KINGS COUNTY REAL ESTATE PROJECT

GROUP 4

- CLYDE OCHIENG
- MATTHEW KIPTOO
- HILARY SIMIYU
- SOLPHINE JOSEPH
- GRACE ROTICH
- DERRICK KIPTOO

BUSINESS UNDERSTANDING

Background

- Our agency faces market fluctuations and increased competition in the real estate sector.
- Pricing volatility presents significant challenges for our agents in developing effective business strategies.
- We are seeking strategic guidance to enhance our purchasing and selling efforts.
- Our priority is informed decision-making to identify areas with maximum returns on investment

Challenges

- The scarcity of available residential units in Kings County has led to increased costs across different grades (e.g., average, good, excellent, luxurious).

- Disparities in the housing market are evident on a sub-regional level within King County

- Properties in sought-after areas, like waterfront or downtown areas, are limited and highly prized

Proposed Solution

By leveraging multiple regression analysis, stakeholders can gain a deeper understanding of the complex factors driving house prices in Kings County and develop informed strategies to address challenges related to scarcity, disparities, and location desirability

OBJECTIVES

- To perform the analysis of the key factors influencing house prices.
- To develop multilinear regression models to predict house prices based on relevant features.
- To use insights from the regression analysis to optimize pricing strategies for both purchasing and selling properties.

DATA UNDERSTANDING

- *Dataset:* House sales in King County, Washington, USA
-
- *Features*: Includes various attributes like bedrooms, bathrooms, square footage, and geographical details

- Represents: Each row corresponds to a specific house sale entry
- *Size*: 21 columns and 21,597 rows

- *Data Types:* Mix of integers (int64), floating-point numbers (float64), and objects (strings)

- The maximum price of a house is $7,700,000 dollars and the minimum price is $78,000 dollars

- The average price of a house is approximately $540,000

DATA PREPARATION

Data preparation entailed a meticulous process ofdealing with the anomalies in the Kc _Dataset before generating the models

- In our analysis, we discovered that the 'yr_renovated' feature has the highest percentage of missing values, accounting for 17% of the data. Due to its substantial amount of missing information and its lack of relevance to our business problem, we have decided to drop this feature from our model.

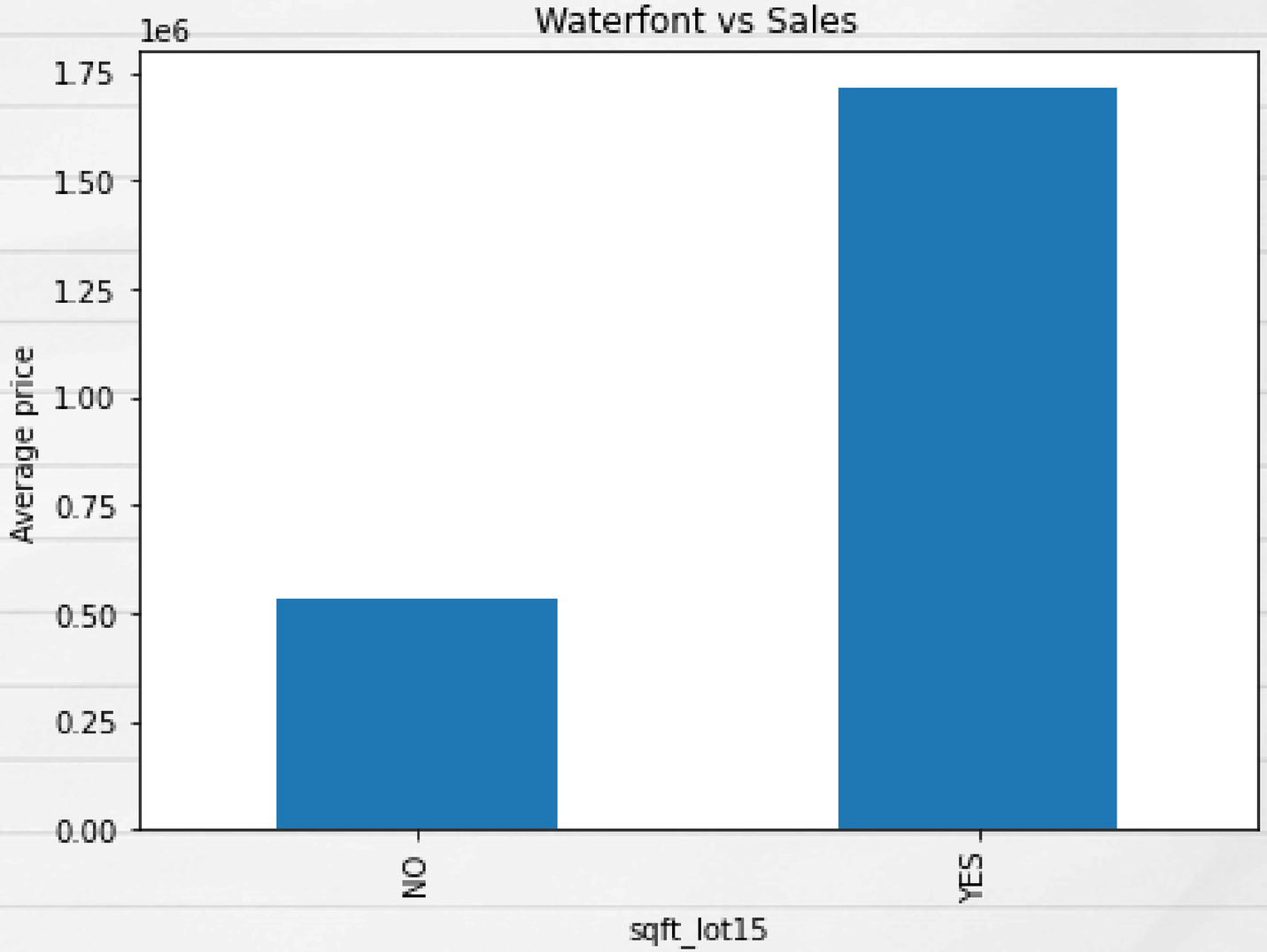
- Additionally, we found that the 'Waterfront' feature has 11% of its values missing. We made the assumption that these missing values indicate properties that do not have waterfront access. Therefore, we filled these missing values with 'NO' to reflect this assumption.

- By addressing missing values in these features, we ensure that our model is based on complete and relevant data, allowing us to make more accurate predictions and insights in line with our business objectives.

EXPLORATORY DATA ANALYSIS

Waterfront Feature:

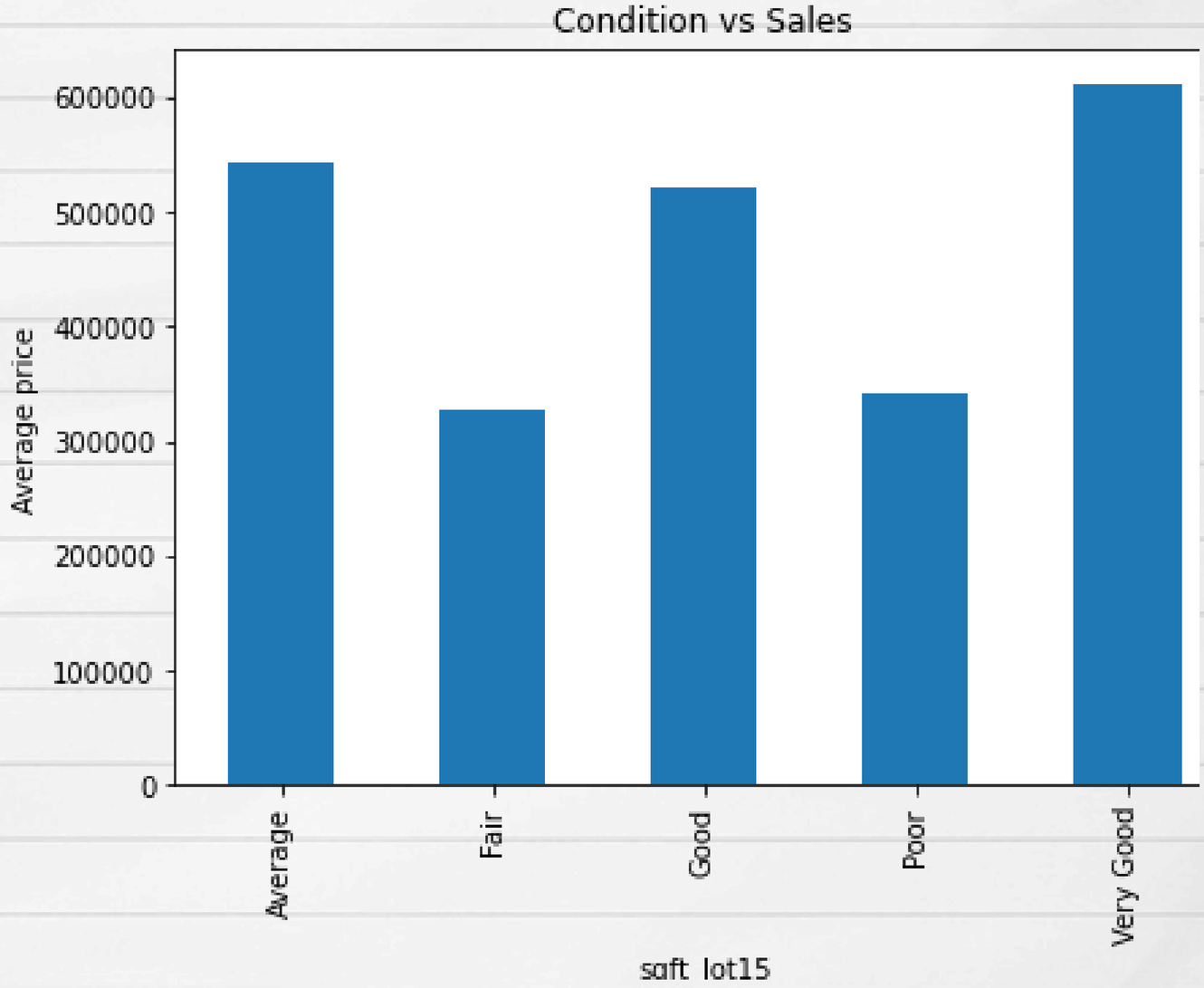
The visualization helps us to understand the impact of the waterfront on the price of the house. This shows that the houses with waterfront tend to have higher prices compared to houses without.



EXPLORATORY DATA ANALYSIS

Condition Feature:

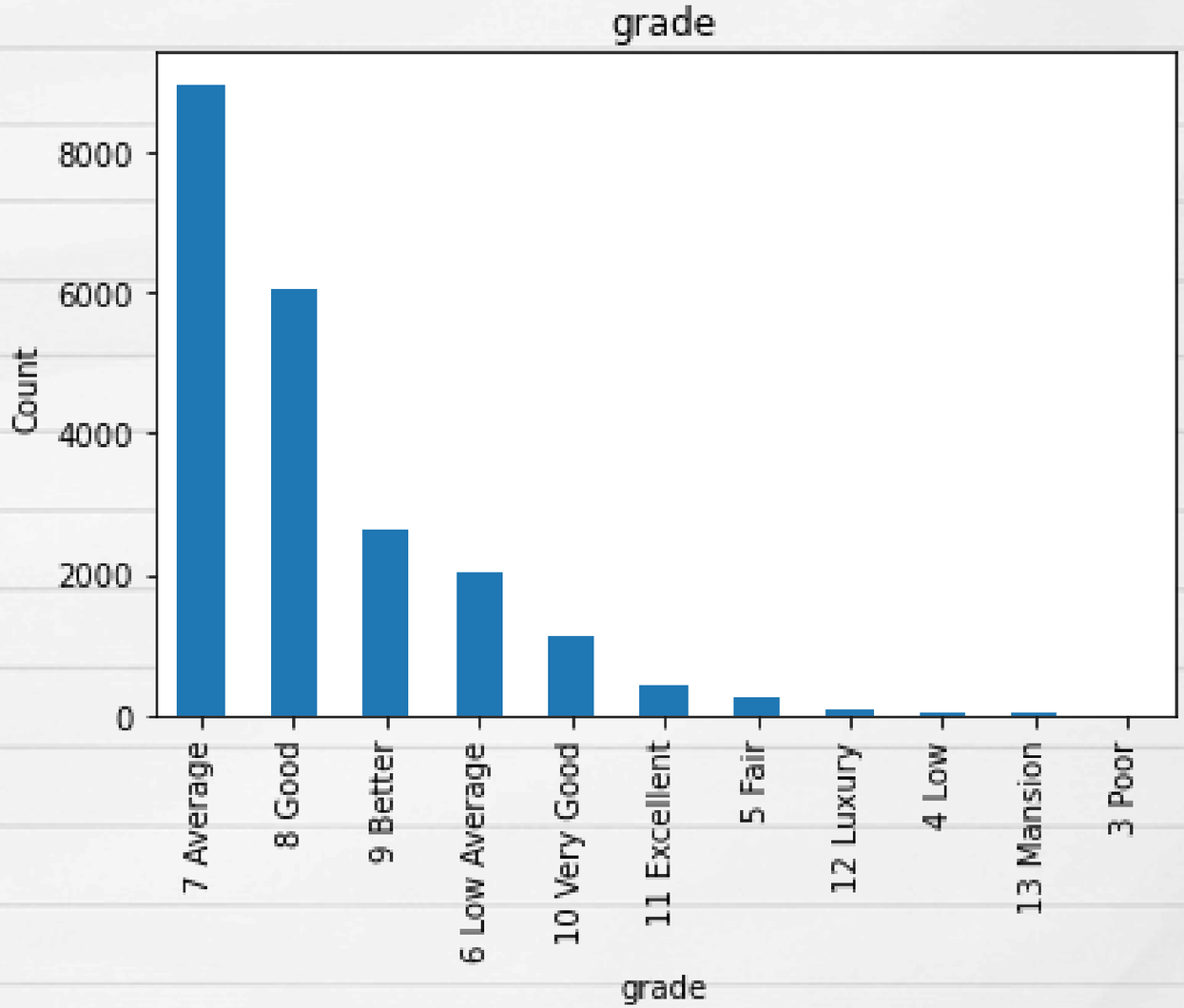
From this it appears houses in poor condition are priced higher than those in fair condition.



EXPLORATORY DATA ANALYSIS

Grade Feature:

The visualization helps us to understand the amount of each grade .



EXPLORATORY DATA ANALYSIS

Correlation:

Heatmap of Correlation Matrix for Numeric Columns

	id	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	sqft_above	yr_built	r_renovated	zipcode	lat	long	sqft_living15	sqft_lot15
id	1.00	-0.02	0.00	0.01	-0.01	-0.13	0.02	-0.01	0.02	-0.01	-0.01	-0.00	0.02	-0.00	-0.14
price	-0.02	1.00	0.31	0.53	0.70	0.09	0.26	0.61	0.05	0.13	-0.05	0.31	0.02	0.59	0.08
bedrooms	0.00	0.31	1.00	0.51	0.58	0.03	0.18	0.48	0.16	0.02	-0.15	-0.01	0.13	0.39	0.03
bathrooms	0.01	0.53	0.51	1.00	0.76	0.09	0.50	0.69	0.51	0.05	-0.20	0.02	0.22	0.57	0.09
sqft_living	-0.01	0.70	0.58	0.76	1.00	0.17	0.35	0.88	0.32	0.06	-0.20	0.05	0.24	0.76	0.18
sqft_lot	-0.13	0.09	0.03	0.09	0.17	1.00	-0.00	0.18	0.05	0.00	-0.13	-0.09	0.23	0.14	0.72
floors	0.02	0.26	0.18	0.50	0.35	-0.00	1.00	0.52	0.49	0.00	-0.06	0.05	0.13	0.28	-0.01
sqft_above	-0.01	0.61	0.48	0.69	0.88	0.18	0.52	1.00	0.42	0.02	-0.26	-0.00	0.34	0.73	0.20
yr_built	0.02	0.05	0.16	0.51	0.32	0.05	0.49	0.42	1.00	-0.23	-0.35	-0.15	0.41	0.33	0.07
r_renovated	-0.01	0.13	0.02	0.05	0.06	0.00	0.00	0.02	-0.23	1.00	0.07	0.03	-0.07	-0.00	0.00
zipcode	-0.01	-0.05	-0.15	-0.20	-0.20	-0.13	-0.06	-0.26	-0.35	0.07	1.00	0.27	-0.56	-0.28	-0.15
lat	-0.00	0.31	-0.01	0.02	0.05	-0.09	0.05	-0.00	-0.15	0.03	0.27	1.00	-0.14	0.05	-0.09
long	0.02	0.02	0.13	0.22	0.24	0.23	0.13	0.34	0.41	-0.07	-0.56	-0.14	1.00	0.34	0.26
sqft_living15	-0.00	0.59	0.39	0.57	0.76	0.14	0.28	0.73	0.33	-0.00	-0.28	0.05	0.34	1.00	0.18
sqft_lot15	-0.14	0.08	0.03	0.09	0.18	0.72	-0.01	0.20	0.07	0.00	-0.15	-0.09	0.26	0.18	1.00

1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75

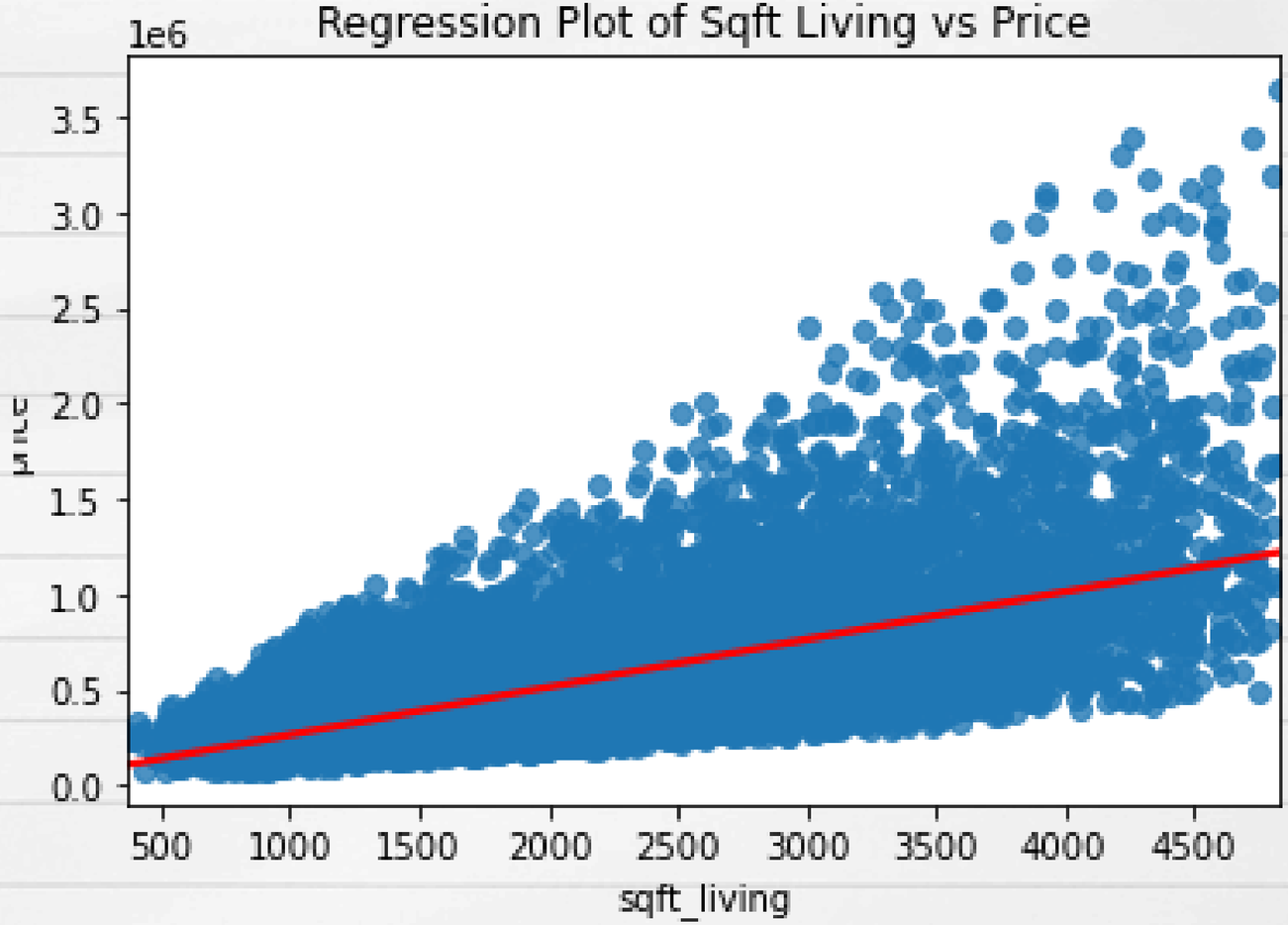
MODEL 1

Simple Linear based model:

The model focused on one predictor variable(sqft_living) and the target variable which is price

For a house with 0 squarefoot of living area, our model would predict a home price of about 15203.833 dollars.

'MODEL R-SQUARED': 0.4302319559899205,

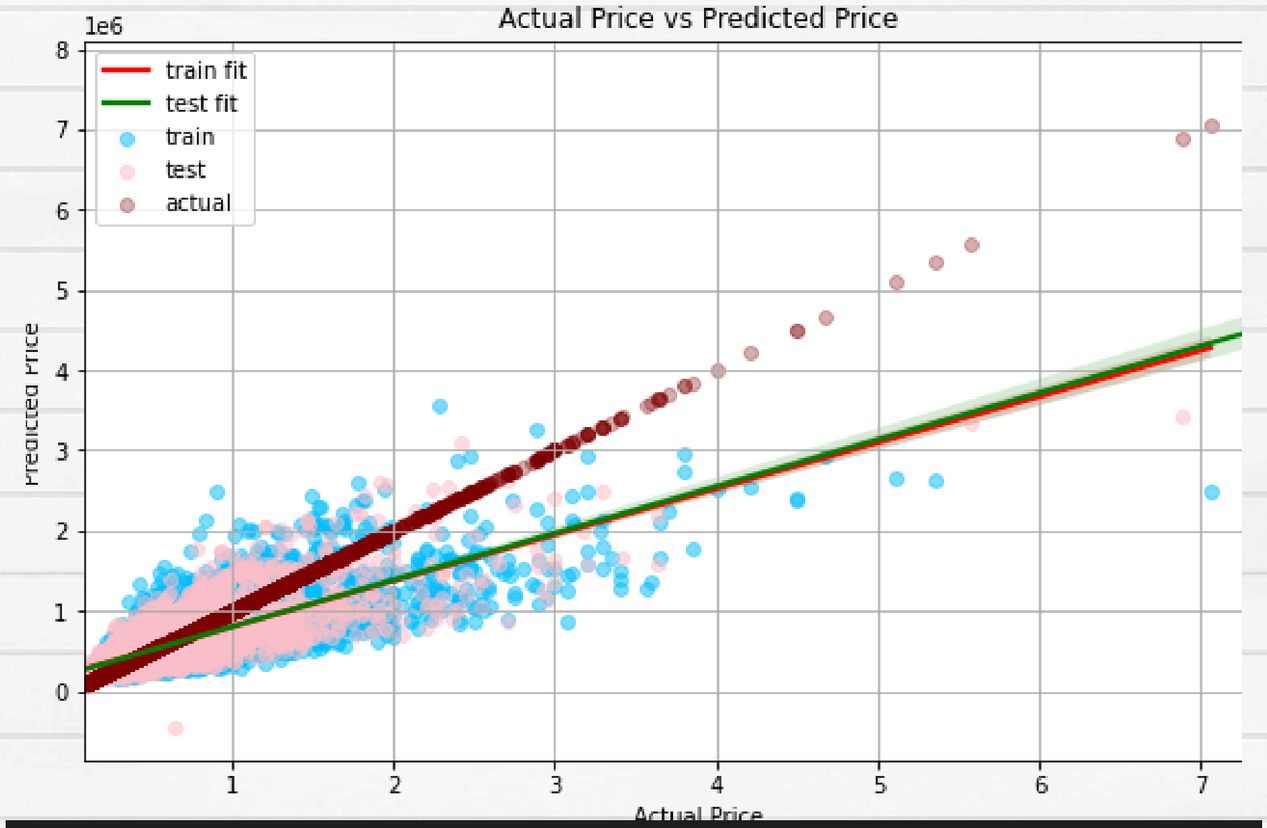


Multilinear Regression

The model focused on several predictor variables predictor variable(sqft_living) including the categorical variables

MODEL R-SQUARED': 0.5831491676288707,

TRAIN-TEST SPLIT MODEL

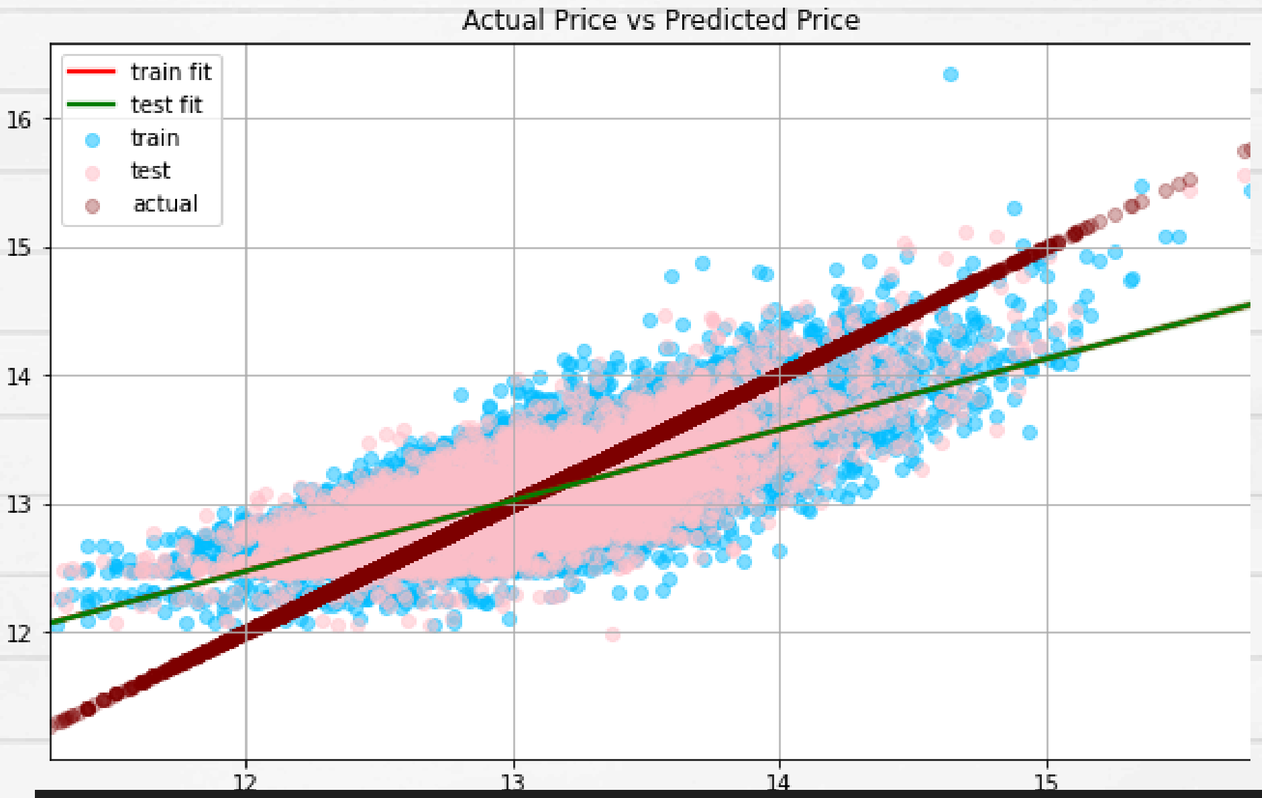


mean_diff: 31.62652023361938

	Actual Values	Predicted Value	Difference	Percentage Difference
19009	1000000.0	6.320870e+05	367913.009595	36.791301
6370	870000.0	1.083277e+06	213277.441830	24.514648
13983	160000.0	3.030799e+05	143079.872314	89.424920
7837	1650000.0	7.936058e+05	856394.226441	51.902680
9491	367500.0	4.820375e+05	114537.536619	31.166677

- The model explains 59% of the variability in the target variable.
- The train MAE and test MAE are 156674.07834 and 154037.816273 respectively, which are quite large values, indicating that the model is not performing well.
- The train RMSE and test RMSE are 237906.0435 and 234825.1832 respectively, which are also quite large.
- Overall, these metrics suggest that the linear regression model is not performing very well in making accurate predictions on the dataset.

LOG- MODEL



mean_diff: 2.1662961050481617

	Actual Values	Predicted Value	Difference	Percentage Difference
19009	13.815511	13.152049	0.663462	4.802298
6370	13.676248	13.816022	0.139773	1.022016
13983	11.982929	12.549739	0.566810	4.730143
7837	14.316286	13.494813	0.821472	5.738027
9491	12.814479	13.059385	0.244907	1.911172

- The model explains 56.8% of the variance in the logarithmic actual values.
- The model's logarithmic predictions are off by about 28.2%and 28.4 respectively from the logarithmic actual values.
- The model's logarithmic predictions are off by the squared value of about 12.28% and 12.56% respectively from the logarithmic actual values.
- Based on our visualizations, we can see that train and actual regression are much closer compared to the initial base model.

Conclusion

- The square footage of the living area is one of the consistent numeric variables
- The parameters R squared increases from one model to another, to indicate that the more predictor variables are used the more refined the model

- Properties in sought-after areas, like waterfront or downtown areas, are limited and highly prized

Recommendation

- Highlight properties with higher grades to appeal to luxury buyers seeking premium features and amenities.
- Emphasize spacious interiors and versatile floor plans to attract buyers looking for ample living space for their needs.
- Capitalize on properties with desirable views by showcasing them prominently in marketing materials and virtual tours.

THANK YOU VERY MUCH!