U

ASSESSMENT COVERSHEET

Attach this coversheet as the cover of your submission. All sections must be completed.

Section A: Submission Details

Programme	:	BACHELOR OF INFORMATION TECHNOLOGY (HONS.) IN SOFTWARE ENGINEERING
Course Code & Name	:	IGB20303 PROBABILITY AND STATISTICS FOR IT
Course Lecturer(s)	:	NURASHIKIN SAALUDIN
Submission Title	:	ARTICLE / REPORT: HOUSE PRICE ANALYSIS
Deadline	:	Day 30 Month 05 Year 2022 Time 11:00PM

Penalties :

- 5% will be deducted per day to a maximum of four (4) working days, after which the submission will **not** be accepted.
- Plagiarised work is an Academic Offence in University Rules & Regulations and will be penalised accordingly.

Section B: Academic Integrity

Tick (√) each box below if you agree:

☑ I have read and understood the UniKL's policy on Plagiarism in University Rules & Regulations.

☑ This submission is my own, unless indicated with proper referencing.

☑ This submission has not been previously submitted or published.

☑ This submission follows the requirements stated in the course.

Section C: Submission Receipt

(must be filled in manually)

Office Receipt of Submission

Date & Time of Submission (stamp)	Student Name(s)	Student ID(s)
30.05.2022, 11:00PM	NUR ALISA ZARINA BINTI NAZMI (L01)	52213121129
	MUHAMMAD AMIN BIN SHAMSUL ANUAR (L01)	52213121098
	MUHAMMAD AMIR QAYYUM BIN SUHAIMI (L01)	52213121254
	ARISA NURFARINA BINTI ISMA ZAKI (L02)	52213121149
	MUHAMMAD AFIQ HAIKAL BIN ISMAIL (L01)	52213121135
	AMAMRA YAHIA MOUNIB (L01)	52213221164

Student Receipt of Submission

This is your submission receipt, the only accepted evidence that you have submitted your work. After this is stamped by the appointed staff & filled in, cut along the dotted lines above & retain this for your record.

Date & Time of Submission (stamp)	Course Code	Submission Title	Student ID(s) & Signature(s)
30.05.2022, 11:00PM	IGB20303	ARTICLE / REPORT: HOUSE PRICE ANALYSIS	52213121129 52213121098 52213121254 52213121149 52213121135 52213221164

House Price Analysis

Name of Authors (with Student ID and Group):

1. NUR ALISA ZARINA BINTI NAZMI (52213121129) - L01 / Leader

2. MUHAMMAD AMIN BIN SHAMSUL ANUAR (52213121098) - L01

3. MUHAMMAD AMIR QAYYUM BIN SUHAIMI (52213121254) - L01

4. ARISA NURFARINA BINTI ISMA ZAKI (52213121149) - L02

5. MUHAMMAD AFIQ HAIKAL BIN ISMAIL (52213121135) - L01

6. AMAMRA YAHIA MOUNIB (52213221164) – L01

Abstract

Buying a house is most often a part of an average person's financial planning. They can be inexperienced amateurs with limited information about the market. In the future, there will be many more people to come who wish to delve and study house prices. To help these people, it is important for us to know how to analyse the data and what could possibly affect the prices. From there we would even be able to see a trend in the house price marketing. The real estate markets are examples of mentioned data analysis opportunities. This paper studies the data of house prices of the overall cities in Washington, USA. The objective of this paper is to provide statistical analysis and predictive inference of house prices. We will also determine the significant differences of the prices between number of bedrooms, number of bathrooms, square feet lot, and square feet living using multiple linear regression. At the end of the study, the F-statistic of the predictive model is large while the p-value is extremely small, resulting in a rejection of the null hypothesis. Therefore, we concluded that the square feet of living space and the amount of bathrooms will affect the house prices.

Keywords: House Price; Correlation; Multiple Linear Regression; Analysis

1 Introduction

1.0 Research Background

A house is one of the essential requirements of life. It keeps us safe from harsh weather such as rain, sunlight and storms and fills us with unforgettable memories. The rising prices of residential properties worry a ton of residents. People pay a fortune to buy their Dream House. Due to a lack of proper framework, prices have surged and thus the development of negative sentiment in the market. This is a concerning issue for many individuals as if not handled, buying a house will become impossible for many citizens.

Although their intentions are different, the parties who want to buy houses enter the real estate market and may benefit from increases in a property's value. Property value is related directly to the housing ownership ratio. Especially in developing countries, whether the high rate of housing ownership is sustainable is discussed in the literature. The main question of the sustainability of the housing market is affordability. Housing cannot be sustainable unless it is low-priced and cost-effective. Whether housing is cost-effective also affects the sustainability of its use as an investment tool. The real estate market is more stable than volatile financial markets such as foreign exchanges, interest rates, and the stock market.

In the real estate sector, which has become a very profitable investment tool, especially in the last 15 years, housing prices determine the profitability of the sector. At this point, the determination of housing prices has been one of the most important subtopics of the sector. This topic has prompted many market players, from residential investors to real estate investment trusts and from individual investors to government officials, to predict the movement of housing prices, and they use a variety of methods for this.

Our study obtains the data from Kaggle.com (https://www.kaggle.com/datasets/shree1992/housedata) which is titled House Price Prediction. The author of this data set is Shree.

The data includes 4 components of a house that we will be using to draw a correlation between these components and how it affects house prices. We will be using several methods with one of them being the Descriptive Statistics analysis which is the type of analysis of data that helps describe, show, or summarize data points in a constructive way such that patterns might emerge that fulfil every condition of the data. We aim to determine the housing factor that mostly affect its prices, and hopefully help potential buyers to make informed purchasing decisions. Thus, eliminating surge gains and promoting a healthy market.

1.1 Research Objectives and Scope

This research aims to provide statistical analysis and predictive inference of house prices.

To achieve the aforementioned research aim, the following research objectives are established as follows:

1. To identify the independent variables influencing house prices.
2. To identify if any correlation exists between the number of bedrooms, number of bathrooms, total square feet of the house lot and total square feet of living space of the house-on-house prices.
3. To determine the significant difference between the number of bedrooms, number of bathrooms, total square feet of the house lot and total square feet of living space of the house-on-house prices.
4. To determine the most significant correlation coefficient in relation to house prices.
5. To construct a predictive model for house prices.

2 Methodology/Materials

2.1 Data Description

This data set whose top contributors are Oh Seok Kim and Yash Patel contains information of more than 4000 houses in Washington, United States of America recorded throughout the three months of 2014 in May, June, and July. The comma-separated values (.csv) file for the data set was accessed using Excel, and it contained 18 columns and 4600 rows of data. Keep in mind that the rows referenced in this study directly reflect the .csv file's rows of data after subtracting the row number with 1. The reason for this is to account for the column names taking up the first row, offsetting the rest of the data rows by 1.

The 18 columns represent the 18 variables to potentially be used in this study, which are as follows:

Table 1: Variable and Description of Data

#	Variable	Description
1	date	Refers to the date, displayed in "DD/MM/YYYY" format, and time the information of the houses was registered. e.g.: *2/5/2014 0:00*
2	price	Refers to the pricing of the houses in USD currency. e.g.: *313000*
3	bedrooms	Refers to the number of bedrooms inside the corresponding house. e.g.: *3*
4	bathrooms	Refers to the number of bathrooms inside the corresponding house. * e.g.: *1.5*, *2.25*
5	sqft_living	Refers to the total square feet of living space of the house, including the above ground and below ground spaces. e.g.: *1340*
6	sqft_lot	Refers to the total square feet of the house lot. e.g.: *7912*
7	floors	Refers to the number of floors of the house. ** e.g.: *1.5*
8	waterfront	Refers to whether the house is situated next to an area of water, such as the sea. There are only two: *0* or *1*
9	view	Refers to the rating of the view seen from the house. e.g.: *4*
10	condition	Refers to the condition of the house, such as its appearance and quality. e.g.: *3*

11	sqft_above	Refers to the total square feet of living space of the above ground of the house. e.g: *1340*
12	sqft_basement	Refers to the total square feet of living space below ground of the house. e.g: *0*, *280*
13	yr_built	Refers to the year the house was first built. e.g.: *1955*
14	yr_renovate	Refers to the year of the house's most recent renovation, if any. e.g.: *2005*
15	street	Refer to the street address of the house. e.g.: *18810 Densmore Ave N*
16	city	Refers to the city the house is located in. e.g.: *Shoreline*
17	statezip	Refers to the state zip of the house. e.g.: *WA 98133*
18	country	Refers to the country the house is located in. e.g.: *USA*

* The decimals in the numbers of bathrooms are determined by the completeness of the bathroom(s). According to Lisa Johnson Mandell (2022) and Allanah Dykes (2020), a full, whole bathroom must contain four key fixtures which are the toilet, sink, bathtub, and a shower. A half bathroom only comes with two of these fixtures, typically a toilet and a sink. A quarter bathroom follows the same principle where it comes with only one, sometimes as a simple shower stall one can find near public swimming pools.

** The decimals in the numbers of floors are determined by a characteristic of the additional floor. For example, Alan F MacDonald (2020) wrote in his article that if a house had its second floor smaller than the main floor, usually half the size, and does not sit inside the roof of the main floor, then the house is considered a 1.5-storey house instead of a 2-storey house and will not be considered to have 2 floors.

However, as detailed and elaborate as the data set is, not everything can be as perfect as one could hope for. There were a few data that lacked in pertinence or reliability, both of which would be omitted from the data set when conducting the analyses. The first to not contribute to the data set is the date. The variable only depicts the date of when the house data was registered into the data set and does not affect nor is affected by any other variables. Since it does not repeat the same houses during different dates to, for example, compare prices, then its purpose leans more towards verifying the integrity of the data by confirming it is accurate to the house data at the specified date and time rather than producing a meaningful conclusion.

The following variables to be omitted are the waterfront, view, and condition as these variables share the same general impertinence. The waterfront variable only has values in 33 data of the set, making up only 0.7% of the total 4600 data. On the other hand, view and condition are vague variables that were not explained anywhere in the source material and, judging from the meaning of their variable names, the values they hold are subjective compared to more concrete and objective variables like the lot's area in square feet or the number of bedrooms. Lastly, the city variable, which has 44 unique values, would also be removed as it is a categorical (nominal) data type, which means that the non-numerical values would be useless in terms of this research's objectives.

Hence, to accurately examine the data set, a table of only necessary variables, therefore the variables which will be used in this research, was prepared to visually represent the data. The following table includes the top five data records, sorted by its original input date, to further elaborate.

Table 2: Example of Records from Chosen Variables

#	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors
1	31300	3	1.50	1340	7912	1.5
2	2384000	5	2.50	3650	9050	2.0
3	342000	3	2.00	1930	11947	1.0
4	420000	3	2.25	2000	8030	1.0
5	550000	4	2.50	1940	10500	1.0

2.2 Data Pre-Processing

Prior to fulfilling this study's aims and objectives, data cleaning was conducted using R to remove any incorrect, duplicate, incomplete, and irrelevant data within the data set. Primarily, the data will be removed of its variables irrelevant to the analysis, including unquantifiable data like city or state. After the removal, the only variables left are price, bedroom, bathroom, sqft_living, sqft_lot, and floors.

A summary of the data was produced using the summary() functionality in R and is represented in the table below.

Table 3: Summary of Original Data Set (House Data) After Removal of Irrelevant Variables

Variables	Mean	Min.	1st Quartile	Median	3rd Quartile	Max.
price	551963	0	322875	460943	654962	26590000
bedrooms	3.401	0.000	3.000	3.000	4.000	9.000
bathrooms	2.161	0.000	1.750	2.250	2.500	8.000
sqft_living	2139	370	1460	1980	2620	13540
sqft_lot	14582	638	5001	7683	11001	1074218
floors	1.512	1.000	1.000	1.500	2.000	3.500

Next, the data will be cleaned before proceeding with any other analyses to ensure it will be free from data irregularities, such as invalid values and outliers. As expressed by Emily Burns (2021), there is no single correct answer to deal with missing data values. However, for this study's situation, one option that would seem the most desirable to solve the problem of missing or invalid values is to remove the offending data rows outright.

The process begins by removing the most inexplicable data, which is the price variable containing values of 0. Evidently, nothing will cost 0 USD, especially not houses. Therefore, the rows with their price at 0 will not be included in any calculations as they are incomplete and unreliable.

There is also a loose definition of "house" in this data set. Therefore, two conditions will be set to properly define it:

1. A house shall not have less than 1 bedroom.
2. A house shall not have less than 0.25 bathrooms.

Based on the data set, there are 2 rows of data that break these conditions, which are row 2366 and row 3210, the former located at 814 E Howe St and the latter at 20418 NE 64th Pl. These houses, both having 0 bedrooms and 0 bathrooms, will be excluded from the analysis.

Another data row shows that it has its price as 7800, which is an unreasonably low price for a house. The three lowest prices after it is 80000, 83000, and 83300, which puts into comparison how low its price is compared to even the cheapest of houses. Ryan West (2021) singlehandedly constructed a fully functional house slightly bigger than a lorry worth, and it was worth $8000. An important distinction to make is that the reason the data row is being removed from the data set is not because it is an outlier, but because it is simply illogical and is considered an invalid value.

As for outliers, they have the potential to heavily skew the results of analysis and any hypothesis test; a thought shared by Pritha Bhandari (2021). For example, one house located in 12005 SE 219th Ct, Kent has the highest price data of 26,590,000, a number 4817% higher than the average price of all the data, 551,963, before the exclusion of invalid values. This is one of the three data rows to be removed for being exceptional outliers, the other two being a house with a price of 12,899,000 and 7,062,500. It should be noted that the number of outliers removed is limited as this research aims to use as much of the data set as possible, without removing potentially important data.

After cleaning the data, the summary of the clean data set is as shown below.

Table 4: Summary of Clean Data Set (New House Data)

Variables	**Mean**	**Min.**	**1st Quartile**	**Median**	**3rd Quartile**	**Max.**
price	547872	80000	326100	46500	657000	4668000
bedrooms	3.396	1.000	3.000	3.000	4.000	9.000
bathrooms	2.156	0.750	1.750	2.250	2.500	8.000
sqft_living	2130	370	1460	1970	2610	13540
sqft_lot	14832	638	5000	7680	10960	1074218
floors	1.512	1.000	1.000	1.500	2.000	3.500

2.3 Hypothesis Testing and Analysis of Variance (ANOVA)

ANOVA is a statistical method for comparing statistical groups based on dependent and independent variables. The technique of analysis of variance (ANOVA) compares the number of means using a sample of data. ANOVA is a statistical test that compares two or more means for groups or variances. The variables that are measured, such as a test score, are referred to as dependent variables, while the variables that are controlled, such as a test paper correction method, are referred to as independent variables. ANOVAs are used by researchers and students depending on their study goals. The three most common versions of ANOVA are one-way ANOVA, two-way ANOVA, and N-way ANOVA. In this research, the one-way ANOVA method is used.

The one-way ANOVA test is a statistical hypothesis test used to assess whether the means of numerous groups are equal. The requirement for a one-way ANOVA test arises from the fact that the t-test cannot be performed when there are more than two groups. The one-way ANOVA compares the means of the groups you are interested in to see if any of them are statistically significantly different from one another. It examines the null hypothesis, which states:

$$\boldsymbol{H_0 : \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_k}$$

Where k = number of groups and μ = group mean. If the one-way ANOVA provides a statistically significant result, then accept the alternative hypothesis (HA), which states that there are at least two statistically significant differences between the group means. It is crucial to remember that the one-way ANOVA is an omnibus test statistic that cannot tell you which specific groups were statistically significantly different from each other, simply that there were at least two of them. A post hoc test is required to establish which individual groups varied from each other.

2.4 Model Implementation

There are various ways to construct a predictive model. In short, predictive modelling is a commonly used statistical technique to predict future behaviour or outcomes by analysing patterns in a given set of input data (Lawton et al., 2022). To fulfill the objective of this research of constructing a predictive model for house prices, a regression model is developed after a brief analysis of the correlation coefficient between prices and other independent variables.

Paraphrasing an introduction to regression models by Bevans (2020), the models describe the relationship between variables by drawing a line between the data to estimate the dependency of a dependent variable on an independent variable. The type of regression to be used in this analysis is the linear regression model to gauge the strength of the relationship between variables and to also determine the value of a dependent variable at a certain value of an independent variable. An example to use from this study is the relationship between the price variable and the total square feet of living area. A natural hypothesis to be made here is that the price will go up as the total square feet of the living area will go up, but a linear regression model will either confirm or reject that and provide an estimation of what the price will be according to the living area, given that they correlate with one another.

Due to the nature of the data set, a multiple linear regression model will be produced. The only difference between multiple linear regression and simple linear regression is that the latter only uses one independent variable whereas multiple uses many, as the dependent variable's value will depend on multiple independent values.

According to Bevans (2022) and Zach (2020), there are four assumptions to be made when producing a multiple linear regression model:

Table 5: Four Assumptions Made Before Producing a Multiple Linear Regression Model

Assumption	**Description**
Homoscedasticity	The residuals have constant variance at every value of independent variables, otherwise known as homogeneity of variance.
Independence of observations	There is no correlation between consecutive residuals in time-series data, and the residuals are independent. There are also no hidden relationships among observations.
Normality	The residuals of the model are normally distributed, following a normal distribution.
Linearity	A linear relationship exists between the independent variable and the dependent variable, and the line of best-fit through the data points is a straight line.

$$\boldsymbol{y = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n + \varepsilon}$$

The above shows the formula for multiple linear regression. Three calculations, which are the regression coefficients that lead to the smallest overall model error, the t-statistic of the model, and its associated p-value will also be calculated to find the best-fit line for each independent variable. Afterward, it will continue to calculate the t-statistic and p-value for each regression in the model.

3 Results and Discussion

3.1 Exploratory Data Analysis (EDA)

According to Komorowski et al. (2016), Explanatory Data Analysis (EDA) is an essential early step after data collection and pre-processing where the data is visualized and plotted without any assumptions to help assess the quality of the data and build models. The majority of EDA techniques are graphical, with a few quantitative techniques. Graphics are used heavily because EDA tools are used primarily to explore data, and graphically displaying information gives analysts unparalleled power to do so. (Kaski & Samuel, 1997). As this research focuses on the price as the dependent variable, this EDA will explore any potential relationships that involves this variable.

3.1.1 Measure of Central Tendency

The research first explores the data set's measure of central tendency to provide a basic understanding of the data. According to Frost (2022), the measure of central tendency is a summary statistic that represents the center point or typical value of a dataset. These measures indicate where most values in a distribution fall and are also referred to as the central location of a distribution. The most common measures of central tendency are the mean, median and mode.

To provide a clearer analysis, the following table is provided to better visualize the spread of the data.

Table 6: Measure of Central Tendency

Variables	Mean	Median	Mode
price	547872	46500	300000
bedrooms	3.396	3.000	3.000
bathrooms	2.156	2.250	2.500
sqft_living	2130	1970	1940, 1720
sqft_lot	14832	7680	5000
floors	1.512	1.500	1.000

In particular, attention should be given to the price variable. According to Table x, the mean or average house price in the dataset is US$547,872. The median or middle value of the house price, on the other hand, is US$46,500 while the mode or most frequent house price is US$300,000.

3.1.2 Box Plot

A box plot is used to visually compare the original data set (House Data) to the clean data set (New House Data) to investigate whether the removed data actually influences the data set significantly. The box plot is also used to check the skewness of data, which in this case focuses on the price variable's data.

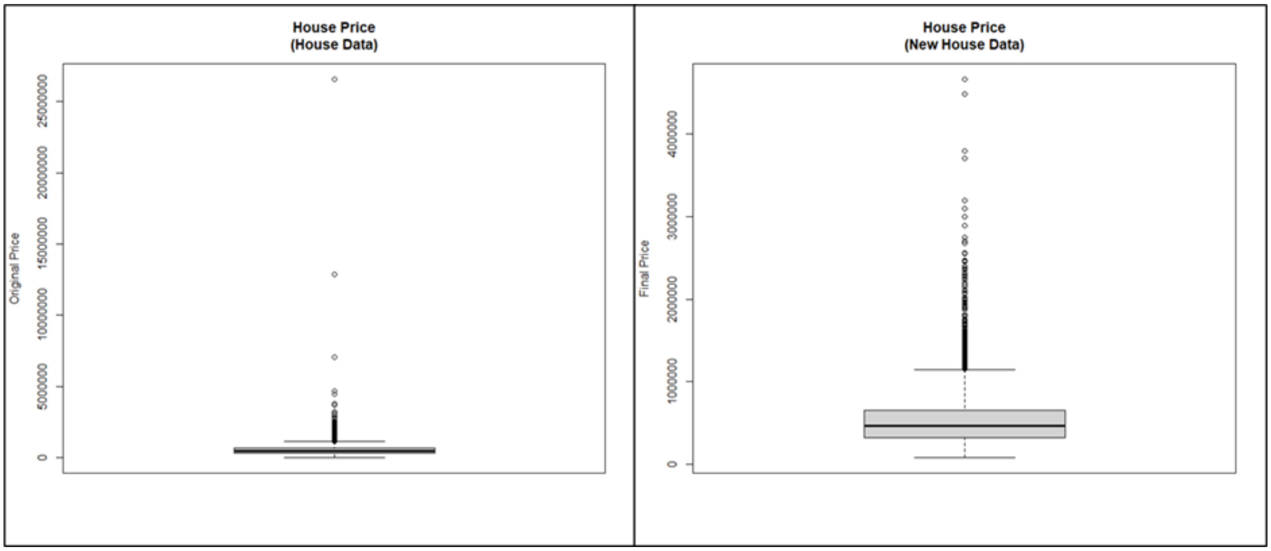


Figure 1: Box Plot of House Price (House Data - New House Data)

Table 7: Summary Table of House Price (House Data - New House Data)

Variables	Min.	1st Quartile	Median (2nd Quartile)	3rd Quartile	Max.	IQR (Q3 – Q1)
House Price (House Data)	0	322875	460943	654962	26590000	332087
House Price (New House Data)	80000	326100	465000	657000	4668000	330900

To begin with, the price values are sorted. Then four equal sized groups are made from the ordered values. That is, 25% of all prices are placed in each group. The lines dividing the groups are called quartiles, and the groups are referred to as quartile groups. As shown in the figure above, there is a significant difference in the shape of the box plot after the data has been filtered or cleaned.

In the case of the data's skewness, both the original data set (House Data) and the clean data set (New House Data) are positively skewed. As the skewness is not entirely clear visually, the values in Table 7 is used to check this and are calculated as in the table below.

Table 8: Calculation of Differences in Quartile Ranges of House Price (House Data - New House Data)

House Price (House Data)	**House Price (New House Data)**
(Q2 – Q1) = 137068	(Q2 – Q1) = 138900
(Q3 – Q2) = 194091	(Q3 – Q2) = 192000
(Q3 – Q2) > (Q2 – Q1) = Positive skew	(Q3 – Q2) > (Q2 – Q1) = Positive skew

However, the difference between (Q2 – Q1) and (Q3 – Q2) decreases once the data is cleaned, from a value of 57023 to 53100. This is good as the distribution is then said to be closer to a normal distribution.

3.1.3 Scatter Plot and Correlation Analysis

A scatter plot is used to visualize if any correlation exists between two variables. For this research, it is used to identify if any correlation exists between the number of bedrooms, number of bathrooms, total square feet of the house lot (sqft_lot) and total square feet of living space of the house (sqft_living) on house prices. The original data (House Data) and the filtered data (New House Data) are compared side by side. The correlation coefficient is also noted as it is a quick indicator of the strength of the linear relationship. The cor() function in R is used to determine the correlation coefficient for each relationship.

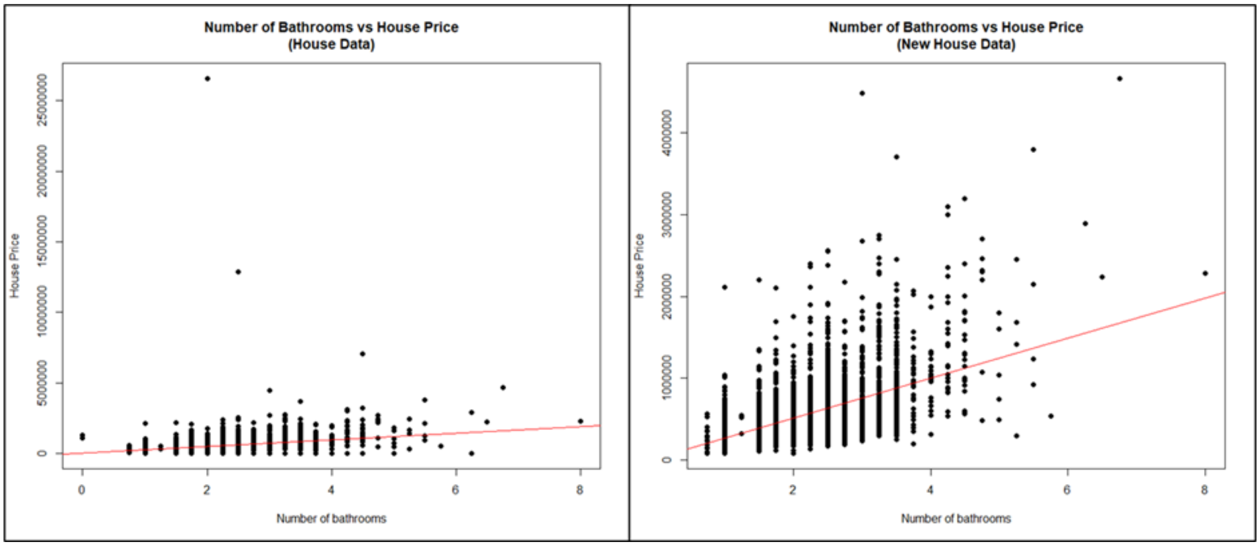


Figure 2: Scatter Plot of Number of Bathrooms vs House Price (House Data - New House Data)

Figure 2 above displays the difference of scatter plot and correlation between the number of bathrooms and the price of houses. The value of the correlation for the original data (House Data) is 0.3271099. It shows a weak, positive, linear association. There are many outliers.

On the right side is the clean data set (New House Data) and 0.5329851 is the correlation coefficient recorded. It shows a moderately strong, positive, linear association. There are a few potential outliers. There is an increase in the value of the correlation when the variable price is cleaned by removing the records with substantially different prices values and removing the records that contain bathroom value equal to 0.

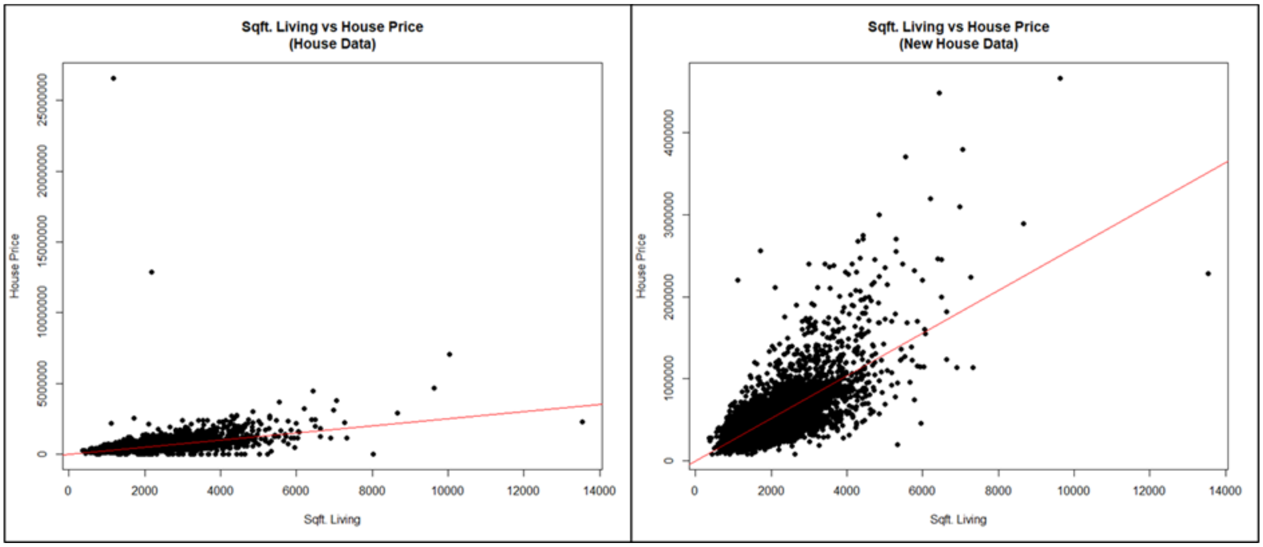


Figure 3: Scatter Plot of Sqft. Living vs House Price (House Data - New House Data)

Next, Figure 3 illustrates the difference of scatter plot and correlation between the total square feet of living space of the house (sqft_living) and the price of houses for two different data set conditions. The value of the correlation for the original data (House Data) is 0.4304100. It shows a weak, positive, linear association. There are a lot of outliers.

On the other hand, the clean data (New House Data) has a correlation coefficient of 0.6941563. It shows a strong, positive, linear association. There are a few potential outliers. Hence, there is a significant increase in the value of the correlation coefficient when the price variable is cleaned by removing the records with substantially different price values.

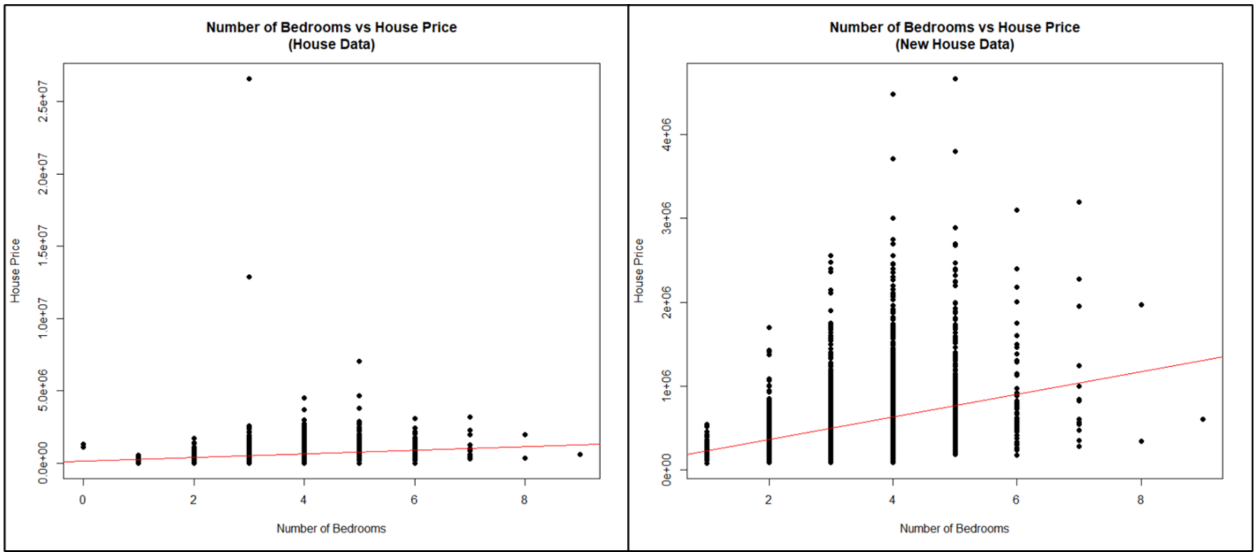


Figure 4: Scatter Plot of Number of Bathrooms vs House Price (House Data - New House Data)

Figure 4 displays the difference of scatter plot and correlation between the total number of bedrooms in the house and the price of houses. The value of the correlation for the original data (House Data) is 0.20033629. It shows a weak, positive, linear association. There are a few potential outliers.

Comparing this to the scatter plot next to it, the clean data (New House Data) has a correlation coefficient of only 0.34103100. It also shows a weak, positive, linear association. There are a few potential outliers. There is a slight increase in the correlation value when the two data are compared.

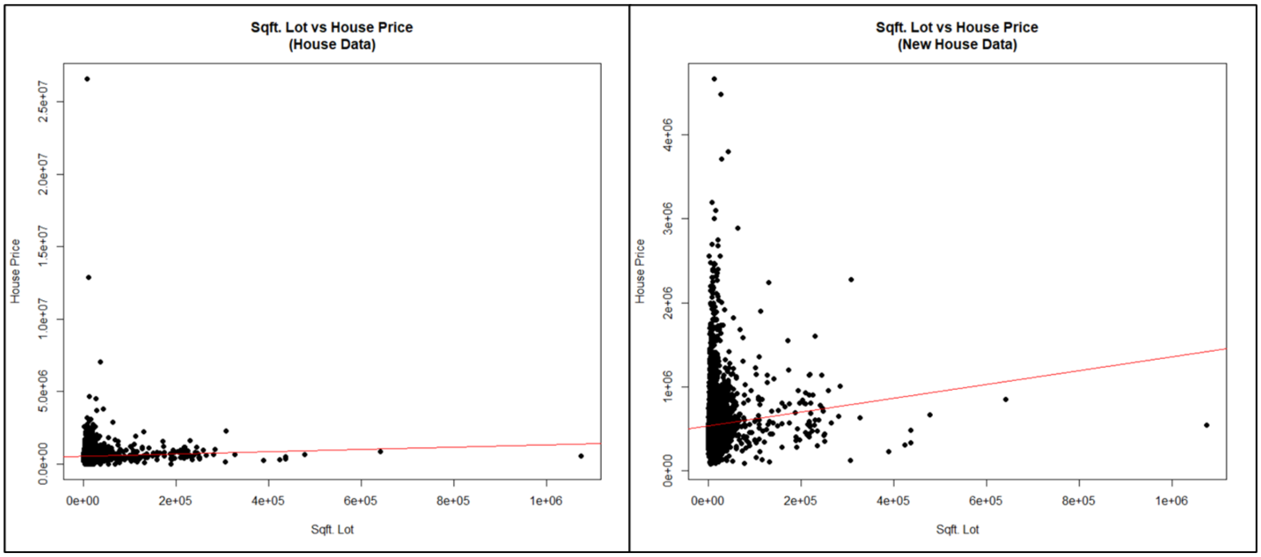


Figure 5: Scatter Plot of Sqft. Lot vs House Price (House Data - New House Data)

Figure 5, on the other hand, visualises the difference of scatter plot and correlation between the total square foot lot of the house and the price of houses. The value of the correlation for the original data (House Data) is 0.0504513. It shows a weak, positive, linear association. There are a few potential outliers. The clean data (New House Data) has a correlation coefficient of only 0.08290145. It also shows a weak, positive, linear association. There are a few potential outliers. There is a very minor increase in the correlation value when comparing the two data.

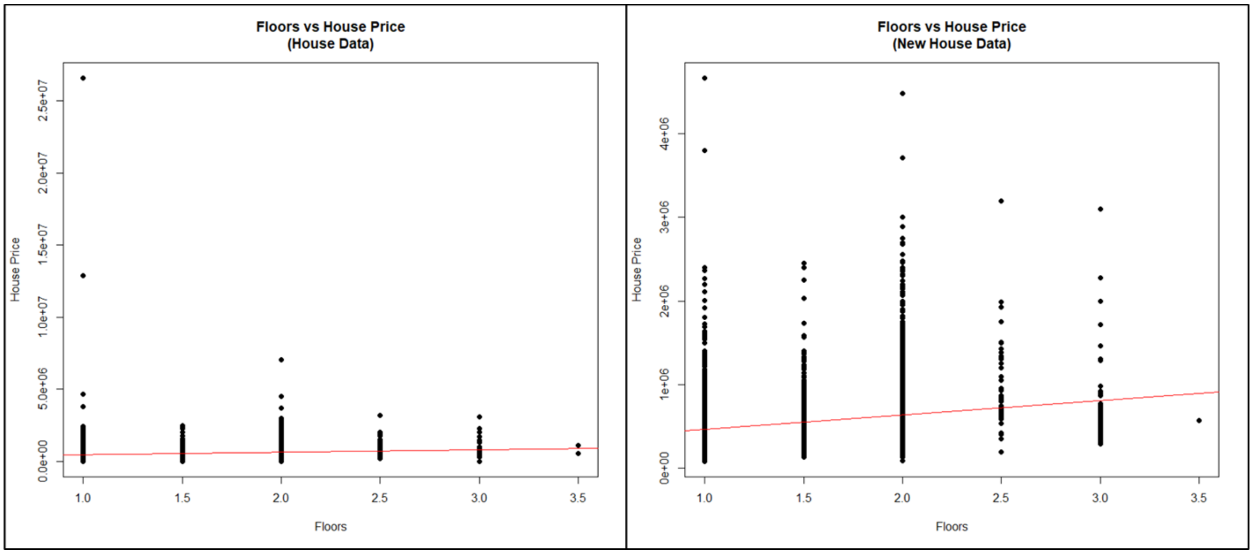


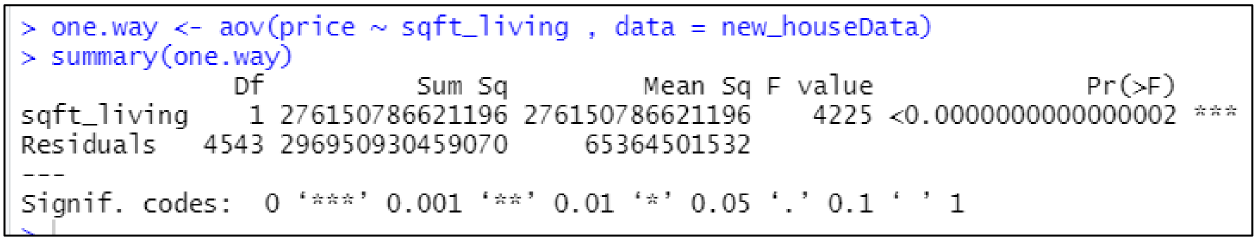
Figure 6: Scatter Plot of Floors vs House Price (House Data - New House Data)

Lastly, Figure 6 displays the difference of scatter plot and correlation between the number of floors in the house and the price of houses. The value of the correlation for the original data (House Data) is 0.1514608. It shows a weak, positive, linear association. There are a few potential outliers. The clean data (New House Data) has a correlation coefficient of only 0.1514608. It also shows a weak, positive, linear association. There are a few potential outliers. There is a very minor increase in the correlation value when comparing the two data.

3.2 Hypothesis Testing and Analysis of Variance (ANOVA)

One-way ANOVA

ANOVA determines whether the groups created by the levels of the independent variable are statistically different by calculating whether the means of the treatment levels are different from the overall mean of the dependent variable. For this, the variable for (New House data) that has the highest correlation value is chosen which is the total square foot living of the house.



```
> one.way <- aov(price ~ sqft_living , data = new_houseData)
> summary(one.way)
              Df          Sum Sq         Mean Sq F value               Pr(>F)    
sqft_living    1 276150786621196 276150786621196    4225 <0.0000000000000002 ***
Residuals   4543 296950930459070     65364501532                                 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

Figure 7: Summary Table of One-Way ANOVA Test

The first column lists the independent variable along with the model residuals (aka the model error).

The **Df column** displays the degrees of freedom for the independent variable (calculated by taking the number of levels within the variable and subtracting 1), and the degrees of freedom for the residuals (calculated by taking the total number of observations minus 1, then subtracting the number of levels in each of the independent variables).

The **Sum Sq** column displays the sum of squares (a.k.a. the total variation) between the group means and the overall mean explained by that variable. The sum of squares for the total square foot living is 276,150,786,621,196 while the sum of squares of the residuals is 296,950,930,459,070.

The **Mean Sq** column is the mean of the sum of squares, which is calculated by dividing the sum of squares by the degrees of freedom.

The **F-value** column is the test statistic from the F test: the mean square of each independent variable divided by the mean square of the residuals. The larger the F-value, the more likely it is that the variation associated with the independent variable is real and not due to chance. The F-value recorded is 4225.

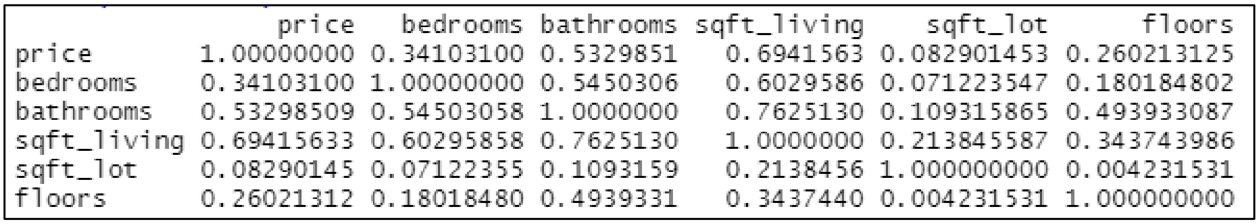
The **Pr(>F)** column is the p-value of the F-statistic. This shows how likely it is that the F-value calculated from the test would have occurred if the null hypothesis of no difference among group means were true.

The **p-value** recorded is 0.0000000000000002.

Because the p-value of the variable, total square foot living, is significant ($p < 0.05$), it is likely that the total square feet of living does have a significant effect on the house price.

3.3 Model Evaluation and Analysis

To further analyse the dataset's house prices, a correlation analysis is conducted. As previously explored in the EDA, the correlation coefficient between prices and other independent variables has improved after data cleaning. Hence, prior to building a predictive model for house prices, the correlation coefficient between each independent variable was examined once more, with a focus on each variable's relationship with the dependent price variable.



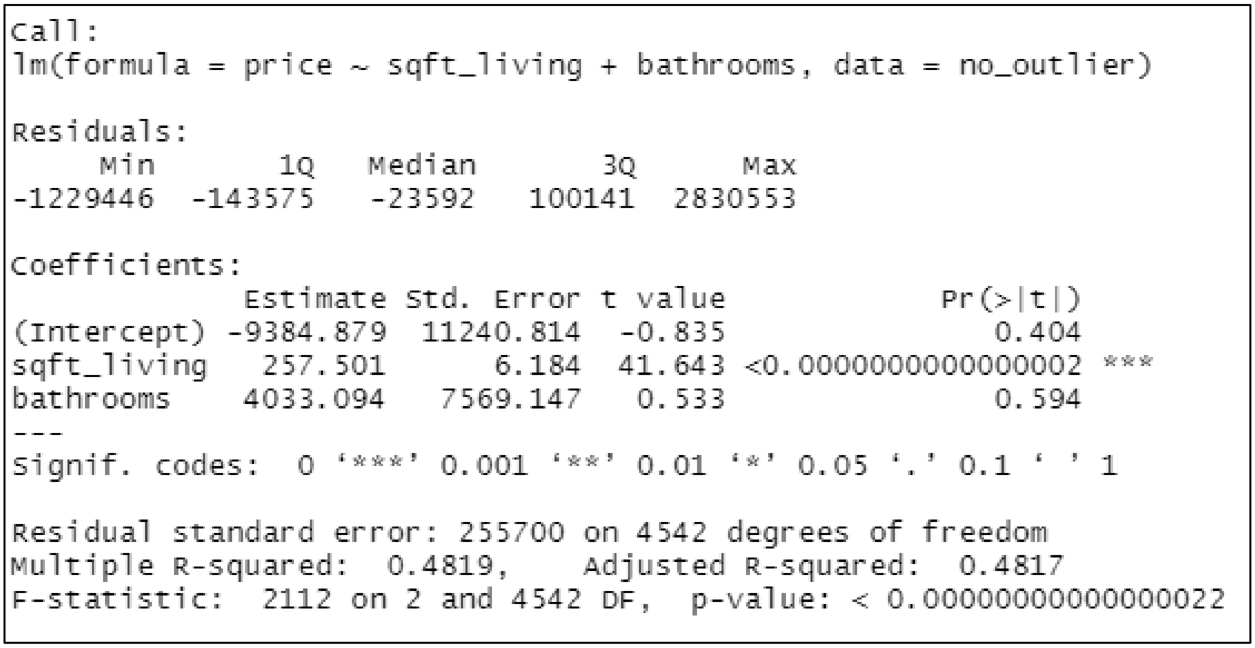
```
                 price   bedrooms bathrooms sqft_living    sqft_lot      floors
price       1.00000000 0.34103100 0.5329851   0.6941563 0.082901453 0.260213125
bedrooms    0.34103100 1.00000000 0.5450306   0.6029586 0.071223547 0.180184802
bathrooms   0.53298509 0.54503058 1.0000000   0.7625130 0.109315865 0.493933087
sqft_living 0.69415633 0.60295858 0.7625130   1.0000000 0.213845587 0.343743986
sqft_lot    0.08290145 0.07122355 0.1093159   0.2138456 1.000000000 0.004231531
floors      0.26021312 0.18018480 0.4939331   0.3437440 0.004231531 1.000000000
```

Figure 8: Correlation Coefficients Between Each Independent Variable

As seen in Figure 8, price and sqft_living has the highest correlation with a magnitude of 0.69415633, which means it is said to be a moderate correlation. However, it should be noted that a high correlation is generally valued between 0.7 to 0.9, which this relationship almost achieves. The other moderately correlated variables are price and bathrooms with a magnitude of 0.53298509.

Unfortunately, price and bedrooms have a low correlation of 0.34103100, while price and floors, as well as price and sqft_lot, garnered a measly magnitude of 0.26021312 and 0.08290145 which means that they have little to no correlation with each other.

After conducting this brief analysis, sqft_living and bathrooms are selected as independent variables to build a multiple linear regression model to predict house prices. Using the summary() function in R, the results are shown in Figure 9.



```
Call:
lm(formula = price ~ sqft_living + bathrooms, data = no_outlier)

Residuals:
     Min       1Q   Median       3Q      Max 
-1229446  -143575   -23592   100141  2830553 

Coefficients:
             Estimate Std. Error t value          Pr(>|t|)    
(Intercept) -9384.879  11240.814  -0.835             0.404    
sqft_living   257.501      6.184  41.643 <0.0000000000000002 ***
bathrooms    4033.094   7569.147   0.533             0.594    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 255700 on 4542 degrees of freedom
Multiple R-squared:  0.4819,	Adjusted R-squared:  0.4817 
F-statistic:  2112 on 2 and 4542 DF,  p-value: < 0.00000000000000022
```

Figure 9: Summary Table of Constructed Regression Model

To analyse the summary table above, **Residuals** describe the difference between the actual values and the predicted values. According to the output, the distribution of this predictive model is not quite symmetrical and may be considered right or positively skewed. This indicates that the model may not predict as well at the higher price range as it might for the low ranges. It should be noted that the data set was originally heavily right-skewed as well, which might have influenced this outcome.

Figure x also includes the coefficients table which contains the coefficients for the regression equation (model), tests of significance for each variable and R-squared value. The **Estimate** column in the coefficients table provides the coefficients for each independent variable in the regression model. Thus, the predictive model for house prices is:

$$\boldsymbol{price\ (y) = \ -9384.879\ +\ 257.501 * (sqft\_living)\ +\ 4033.094 * (bathrooms)}$$

As such, there is a 257.501 increase in sqft_living for each extra total square feet of living space of the house. Similarly, for each US$ increase in price, the number of bathrooms increases by 4033.094.

The last column of the coefficients table, **Pr(>|t|)** contains the p-values for each of the independent variables. The p-value, in association with the t-statistic, explains how significant the coefficient is to the model. In practice, any p-value below 0.05 is usually deemed as significant, assuming that α = 0.05. This means that the coefficient is not zero, which means that the coefficient does add value to the model by helping to explain the variance within the dependent variable. In the case of this regression model, the p-value for sqft_living is extremely small at <0.0000000000000002 which means that there is convincing evidence that this coefficient is not zero. The number of asterisks next to this column represents the **Significant codes**, whereby the more asterisks there are, the more significant the coefficient.

To assess how well the regression model fits the dataset, the Residual standard error, Multiple R-squared, Adjusted R-squared, F-statistic and p-value should be analysed. In the case of **Residual standard error**, it explains the average distance that the observed values fall from the regression line whereby the smaller the value, the better the regression model able to fit the data. For this model, the Residual standard error is 255700 with 4542 degrees of freedom.

The **Multiple R-squared** is most often used for simple linear regression (one predictor) as the value increases with the number of independent variables so it is generally recommended to use the **Adjusted R-squared** value for multiple linear regressions. In the case of this predictive model, **Adjusted R-squared** indicates that 48.17% of the variation in price can be explained by the model containing sqft_living and bathrooms.

When running a regression model on R, the null hypothesis (H_0) indicates that there is no relationship between the dependent variable and the independent variables. Conversely, the alternative hypothesis (H_1) suggests that there is a relationship. Hence, as the **F-statistic** of the predictive model is quite large while the **p-value** is so small it is basically zero (<0.00000000000000022), this would lead to a rejection of the null hypothesis. Thus, there is strong evidence that a relationship does exist between price, sqft_living and bathrooms.

References

Analysis Of Variance (ANOVA) (2019, August 28). Retrieved May 28, 2022, from https://statswork.com/blog/analysis-of-variance-anova/

Bevans, R. (2022, May 6). *An introduction to multiple linear regression*. Scribbr. Retrieved May 28, 2022, from https://www.scribbr.com/statistics/multiple-linear-regression/

Bevans, R. (2022, May 6). *An introduction to simple linear regression*. Scribbr. Retrieved May 28, 2022, from https://www.scribbr.com/statistics/simple-linear-regression/

Calkins, K. G. (2015, July 18). *Correlation Coefficients*. Correlation coefficients. Retrieved May 27, 2022, from https://www.andrews.edu/~calkins/math/edrm611/edrm05.htm

Chen, R., Gan, C., Hu, B., & Cohen, D. A. (2013). *An empirical analysis of house price bubble: A case study of beijing ...Ryan Chen*. Retrieved May 28, 2022, from https://www.researchgate.net/publication/314570230_An_Empirical_Analysis_of_House_Price_Bubble_A_Case_Study_of_Beijing_Housing_Market

Dubin, R. (1998). *Predicting House prices using multiple listings data*. Retrieved May 29, 2022, from https://www.researchgate.net/publication/5151497_Predicting_House_Prices_Using_Multiple_Listings_Data

Frost, J. (2022, March 23). *Measures of central tendency: Mean, median, and mode*. Statistics By Jim. Retrieved May 28, 2022, from https://statisticsbyjim.com/basics/measures-central-tendency-mean-median-mode/

Ganeson, C., & Abdul Muin, I. M. (2015). *An analysis of the factors affecting house prices in Malaysia an ...* Retrieved May 30, 2022, from http://eprints.usm.my/37601/1/sspis_2015_ms224_-_235.pdf

Kenton, W. (2022, May 20). *How analysis of variance (ANOVA) works. Investopedia*. Retrieved May 30, 2022, from https://www.investopedia.com/terms/a/anova.asp#:~:text=Analysis%20of%20variance%20(ANOVA)%20is,the%20random%20factors%20do%20not

Komorowski, M., Marshall, D. C., Salciccioli, J. D., & Crutain, Y. (2016). Exploratory Data Analysis. *Secondary Analysis of Electronic Health Records*, 185–203. https://doi.org/10.1007/978-3-319-43742-2_15

Lawton, G., Carew, J. M., & Burns, E. (2022, January 21). *What is predictive modeling?* SearchEnterpriseAI. Retrieved May 27, 2022, from https://www.techtarget.com/searchenterpriseai/definition/predictive-modeling#:~:text=Predictive%20modeling%20is%20a%20mathematical,forecast%20activity%2C%20behavior%20and%20trends.

Macdonald, A. F. (2020, November 5). *What is a 1.5 storey house?: Real estate definition*. GimmeShelter. Retrieved May 25, 2022, from https://www.gimme-shelter.com/what-is-a-1-5-storey-house-50104/

Mahale, A., Bhistannavar, V., Chauhan, N., & Matey, V. (2022). *(PDF) housing price prediction using supervised learning*. Retrieved May 28, 2022, from https://www.researchgate.net/publication/359369961_Housing_Price_Prediction_Using_Supervised_Learning

Mandell, L. J. (2022, May 20). *What is a half-Bath? or a quarter bath or three-quarter bath, for that matter?* Real Estate News & Insights | realtor.com®. Retrieved May 25, 2022, from https://www.realtor.com/advice/buy/what-is-a-half-bath/

Nuuter, T., Lill, I., & Tupenaite, L. (2014). *Ranking of Housing Market Sustainability in Selected European Countries.* Retrieved May 26, 2022, from http://www.ceneast.com/wp-content/uploads/2020/09/Dissemination%20publications/P-4_Nuuter_Lill_RAnking.pdf

One-way ANOVA (2022, May 30). Retrieved from https://statistics.laerd.com/statistical-guides/one-way-anova-statistical-guide.php

Shree. (2018, August 26). *House price prediction*. Kaggle. Retrieved May 24, 2022, from https://www.kaggle.com/datasets/shree1992/housedata

Wang, Z., Hoon, J., & Lim, B. (2013). *The impacts of housing affordability on social and economic sustainability in Beijing 20.* Australasian Journal of Construction Economics and Building - Conference Series. Retrieved May 30, 2022, from https://epress.lib.uts.edu.au/journals/index.php/AJCEB-Conference-Series/article/view/3154

What is a 1/4 bath? Hunker. (2020, October 3). Retrieved May 25, 2022, from https://www.hunker.com/13413443/what-is-a-14-bath

Zach. (2020, January 8). *The four assumptions of linear regression*. Statology. Retrieved May 28, 2022, from https://www.statology.org/linear-regression-assumptions/

Appendix

RStudio code

```
attach(data)
#NOTES
#1) Data.csv = raw/original data
#2) houseData = Data.csv - 12 unused columns
#3) new_houseData = houseData - (4 high diff. price value + price=0 + bed/bathrooms =0)
################################################################################
#1# CREATE new data set, remove useless columns;
## (Data.csv) Rows and columns;
print(paste("Number of records: ", nrow(data)))         #4600
print(paste("Number of features: ", ncol(data)))        #18
## Filter only columns used;
houseData <- data[,c("price", "bedrooms", "bathrooms",
                    "sqft_living", "sqft_lot", "floors")]
## DISPLAY (houseData) Rows and columns;
print(paste("Number of records: ", nrow(houseData)))    #4600
print(paste("Number of features: ", ncol(houseData)))   #6
#################################################################################
#2# CLEAN the data (housedata -> new_housedata);
## IDENTIFY price range and values
range(houseData$price)
houseData[with(houseData,order(-price)),]
## REMOVE the records with substantially different prices value;
## -c(26,590,000.0 / 12,899,000.0 / 7,800.0 / 7,062,500)
```

```
houseData_sdf <- houseData[-c(4351, 4347, 4352, 2287 ), ]
## REMOVE records with features value = 0;
new_houseData <- houseData_sdf
which(new_houseData$price == 0)        # -49 rows
which(new_houseData$bedrooms == 0)     #  -2 rows same
which(new_houseData$bathrooms == 0)    #  -2 rows same
new_houseData <-houseData_sdf[houseData_sdf$price != 0
                        & houseData_sdf$bedrooms != 0
                        & houseData_sdf$bathrooms != 0, ]
## DISPLAY (new_houseData) Rows and columns
print(paste("Number of records: ", nrow(new_houseData)))    #4545
print(paste("Number of features: ", ncol(new_houseData)))   #6
################################################################################
#3# DISPLAY Data summary;
summary(houseData)
summary(new_houseData)
################################################################################
#4# GRAPH Box plot - price
boxplot(houseData$price, ylab = "Original Price", main = "House Price\n(House Data)")
boxplot(new_houseData$price, ylab = "Final Price", main = "House Price\n(New House Data)")
summary(houseData$price)
summary(new_houseData$price)
################################################################################
#5# FIND Correlation value with price
## houseData
cor(houseData)
cor(new_houseData)

#################################################################################

#6# Scatter Plot - (houseData) vs price

# bathrooms

plot(houseData$bathrooms, houseData$price,

    main = "Number of Bathrooms vs House Price\n(House Data)",

    xlab = "Number of bathrooms", ylab = "House Price",

    pch = 19, frame = TRUE)

abline(lm(houseData$price ~ houseData$bathrooms, data = houseData), col = "red")

# sqft living

plot(houseData$sqft_living, houseData$price,

    main = "Sqft. Living vs House Price\n(House Data)",

    xlab = "Sqft. Living", ylab = "House Price",

    pch = 19, frame = TRUE)

abline(lm(houseData$price ~ houseData$sqft_living, data = houseData), col = "red")

# bedrooms

plot(houseData$bedrooms, houseData$price,

    main = "Number of Bedrooms vs House Price\n(House Data)",

    xlab = "Number of Bedrooms", ylab = "House Price",

    pch = 19, frame = TRUE)

abline(lm(houseData$price ~ houseData$bedrooms, data = houseData), col = "red")

# sqft_lot

plot(houseData$sqft_lot, houseData$price,

    main = "Sqft. Lot vs House Price\n(House Data)",

    xlab = "Sqft. Lot", ylab = "House Price",

    pch = 19, frame = TRUE)
```

```
abline(lm(houseData$price ~ houseData$sqft_lot, data = houseData), col = "red")

# floors

plot(houseData$floors, houseData$price,

   main = "Floors vs House Price\n(House Data)",

   xlab = "Floors", ylab = "House Price",

   pch = 19, frame = TRUE)

abline(lm(houseData$price ~ houseData$floors, data = houseData), col = "red")

#######################################################################

#7# Scatter Plot - (new_houseData) vs price

# bathrooms

plot(new_houseData$bathrooms, new_houseData$price,

   main = "Number of Bathrooms vs House Price\n(New House Data)",

   xlab = "Number of bathrooms", ylab = "House Price",

   pch = 19, frame = TRUE)

abline(lm(new_houseData$price ~ new_houseData$bathrooms, data = new_houseData), col = "red")

# sqft living

plot(new_houseData$sqft_living, new_houseData$price,

   main = "Sqft. Living vs House Price\n(New House Data)",

   xlab = "Sqft. Living", ylab = "House Price",

   pch = 19, frame = TRUE)

abline(lm(new_houseData$price ~ new_houseData$sqft_living, data = new_houseData), col = "red")

# bedrooms

plot(new_houseData$bedrooms, new_houseData$price,

   main = "Number of Bedrooms vs House Price\n(New House Data)",

   xlab = "Number of Bedrooms", ylab = "House Price",

   pch = 19, frame = TRUE)
```

```
abline(lm(new_houseData$price ~ new_houseData$bedrooms, data = new_houseData), col = "red")
# sqft_lot
plot(new_houseData$sqft_lot, new_houseData$price,
    main = "Sqft. Lot vs House Price\n(New House Data)",
    xlab = "Sqft. Lot", ylab = "House Price",
    pch = 19, frame = TRUE)
abline(lm(new_houseData$price ~ new_houseData$sqft_lot, data = new_houseData), col = "red")
# floors
plot(new_houseData$floors, new_houseData$price,
    main = "Floors vs House Price\n(New House Data)",
    xlab = "Floors", ylab = "House Price",
    pch = 19, frame = TRUE)
abline(lm(new_houseData$price ~ new_houseData$floors, data = new_houseData), col = "red")
########################################################################
#8# ANOVA (houseData)
#one way
options(scipen = 999)
one.way <- aov(price ~ bedrooms , data = new_houseData)
summary(one.way)
########################################################################
```