KING COUNTY HOUSE SALES PREDICTION

Authors:

1. JIMCOLLINS WAMAE
2. LEO KARIUKI
3. MAUREEN KITANGA
4. PRISCILA KAMIRI
5. SAMUEL KYALO
6. STEVE GITHINJI

1. BUSINESS OVERVIEW

1.1 INTRODUCTION

King County, located in the state of Washington in the United States, is a highly desirable place to live due to its proximity to major metropolitan areas such as Seattle, Bellevue, and Redmond. As a result, the real estate market in King County is highly competitive, with many buyers vying for a limited number of properties.

The domain of house sales in King County is complex and ever-changing, with a wide range of factors affecting the market. These challenges include **a limited supply of housing**, especially in **desirable neighborhoods** close to employment centers. This has led to an increase in prices due to increased demand.

This project seeks to understand the current state of the King County real estate market by evaluating data from the King County House Sales dataset in order to provide insights to a Real Estate Developer in the northwestern county that identifies and acquires land for new development projects. As the developer seeks to maximize their income, the project seeks to provide insights as to how the size and quality of the property affects the sale price, how the house neighborhood/location affects the sale price and how accurately the model can predict sale prices based on the available features, which will assist the developer to make data driven decisions on the type of homes to develop.

In conclusion, the project seeks to provide a model that provides insights on the current state of the market as well as predict the sale price of homes based on their features.

1.2 PROBLEM STATEMENT

In King County, finding suitable areas for new housing developments is becoming ever more difficult. It is challenging to pinpoint neighborhoods with enough demand and affordability to sustain new development due to the low supply of housing and the high costs in many regions.

This project seeks to provide insights into the available data on house sales in the area and provide recommendations to the real estate developer. The project will leverage a wide range of data points including the size of the houses, their quality, their age, and additional features such as the views and whether they have a waterfront, in order to establish the relationships between these features and the sale price of those houses.

1.3 OBJECTIVES

The main objective of this project is to identify the main features that affect home prices in King County Washington and predict the sale prices of the homes, based on these features, with a certain level of accuracy. The specific objectives were:

1. To find the relationship between the size and quality of the property to the price of homes in King County.
2. To find the relationship between the neighborhood/ location of the property to the house price.
3. To build a model and establish how accurately the price of homes in King County can be predicted using the available features.

2. DATA UNDERSTANDING

The King County House Sales dataset contains information on over 21,000 home sales in King County, Washington, USA between May 2014 and May 2015. It includes information about:

1. The **size of the homes** which is described by the area in square feet of: The lot, the living area, the basement, the area above ground and the number of bedrooms and bathrooms. It also includes the number of floors of the homes sold.
2. The **quality of the homes** is described by the features: condition, grade, the year built, and the year renovated.
3. The **neighborhood** around which the property stands is described by the zicode, the latitude and longitude co-ordinates and the size of 15 properties around it.
4. **Additional features** such as: the view and whether the house is on a waterfront.

3. DATA CLEANING

3.1 Missing Values

The dataset contained missing values in three columns:

1. Waterfront

The column Describes whether the house is on a waterfront. It is a categorical column with two unique values, “YES” and “NO”.

There were 2376 missing values in this column. This is 11% of the total data in the dataset.

The missing values were replaced by the mode of the column since 99.2% of the homes in the existing data had no waterfront. **The missing values were therefore replaced with “NO”**

2. View

This is a categorical column which describes the quality of view from the house. It has 5 unique values: NONE, AVERAGE, GOOD, FAIR and EXCELLENT.

There were 63 missing values in the column which represents 0.29% of the total data.

The missing records were replaced with ‘NONE’.

3. Yr_renovated

This is a numerical column which states the year when the house was renovated. It contained 3,842 missing values.

These missing values could either suggest that the houses have never been renovated or that the data was erroneous.

This project did not utilize this feature in its analysis and therefore, the column was **dropped.**

3.2 Invalid Data

The dataset contained no duplicate values. The `sqft_basement` column contained some rows with **‘?’** as a value. These were **replaced with 0.0.**

4. EXPLORATORY DATA ANALYSIS

4.1 Univariate Analysis

In this section, each column in the dataset was explored and the distributions of features described. The columns were categorized into:

1. Categorical Columns.
2. Numerical Columns.

4.1.1 Categorical Columns

There were five categorical columns in the dataset that were analyzed.

a) waterfront
b) view
c) condition
d) grade
e) zipcode

a.) Waterfront

This column describes whether the house has a waterfront or does not. It has two values, 'YES' and 'NO'.

b.) The View

The View describes the quality of the view of the home, on a scale from 'None', meaning the home has no view at all, to 'FAIR', 'AVERAGE', GOOD' and 'EXCELLENT', starting from the worst to the best view respectively.

Several factors were used to grade the quality including views of Mt. Rainier, Olympics, Cascades, Territorial, Seattle Skyline, Puget Sound, Lake Washington, Lake Sammamish, small lake / river / creek, and others.

c.) Condition

The column identifies the condition of the house. It is relative to the age and grade of the house.

It has 5 values that describe the condition on a scale described by: 'Poor', 'Fair', 'Average', 'Good' and 'Very Good', from the worst condition to the best respectively.

d.) Grade

The grade column identifies the quality of construction and design of the house in terms of the construction quality of improvements. Grades run from grade 1 to 13.

e.) Zipcode

The column identifies the zipcode area the house is in.

4.1.2 Numeric Columns

There were 11 numerical columns in the dataset that were analyzed:

a) Price
b) Bedrooms
c) Bathrooms
d) Sqft_living
e) Sqft_loft
f) Floors
g) Sqft_above
h) Sqft_basement
i) Yr_built
j) Lat
k) Long

a.) Price

The price column describes the price of the house.

b.) Bedrooms

This column describes the number of bedrooms in a home.

c.) Bathrooms

This column identifies the number of bathrooms in the house.

d.) Sqft Living

The column describes the area of the house in square feet.

e.) Sqft_Lot

This column identifies the square footage of the lot that the home stands in.

f.) Floors

This column describes the number of floors in the house.

g.) Sqft Above

The column describes the area of the house above the ground in square feet.

h.) Sqft Basement

This column describes the area of the basement of the house in square feet.

i.) Yr Built

This column describes the year that the house was built.

j.) Sqft Living 15

This column describes the area of interior housing living space of the nearest 15 neighbors.

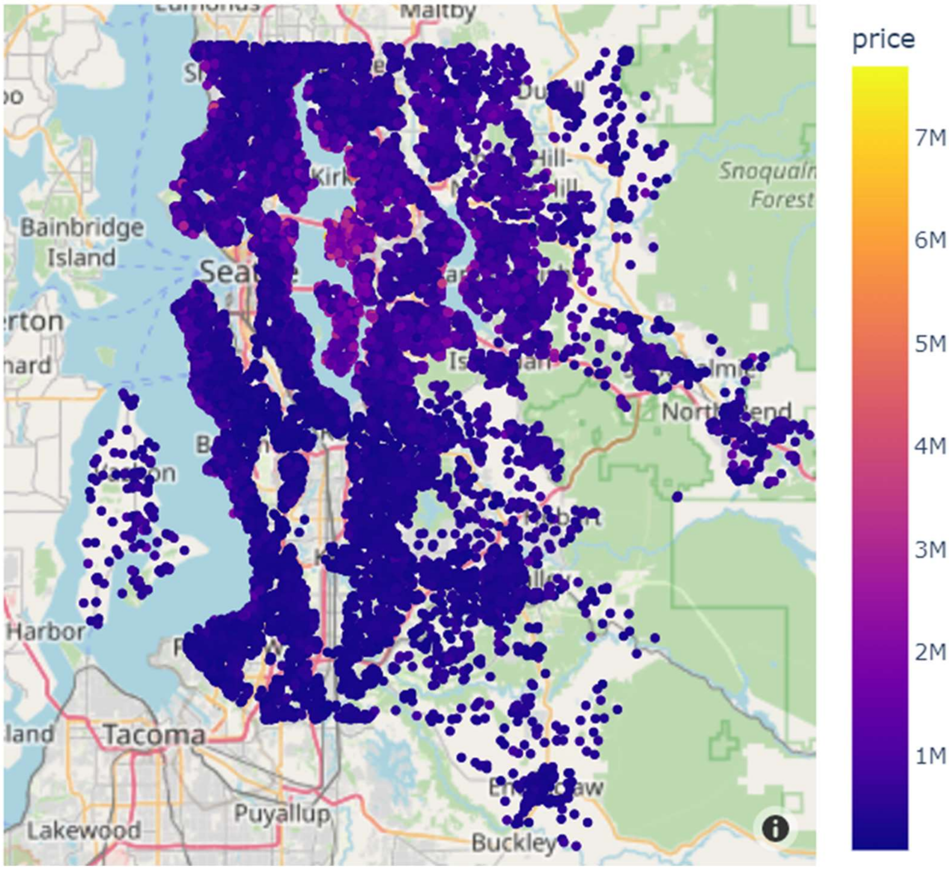
k.) Sqft Lot15

This column describes the area of the land for the nearest 15 neighbors in square feet.

l.) Lat and Long

These columns describe the latitude and longitude coordinates, specifying the location where the house is located.

To visualize the data, a map box scatter plot was created that shows the location of the houses on a map.

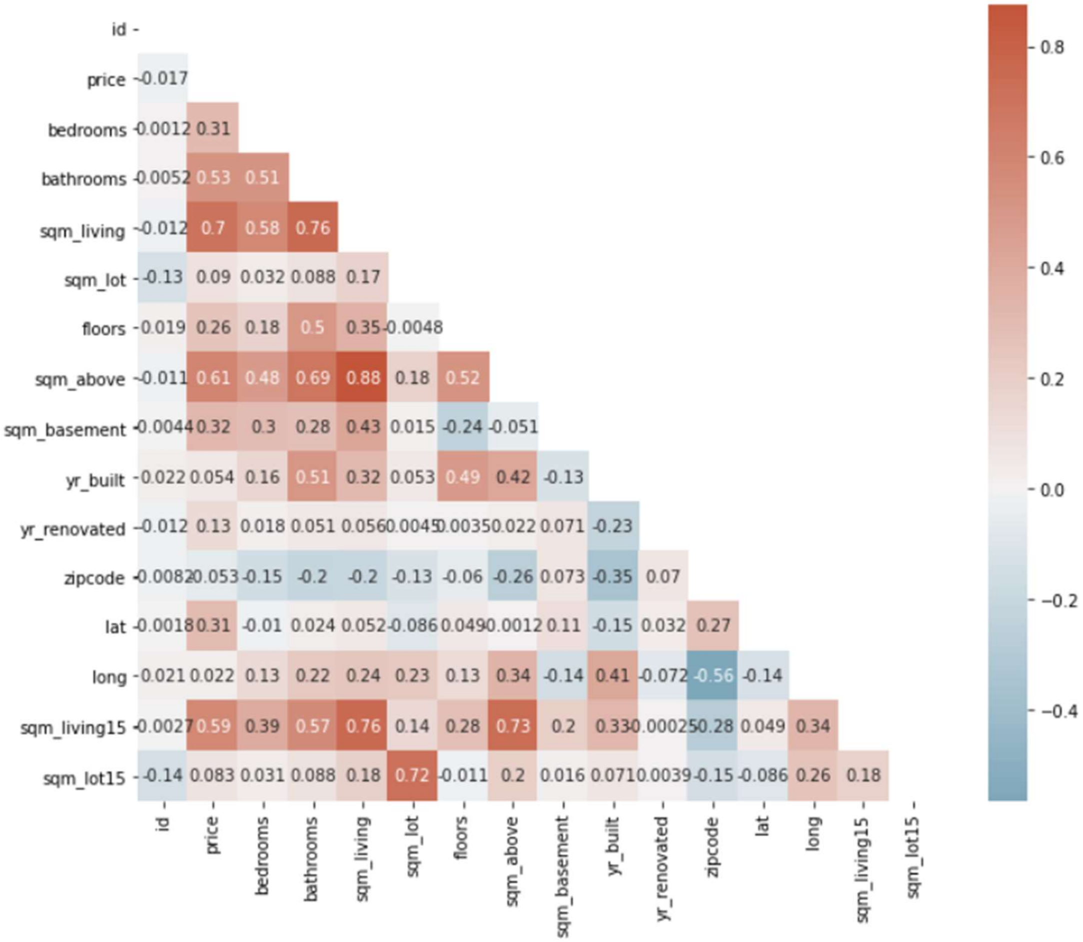


4.1.3 Linear Transformations

Considering that most countries use the metric system of measurement, linear transformations were performed to transform the units of area from square feet to square meters. This was done to make them more interpretable to our stakeholders.

4.1.4 Correlation Matrix

A correlation matrix was created to visualize the correlations between the different features. A higher figure describes a higher correlation.



Based on the correlation matrix generated from the dataset, we can see that the most strongly correlated feature with the target column `price` is `sqft_living` with a correlation coefficient of `0.7`. This suggests that there is a strong positive linear relationship between the living area of a house and its price. Houses with larger living areas are likely to have higher prices than those with smaller living areas.

5. DATA MODELING

A regression technique was used in this section.

Regression is a statistical method used to estimate the relationship between a dependent variable and one or more independent variables. The goal of regression analysis is to model the relationship between the variables and to use the model to make predictions or to understand the underlying factors that affect the dependent variable. In this case, it was used to **estimate the effect that the different features of the homes have on the dependent variable, the price of the homes.**

Furthermore, due to the multiple features in the dataset, a multiple linear regression was performed. Multiple linear regression is a regression algorithm that is used to predict the value of a dependent variable based on the value of multiple independent variables (unlike simple linear regression which only uses one independent variable).

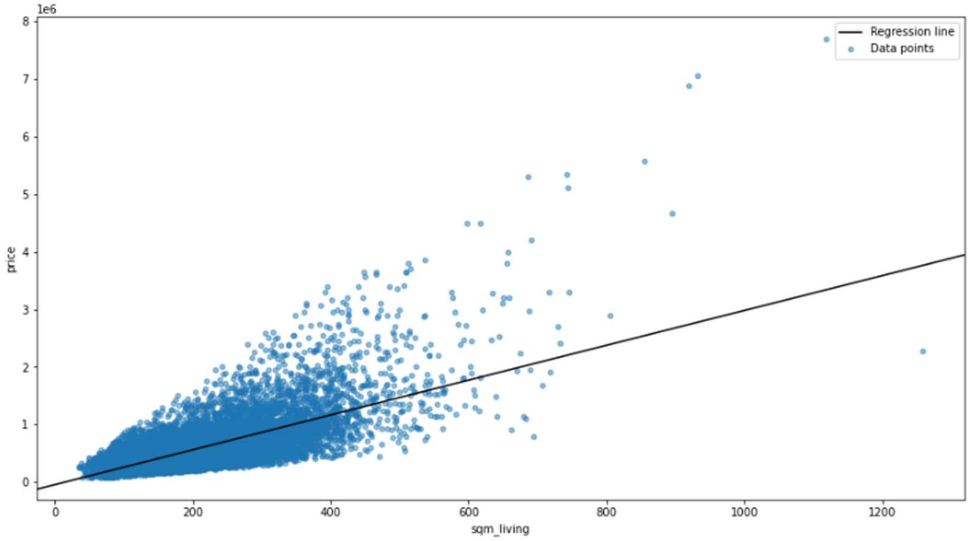
5.1 Simple Linear Regression

The baseline model was built by simple linear regression.

It was performed between two variables: The sale price as the dependent variable and the **size of living in square meters as the independent variable**.

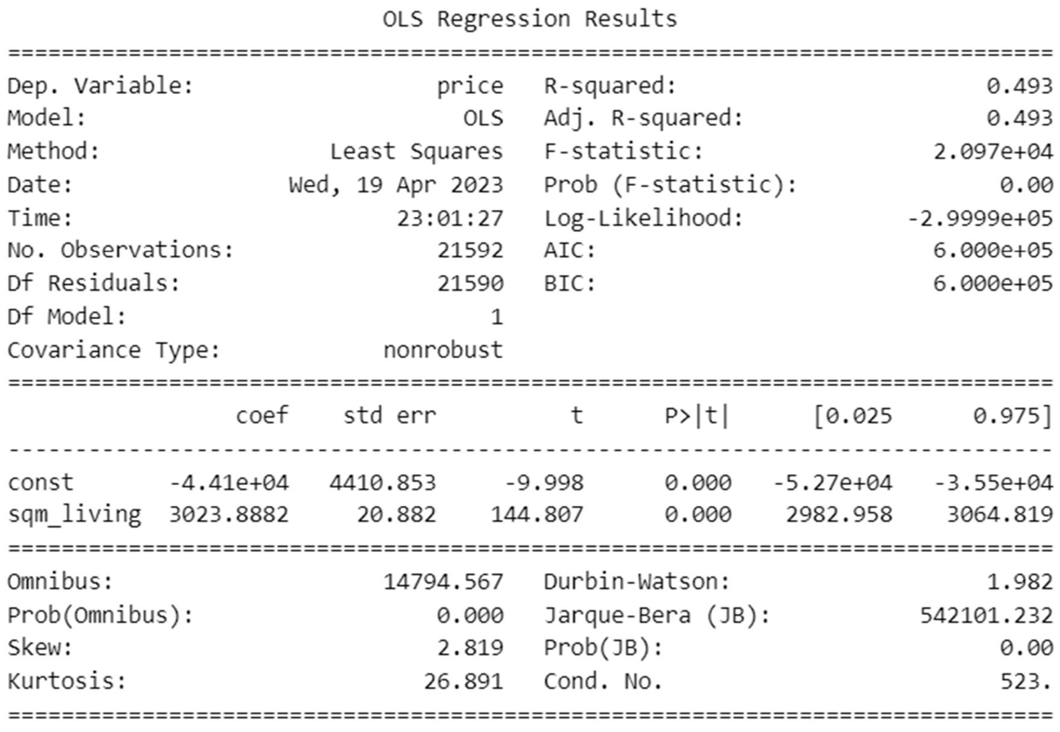
The variable **sqm_living** was selected for this as it has the highest correlation with the price column. This is expected as the size of the house is a major factor in determining the price of the house.

In order to test whether the relationship between sqm_living and price is **linear,** a scatter plot was plotted as shown below.



The scatter plot shows a **linear relationship**. This means that an increase in the area of living space results in an increase in the price.

A simple linear regression was then performed, and the results were obtained as shown below:



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.493
Model:                            OLS   Adj. R-squared:                  0.493
Method:                 Least Squares   F-statistic:                 2.097e+04
Date:                Wed, 19 Apr 2023   Prob (F-statistic):               0.00
Time:                        23:01:27   Log-Likelihood:            -2.9999e+05
No. Observations:               21592   AIC:                         6.000e+05
Df Residuals:                   21590   BIC:                         6.000e+05
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       -4.41e+04   4410.853     -9.998      0.000   -5.27e+04   -3.55e+04
sqm_living  3023.8882     20.882    144.807      0.000    2982.958    3064.819
==============================================================================
Omnibus:                    14794.567   Durbin-Watson:                   1.982
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           542101.232
Skew:                           2.819   Prob(JB):                         0.00
Kurtosis:                      26.891   Cond. No.                         523.
==============================================================================
```

The **Mean Absolute Error (MAE)** for the baseline model was calculated to be 173,829.544. This means that the model is off by about $173,829.544 in the prediction.

MAE is a measure of the average absolute difference between the predicted and actual values. In other words, it quantifies how far off the model's predictions are from the true values, on average.

This was selected to minimize the difference between the predicted and actual prices, and MAE provides a good measure of how well the model is performing.

In general, a lower MAE indicates that the model is making more accurate predictions, whereas a higher MAE suggests that the model's predictions are less reliable.

Baseline Model Evaluation and Interpretation

The baseline model is statistically significant, as it has a P value of less than 0.05 and explains about 49.2% of the variance in price. The model is off by about $173,829.

The coefficients for the intercept, `sqm_living` is statistically significant.

The results show that every increase of 1 in square footage of living space in the home is associated with an increase of $3,023 in the Sale price.

The model is statistically significant except it only explains about 49% variance of our target. However, we target to achieve `R-squared` of about 70% and a lower mean absolute error. This informed the decision to build another model.

5.2 Multiple Linear Regression

The baseline model was iterated by building a multiple linear regression model with more than one independent variable.

A new data frame that contained all the features to be included in the model was created.

In the selection of variables to keep in the model, a correlation matrix was created. This was done in order to reduce **multicollinearity.** Multicollinearity is a situation in which two or more independent variables are highly correlated. This would cause problems in the model as it can lead to unstable estimates of the regression coefficients. Therefore, all the variables that are highly correlated with one another were removed.

Ordinal Encoding

Ordinal encoding converts each label into integer values and the encoded data represents the sequence of labels.

Using the official King County Assessor Website, which is linked below: (https://info.kingcounty.gov/assessor/esales/Glossary.aspx?type=r), it showed that the values in the `condition` and `grade` columns are ordinal, and have been assigned a value based on the quality of the feature. Therefore, these columns were ordinal encoded.

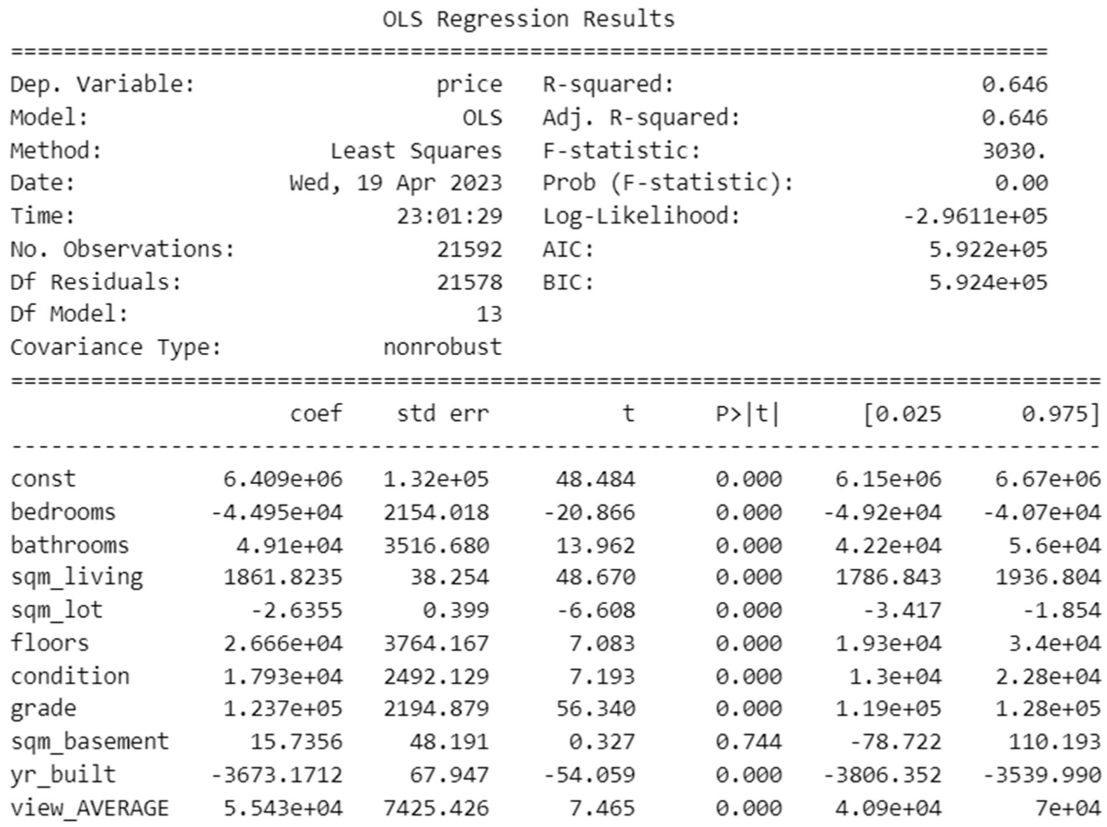
One Hot Encoding

In order to use categorical variables in our model, multiple dummy variables were created, one for each category of the categorical variable.

The data was then checked for Multicollinearity using 0.75 as the cutoff.

Two Multiple Regression Models were performed.

The first multiple Linear regression model was obtained below:



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.646
Model:                            OLS   Adj. R-squared:                  0.646
Method:                 Least Squares   F-statistic:                     3030.
Date:                Wed, 19 Apr 2023   Prob (F-statistic):               0.00
Time:                        23:01:29   Log-Likelihood:            -2.9611e+05
No. Observations:               21592   AIC:                         5.922e+05
Df Residuals:                   21578   BIC:                         5.924e+05
Df Model:                          13
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
const          6.409e+06   1.32e+05     48.484      0.000    6.15e+06    6.67e+06
bedrooms      -4.495e+04   2154.018    -20.866      0.000   -4.92e+04   -4.07e+04
bathrooms        4.91e+04  3516.680     13.962      0.000    4.22e+04     5.6e+04
sqm_living     1861.8235     38.254     48.670      0.000    1786.843    1936.804
sqm_lot          -2.6355      0.399     -6.608      0.000      -3.417      -1.854
floors         2.666e+04   3764.167      7.083      0.000    1.93e+04     3.4e+04
condition      1.793e+04   2492.129      7.193      0.000     1.3e+04    2.28e+04
grade          1.237e+05   2194.879     56.340      0.000    1.19e+05    1.28e+05
sqm_basement     15.7356     48.191      0.327      0.744     -78.722     110.193
yr_built      -3673.1712     67.947    -54.059      0.000   -3806.352   -3539.990
view_AVERAGE   5.543e+04   7425.426      7.465      0.000    4.09e+04       7e+04
```

The **Mean Absolute Error (MAE)** was calculated and found to be: **140,692.2045.**

This means that the model is off by about **$140,692.2045** in the prediction.

Comparing these results to those obtained in the baseline model:

Baseline Model Mean Absolute Error: 173829.54414048957

Iterated Model Mean Absolute Error: 140692.20451544298

Baseline Model Adjusted R-squared: 0.4926808087361043

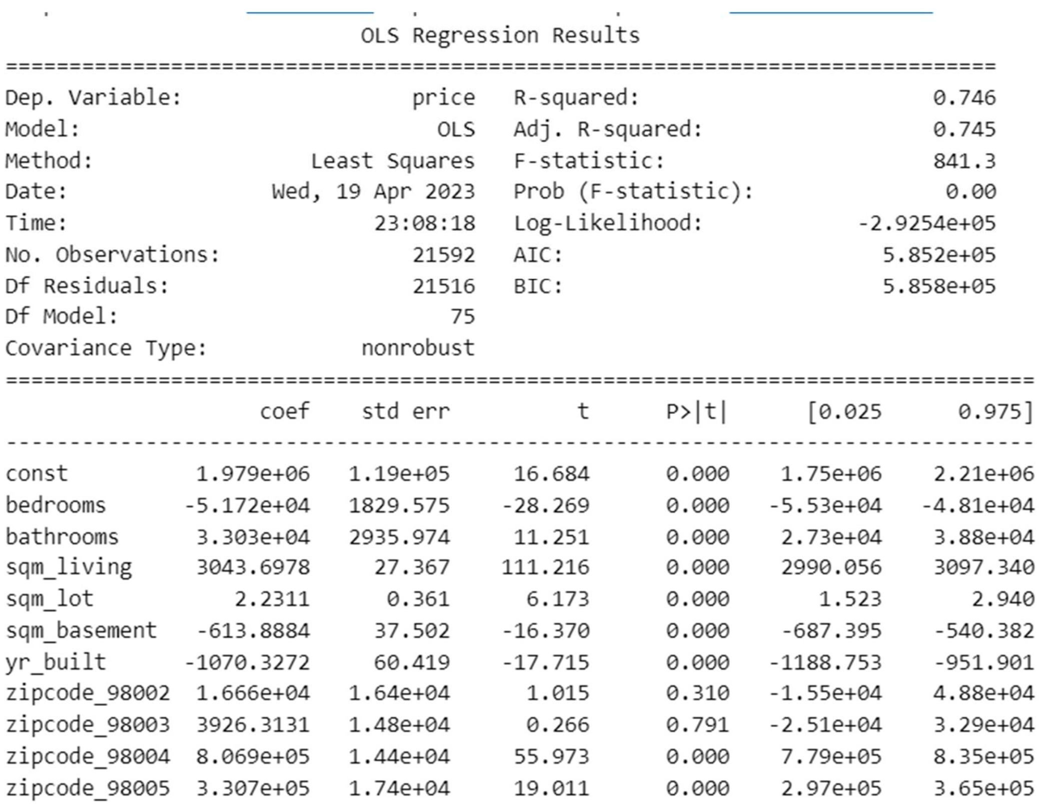
Iterated Model Adjusted R-squared: 0.6458660832294416

Overall, the model performed better.

From the model results, we can see that the model is statistically significant, and it explains 64.5% of the variance in the data compared to the 49.2% in the baseline model.

Furthermore, the model is off by about $140,694 compared to the $173,824 in the baseline model. This is a significant improvement.

Second Multiple Linear Regression Model results were obtained as shown below:



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.746
Model:                            OLS   Adj. R-squared:                  0.745
Method:                 Least Squares   F-statistic:                     841.3
Date:                Wed, 19 Apr 2023   Prob (F-statistic):               0.00
Time:                        23:08:18   Log-Likelihood:            -2.9254e+05
No. Observations:               21592   AIC:                         5.852e+05
Df Residuals:                   21516   BIC:                         5.858e+05
Df Model:                          75
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
const          1.979e+06   1.19e+05     16.684      0.000    1.75e+06    2.21e+06
bedrooms      -5.172e+04   1829.575    -28.269      0.000   -5.53e+04   -4.81e+04
bathrooms      3.303e+04   2935.974     11.251      0.000    2.73e+04    3.88e+04
sqm_living     3043.6978     27.367    111.216      0.000    2990.056    3097.340
sqm_lot           2.2311      0.361      6.173      0.000       1.523       2.940
sqm_basement   -613.8884     37.502    -16.370      0.000    -687.395    -540.382
yr_built      -1070.3272     60.419    -17.715      0.000   -1188.753    -951.901
zipcode_98002  1.666e+04   1.64e+04      1.015      0.310   -1.55e+04    4.88e+04
zipcode_98003  3926.3131   1.48e+04      0.266      0.791   -2.51e+04    3.29e+04
zipcode_98004  8.069e+05   1.44e+04     55.973      0.000    7.79e+05    8.35e+05
zipcode_98005  3.307e+05   1.74e+04     19.011      0.000    2.97e+05    3.65e+05
```

The **Mean Absolute Error (MAE)** was calculated and found to be: **108993.5448**.

This means that the model is off by about **$108,993.5448** in the prediction.

6. REGRESSION RESULTS

6.1 Simple Linear Model

The baseline model is statistically significant overall and explains about 49.2% of the variance in price. Each prediction is off by about $173,829.

The coefficient for the intercept is statistically significant.

Every increase of 1 square meter of living space in the home is associated with an increase of $3023 in the Sale price.

6.2 Multiple Linear Model 1

These results can be interpreted to mean that:

1. A **one-unit increase in the number of** bedrooms is associated with a decrease of **$ 44,945.2** in home price.
2. A **one-unit increase in the number of bathrooms** is associated with an increase of **$49,101.5** in home prices.
3. A **one-unit increase in the area of living space, i.e by 1 square meter,** is associated with an increase of **$1,861.8** in home price.
4. A **one-unit increase in the area of the lot, i.e by 1 square meter,** is associated with a decrease of **$2.635** in home price.
5. A **one-unit increase in the number of floors** is associated with an increase of **$26,661.7** in home prices.
6. A one-unit increase in the **condition rating** of the home is associated with an increase of **$17,925.4** in home price.
7. A one-unit increase in the **grade rating** of the home is associated with an increase of **$123,659.5** in home price.
8. A one-unit increase in the **area of the basement,** i.e by 1 square meter, is associated with an increase of **$15.736** in home price.
9. A one-unit increase in the **year the home was built** is associated with a decrease of **$3,673.2** in home price.
10. Having an **EXCELLENT view** is associated with an increase of **$490,956.0** in home price. This suggests that excellent views are highly desirable and tend to increase the price of a home.
11. Having a **GOOD view** is associated with an increase of **$126,904.6** in home price.
12. Having a **FAIR view** is associated with an increase of **$118,657.0** in home price.
13. Having an **AVERAGE view** is associated with an increase of **$55,4282.3** in home price.

There was an increase of adjusted R-Squared from about 49% in the Baseline Model to about 65% in this model. This was a significant increase, but it did not achieve our target of 70%.

6.3 Multiple Linear Model 2

The model explains about 74% of the variance in price. Each prediction is off by about $ 109,000.

The results obtained are interpreted to mean that:

1. A **one-unit increase in the number of bedrooms** is associated with an increase of **$ 16,410** in home price.
2. A **one-unit increase in square meters of living** space is associated with an increase of **$ 2,813.50 in home price.**
3. A **one-unit increase in square meter of the lot size** is associated with an increase of **$ 2.9292 in home price**.
4. A **one-unit increase in the square meter of the basement** is associated with a decrease of **$ 838.71 in home price.**
5. A **one-unit increase in the year the home was built** is associated with a decrease of **$ 1070.3272 in home price.**
6. As for **the zipcode:** Compared to `zipcode_98103`, `zipcode_98004` has the highest increase of $462,900 in home price and also, Compared to `zipcode_98103`, `zipcode_98092` has the highest decrease of $381,700 in home price.

7. CONCLUSION AND DRAWBACKS

The **second Multiple Linear Model** was chosen as the final model. This is because it explained about 74 % of the variance in price, about **10% more than Multiple Linear Model 1**. It also had a **lower Mean Absolute Error**, by about $ 32,000.

From the final model, `bathroom` is associated with bringing the highest increase in sale price.

An increase in `sqm_living` count by 1 unit had the second highest associated increase in price.

Compared to `zipcode_98103`, `zipcode_98004` has the highest increase of $462,900 in home price.

When building new houses, The Real Estate Developer should therefore prioritize:

1. Increasing the number of bathrooms,
2. Consider the size (square meters) of the living space.
3. Consider building houses in the postal area of `zipcode_98004`

Our final model prediction of the house prices is off by about $109,000.

The study had drawbacks in that it had many missing values.

A further study may be required with a larger dataset for better insights.