PHASE-TWO PROJECT

Done by : GROUP 17

MEMBERS : Lucy Munge, Frank Kiptoo and Curtis Kariuki

PART – TIME STUDENTS

Table of Contents

The table of contents of this presentation are:

- Purpose for the presentation
- Visualization
- Model presentation
  - ✓Linear Regression (Baseline Model)
  - ✓Simple Linear Regression Visual (2nd Models)
  - ✓3rd Model
- Conclusion
- Recommendation
- Summary

PURPOSE FOR THE PRESENTATION

This presentation was constructed for real estate agency Amani is facing a challenge in providing valuable advice to homeowners regarding home renovations. Homeowners often inquire about the potential increase in the estimated value of their homes after making specific renovations or improvements. The agency needs to develop a predictive model that can accurately estimate the impact of various renovation projects on a home's market value within the northwestern county.

The goal is to offer data-driven recommendations to homeowners, enabling them to make informed decisions about which renovations to undertake and how these renovations will affect the resale value of their homes.

The business questions to be answered are:

- How does the number of bedrooms, bathrooms, grade and square footage of a house correlate with its sale price in King County?(This acts as a guidance to home owners on selling or buying or renovation of a house will affect the price.)

- How much can a homeowner expect the value of their home to increase after a specific renovation project?

- Which renovation projects have the most significant impact on a home's market value in the northwestern county?

- Are there specific combinations of renovation projects that provide an interdependent effect on a home's market value?

This project uses the King County House Sales dataset. The file contains information on over 21,000 housing units.

The columns used in the analysis are:

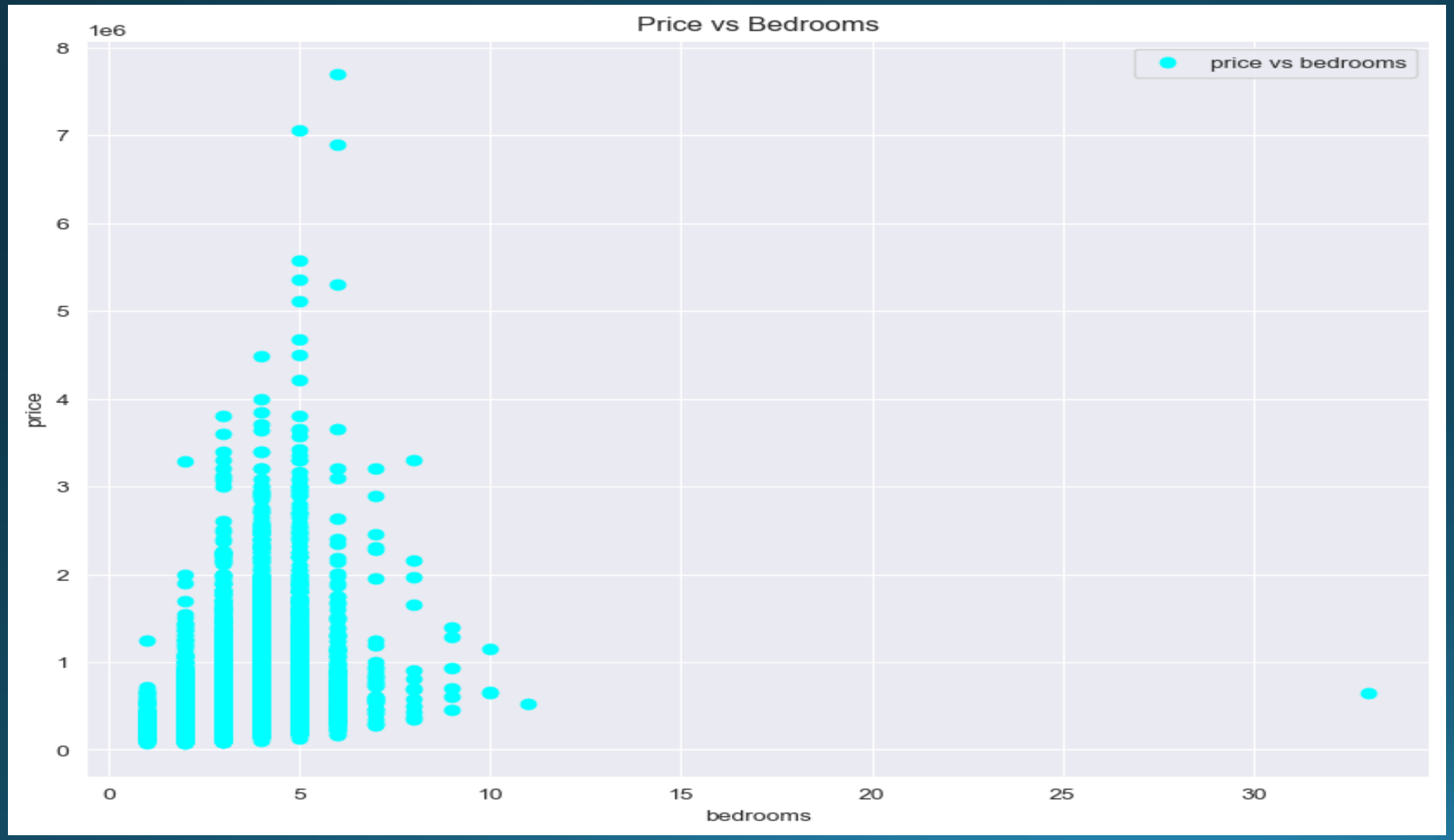
- ✓Price - Sale price (prediction target)
- ✓Condition - How good the overall condition of the house is. Related to maintenance of house
- ✓Bedrooms - Number of bedrooms
- ✓Bathrooms - Number of bathrooms
- ✓Sqft_living - Square footage of living space in the home
- ✓Floors - Number of floors (levels) in house

The data set underwent analysis so that meaningful might be drawn. The processes that were done in data analysis were data cleaning(which included handling missing data and outliers), visualization(that show findings within the dataset) and use of models that were used to draw conclusions and recommendations.

VISUALIZATION

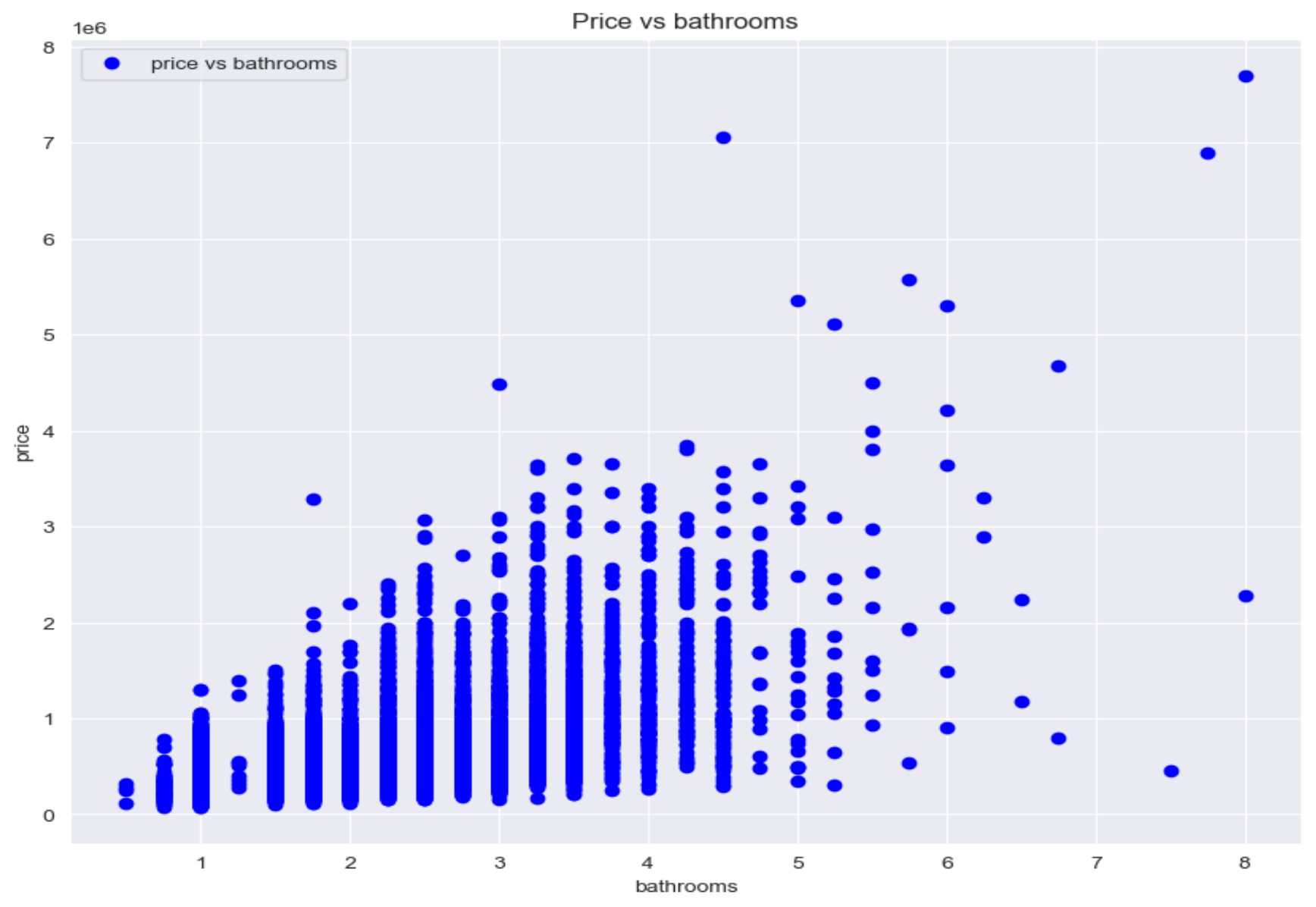
Now we are going to go through the different visuals that were prepared to show the findings of the data set. This visuals were plotted in graph format:

1. PRICE AGAINST BEDROOMS



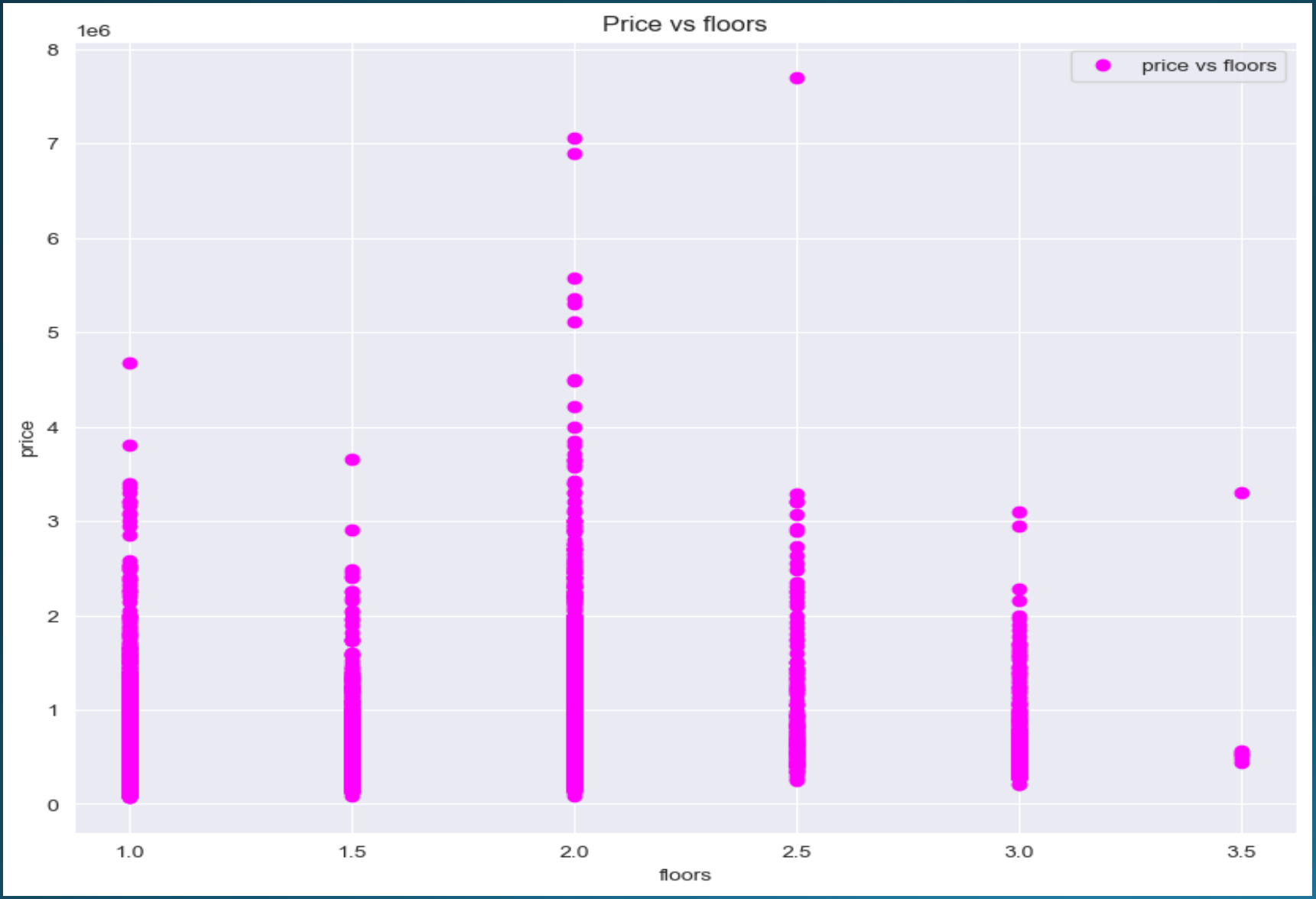
- We can see in the graph 4 – 8 bedrooms seem to fetch a higher price as the number of bedrooms increase.

2. PRICE VS BATHROOMS



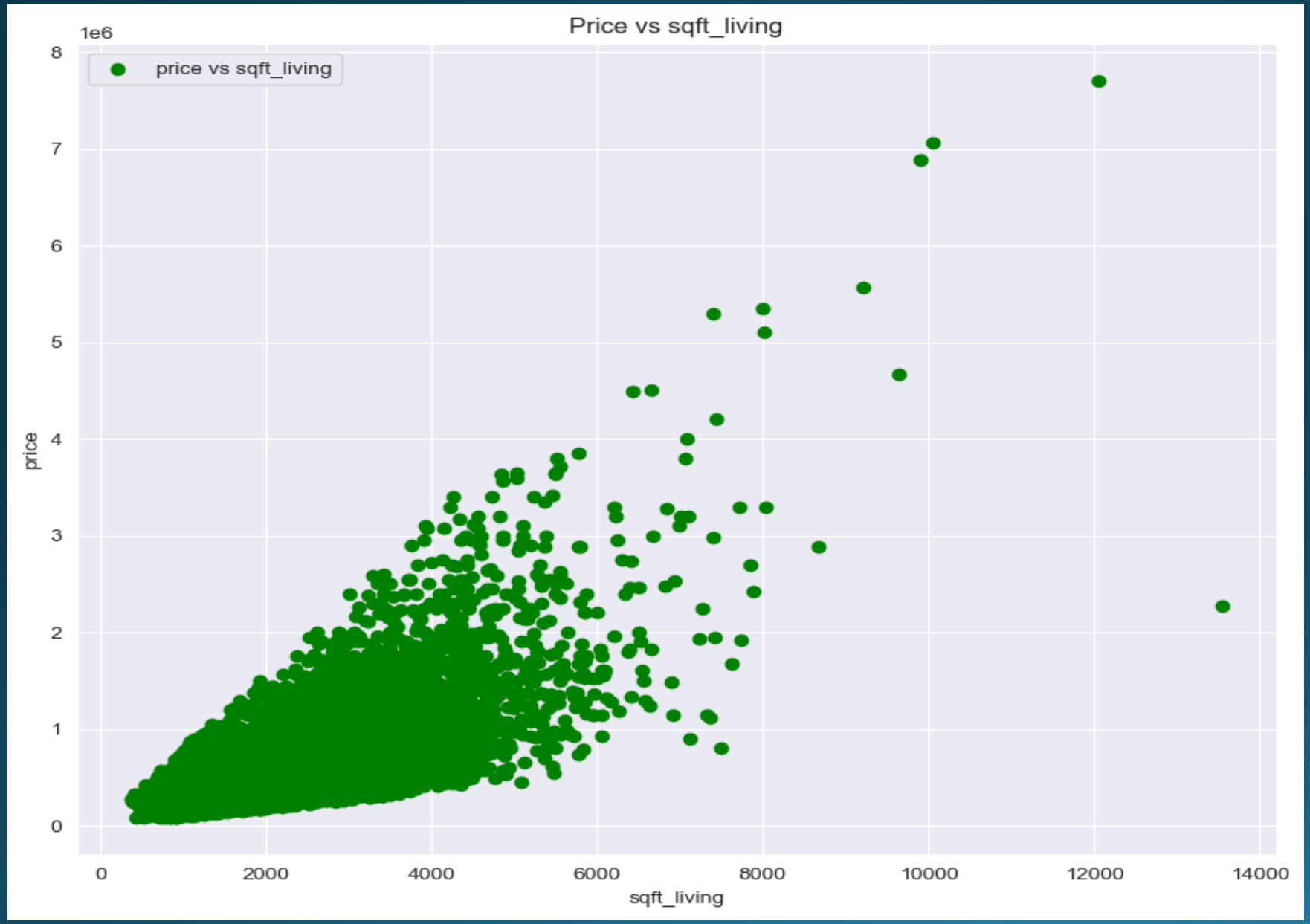
- We can see in the graph that as the number of bathrooms increase the price of property also increases.

3. PRICE VS FLOORS



- We can see that houses with floors of 1.0 to 3.0 seem to increasing value of the property thus indicating it as high preference.

4. PRICE VS SQFT_LIVING

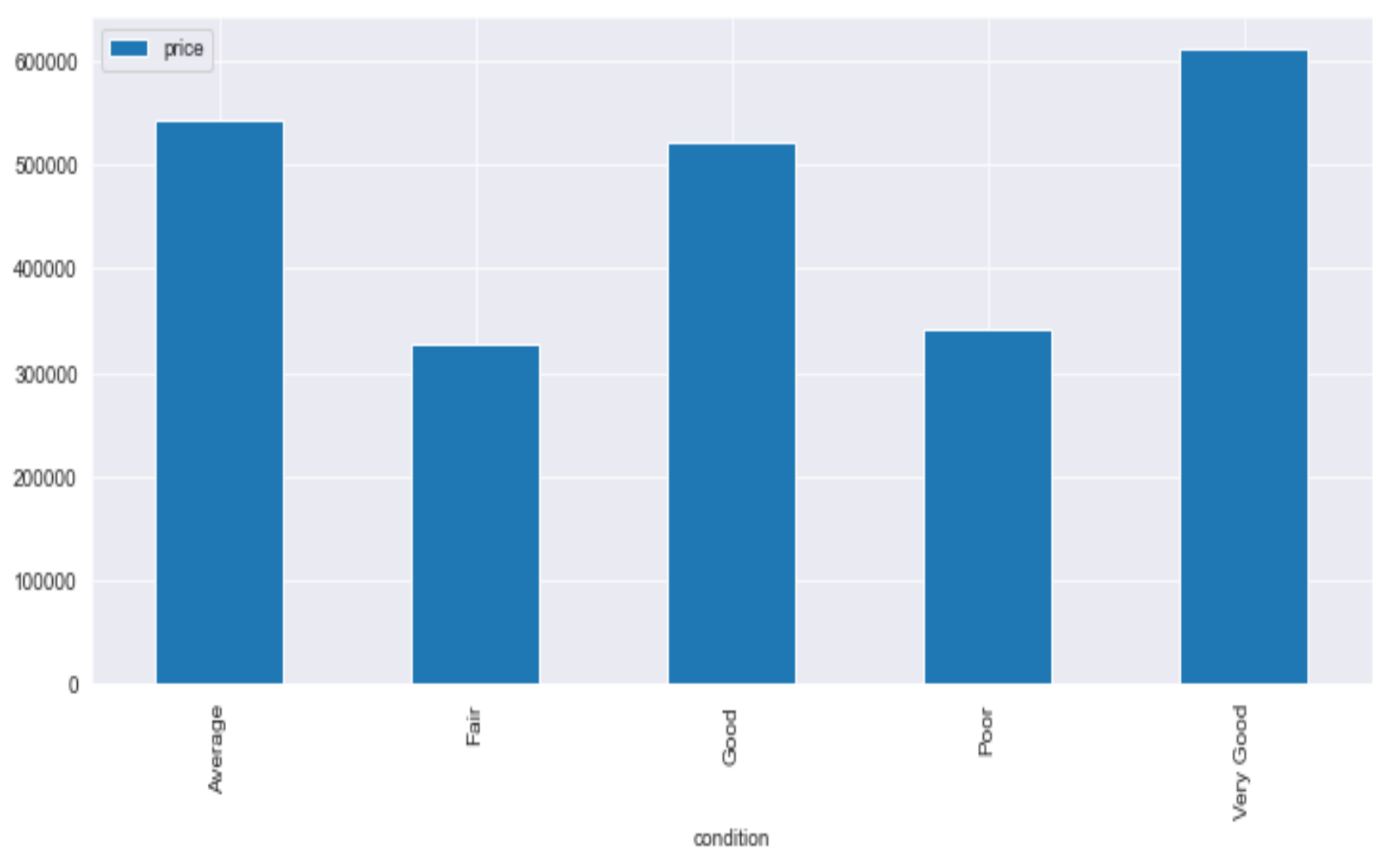


- This graph is used to demonstrate a positive correlation in that as the sqft_living increases the price increases.

MODELING

In this section models were used to find the correlation for each feature in relation to price. A positive correlation will suggest that as the feature increases the price of the property increases was well and that also applies to a negative correlation in that as the feature increases the price of the property tends to decrease.

From the bar plot below we observe that houses in very good condition are the most expensive, while the ones in fair and poor condition are the most affordable, therefore the better the condition of the houses the higher the



This output will display the total number of each bar in the graph which also represents all the houses in that certain condition :

- Average 14020
- Good 5677
- Very Good 1701
- Fair 170
- Poor 29

Simple Linear Regression

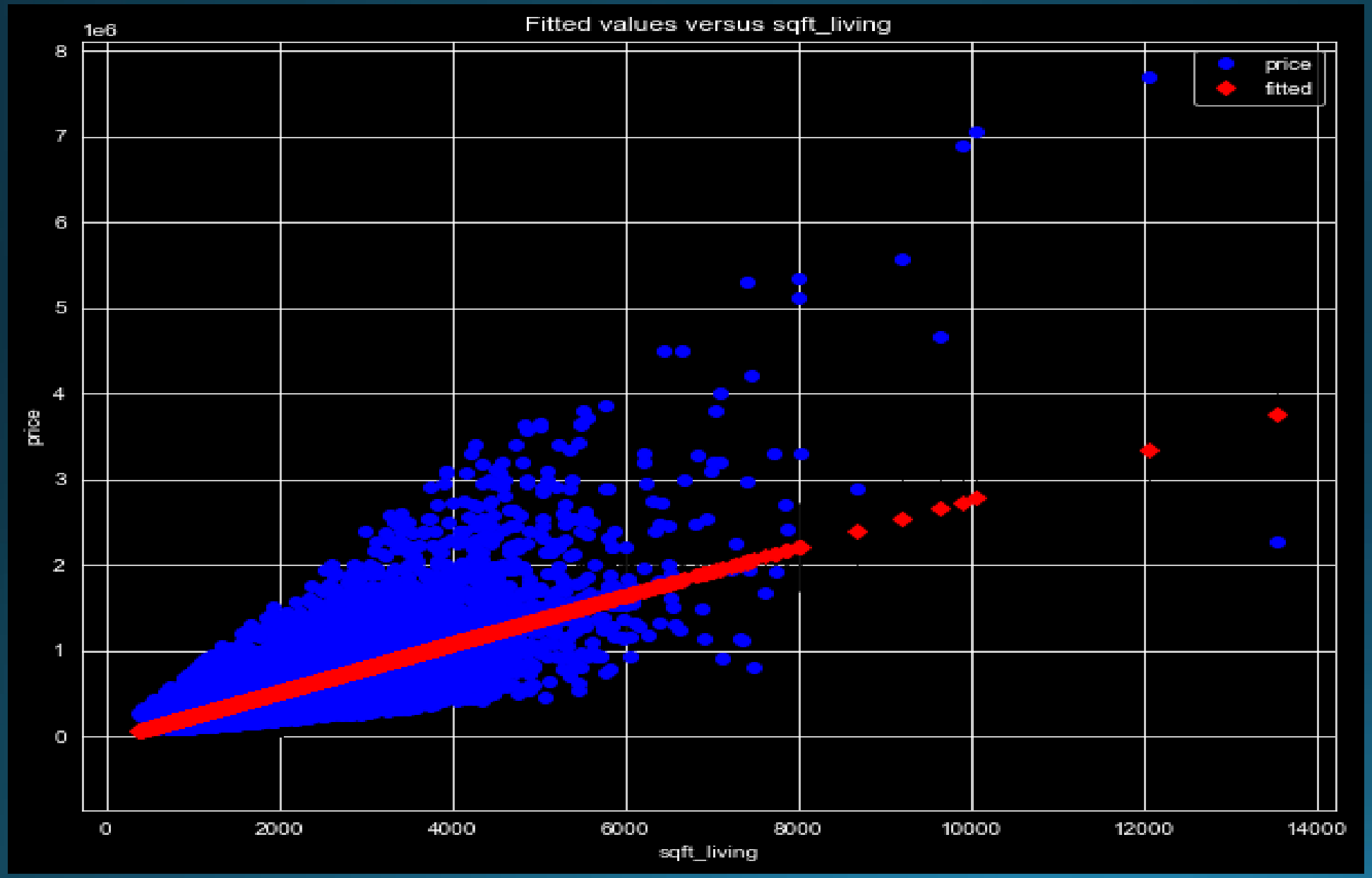
- **Baseline Model**

Linear regression is the statistical and machine learning technique used for modelling the relationship between a dependent variable which in our case is 'Price' and one or more independent variables which in our case are the features(bathrooms, bedrooms, floors…etc.) by fitting a linear equation to the observed data.
In order to evaluate our model a simple linear regression was built .

- From our model, for each additional square foot of living space, the 'price' is expected to increase by 280.863 units when all other factors remain constant

Simple Linear Regression Visualization

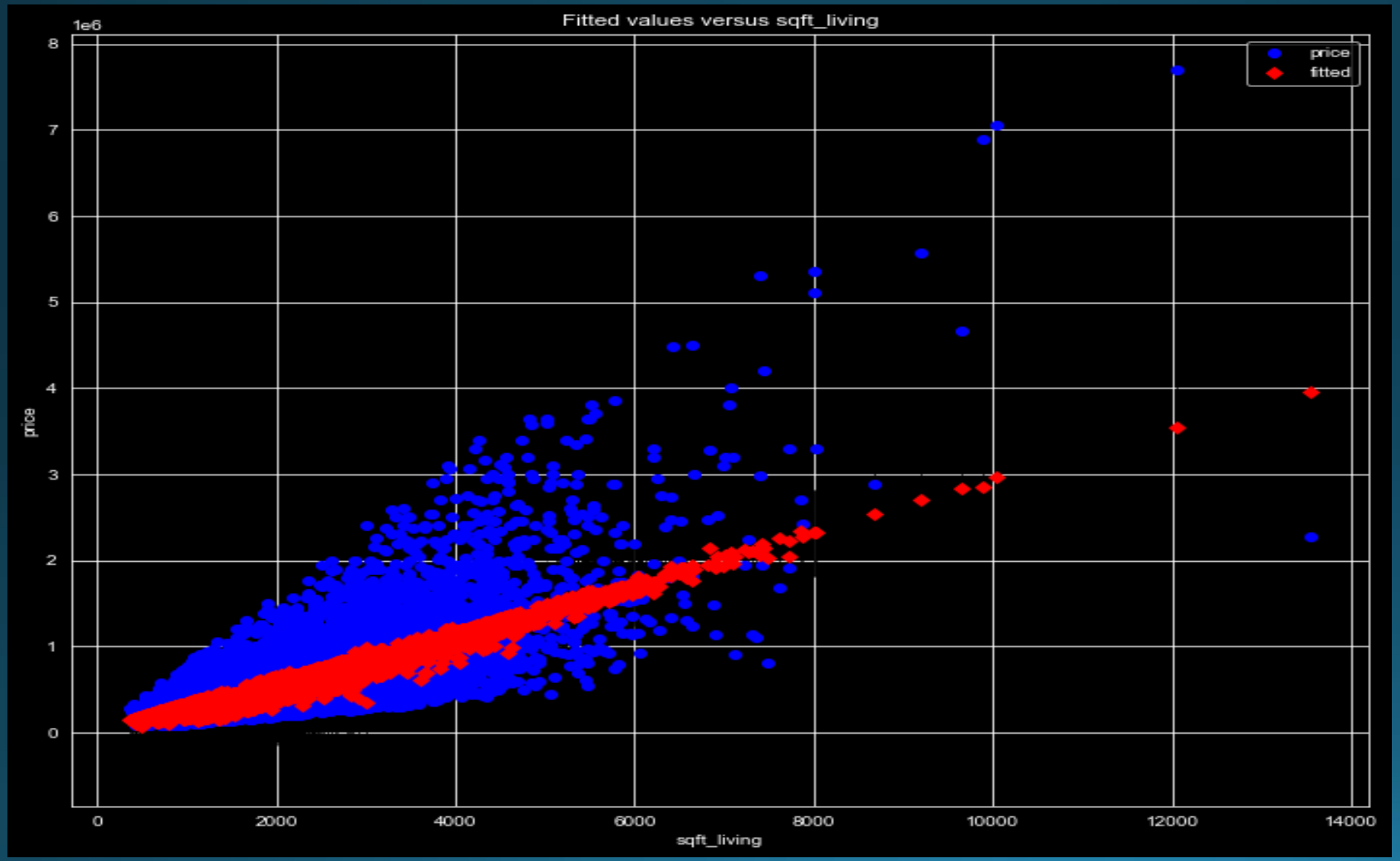
- 2nd Model



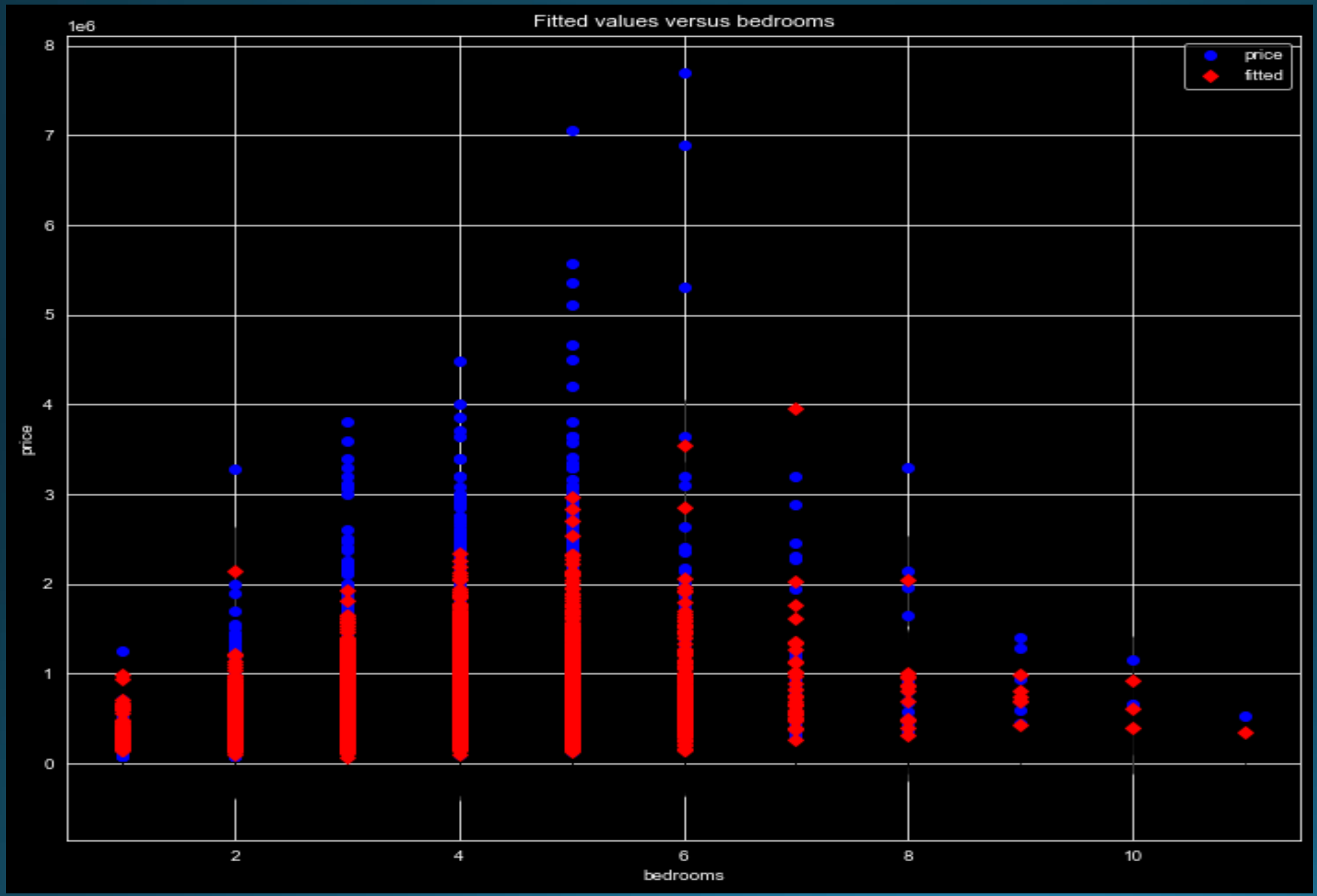
- The plot shows the actual values vs predicted.

- Inserting another Independent variable model

The second model was a multiple linear regression model (Model: OLS) with 'sqft_living' and 'bedrooms' as independent variables

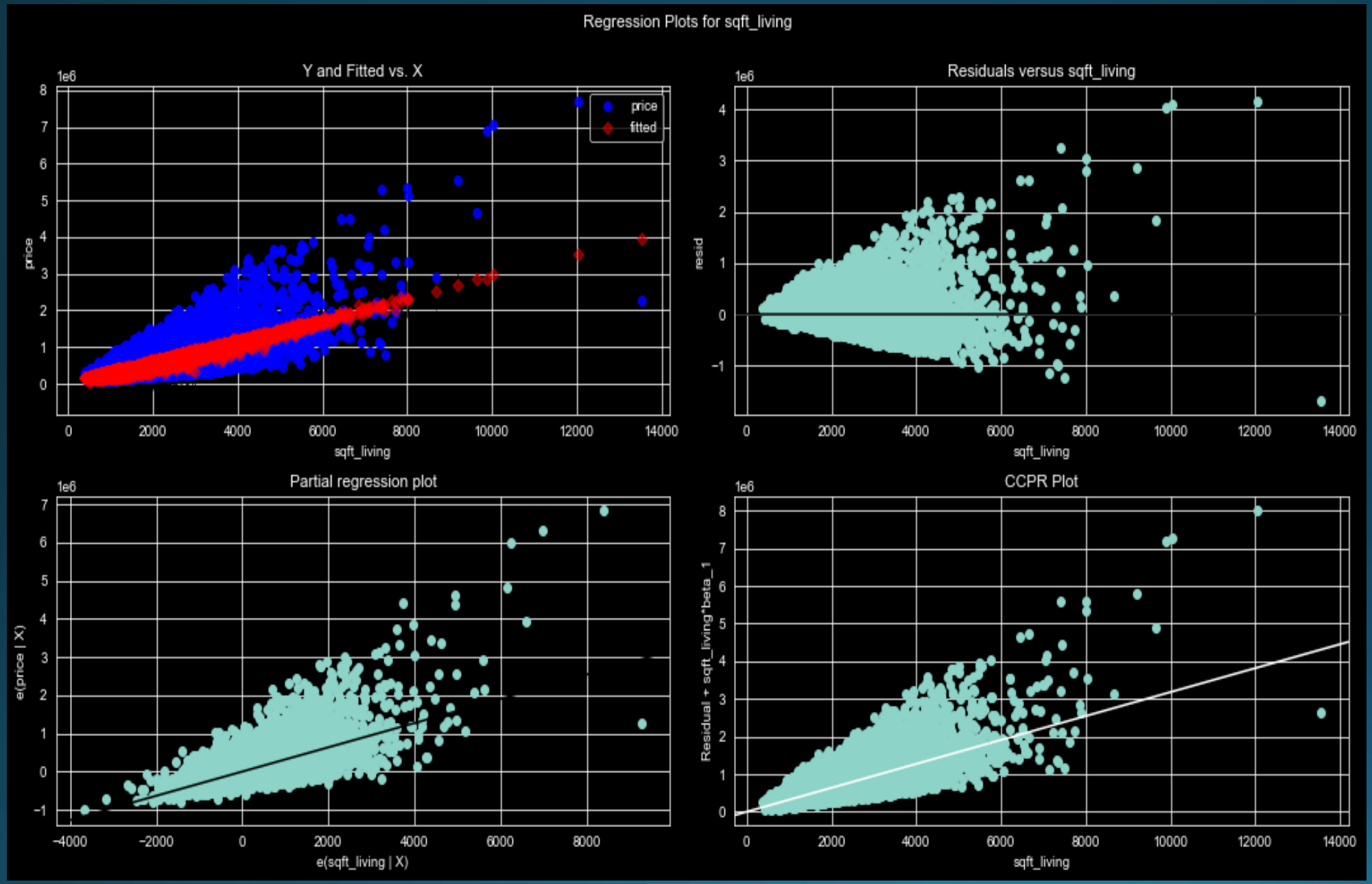


- Each coefficient represents the change in the 'price' associated with a one-unit change in the respective independent variable, holding all other variables constant.
- The plot shows the true (blue) vs. predicted (red) values, with the particular predictor (in this case, sqft_living) along the x-axis.

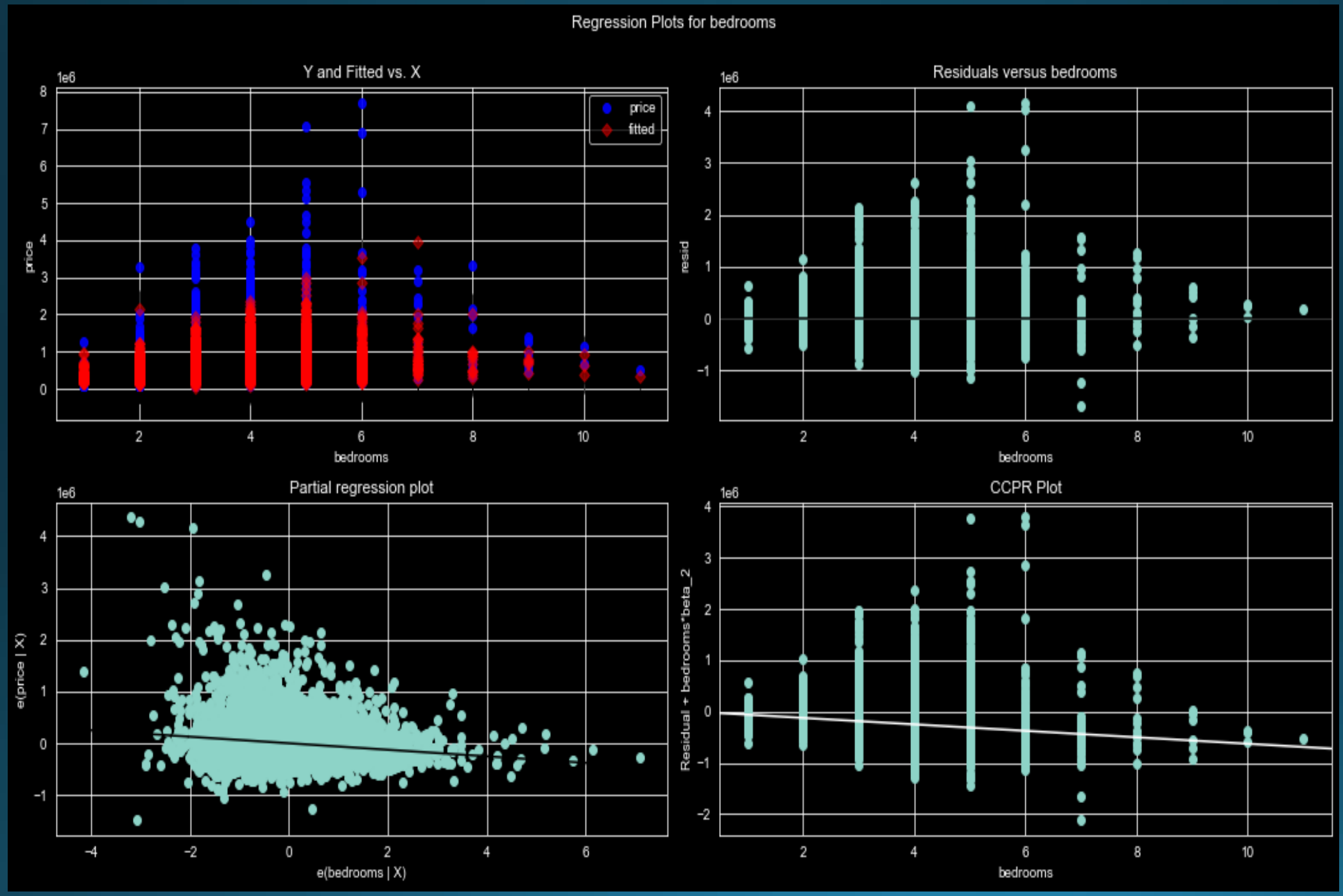


- The plot above shows the fit for the other predictor, bedrooms

- Partial Regression Plot / Plotting Residuals



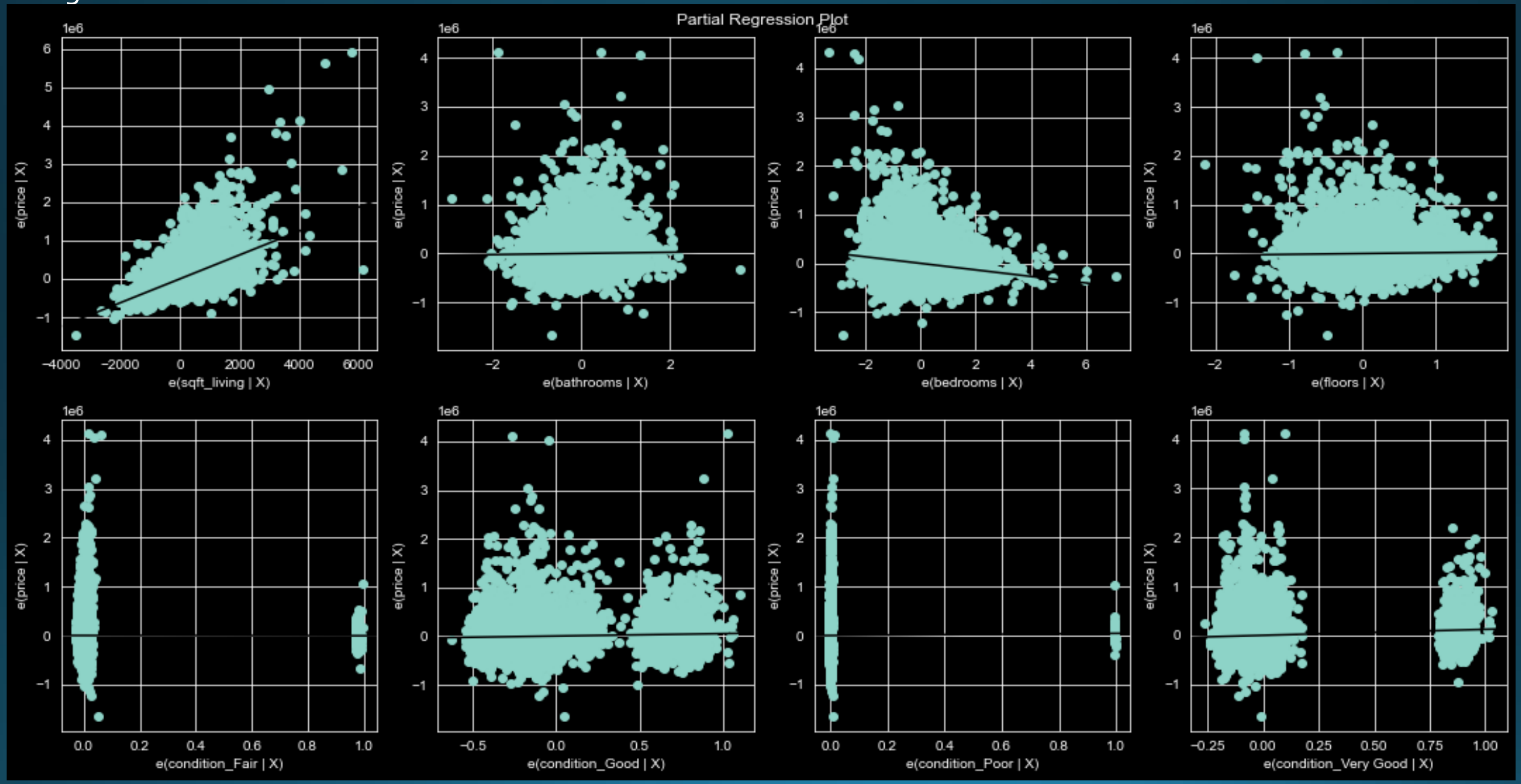
- The image shows a plot of the regression with the "sqft_living" as the exogenous variable



- The above image shows a plot regression of exogenous variable against bedrooms.

Multiple Regression with Many Features

- 3rd Model



First Model

*price = -43,990 + 280.863 * sqft_living*

Second Model

*price = 91,770 + (317.6347 * sqft_living) - (62,950 * bedrooms)*

Third Model

The model is a multiple linear regression model (Model: OLS) with eight independent variables.

*price = 44,920 + (311.64 * sqft_living) + (11,990 * bathrooms) - (67,170 * bedrooms) + (17,060 * floors) - (4768.5025 * condition_Fair) + (49,570 * condition_Good) + (44,370 * condition_Poor) + (122,500 * condition_Very Good)*

Each coefficient represents the change in the 'price' associated with a one-unit change in the respective independent variable, holding all other variables constant.

Condition_fair means the difference associated with a house condition being fair vs an average house. In other words, compared to an average house, we see an associated decrease of about -4,590.8019 USD for fairly conditioned house.

Condition_good is also comparing to an average house. We see an associated increase of about 48,930 USD for a house in good condition compared to a house in average condition.

CONCLUSION

The best model chose out of the three is the third model because of:

- Model Performance: The third model, which incorporates 'condition,' 'bedrooms,' 'bathrooms,' 'sqft_living,' and 'floors,' exhibits the best overall performance among the models considered.
- Prediction Accuracy: The third model has the lowest Root Mean Squared Error (RMSE) of approximately 255,588.39 USD, indicating the highest prediction accuracy among the models.
- Explanatory Power: The third model also has the highest R-squared (R^2) value of approximately 0.5174, signifying the greatest explanatory power. It explains about 51.74% of the variance in home prices, suggesting that it provides the most comprehensive understanding of the factors influencing sale values

RECOMMENDATIONS

❖ Estimating the Impact of Specific Renovation Projects:

The agency can use the Third Model to provide homeowners with estimates of how specific renovation projects will impact the resale value of their homes.

Homeowners can make informed decisions about which renovation projects to prioritize, based on their expected return on investment (ROI). This will empower homeowners to invest in renovations that will maximize their property's resale value.

❖ Identifying Renovation Projects with the Most Impact:

Amani can utilize the third model to identify which specific renovation projects or features have the most significant impact on a home's market value in the northwestern county.

❖Correlation of Bedrooms, Bathrooms, Grade, and Square Footage with Sale Price:

They can leverage the third model to explain how the number of bedrooms, bathrooms, the grade of a house, and its square footage correlate with its sale price in King County. They can utilize the model's coefficients and feature importance analysis to explain the correlations between these variables and sale price.

❖ Identifying Combinations of Renovation Projects:

The third model will help identify specific combinations of renovation projects that provide an interdependent effect on a home's market value.

SUMMARY

In summary, the project study suggests that the number of bedrooms, square footage of living area, condition, number of bedrooms, bathrooms and floors are important factors to consider when determining the price of a home. However, it is essential to consider other market factors and property-specific attributes in conjunction with the findings of this analysis to arrive at an accurate and competitive listing price such as architectural style, lot size and landscaping, upgrades and amenities, historical sales data, market trends, school district, crime rate, zoning and regulations.

THANK-YOU

WE HAVE COME TO THE END OF THE PRESENTATION