King County Real Estate Agency

By: Patricia, Joakim, John & Hanan

Project Overview

01 -------- Explore the Relationship Between the Square Footage of the Home and Housing Prices

02 -------- Assess the Impact of Rennovations on the Housing Price

03 -------- Develop a Linear Regression Model to Predict Housing Prices

Business Problem

This project will address the core business challenge within the King County, USA real estate market. Key stakeholders such as homeowners, real estate agencies, and data science professionals are focused on gaining insights into the factors that impact house prices, enabling them to make informed, data-driven decisions in this dynamic market.

The aim of this project is to conduct a thorough analysis of house sales data in King County, USA, utilizing multiple linear regression modeling techniques. The main goal is to uncover insights into the factors influencing house prices in the region and offer data-driven recommendations for homeowners, real estate agencies, and other stakeholders.

The Data

The dataset used is the King County House Sales dataset. This data represents houses with information on price, bedrooms, bathrooms, sqft living, sqft lot, floors, view, and year built.

Properties of variables of interest:

1. Price: Continuous numeric (float). Represents the sale price of houses in the dataset.
2. Bedrooms: Discrete numeric (integer). Represents the number of bedrooms in each house.
3. Bathrooms: Discrete numeric (integer).Represents the number of bathrooms in each house.
4. Sqft living: Continuous numeric (integer). Represents the total square footage of the living space in each house.
5. Floors: Discrete (integer). Represents the number of floors in each house.
5. View: Categorical (object). Represents the view rating of the property.
6. Year built: Discrete numeric (integer). Represents the year each house was built.

Analysis Steps

01 Data loading & Cleaning

- Importing the necessary Python libraries, including Pandas, NumPy, Matplotlib, and Seaborn, which are commonly used for data manipulation and visualization.
- Reading the house data from a CSV file (“kc_house_data.csv“) into a Pandas DataFrame

02 Feature Selection

Identify the independent variables that most affect our dependent variable, which is the price.

03 Data Analysis

Various visualizations were created to explore the relationships between variables (i.e., scatter plots and box plots), to understand how features such as square footage, the number of bathrooms, bedrooms, floors, and year built relate to house prices.

04 Modelling & Validation

Investigate polynomial relationships and interactions between variables in greater details using regression modelling.

Data Cleaning

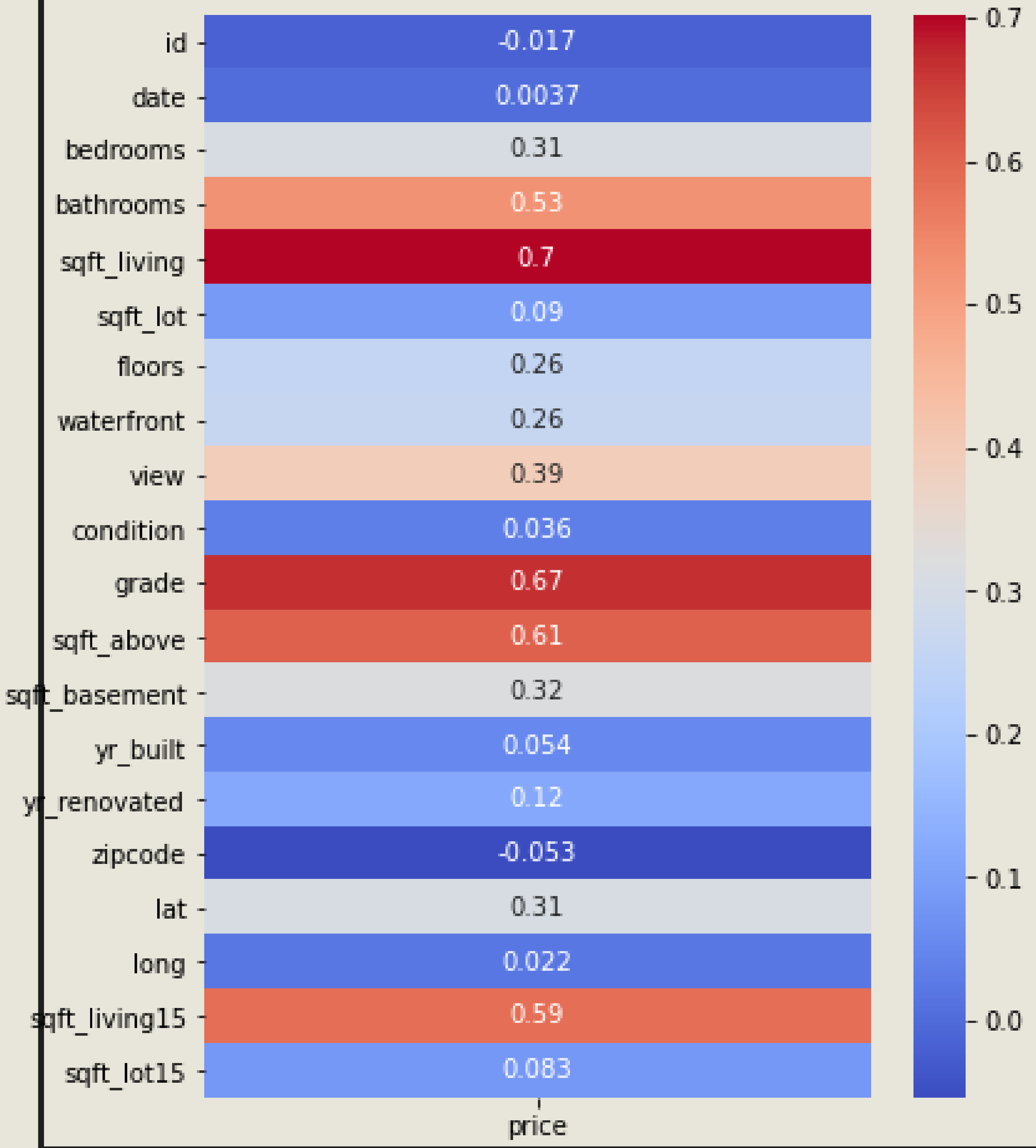
To clean the data we used house_data_df.head() to:

- inspect the first few rows of the DataFrame.
- Checked the shape of the DataFrame using house_data_df.shape to determine the number of rows and columns.
- Used house_data_df.info() to get information about the data types and missing values in each column.
- Checked for duplicate rows using house_data_df.duplicated().

We then cleaned up the data by:

- Addressing the null values in yr_renovated column by replacing the null values with zero as 78% of the houses have not been renovated;
- Replacing the null values in waterfront and view columns with 0 seeing as that is the mode;
- Convert all non-numeric data values to numeric values for analysis i.e., the date column contains values of type string therefore, before any numerical analysis, this should be converted to a numerical data type

Feature Selection



The purpose of feature selection is to identify the independent variables that most affect our dependent variable, which in this case, is the price.

This was done by generating a heatmap of the correlation matrix using Seaborn. Each cell in the matrix shows the correlation coefficient between price and the other features.

Feature Selection

From the heatmap, it is clear that the most impactful features to the house price are:

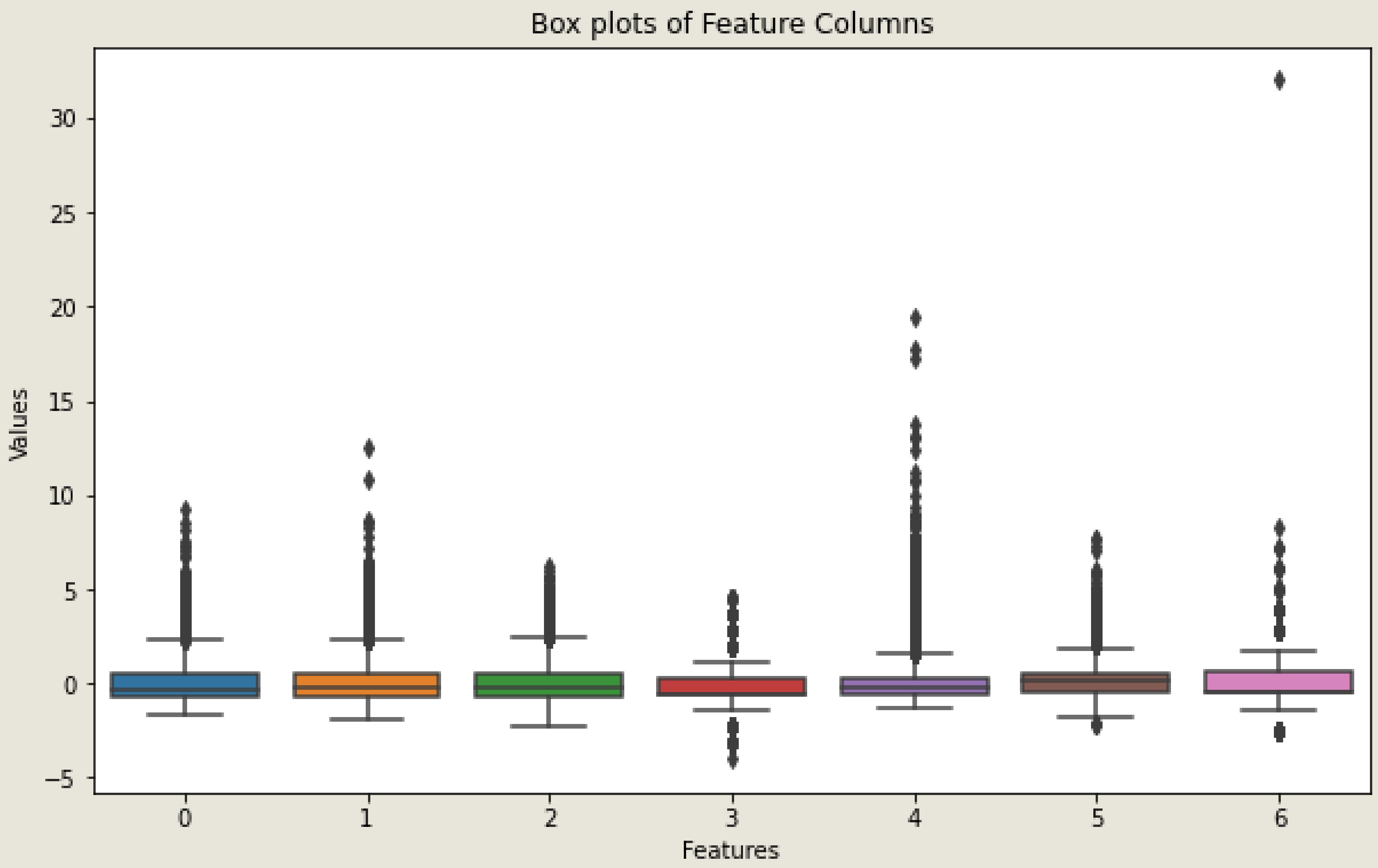
- Square footage of the home(sqft_living);
- Overall grade given to the housing unit based on King County grading system(grade);
- Square footage of the house apart from the basement(sqft_above);
- The square footage of interior housing living space for the nearest 15 neighbors(sqft_living15)

Feature Selection

Feature Cleaning

We started by looking at the selected features' columns for any outliers.

The boxplot here shows that all the columns have outliers.

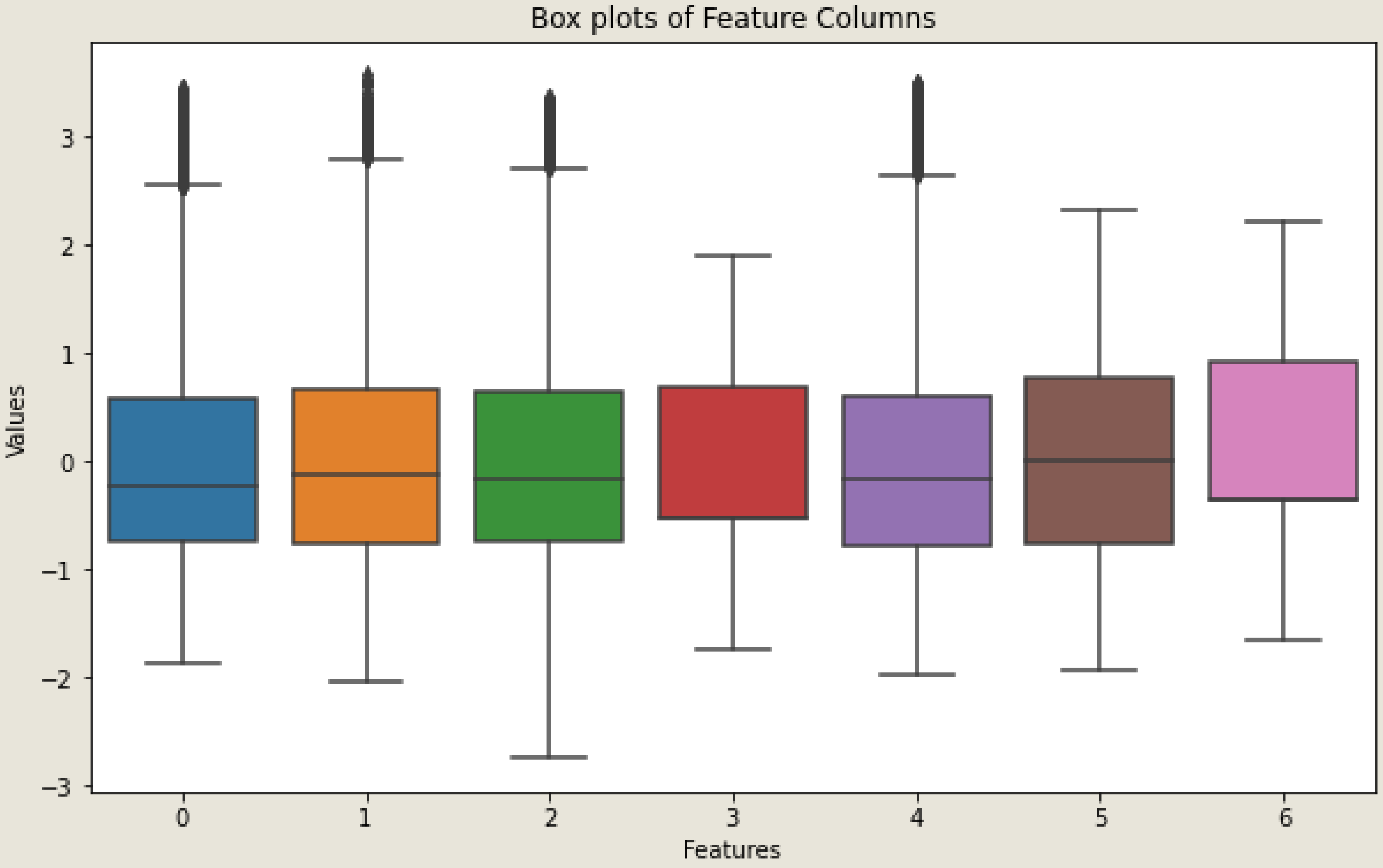


Feature Selection

Feature Cleaning

We then defined the upper and lower bounds to identify and filter out the outliers from the DataFrame.

The boxplot here shows the features without the outliers.

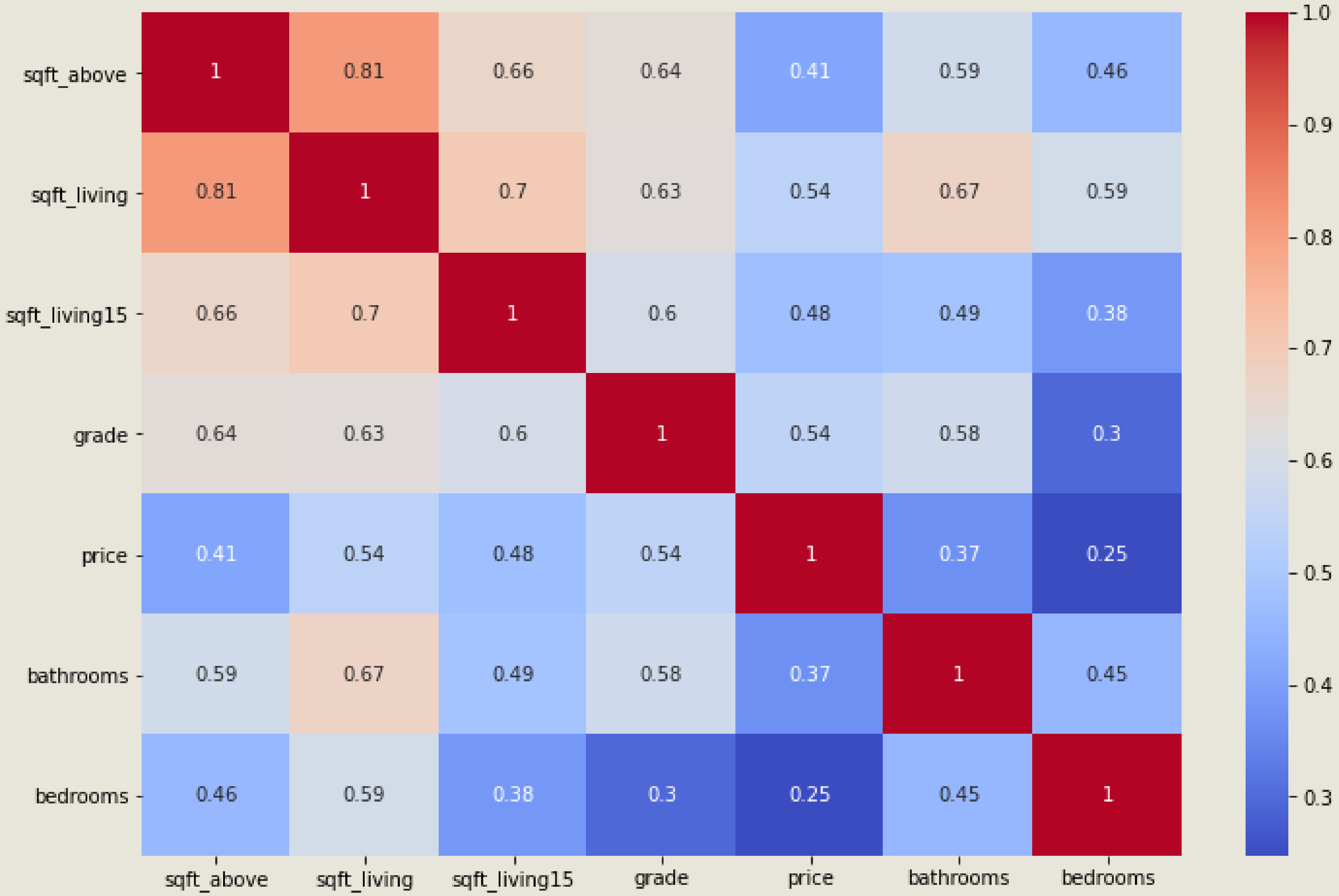


Feature Selection

Feature Cleaning

Next, we check feature collinearity using a correlation heatmap for better selection.

From the heatmap, Sqft_living and sqft_above have a high correlation score indicating a high collinearity. Drop sqft_living for the feature list for better modelling with low collinearity.

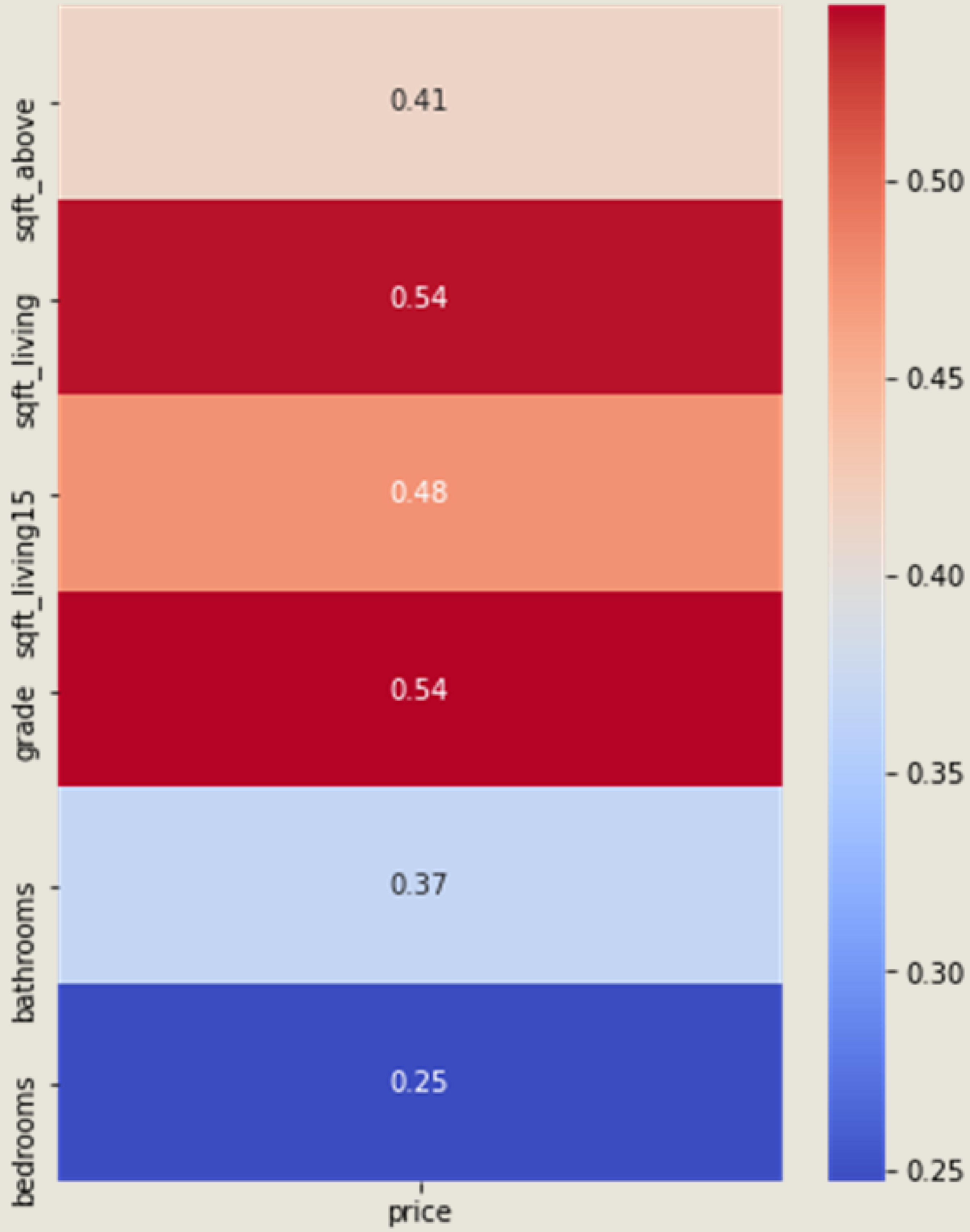


Data Analysis 1

Relationship between the square foot living area and housing prices

From the heatmap of the selected features against the Price, square footage of the home(sqft_living) has the highest correlation to the house price i.e.,0.54.

This shows that square footage area of the home has a very high co-efficient to the price of the house.

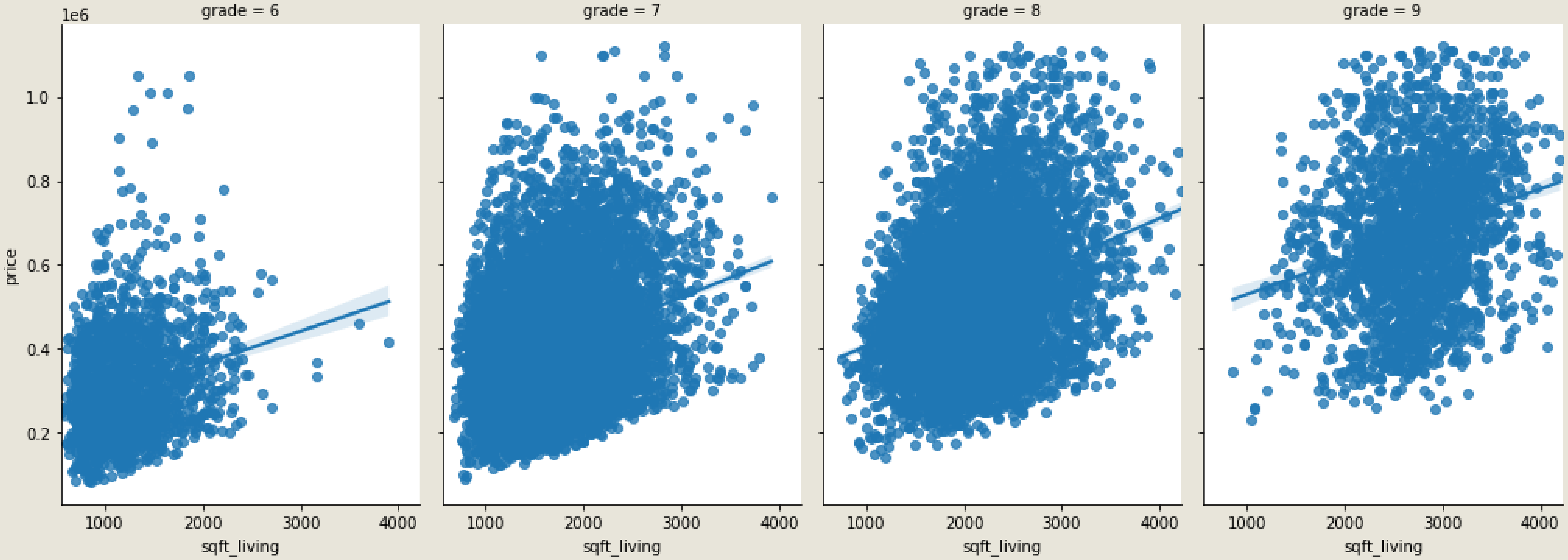


Data Analysis 1

Relationship between the square foot living area and housing prices

As seen from the regression plot (see next slide), as the square footage of the home increases so does the price of the home across different grades of houses further confirming that square footage area of the home has a very high co-efficient to the price of the house.

Regression Plot 1



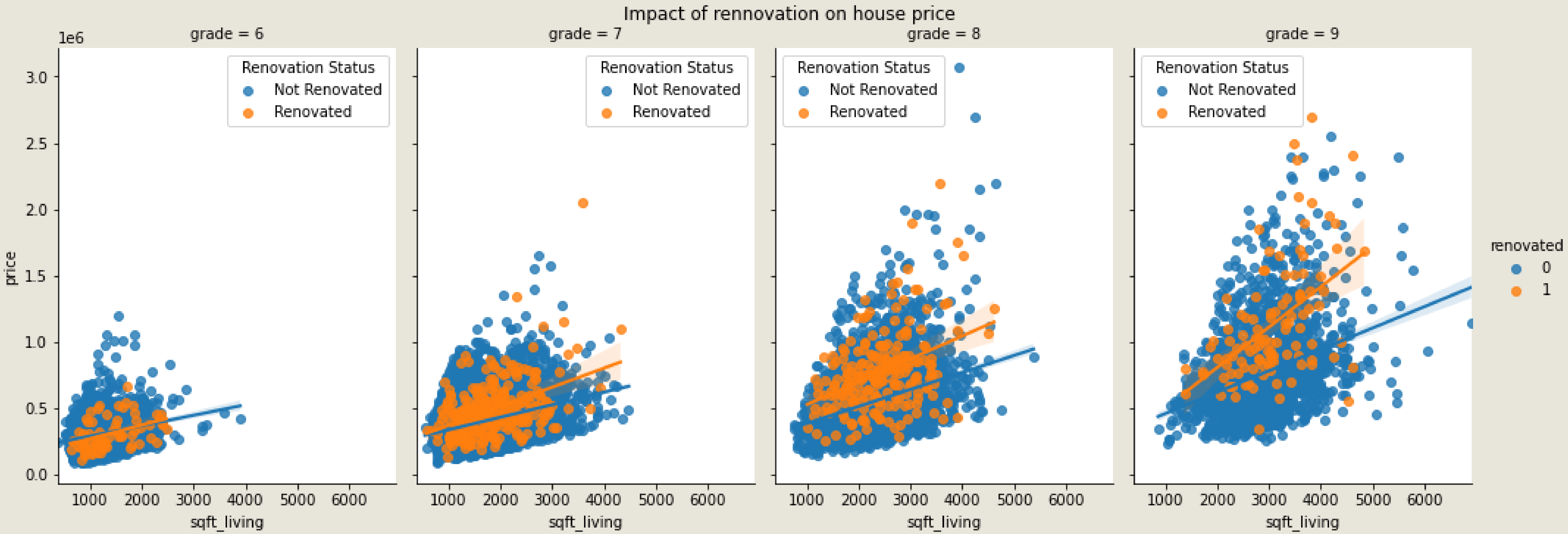
Data Analysis 2

Assess the impact of house rennovations to the house price

In this analysis, we generate a regression plot to show the impact of house renovations on Price, while keeping the other independent variables constant.

The regression plot generated (see next slide), shows the price increase in houses that have been rennovated in comparison to the houses without rennovations. The linear regression model also takes into account the other independent variables with a high correlation to price(i.e., grade and square foot living space)

Regression Plot 2



Modelling

The following steps were taken while modelling:

1. Start with the data with outliers and create the baseline model.

2. Add one predictor (independent) variable.

3. Check the performance.

4. Add a categorical variable.

5. Repeat steps 2 - 4 until adequate performance is reached.

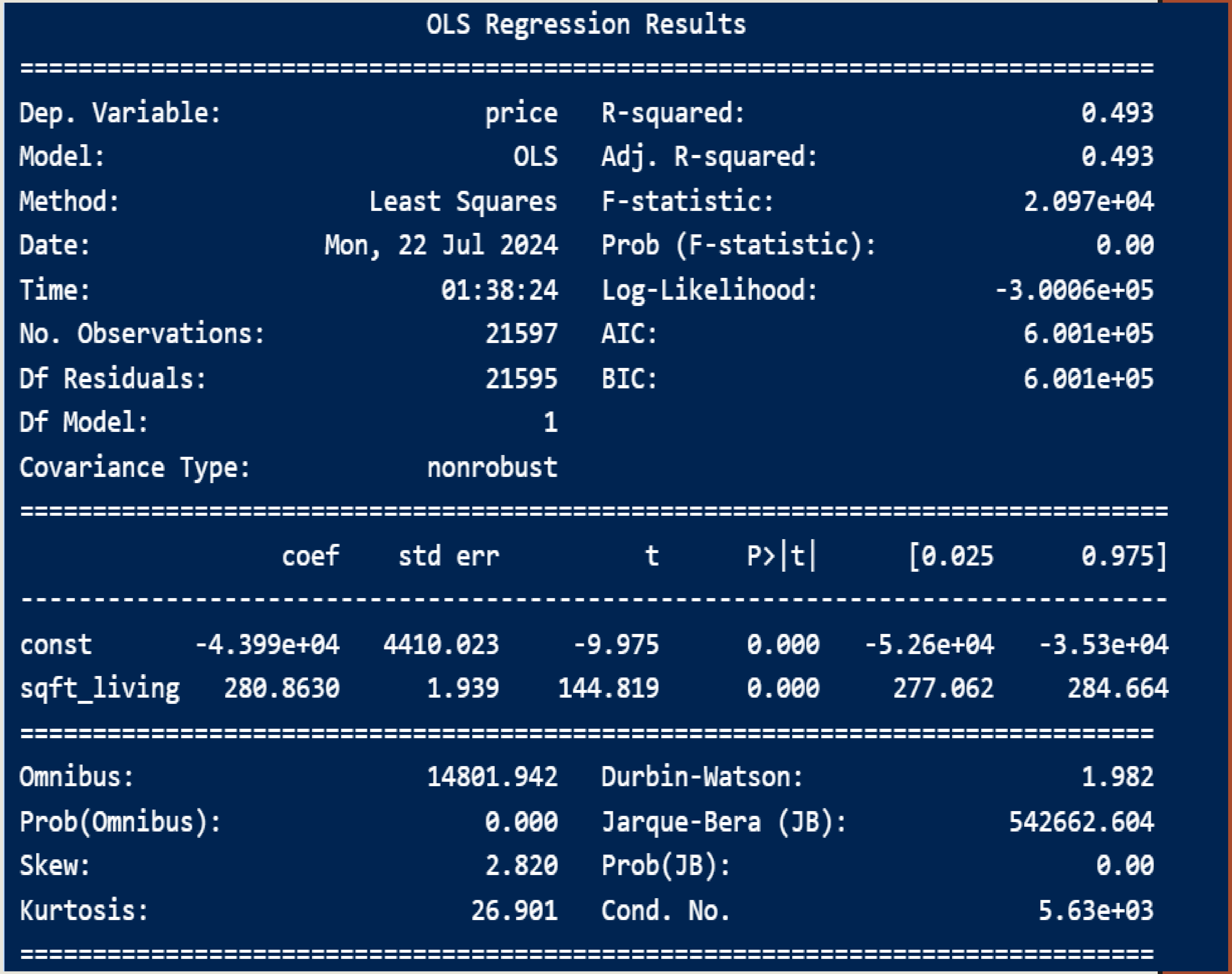
6. Repeat these steps for the data without outliers and choose the best model.

Modelling

1. Base Model

From the heatmap in Data Analysis 1, it was clear that the square footage of the home(sqft_living), has the highest correlation to the house price. The OLS regression results confirms the results of the heatmap i.e.,

- The model is statistically significant overall, with an F-statistic p-value well below 0.05
- The model explains about 49.6% of the variance in price meaning, that 49.6% of the variance in house prices is explained by the square footage of living space ("sqft_living").
- The model coefficients (const and sqft_living) are both statistically significant, with t-statistic p-values well below 0.05.



```
                            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:                Mon, 22 Jul 2024   Prob (F-statistic):               0.00
Time:                        01:38:24   Log-Likelihood:            -3.0006e+05
No. Observations:               21597   AIC:                         6.001e+05
Df Residuals:                   21595   BIC:                         6.001e+05
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                  coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       -4.399e+04   4410.023     -9.975      0.000   -5.26e+04   -3.53e+04
sqft_living   280.8630      1.939    144.819      0.000     277.062     284.664
==============================================================================
Omnibus:                    14801.942   Durbin-Watson:                   1.982
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           542662.604
Skew:                           2.820   Prob(JB):                         0.00
Kurtosis:                      26.901   Cond. No.                     5.63e+03
==============================================================================
```

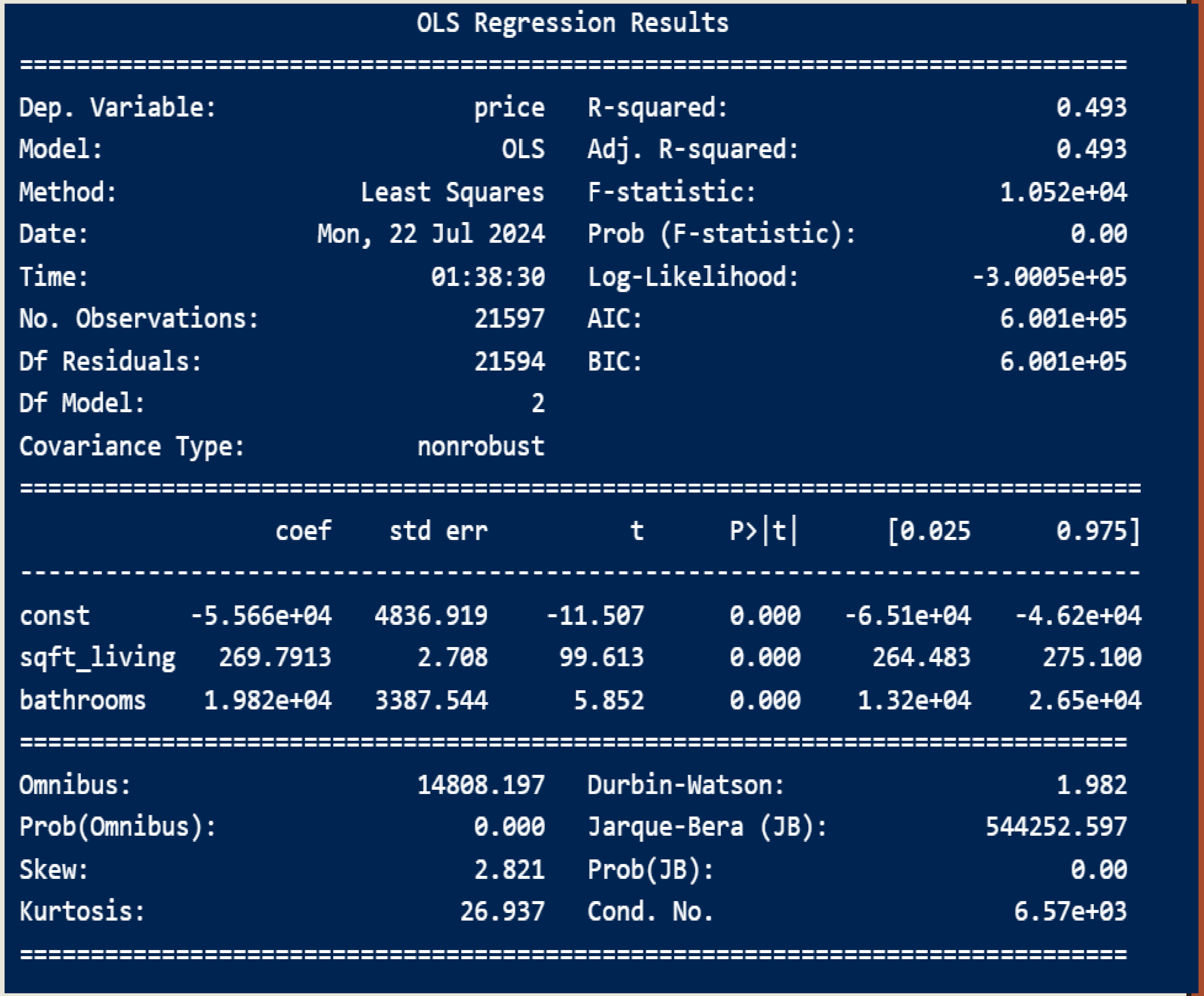
Modelling

2. Adding another independent variable

To improve the overall model performance, we add in another independent variable i.e., bathrooms.

The OLS regression results indicates:

- The model is statistically significant overall, with an F-statistic p-value well below 0.05
- The model exhibits an R-squared value of 49.7% indicating that approximately 49.7% of the variance in house prices is explained by the square footage of living space ("sqft_living") and the number of bathrooms in the houses. This shows a slight improvement of 0.1%.
- The model coefficients (const, sqft_living and bathrooms) are statistically significant, with t-statistic p-values well below 0.05.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.493
Model:                            OLS   Adj. R-squared:                  0.493
Method:                 Least Squares   F-statistic:                 1.052e+04
Date:                Mon, 22 Jul 2024   Prob (F-statistic):               0.00
Time:                        01:38:30   Log-Likelihood:            -3.0005e+05
No. Observations:               21597   AIC:                         6.001e+05
Df Residuals:                   21594   BIC:                         6.001e+05
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const      -5.566e+04   4836.919    -11.507      0.000   -6.51e+04   -4.62e+04
sqft_living  269.7913      2.708     99.613      0.000     264.483     275.100
bathrooms   1.982e+04   3387.544      5.852      0.000    1.32e+04    2.65e+04
==============================================================================
Omnibus:                    14808.197   Durbin-Watson:                   1.982
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           544252.597
Skew:                           2.821   Prob(JB):                         0.00
Kurtosis:                      26.937   Cond. No.                     6.57e+03
==============================================================================
```

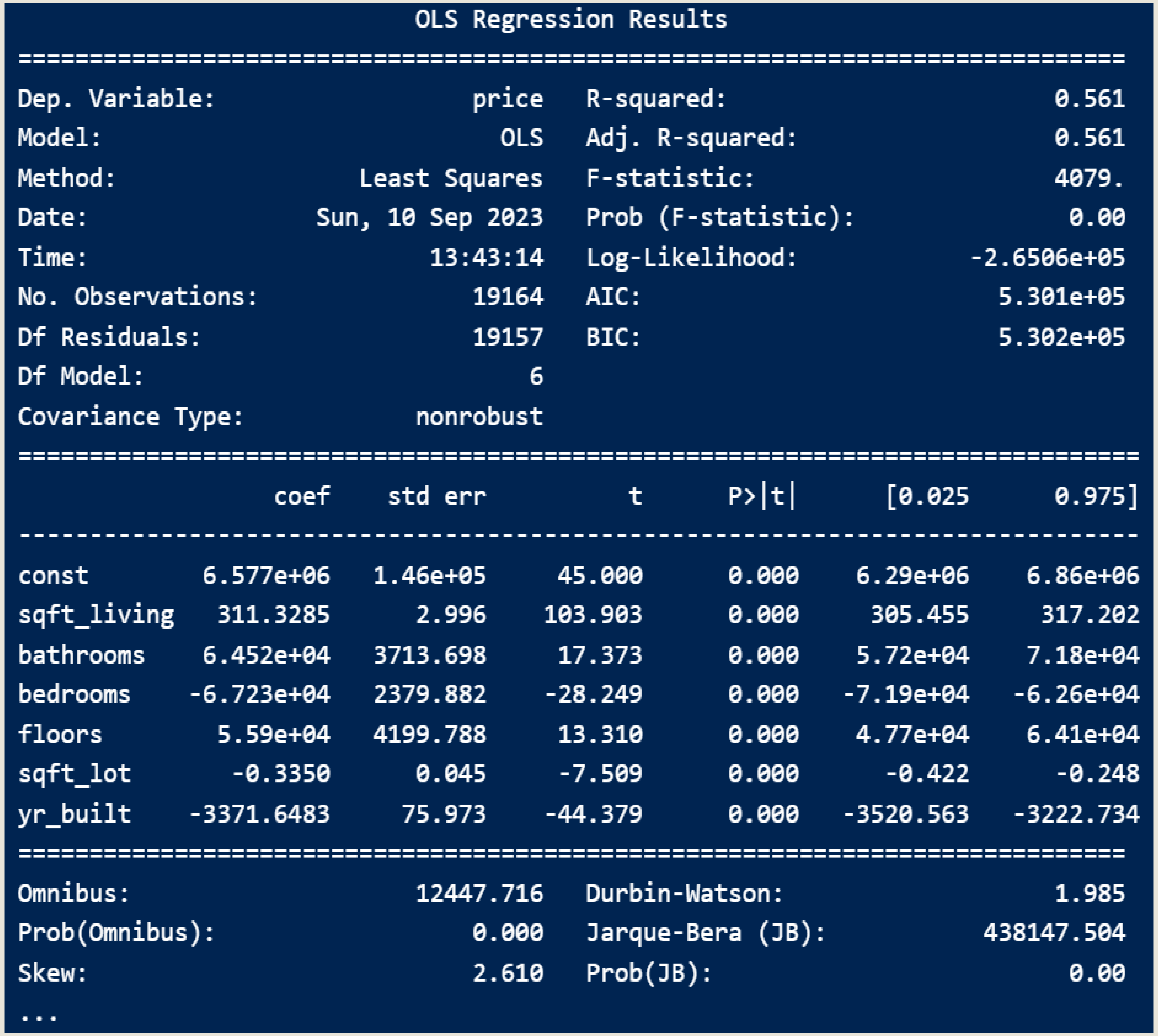
Modelling

3. Adding all the features

To improve the overall model performance further, we then add in all the other independent variables i.e., bedrooms, floors, sqft_lot & yr_built.

The OLS regression results indicates:

- The model is statistically significant overall, with an F-statistic p-value well below 0.05.
- The model exhibits an R-squared value of 56.1% of the variance in price, indicating that approximately 56.1% of the variance in house prices is explained by the model in this iteration and its predictor variables.
- This is an overall improvement of 6.4% from the last model.
- The model coefficients are statistically significant, with t-statistic p-values well below 0.05.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.561
Model:                            OLS   Adj. R-squared:                  0.561
Method:                 Least Squares   F-statistic:                     4079.
Date:                Sun, 10 Sep 2023   Prob (F-statistic):               0.00
Time:                        13:43:14   Log-Likelihood:            -2.6506e+05
No. Observations:               19164   AIC:                         5.301e+05
Df Residuals:                   19157   BIC:                         5.302e+05
Df Model:                           6
Covariance Type:            nonrobust
==============================================================================
                  coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        6.577e+06   1.46e+05     45.000      0.000    6.29e+06    6.86e+06
sqft_living   311.3285      2.996    103.903      0.000     305.455     317.202
bathrooms    6.452e+04   3713.698     17.373      0.000    5.72e+04    7.18e+04
bedrooms    -6.723e+04   2379.882    -28.249      0.000   -7.19e+04   -6.26e+04
floors        5.59e+04   4199.788     13.310      0.000    4.77e+04    6.41e+04
sqft_lot       -0.3350      0.045     -7.509      0.000      -0.422      -0.248
yr_built    -3371.6483     75.973    -44.379      0.000   -3520.563   -3222.734
==============================================================================
Omnibus:                    12447.716   Durbin-Watson:                   1.985
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           438147.504
Skew:                           2.610   Prob(JB):                         0.00
...
```

Modelling

4. Adding a categorical variable

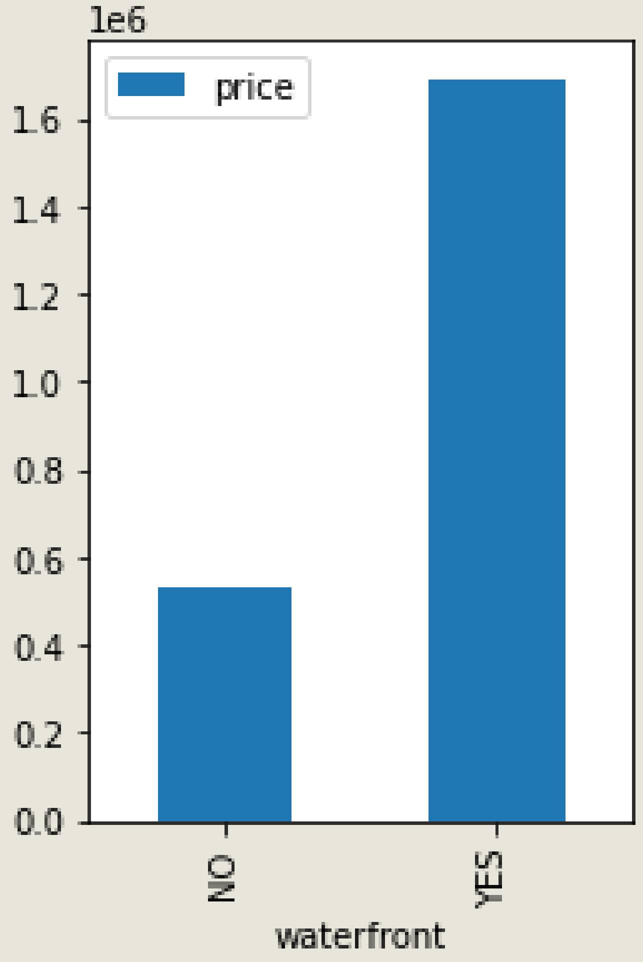
Our dataset has 4 categorical variables i.e., waterfront, view, condition and grade.

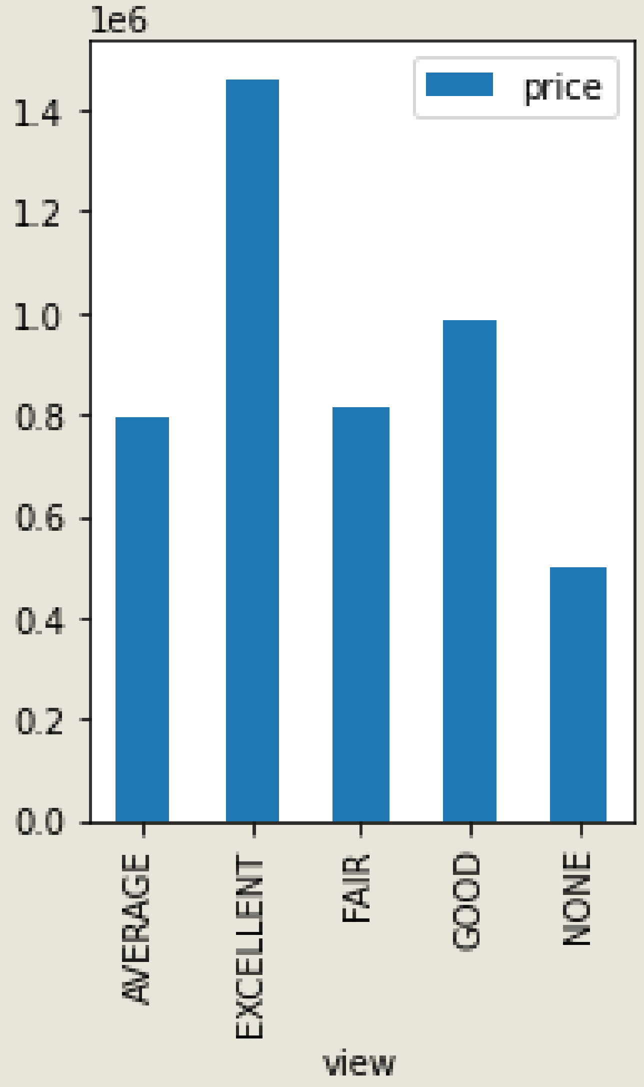
In a bid to improve the model further, we pick one categorical variable based on which one categorizes the houses' condition the best.

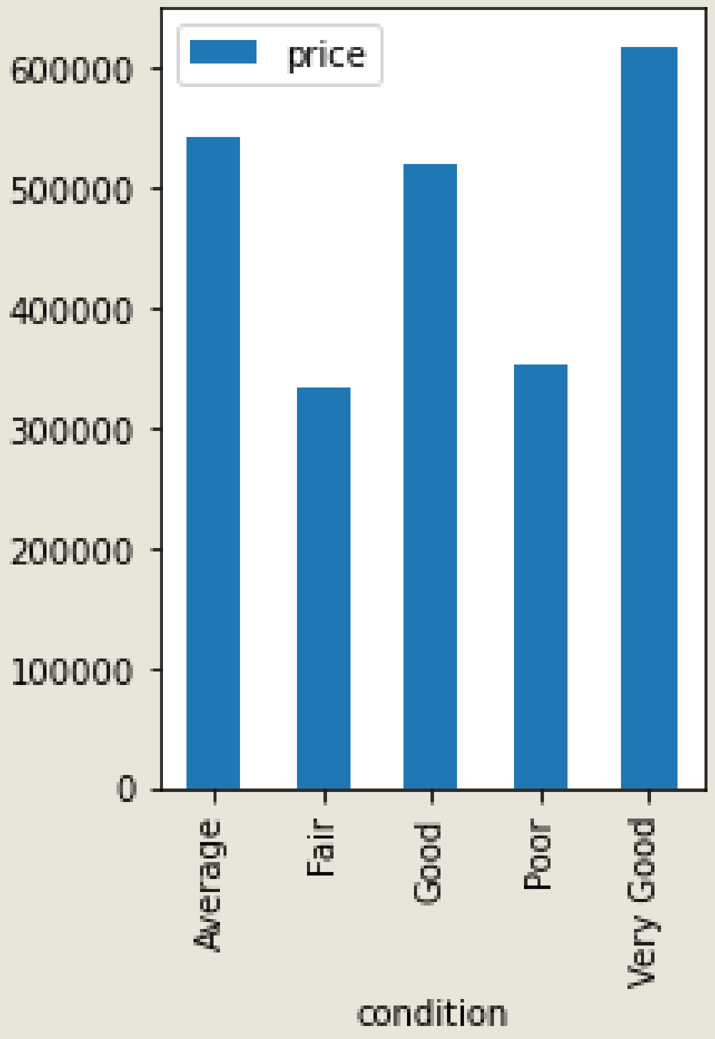
Based on the graphs (see next slide), our best option is view as it categorizes the condition of the house from NONE to EXCELLENT.

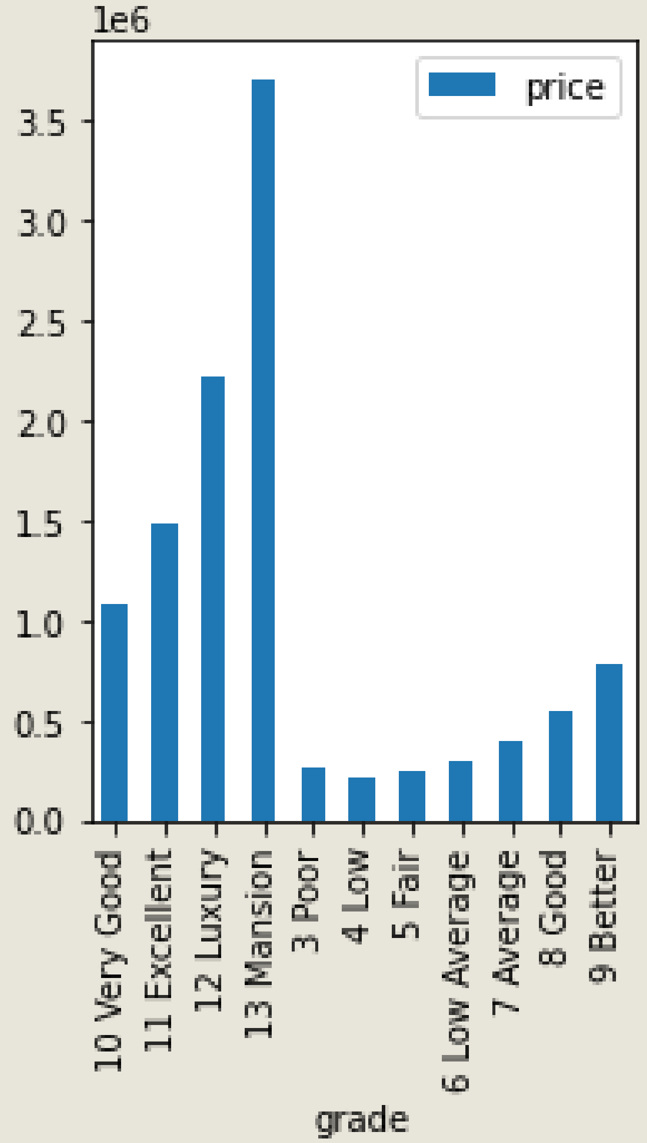
Modelling

4. Adding a categorical variable









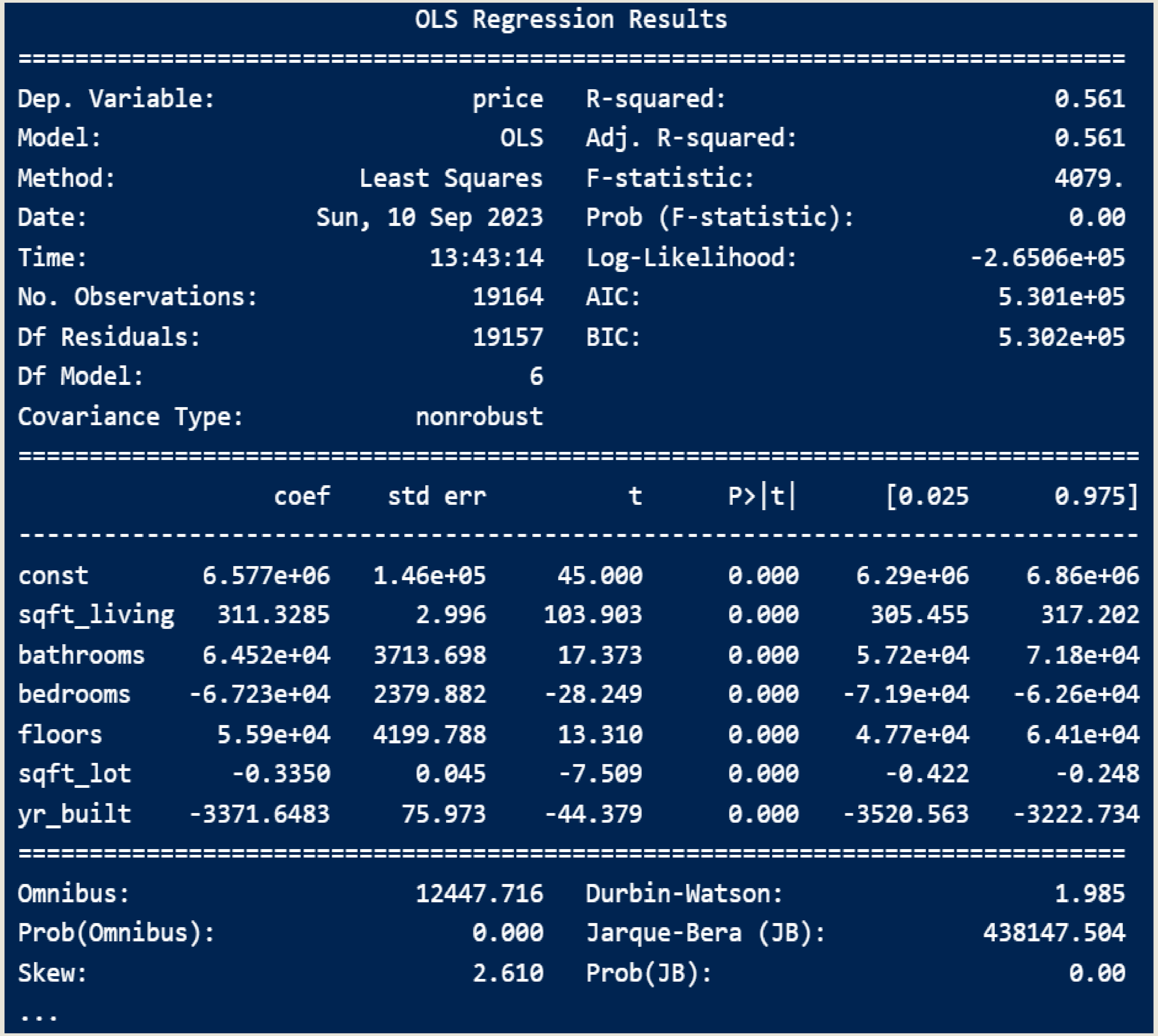
Modelling

4. Adding a categorical variable

The OLS regression results indicates:

- The model is statistically significant overall, with an F-statistic p-value well below 0.05

- The model exhibits an R-squared value of 59.5% of the variance in price, meaning 59.5% of the variance in house prices is explained by the model in this iteration and its predictor variables.
- There is also an overall improvement of 3.4% from the last model.
- The model coefficients are statistically significant, with t-statistic p-values well below 0.05



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.561
Model:                            OLS   Adj. R-squared:                  0.561
Method:                 Least Squares   F-statistic:                     4079.
Date:                Sun, 10 Sep 2023   Prob (F-statistic):               0.00
Time:                        13:43:14   Log-Likelihood:            -2.6506e+05
No. Observations:               19164   AIC:                         5.301e+05
Df Residuals:                   19157   BIC:                         5.302e+05
Df Model:                           6
Covariance Type:            nonrobust
==============================================================================
                  coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        6.577e+06   1.46e+05     45.000      0.000    6.29e+06    6.86e+06
sqft_living   311.3285      2.996    103.903      0.000     305.455     317.202
bathrooms    6.452e+04   3713.698     17.373      0.000    5.72e+04    7.18e+04
bedrooms    -6.723e+04   2379.882    -28.249      0.000   -7.19e+04   -6.26e+04
floors        5.59e+04   4199.788     13.310      0.000    4.77e+04    6.41e+04
sqft_lot       -0.3350      0.045     -7.509      0.000      -0.422      -0.248
yr_built    -3371.6483     75.973    -44.379      0.000   -3520.563   -3222.734
==============================================================================
Omnibus:                    12447.716   Durbin-Watson:                   1.985
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           438147.504
Skew:                           2.610   Prob(JB):                         0.00
...
```

Conclusion 1

From exploring the relationship between the square foot living area and housing prices, it is clear that there is a strong positive correlation (r = 0.7) between square foot living area(sqft_living) and housing prices(price). This indicates that larger properties tend to command higher prices in the Kings County housing market.

The high correlation indicates that the property size significantly influences housing prices, guiding decisions for real estate investors and homebuyers.

Conclusion 2

From assessing the impact of house rennovations to the house price, we fount that there is a clear price increase in houses that have been rennovated in comparison to houses without rennovations. The linear regression model and also takes into account the other independent variables with a high correlation to price(grade and square foot living space)

This suggests that upgrading the overall grade and quality of your home can lead to higher market prices.

Recommendation 1

Focus on Key Upgrades

Upgrading the overall grade and quality of your home, such as high-end finishes, better materials, and improved aesthetics, can lead to higher market prices. Focus on renovations that improve the grade of your home.

Recommendation 2

Enhance Living Space

Increasing the living space (sqft_living) is a valuable investment. This could include adding extensions or converting unused areas (like basements or attics) into livable spaces.

Recommendation 3

Have knowledge of Kings County House Grading System

Incorporate insights from Kings County on how they grade houses. The house grade has a high impact on the price, hence having knowledge of that grading system is important.