes |
| SQUARE FEET | No | No | Right skewed with normal and extreme outliers | Yes |
| LOT SIZE | No | No | Right skewed with normal and extreme outliers | Yes |
| AGE | No | No | Left skewed with normal and extreme outliers | Yes |

## Z-Scores Method

It is important to identify inconsistent or unusual measurement in a data set. Observations that are unusually large or small relative to the data values are considered outliers.

The following z-score plot for each Quantitative Ratio variable. A list of observation that have been identified as normal or extreme outlier with their corresponding value and z-score will be provided. The observation identified as outliers are consistent with the interval method as expected. Observations with z-scores greater than 3 in absolute value are considered extreme outliers. For highly skewed data sets such as this, observations with z-scores greater than 2 in absolute value are considered normal outliers.

According to this scatter plot there are 2 extreme outliers and 2 normal outliers.

|  |  |  |  |
| --- | --- | --- | --- |
| *Observation* | *Value* | *Z-Score* | *Note* |
| 137 | 15000000 | -7.1676 | *Extreme* |
| 114 | 9988000 | -4.39414 | *Extreme* |
| 131 | 6888000 | -2.67871 | *Normal* |
| 57 | 5980000 | -2.17626 | *Normal* |

According to this scatter plot there is 1 extreme outlier and 4 normal outliers.

Figure 7 BEDS Z-Scores

Table 10 BEDS Z-Score Outliers

|  |  |  |  |
| --- | --- | --- | --- |
| *Observation* | *Value* | *Z-Score* | *Note* |
| 2 | 1 | -2.10634 | *Normal* |
| 31 | 1 | -2.10634 | *Normal* |
| 87 | 8 | 2.532339 | *Normal* |
| 119 | 8 | 2.532339 | *Normal* |
| 56 | 13 | 5.845679 | *Extreme* |

According to this scatter plot there is 1 extreme outlier and 3 normal outliers.

Figure 8 BATHS Z-Scores

Table 11 BATHS Z-Score Outliers

|  |  |  |  |
| --- | --- | --- | --- |
| *Observation* | *Value* | *Z-Score* | *Note* |
| 64 | 6.25 | 2.13925 | *Normal* |
| 131 | 6.5 | 2.321535 | *Normal* |
| 137 | 6.75 | 2.503821 | *Normal* |
| 56 | 9.25 | 4.326675 | *Extreme* |

According to this histogram there are 3 extreme outliers and 2 normal outliers.

Figure 9 SQUARE FEET Z-Scores

Table 12 SQUARE FEET Z-Score Outliers

|  |  |  |  |
| --- | --- | --- | --- |
| *Observation* | *Values* | *Z-Score* | *Note* |
| 112 | 9116 | 2.126689 | *Normal* |
| 131 | 10088 | 2.513686 | *Normal* |
| 114 | 14140 | 4.126972 | *Extreme* |
| 56 | 15360 | 4.61271 | *Extreme* |
| 137 | 15975 | 4.857569 | *Extreme* |

According to this histogram there is 4 extreme outlier.

Figure 10 LOT SIZE Z-Scores

Table 13 LOT SIZE Z-Score Outliers

|  |  |  |  |
| --- | --- | --- | --- |
| *Observation* | *Values* | *Z-Score* | *Note* |
| 71 | 435710 | 3.417679 | *Extreme* |
| 14 | 468270 | 3.702377 | *Extreme* |
| 7 | 497890 | 3.961367 | *Extreme* |
| 89 | 497890 | 3.961367 | *Extreme* |

According to this histogram there is 1 extreme outlier and 2 normal outliers.

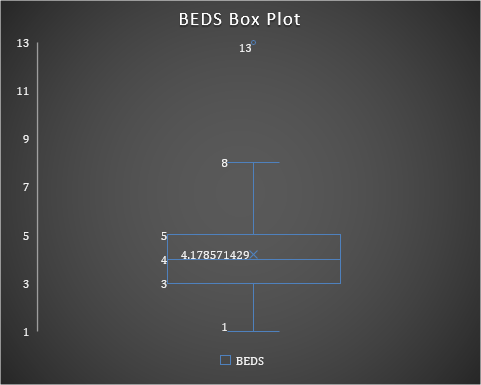
Figure 11 AGE Z-Scores

Table 14 AGE Z-Score Outliers

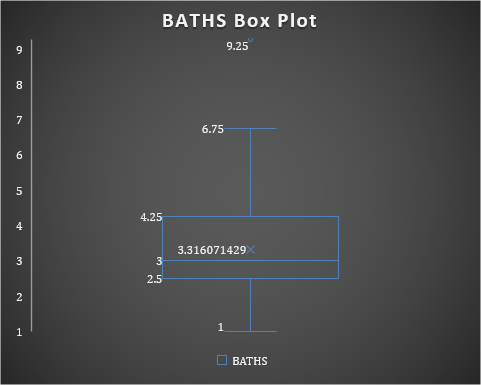
|  |  |  |  |
| --- | --- | --- | --- |
| *Observation* | *Values* | *Z-Score* | *Note* |
| 135 | 88 | 2.50314 | *Normal* |
| 46 | 88 | 2.50314 | *Normal* |
| 26 | 106 | 3.304959 | *Extreme* |

## box plots (iqr method)

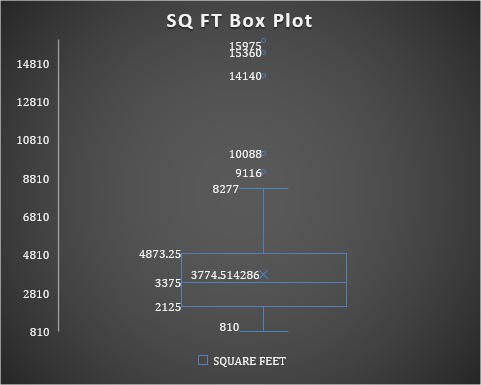
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | PRICE | | BEDS | BATHS | SQUARE FEET | LOT SIZE | AGE |
| Lower Outer fence | -4601000 | | -3 | | -2.75 | -6033.25 | -58008.8 | -103.75 |
| Lower Inner fence | -1851875 | | 0 | | -0.125 | -1949.13 | -24509.6 | -46.375 |
| Upper Inner Fence | 5479125 | | 8 | | 6.875 | 8941.875 | 64821.38 | 106.625 |
| Upper Outer fence | 8228250 | | 11 | | 9.5 | 13026 | 98320.5 | 197 |



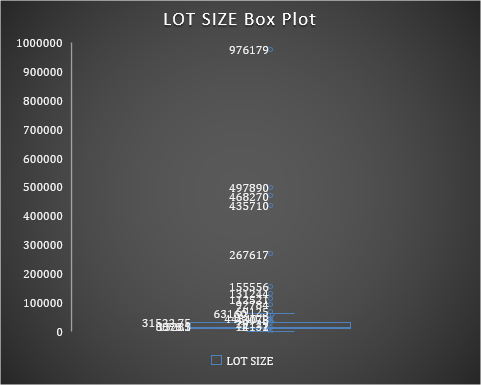
|  |  |
| --- | --- |
| *Observation* | *Value* |
| 56 | 13 |
| 87 | 8 |
| 119 | 8 |



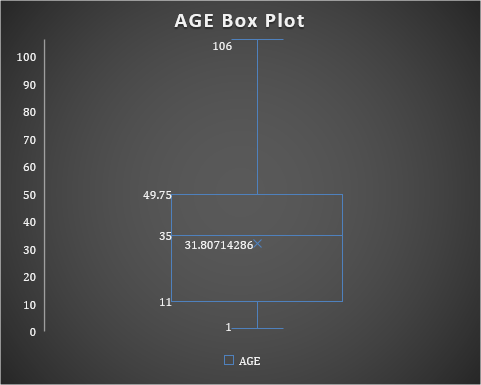
|  |  |
| --- | --- |
| *Observation* | *Value* |
| 56 | 9.25 |
| 137 | 6.75 |



|  |  |
| --- | --- |
| *Observation* | *Value* |
| 183 | 15,975 |
| 70 | 15,360 |
| 146 | 14,140 |
| 175 | 10,088 |
| 144 | 9,116 |



|  |  |
| --- | --- |
| *Observation* | *Value* |
| 12 | 976,179 |
| 9 | 497,890 |
| 113 | 497,890 |
| 16 | 468,270 |
| 89 | 435,710 |



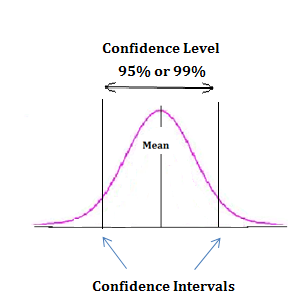
|  |  |
| --- | --- |
| *Observation* | *Values* |
| 35 | 106 |

## OUTLIER SUMMARY

The Z-Scores and Box Plot (IQR) method yield different results. Z-Score method is appropriate for data sets that are close to a normal distribution. IQR method is appropriate for data sets are highly skewed like it is the case of this data set.

## confidence interval

The width of the confidence interval tells us more about how certain we are about the true mean in the population. The confidence interval, the width, is stated as a mean plus or minus the confidence level. The confidence interval is a range of plausible values for the population average. As the confidence level increase, the confidence interval gets wider.



|  |  |
| --- | --- |
| *PRICE* | |
| Mean | 2,047,228.71 |
| Confidence Level(80.0%) | 196666.2495 |
| Lower Limit | 1,850,562.46 |
| Upper Limit | 2,243,894.96 |

We are 80% confident that the true mean falls in between this $1,850,562.46 and $2,243,894.96.

|  |  |
| --- | --- |
| *PRICE* | |
| Mean | 2,047,228.71 |
| Confidence Level(90.0%) | 252904.244 |
| Lower Limit | 1,794,324.46 |
| Upper Limit | 2,300,132.95 |

We are 90% confident that the true mean falls in between this $1,794,324.46 and $2,300,132.95.

|  |  |
| --- | --- |
| *PRICE* | |
| Mean | 2,047,228.71 |
| Confidence Level(99.0%) | 398879.4231 |
| Lower Limit | 1,648,349.28 |
| Upper Limit | 2,446,108.13 |

We are 99% confident that the true mean falls in between this $1,648,349.28 and $2,446,108.13.

# Modeling

## Modeling Techniques

For this project simple and multiple linear regression is the modeling technique will be used. Several explanatory variables will help predict the value of the response variable (PRICE).

## Simple Regression Modeling of Independent Variables

### Scatter Plots

A scatter plot will be used to describe bivariate data (x,y) relationship graphically. The coefficient of determination, R2, represents the proportion of the total sample variability that is explained by the linear relationship between x and y.

The following are scatter plots for the independent variables: BEDS, BATHS, SQUARE FEET, LOT SIZE and AGE.

Figure 12 Simple Linear Regression Model: BEDS versus PRICE

For every 1-unit increase in BEDS, the home PRICE increases by $551,486 dollars when other variables are held fixed. When BEDS is zero, the PRICE is -$257,194. The coefficient of determination is 0.21, thus this regression model explains 21% of the total variation from the original data. The relationship between BEDS and PRICE is moderately strong positive. The correlation coefficient is 0.46.

Figure 13 Simple Linear Regression Model: BATHS versus PRICE

For every 1-unit increase in BATHS, the home PRICE increases by $962,659 dollars when other variables are held fixed. When BATHS is zero, the PRICE is -$0. The coefficient of determination is 0.53, thus this regression model explains 53% of the total variation from the original data. The relationship between BATHS and PRICE is strong to very strong positive. The correlation coefficient is 0.73.

Figure 14 Simple Linear Regression Model: SQUARE FEET versus PRICE

For every 1-unit increase in SQUARE FEET, the home PRICE increased by $636.25 dollars when other variables are held fixed. When SQUARE FEET is zero, the PRICE is -$354,289. The coefficient of determination is 0.78, thus this regression model explains 78% of the total variation from the original data. The correlation coefficient is 0.88.

For every 1-unit increase in LOT SIZE, the home PRICE decreased by $1.61 dollars when other variables are held fixed. When SQUARE FEET is zero, the PRICE is $2,000,000. The coefficient of determination is 0.01, thus this regression model explains 1% of the total variation from the original data. The correlation coefficient is -0.10.

For every 1-unit increase in AGE, the home PRICE decreased by $27,608 dollars when other variables are held fixed. When AGE is zero, the PRICE is $3,000,000. The coefficient of determination is 0.11, thus this regression model explains 11% of the total variation from the original data. The correlation coefficient is -0.34.

The independent variable SQUARE FEET is the best based on R². The coefficient of determination is 0.78, thus this regression model explains 78% of the total variation from the original data.

|  |  |  |  |
| --- | --- | --- | --- |
|  | EQUATION | R² | RANK |
| BEDS | PRICE = 551486 \* BEDS - 257194 | 0.2121 | 3 |
| BATHS | PRICE = 962659 \* BATHS - 1E+06 | 0.5338 | 2 |
| SQUARE FEET | PRICE = 636.25 \* SQ FT - 354289 | 0.782 | 1 |
| LOT SIZE | PRICE = -1.6121 \* LOT SIZE + 2E+06 | 0.0104 | 5 |
| AGE | PRICE = -27608 \* AGE + 3E+06 | 0.1176 | 4 |

### Coefficient of correlation

**Correlations** measure the strength of the relationship, or association, between two variables. The coefficient of correlation, r, is a measurement of strength of the linear relationship between two variables x and y.

Table 16 Correlation Matrix

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | PRICE | BEDS | BATHS | SQUARE FEET | LOT SIZE | AGE |
| PRICE | 1 |  |  |  |  |  |
| BEDS | 0.460521 | 1 |  |  |  |  |
| BATHS | 0.730587 | 0.785073\* | 1 |  |  |  |
| SQUARE FEET | 0.884289 | 0.677023 | 0.859328\* | 1 |  |  |
| LOT SIZE | -0.10203 | -0.21596 | -0.17592 | -0.10598 | 1 |  |
| AGE | -0.34296 | -0.23916 | -0.53527 | -0.39192 | 0.078575 | 1 |

The strongest correlation is between SQUARE FEET and BATHS. The 0.86 value of exceed the recommend threshold of 0.7, so we can assume multicollinearity exist. The second strongest correlation is between BATHS and BEDS. The 0.79 value of exceed the recommend threshold of 0.7, so we can assume multicollinearity exist.

|  |  |  |  |
| --- | --- | --- | --- |
|  | PRICE | CORRELATION INTERPRETATION | RANK |
| BEDS | 0.460521 | Moderately strong positive | 3 |
| BATHS | 0.730587 | Strong to very strong positive | 2 |
| SQUARE FEET | 0.884289 | Very strong positive | 1 |
| LOT SIZE | -0.10203 | No correlation to “negligible” negative | 5 |
| AGE | -0.342959 | Weak to moderately strong negative | 4 |

### Regression summary for the independent variables

## Regression: Model 1

This multiple linear regression, including the dependent variable (PRICE) and ALL independent variables.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *1* | SUMMARY OUTPUT |  |  |  |  |  |  |  |  |
| *2* |  |  |  |  |  |  |  |  |  |
| *3* | *Regression Statistics* | |  |  |  |  |  |  |  |
| *4* | Multiple R | 0.907684894 |  |  |  |  |  |  |  |
| *5* | R Square | 0.823891866 |  |  |  |  |  |  |  |
| *6* | Adjusted R Square | 0.817320667 |  |  |  |  |  |  |  |
| *7* | Standard Error | 772384.2239 |  |  |  |  |  |  |  |
| *8* | Observations | 140 |  |  |  |  |  |  |  |
| *9* |  |  |  |  |  |  |  |  |  |
| *10* | ANOVA |  |  |  |  |  |  |  |  |
| *11* |  | *df* | *SS* | *MS* | *F* | *Significance F* |  |  |  |
| *12* | Regression | 5 | 3.73992E+14 | 7.47984E+13 | 125.3792 | 9.35E-49 |  |  |  |
| *13* | Residual | 134 | 7.99414E+13 | 5.96577E+11 |  |  |  |  |  |
| *14* | Total | 139 | 4.53933E+14 |  |  |  |  |  |  |
| *15* |  |  |  |  |  |  |  |  |  |
| *16* |  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
| *17* | Intercept | 176533.9502 | 300940.3583 | 0.586607762 | 0.558454 | -418673.642 | 771741.5 | -418674 | 771741.5 |
| *18* | BEDS | -404250.142 | 75393.68316 | -5.361856923 | 3.5E-07 | -553365.711 | -255135 | -553366 | -255135 |
| *19* | BATHS | 238137.8466 | 130737.6282 | 1.821494316 | 0.070762 | -20438.402 | 496714.1 | -20438.4 | 496714.1 |
| *20* | SQUARE FEET | 702.2042234 | 51.91558879 | 13.52588384 | 8.02E-27 | 599.5242347 | 804.8842 | 599.5242 | 804.8842 |
| *21* | LOT SIZE | -0.70064854 | 0.589221859 | -1.189108194 | 0.2365 | -1.86602671 | 0.46473 | -1.86603 | 0.46473 |
| *22* | AGE | 4751.700578 | 3734.003385 | 1.272548546 | 0.205382 | -2633.5075 | 12136.91 | -2633.51 | 12136.91 |

Table 17 Regression Summary All Independent Variables

### Summary

From the *Coefficients* section, we see that the model is:

β1: -404250.14: The mean home price *PRICE*(y) decreases by $404,250.14 for every beds unit increase in *BEDS*(X1) when other variables are held fixed.

β2: 238137.85: The mean home price *PRICE*(y) increases by $238,137.85 for every baths unit increase in *BATHS*(X2) when other variables are held fixed.

β3: 702.20: The mean home price *PRICE*(y) increases by $702.20 for every square feet unit increase in *SQUARE FEET*(X3) when other variables are held fixed.

β4: -0.70: The mean home price *PRICE*(y) decrease by $ 0.70 for every lot size unit increase in *LOT* *SIZE*(X4) when other variables are held fixed.

β5: 4751.70: The mean home price *PRICE*(y) increases by $4751.70 for every age unit increase in *AGE*(X5) when other variables are held fixed.

The p-value of the F-Test is 9.35E-49 which is less than 0.05. The model is significant. The coefficient of determination (Adjusted R-Square) is 0.817, thus this regression model explains 81.7% of the total variation from the original data. The remaining variation is due to other factors that were not included in the model. The model error is 596577389388.323.

The p-value for *BEDS* and *SQUARE FEET* are significant. The p-value for *BATHS*, *LOT SIZE* and *AGE* are not significant.

Model 1 the p-value for y-intercept, *BATHS*, *LOT* *SIZE* and *AGE* is greater than 0.05; therefore, these explanatory variables are not significant. The recommendation is to drop *LOT* *SIZE* variable.

Model 1 *BATHS* slope p-value is 0.0707 which is more than 0.05, therefore we fail to reject H0. There is not enough evidence to conclude the β1 ≠ 0. We cannot conclude that the explanatory variable *BATHS* has a linear relationship with outcome variable *PRICE*. The slope is not significant.

Model 1 *LOT SIZE* slope p-value is 0.2365 which is more than 0.05, therefore we fail to reject H0. There is not enough evidence to conclude the β1 ≠ 0. We cannot conclude that the explanatory variable *LOT SIZE* has a linear relationship with outcome variable *PRICE*. The slope is not significant.

Model 1 *AGE* slope p-value is 0.2053 which is more than 0.05, therefore we fail to reject H0. There is not enough evidence to conclude the β1 ≠ 0. We cannot conclude that the explanatory variable *AGE* has a linear relationship with outcome variable *PRICE*. The slope is not significant.

Model 1 *BEDS* slope p-value is 3.50288E-07 which is less than 0.05, therefore we reject H0, and conclude the β1 ≠ 0. The explanatory variable BEDS has a linear relationship with outcome variable *PRICE*. The slope is significant.

Model 1 SQUARE FEET slope p-value is 8.02198E-27 which is less than 0.05, therefore we reject H0, and conclude the β1 ≠ 0. The explanatory variable *SQUARE* *FEET* has a linear relationship with outcome variable *PRICE*. The slope is significant.

## Regression: Model 2

This multiple linear regression, including the dependent variable (PRICE) and BEDS, BATHS, SQUARE FEET and AGE independent variables. The independent variable LOT SIZE was removed.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| SUMMARY OUTPUT |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| *Regression Statistics* | |  |  |  |  |  |  |  |
| Multiple R | 0.906661 |  |  |  |  |  |  |  |
| R Square | 0.822034 |  |  |  |  |  |  |  |
| Adjusted R Square | 0.81676 |  |  |  |  |  |  |  |
| Standard Error | 773567.6 |  |  |  |  |  |  |  |
| Observations | 140 |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |  |  |  |
| Regression | 4 | 3.73E+14 | 9.32871E+13 | 155.8925 | 1.41E-49 |  |  |  |
| Residual | 135 | 8.08E+13 | 5.98407E+11 |  |  |  |  |  |
| Total | 139 | 4.54E+14 |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
| Intercept | 91789.96 | 292828.1 | 0.313460252 | 0.754415 | -487334 | 670913.8 | -487334 | 670913.8 |
| BEDS | -393032 | 74915.67 | -5.24632063 | 5.86E-07 | -541192 | -244871 | -541192 | -244871 |
| BATHS | 246811.7 | 130734 | 1.887892842 | 0.061187 | -11739.8 | 505363.3 | -11739.8 | 505363.3 |
| SQUARE FEET | 696.7583 | 51.79241 | 13.45290282 | 1.06E-26 | 594.3288 | 799.1877 | 594.3288 | 799.1877 |
| AGE | 4696.433 | 3739.434 | 1.255920576 | 0.211314 | -2699.02 | 12091.88 | -2699.02 | 12091.88 |

### Summary

From the *Coefficients* section, we see that the model is:

β1: -393032: The mean home price *PRICE*(y) decreases $393,032.00 for every beds unit increase in *BEDS*(X1) when other variables are held fixed.

β2: 246811.7: The mean home price *PRICE*(y) increases $246,811.70 for every baths unit increase in *BATHS*(X2) when other variables are held fixed.

β3: 696.7583: The mean home price *PRICE*(y) increases $696.76 for every square feet unit increase in *SQUARE FEET*(X3) when other variables are held fixed.

β4: 4696.43: The mean home price *PRICE*(y) increases $4,696.43 for every age unit increase in *AGE*(X4) when other variables are held fixed.

The p-value of the F-Test is 1.41E-49 which is less than 0.05. The model is significant. The coefficient of determination (Adjusted R-Square) is 0.816, thus this regression model explains 81.6% of the total variation from the original data. The remaining variation is due to other factors that were not included in the model. The model error is 598406797475.385.

The p-value for *BEDS* and *SQUARE FEET* are significant. The p-value for *BATHS* and *AGE* are not significant.

Model 2 the p-value for y-intercept, *BATHS* and *AGE* is greater than 0.05; therefore, these explanatory variables are not significant. The recommendation is to drop *AGE* variable.

Model 2 *BATHS* slope p-value is 0.061187 which is more than 0.05, therefore we fail to reject H0. There is not enough evidence to conclude the β1 ≠ 0. We cannot conclude that the explanatory variable *BATHS* has a linear relationship with outcome variable *PRICE*. The slope is not significant.

Model 2 *AGE* slope p-value is 0.211314 which is more than 0.05, therefore we fail to reject H0. There is not enough evidence to conclude the β1 ≠ 0. We cannot conclude that the explanatory variable *AGE* has a linear relationship with outcome variable *PRICE*. The slope is not significant.

Model 2 *BEDS* slope p-value is 5.86E-07 which is less than 0.05, therefore we reject H0, and conclude the β1 ≠ 0. The explanatory variable BEDS has a linear relationship with outcome variable *PRICE*. The slope is significant.

Model 2 SQUARE FEET slope p-value is 1.06E-26 which is less than 0.05, therefore we reject H0, and conclude the β1 ≠ 0. The explanatory variable *SQUARE* *FEET* has a linear relationship with outcome variable *PRICE*. The slope is significant.

## Regression: Model 3

This multiple linear regression, including the dependent variable (PRICE) and BEDS, BATHS, and SQUARE FEET independent variables. The independent variable AGE was removed.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| SUMMARY OUTPUT |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| *Regression Statistics* | |  |  |  |  |  |  |  |
| Multiple R | 0.905513 |  |  |  |  |  |  |  |
| R Square | 0.819954 |  |  |  |  |  |  |  |
| Adjusted R Square | 0.815983 |  |  |  |  |  |  |  |
| Standard Error | 775207.8 |  |  |  |  |  |  |  |
| Observations | 140 |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |  |  |  |
| Regression | 3 | 3.72E+14 | 1.24068E+14 | 206.4545 | 1.97E-50 |  |  |  |
| Residual | 136 | 8.17E+13 | 6.00947E+11 |  |  |  |  |  |
| Total | 139 | 4.54E+14 |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
| Intercept | 349950.1 | 208998.2 | 1.674416489 | 0.096347 | -63356.6 | 763256.8 | -63356.6 | 763256.8 |
| BEDS | -360171 | 70346.96 | -5.11992362 | 1.02E-06 | -499286 | -221056 | -499286 | -221056 |
| BATHS | 160374 | 111386.9 | 1.439791974 | 0.152224 | -59900.4 | 380648.4 | -59900.4 | 380648.4 |
| SQUARE FEET | 707.4996 | 51.18968 | 13.82113744 | 1.09E-27 | 606.2689 | 808.7304 | 606.2689 | 808.7304 |

### Summary

From the *Coefficients* section, we see that the model is:

β1: -360171: The mean home price *PRICE*(y) decreases $36,0171.00 for every beds unit increase in *BEDS*(X1) when other variables are held fixed.

β2: 160374: The mean home price *PRICE*(y) increases $160,374.00 for every baths unit increase in *BATHS*(X2) when other variables are held fixed.

β3: 707.4996: The mean home price *PRICE*(y) increases $707.50 for every square feet unit increase in *SQUARE FEET*(X3) when other variables are held fixed.

The p-value of the F-Test is 1.97E-50 which is less than 0.05. The model is significant. The coefficient of determination (Adjusted R-Square) is 0.815, thus this regression model explains 81.5% of the total variation from the original data. The remaining variation is due to other factors that were not included in the model. The model error is 600947106902.356.

The p-value for *BEDS* and *SQUARE FEET* are significant. The p-value for *BATHS* is not significant.

Model 3 the p-value for y-intercept and *AGE* is greater than 0.05; therefore, these explanatory variables are not significant. The recommendation is to drop *BATHS* variable.

Model 3 *BATHS* slope p-value is 0.152224 which is more than 0.05, therefore we fail to reject H0. There is not enough evidence to conclude the β1 ≠ 0. We cannot conclude that the explanatory variable *BATHS* has a linear relationship with outcome variable *PRICE*. The slope is not significant.

Model 3 *BEDS* slope p-value is 1.02E-06 which is less than 0.05, therefore we reject H0, and conclude the β1 ≠ 0. The explanatory variable BEDS has a linear relationship with outcome variable *PRICE*. The slope is significant.

Model 3 SQUARE FEET slope p-value is 1.09E-27 which is less than 0.05, therefore we reject H0, and conclude the β1 ≠ 0. The explanatory variable *SQUARE* *FEET* has a linear relationship with outcome variable *PRICE*. The slope is significant.

## Regression: Model 4

This multiple linear regression, including the dependent variable (PRICE) and BEDS and SQUARE FEET independent variables. The independent variable BATHS was removed.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| SUMMARY OUTPUT | |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| *Regression Statistics* | |  |  |  |  |  |  |  |
| Multiple R | 0.903997 |  |  |  |  |  |  |  |
| R Square | 0.81721 |  |  |  |  |  |  |  |
| Adjusted R Square | 0.814541 |  |  |  |  |  |  |  |
| Standard Error | 778237.6 |  |  |  |  |  |  |  |
| Observations | 140 |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |  |  |  |
| Regression | 2 | 3.71E+14 | 1.85479E+14 | 306.2466 | 2.78E-51 |  |  |  |
| Residual | 137 | 8.3E+13 | 6.05654E+11 |  |  |  |  |  |
| Total | 139 | 4.54E+14 |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
| Intercept | 453124.6 | 197096.7 | 2.298996293 | 0.023018 | 63379.41 | 842869.7 | 63379.41 | 842869.7 |
| BEDS | -305467 | 59435.43 | -5.13947631 | 9.3E-07 | -422996 | -187938 | -422996 | -187938 |
| SQUARE FEET | 760.5004 | 35.71011 | 21.29649972 | 2.65E-45 | 689.8861 | 831.1147 | 689.8861 | 831.1147 |

### Summary

From the *Coefficients* section, we see that the model is:

β1: -305467: The mean home price *PRICE*(y) decreases $305467for every beds unit increase in *BEDS*(X1) when other variables are held fixed.

β2: 760.5004: The mean home price *PRICE*(y) increases $760.50 for every square feet unit increase in *SQUARE FEET*(X2) when other variables are held fixed.

The p-value of the F-Test is 2.78E-51 which is less than 0.05. The model is significant. The coefficient of determination (Adjusted R-Square) is 0.815, thus this regression model explains 81.5% of the total variation from the original data. The remaining variation is due to other factors that were not included in the model. The model error is 605653798897.606.

The p-value for *BEDS* and *SQUARE FEET* are significant.

Model 4 *BEDS* slope p-value is 9.3E-07 which is less than 0.05, therefore we reject H0, and conclude the β1 ≠ 0. The explanatory variable BEDS has a linear relationship with outcome variable *PRICE*. The slope is significant.

Model 4 SQUARE FEET slope p-value is 2.65E-45 which is less than 0.05, therefore we reject H0, and conclude the β1 ≠ 0. The explanatory variable *SQUARE* *FEET* has a linear relationship with outcome variable *PRICE*. The slope is significant.

### Outliers removal

For Model 4, it has 4 Normal outliers,3 Extreme outliers and 140 Observations. Those outliers were removed, and the model regression was repeated.

|  |  |  |
| --- | --- | --- |
| *Observation* | *Standard Residuals* | *Note* |
| 56 | -3.34362 | *Extreme* |
| 112 | -3.31407 | *Extreme* |
| 126 | -2.98841 | *Normal* |
| 64 | 2.061524 | *Normal* |
| 54 | 2.432643 | *Normal* |
| 48 | 2.592823 | *Normal* |
| 137 | 5.080406 | *Extreme* |

## Regression: Model 5

This multiple linear regression, including the dependent variable (PRICE) and BEDS and SQUARE FEET independent variables. The outliers from Model 4 have been removed.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| SUMMARY OUTPUT | |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| *Regression Statistics* | |  |  |  |  |  |  |  |
| Multiple R | 0.926676 |  |  |  |  |  |  |  |
| R Square | 0.858729 |  |  |  |  |  |  |  |
| Adjusted R Square | 0.856555 |  |  |  |  |  |  |  |
| Standard Error | 525374.1 |  |  |  |  |  |  |  |
| Observations | 133 |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |  |  |  |
| Regression | 2 | 2.18E+14 | 1.09057E+14 | 395.1078 | 5.67E-56 |  |  |  |
| Residual | 130 | 3.59E+13 | 2.76018E+11 |  |  |  |  |  |
| Total | 132 | 2.54E+14 |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
| Intercept | 32267.44 | 149477.2 | 0.21586864 | 0.829428 | -263455 | 327990.2 | -263455 | 327990.2 |
| BEDS | -154666 | 45746.7 | -3.38091804 | 0.000954 | -245170 | -64161.5 | -245170 | -64161.5 |
| SQUARE FEET | 700.8764 | 30.03259 | 23.33719837 | 2.55E-48 | 641.4605 | 760.2923 | 641.4605 | 760.2923 |

### Summary

From the *Coefficients* section, we see that the model is:

β1: -154666: The mean home price *PRICE*(y) decreases $154,666.00 for every beds unit increase in *BEDS*(X1) when other variables are held fixed.

β2: 700.8764: The mean home price *PRICE*(y) increases $700.88 for every square feet unit increase in *SQUARE FEET*(X2) when other variables are held fixed.

The p-value of the F-Test is 5.67E-56 which is less than 0.05. The model is significant. The coefficient of determination (Adjusted R-Square) is 0.857, thus this regression model explains 85.7% of the total variation from the original data. The remaining variation is due to other factors that were not included in the model. The model error is 276017899964.997.

The p-value for *BEDS* and *SQUARE FEET* are significant. The p-value for *y-intercept* is not significant.

Model 5 *BEDS* slope p-value is 0.000954 which is less than 0.05, therefore we reject H0, and conclude the β1 ≠ 0. The explanatory variable BEDS has a linear relationship with outcome variable *PRICE*. The slope is significant.

Model 5 SQUARE FEET slope p-value is 2.55E-48 which is less than 0.05, therefore we reject H0, and conclude the β1 ≠ 0. The explanatory variable *SQUARE* *FEET* has a linear relationship with outcome variable *PRICE*. The slope is significant.

## Model 4 AND Model 5 COMPARISON

Model 4 is significant. The coefficient of determination (Adjusted R-Square) is 0.81, thus this regression model explains 81% of the total variation from the original data. The model 5 is significant. The coefficient of determination (Adjusted R-Square) is 0.85, thus this regression model explains 85% of the total variation from the original data. Removing outliers did affect the model as the Adjusted R-Square increased. It is not always acceptable to drop an observation just because it is an outlier.

## Multiple Regression Modeling comparison

### Models

The following is a list of all models. Independent variables are identified as X and dependent variables are identified as Y for each model.

Table 18 Multiple Regression Models

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Possible Models | | | | | | | |
|  |  | | | |  |  |  |
|  | PRICE | BEDS | BATHS | SQUARE FEET | LOT SIZE | AGE |  |
| Model 1 | Y | X | X | X | X | X | All explanatory variables |
| Model 2 | Y | X | X | X |  | X | Explanatory variable LOT SIZE removed. |
| Model 3 | Y | X | X | X |  |  | Explanatory variable AGE removed. |
| Model 4 | Y | X |  | X |  |  | Explanatory variable BATHS removed. |
| Model 5 | Y | X |  | X |  |  | Outlier from Model 4 removed. |

### Comparison Tables for Model Significance, Parameter Significance, model goodness of fit

Output 1 compares the model significance for each of the models. A model is significant if the p-value is less than 0.05

Table 19 Models Significance Comparison

|  |  |  |  |
| --- | --- | --- | --- |
| Output 1: Model Significance | | | |
|  | F-Test | P-Value | Significant |
| Model 1 | 125.3792 | 9.35E-49 | Yes |
| Model 2 | 155.8925 | 1.41E-49 | Yes |
| Model 3 | 206.4545 | 1.97E-50 | Yes |
| Model 4 | 306.2466 | 2.78E-51 | Yes |
| Model 5 | 395.1078 | 5.67E-56 | Yes |

Model 1 significance (F) p-value is 9.34E-49 which is less than 0.05, the model is significant.

Model 2 significance (F) p-value is 1.41E-49 which is less than 0.05, the model is significant.

Model 3 significance (F) p-value is 1.97E-50 which is less than 0.05, the model is significant.

Model 4 significance (F) p-value is 2.78E-51 which is less than 0.05, the model is significant.

Model 5 significance (F) p-value is 5.67E-56 which is less than 0.05, the model is significant.

Output 2 compares the parameter significance for each of the models. A parameter is significant if the p-value is less than 0.05.

Table 20 Models Parameter Significance Comparison

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Output 2: Parameter Significance | | | | | | |
|  | Y-int | BEDS | BATHS | SQUARE FEET | LOT SIZE | AGE |
| Model 1 | 0.558454099 | 3.50288E-07 | 0.070762 | 8.02198E-27 | 0.2365 | 0.205382 |
| Model 2 | 0.754415 | 5.86E-07 | 0.061186573 | 1.06E-26 |  | 0.211314 |
| Model 3 | 0.096347 | 1.02E-06 | 0.152224 | 1.09E-27 |  |  |
| Model 4 | 0.023018 | 9.3E-07 |  | 2.65E-45 |  |  |
| Model 5 | 0.829428 | 0.000954 |  | 2.55E-48 |  |  |

Output 3 compares the goodness of fit for each of the models. The models are ranked according to the adjusted R Square. The goodness of fit is ranked from 1 to 5, where 5 is the lowest value and 1 is the highest.

Table 21 Models Goodness of Fit Comparison

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Output 3: Adjusted R-Square | | | | |
|  | R-Square | Adj. R-Square | Highest |  |
| Model 1 | 0.823891866306518 | 0.817320667288105 | 2 |  |
| Model 2 | 0.821763612913881 | 0.816482534777996 | 3 | The adjusted R-Square decrease slightly |
| Model 3 | 0.819954204481669 | 0.815982606051118 | 4 | The adjusted R-Square decrease slightly |
| Model 4 | 0.81720982873769 | 0.814541359084225 | 5 | The adjusted R-Square decrease slightly |
| Model 5 | 0.8587287442152 | 0.856555340280049 | 1 | The adjusted R-Square increased. |

### Best Model Selection

Table 22 Best Model

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Best Model | | | | |
|  | Output 1: Model Significance | Output 2: Parameter Significance | Output 3: Adjusted R-Square | Notes |
| Model 1 | Yes | BEDS and SQUARE FEET | 0.817320667288105 | Only two explanatory variables are significant. |
| Model 2 | Yes | BEDS and SQUARE FEET | 0.816482534777996 | Slight decrease for adjusted R-Square. Only two explanatory variables are significant. |
| Model 3 | Yes | BEDS and SQUARE FEET | 0.815982606051118 | Slight decrease for adjusted R-Square. Only two explanatory variables are significant. |
| Model 4 | Yes | BEDS and SQUARE FEET | 0.814541359084225 | Slight decrease for adjusted R-Square. All explanatory variables and y-intercept are significant. |
| Model 5 | Yes | BEDS and SQUARE FEET | 0.856555340280049 | Increase for adjusted R-Square. BEDS and SQUARE FEET variables are significant, but y-intercept is no significant. |

We selected Model 5 as the best model because the model and explanatory parameters are significant and the Adjusted R Squared is the highest.

Table 23 Model 4 Outliers

## Discussion for Final Model

Figure 15 PRICE Normal Probability Plot

Figure 16 BEDS Normal Probability Plot

Figure 17 SQUARE FEET Normal Probability Plot

Figure 18 Residuals Plot

The mean of zero assumption has been meet. The average is approximately zero.

The variance constancy and independence assumption has not been meet. As the predicted value increases so does the residual.

Figure 19 Residual Plots

Figure 20 Final Model BEDS Residual Plot

Figure 21 Final Model SQUARE FEET Residual Plot

In the case of normality, there appears to be a straight linear line, but we see some slight formations of tails on one end. Therefore, we conclude that for the most part our model appears to be valid. Our model appears to be partially valid. The assumption check reveals that for the most part the requirements of mean zero and normality are met.

## Conclusion

Based on the modeling I was able to conclude that the most parsimonious model had both BEDS and SQUARE FEET as the independent variables. This model had the highest F-test value and the lowest p-value. The model and all the independent variables are significant, and the adjusted r squared is the highest.

This model can be beneficial to prospective home buyers/sellers, real estate agents, and investors just to name a few. One of the benefit of this model is its simplicity. Anyone can roughly estimate the price of a home that is selling by just specifying the number of bedrooms and square feet. One of the drawback of this model is the number of independent variables. Other factors such as zip code, schools, neighborhood, views, etc are not being considered. Another drawback is the fact that the listed PRICE is the value the home is selling for. This number is subjective based on what the homeowner selling his/her own thinks his home is worth. Some potential future areas of study that I am interested is in adding a few more variables such as proximity to public schools, major medical centers, parks and green areas, neighbor, etc.

This project was beneficial in helping continue to master the concept of simple and multiple linear regression. One of the difficulty I faced was figuring out the best way to document the analysis, so it was accurate and understandable by any audience.

# Appendix

The following table contains the data set used to

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ID | PRICE | BEDS | BATHS | SQUARE FEET | LOT SIZE | AGE |
| 1 | 684500 | 2 | 2 | 1730 | 2912 | 39 |
| 2 | 449000 | 1 | 1 | 810 | 74992 | 38 |
| 3 | 2298000 | 5 | 3 | 3530 | 11984 | 63 |
| 4 | 2848000 | 3 | 3.25 | 3600 | 16259 | 26 |
| 5 | 649800 | 3 | 1.75 | 1380 | 9600 | 61 |
| 6 | 2500000 | 4 | 3 | 2857 | 37445 | 75 |
| 7 | 369900 | 2 | 1 | 1105 | 497890 | 43 |
| 8 | 1400000 | 3 | 2 | 2460 | 9799 | 63 |
| 9 | 949950 | 4 | 2.5 | 2740 | 11160 | 44 |
| 10 | 750000 | 3 | 3 | 2437 | 976179 | 45 |
| 11 | 2689950 | 5 | 4 | 4304 | 9450 | 1 |
| 12 | 850000 | 3 | 2.25 | 2120 | 4192 | 19 |
| 13 | 2350000 | 6 | 4 | 5780 | 20908 | 30 |
| 14 | 359000 | 2 | 1.75 | 1001 | 468270 | 39 |
| 15 | 888000 | 3 | 1.75 | 1919 | 8297 | 50 |
| 16 | 655000 | 2 | 1.75 | 1351 | 155556 | 20 |
| 17 | 799000 | 3 | 1.75 | 1640 | 10399 | 51 |
| 18 | 549000 | 2 | 2.25 | 1750 | 2634 | 32 |
| 19 | 1498000 | 5 | 3.25 | 4560 | 18750 | 51 |
| 20 | 999950 | 4 | 3 | 3400 | 13062 | 36 |
| 21 | 899000 | 3 | 2.5 | 1970 | 3502 | 18 |
| 22 | 949950 | 6 | 3 | 2300 | 13503 | 63 |
| 23 | 1490000 | 3 | 1.75 | 2020 | 8515 | 57 |
| 24 | 998000 | 4 | 2.5 | 2570 | 11375 | 45 |
| 25 | 950000 | 3 | 2.75 | 2330 | 36671 | 52 |
| 26 | 649950 | 3 | 1.75 | 1400 | 15000 | 106 |
| 27 | 835000 | 3 | 2 | 1440 | 8400 | 50 |
| 28 | 2836000 | 4 | 3.5 | 4800 | 10634 | 9 |
| 29 | 685000 | 3 | 2 | 1430 | 7464 | 55 |
| 30 | 765000 | 3 | 1.75 | 1250 | 8190 | 52 |
| 31 | 725000 | 1 | 1.5 | 1008 | 102684 | 10 |
| 32 | 979000 | 4 | 2.5 | 2570 | 9760 | 49 |
| 33 | 2895000 | 5 | 4.25 | 5050 | 15120 | 1 |
| 34 | 850000 | 5 | 2.75 | 2690 | 15908 | 50 |
| 35 | 1325000 | 6 | 3.5 | 4140 | 16491 | 19 |
| 36 | 2450000 | 4 | 4 | 4850 | 9123 | 16 |
| 37 | 829950 | 3 | 2.5 | 2200 | 7000 | 44 |
| 38 | 950000 | 4 | 2.75 | 2270 | 7700 | 52 |
| 39 | 1988000 | 5 | 4 | 4381 | 7890 | 1 |
| 40 | 1998000 | 5 | 3.25 | 3770 | 16362 | 42 |
| 41 | 1800000 | 6 | 3.5 | 4600 | 8610 | 22 |
| 42 | 1288000 | 5 | 4.25 | 3770 | 10949 | 38 |
| 43 | 769990 | 3 | 3.25 | 2002 | 1 | 1 |
| 44 | 2785950 | 5 | 4.5 | 4586 | 11408 | 1 |
| 45 | 1498000 | 4 | 3.25 | 4380 | 9186 | 19 |
| 46 | 2354000 | 4 | 2.75 | 2288 | 10347 | 88 |
| 47 | 892000 | 3 | 2.75 | 2040 | 8636 | 39 |
| 48 | 4800000 | 5 | 3.75 | 5090 | 30539 | 36 |
| 49 | 1374950 | 4 | 2.75 | 2260 | 14780 | 48 |
| 50 | 3588888 | 6 | 6 | 5701 | 10230 | 1 |
| 51 | 1208000 | 2 | 2.5 | 1751 | 105864 | 10 |
| 52 | 1088000 | 3 | 2.5 | 3180 | 10497 | 33 |
| 53 | 549950 | 2 | 1.75 | 1320 | 2175 | 32 |
| 54 | 3500000 | 3 | 3 | 2740 | 52307 | 34 |
| 55 | 3988800 | 5 | 5.5 | 5489 | 15666 | 1 |
| 56 | 5580000 | 13 | 9.25 | 15360 | 82328 | 18 |
| 57 | 5980000 | 5 | 5.75 | 7594 | 69125 | 25 |
| 58 | 1998888 | 3 | 1.5 | 2520 | 11273 | 61 |
| 59 | 2460000 | 4 | 2.5 | 4130 | 112521 | 40 |
| 60 | 1195000 | 4 | 2.25 | 2812 | 17411 | 39 |
| 61 | 2499800 | 4 | 4.25 | 5360 | 24515 | 21 |
| 62 | 339995 | 2 | 1.75 | 1254 | 22389 | 37 |
| 63 | 975000 | 2 | 1.75 | 1474 | 43078 | 16 |
| 64 | 3900000 | 6 | 6.25 | 4848 | 8548 | 1 |
| 65 | 2949995 | 6 | 4.5 | 5761 | 16511 | 5 |
| 66 | 1650000 | 4 | 3.75 | 3914 | 41444 | 19 |
| 67 | 2198800 | 5 | 3.25 | 3850 | 23172 | 17 |
| 68 | 2988000 | 6 | 4.25 | 5130 | 9045 | 1 |
| 69 | 3800000 | 5 | 5 | 4568 | 8570 | 1 |
| 70 | 3078950 | 5 | 5 | 4998 | 13500 | 1 |
| 71 | 305000 | 2 | 1 | 889 | 435710 | 40 |
| 72 | 2649950 | 5 | 5 | 4309 | 9137 | 1 |
| 73 | 707990 | 2 | 1.75 | 1360 | 1 | 1 |
| 74 | 3198000 | 5 | 5 | 6200 | 20056 | 17 |
| 75 | 1988000 | 3 | 3 | 3390 | 19833 | 47 |
| 76 | 1198888 | 4 | 3 | 2620 | 17418 | 50 |
| 77 | 675000 | 4 | 1.75 | 1700 | 8640 | 63 |
| 78 | 3588888 | 5 | 4.5 | 5102 | 15300 | 1 |
| 79 | 3498000 | 6 | 6 | 6389 | 9491 | 1 |
| 80 | 4588000 | 6 | 5.25 | 8277 | 45100 | 9 |
| 81 | 2788880 | 5 | 4.25 | 4397 | 9572 | 1 |
| 82 | 689888 | 2 | 2 | 1558 | 4930 | 39 |
| 83 | 2350000 | 5 | 4.5 | 3800 | 10000 | 1 |
| 84 | 985000 | 4 | 2.5 | 2540 | 8712 | 35 |
| 85 | 1098888 | 5 | 2.5 | 2420 | 8395 | 60 |
| 86 | 950000 | 4 | 3.75 | 2750 | 9315 | 40 |
| 87 | 1549000 | 8 | 4.75 | 3470 | 9800 | 51 |
| 88 | 1799000 | 4 | 2.5 | 3560 | 35719 | 42 |
| 89 | 365000 | 2 | 1 | 1012 | 497890 | 43 |
| 90 | 2495000 | 5 | 4.5 | 4645 | 8602 | 1 |
| 91 | 4988000 | 5 | 4.75 | 6500 | 10500 | 43 |
| 92 | 3499000 | 5 | 4.5 | 7950 | 26729 | 13 |
| 93 | 2725000 | 5 | 3.5 | 5030 | 20007 | 54 |
| 94 | 750000 | 6 | 2.75 | 2480 | 8732 | 54 |
| 95 | 1695000 | 4 | 3.5 | 4987 | 31623 | 19 |
| 96 | 908000 | 4 | 2.75 | 2270 | 8666 | 40 |
| 97 | 530000 | 2 | 2 | 978 | 102355 | 29 |
| 98 | 1999000 | 5 | 2.5 | 3360 | 10311 | 36 |
| 99 | 1690000 | 4 | 2.25 | 2340 | 11275 | 62 |
| 100 | 698000 | 5 | 2.75 | 3006 | 267617 | 35 |
| 101 | 2698000 | 5 | 4.5 | 4400 | 15580 | 15 |
| 102 | 1399988 | 4 | 3 | 3580 | 9845 | 11 |
| 103 | 1500000 | 4 | 2.75 | 2140 | 21930 | 57 |
| 104 | 3250000 | 5 | 5 | 6494 | 11945 | 11 |
| 105 | 1998000 | 3 | 2.75 | 3580 | 17182 | 48 |
| 106 | 379800 | 2 | 1.5 | 1018 | 92784 | 48 |
| 107 | 1479800 | 5 | 4.5 | 3770 | 5667 | 1 |
| 108 | 2750000 | 4 | 4.25 | 6340 | 8740 | 10 |
| 109 | 1350000 | 3 | 2.5 | 3030 | 8662 | 31 |
| 110 | 3699000 | 5 | 5.25 | 6348 | 12210 | 9 |
| 111 | 891990 | 3 | 3 | 1796 | 1 | 1 |
| 112 | 3298000 | 5 | 4.5 | 9116 | 48787 | 63 |
| 113 | 1049000 | 4 | 2.5 | 2150 | 7707 | 43 |
| 114 | 9988000 | 5 | 5.75 | 14140 | 71936 | 15 |
| 115 | 2490000 | 6 | 4.5 | 5400 | 10500 | 8 |
| 116 | 3898000 | 5 | 4.25 | 5450 | 14132 | 11 |
| 117 | 799000 | 3 | 2 | 1250 | 7585 | 53 |
| 118 | 1649995 | 6 | 3.25 | 4194 | 8823 | 1 |
| 119 | 1239000 | 8 | 3.25 | 3820 | 9223 | 60 |
| 120 | 4500000 | 4 | 3 | 5060 | 47153 | 34 |
| 121 | 1649999 | 5 | 3.75 | 3982 | 34830 | 56 |
| 122 | 2100000 | 4 | 3.25 | 4881 | 47916 | 35 |
| 123 | 1649000 | 5 | 3 | 3970 | 24011 | 38 |
| 124 | 625000 | 3 | 1.75 | 1180 | 12400 | 52 |
| 125 | 2398000 | 5 | 4.5 | 5300 | 33150 | 48 |
| 126 | 2745000 | 5 | 5.25 | 8058 | 63160 | 42 |
| 127 | 1850000 | 4 | 3.25 | 4070 | 9768 | 20 |
| 128 | 1250000 | 5 | 4.25 | 2570 | 31222 | 53 |
| 129 | 1575000 | 4 | 3.5 | 2768 | 20971 | 28 |
| 130 | 769000 | 3 | 1.75 | 1680 | 18000 | 59 |
| 131 | 6888000 | 6 | 6.5 | 10088 | 131244 | 17 |
| 132 | 1495555 | 4 | 3.25 | 2700 | 4325 | 1 |
| 133 | 3488888 | 5 | 5 | 6369 | 35718 | 1 |
| 134 | 3188888 | 5 | 4 | 5927 | 34939 | 1 |
| 135 | 1898000 | 3 | 2.5 | 2750 | 29129 | 88 |
| 136 | 1300000 | 4 | 4 | 3240 | 22651 | 22 |
| 137 | 15000000 | 5 | 6.75 | 15975 | 50397 | 13 |
| 138 | 1988888 | 4 | 2 | 2000 | 41831 | 62 |
| 139 | 2880000 | 6 | 5.5 | 5970 | 16206 | 3 |
| 140 | 949000 | 4 | 3 | 2906 | 8263 | 34 |