KING COUNTY HOUSE PRICE PREDICTIONS

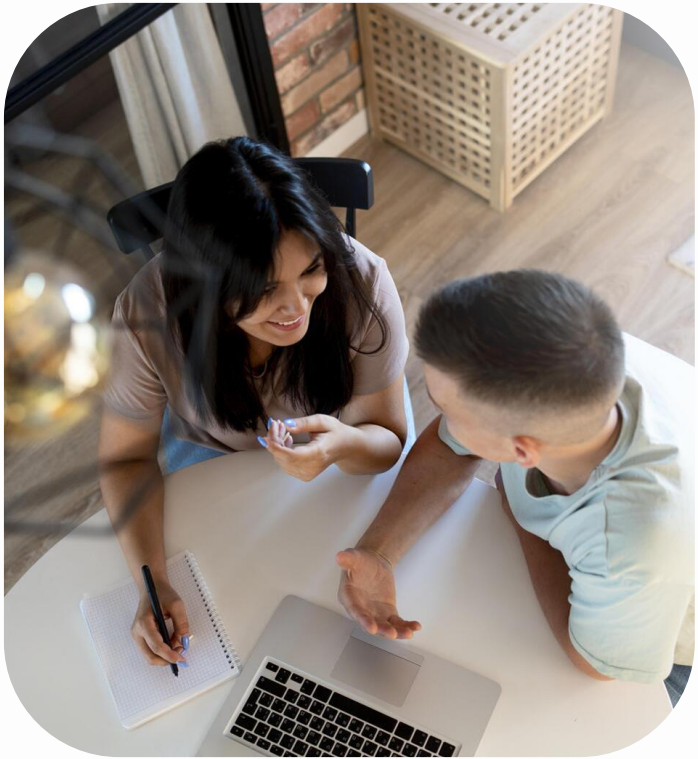
Table of contents

Business Problem

BUSINESS PROBLEM

The real estate agency wants to provide homeowners with advice on how home renovations can potentially increase the estimated value of their homes and by what amount. The agency aims to offer valuable insights to homeowners, helping them make informed decisions about renovation projects that can maximize their return on investment when selling their properties.

Introduction



As an employee for a real estate agency, I am analyzing information from the King County House Sales dataset. Given several factors from the dataset, I aim to provide advice to my clients on how home features will increase the values of homes in King County; more specifically, I aim to determine which home factors are the most lucrative. By understanding which factors increase a home's value, my agency will be able to successfully help homeowners sell their homes for a maximized profit.

03

OBJECTIVES

OBJECTIVES

1. To determine how much would adding an extension to the lot area of the home likely increase sale price?

2. To examine how much would adding an additional bathroom likely increase sale price?

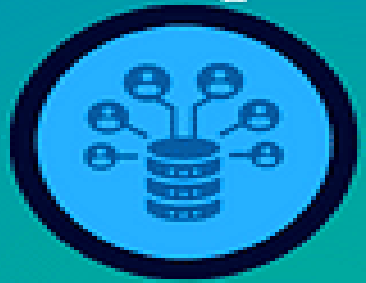
OBJECTIVES

3. To determine how much would adding an extension to the living area of the home likely increase sale price?

4. To determine how much would adding an additional floor to a house likely increase sale price?

Data Analysis Process

Business Problem Definition



Inventory and Data Collection

Data Clea

Data Analysis

Result Communication & Eventual Readjustment

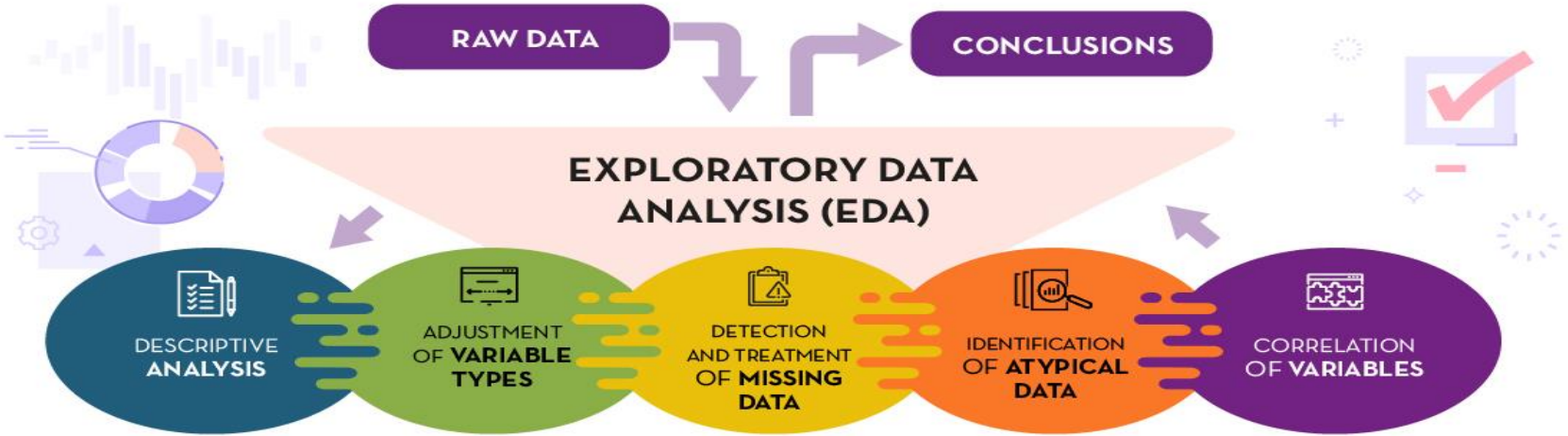
Choose The Rig

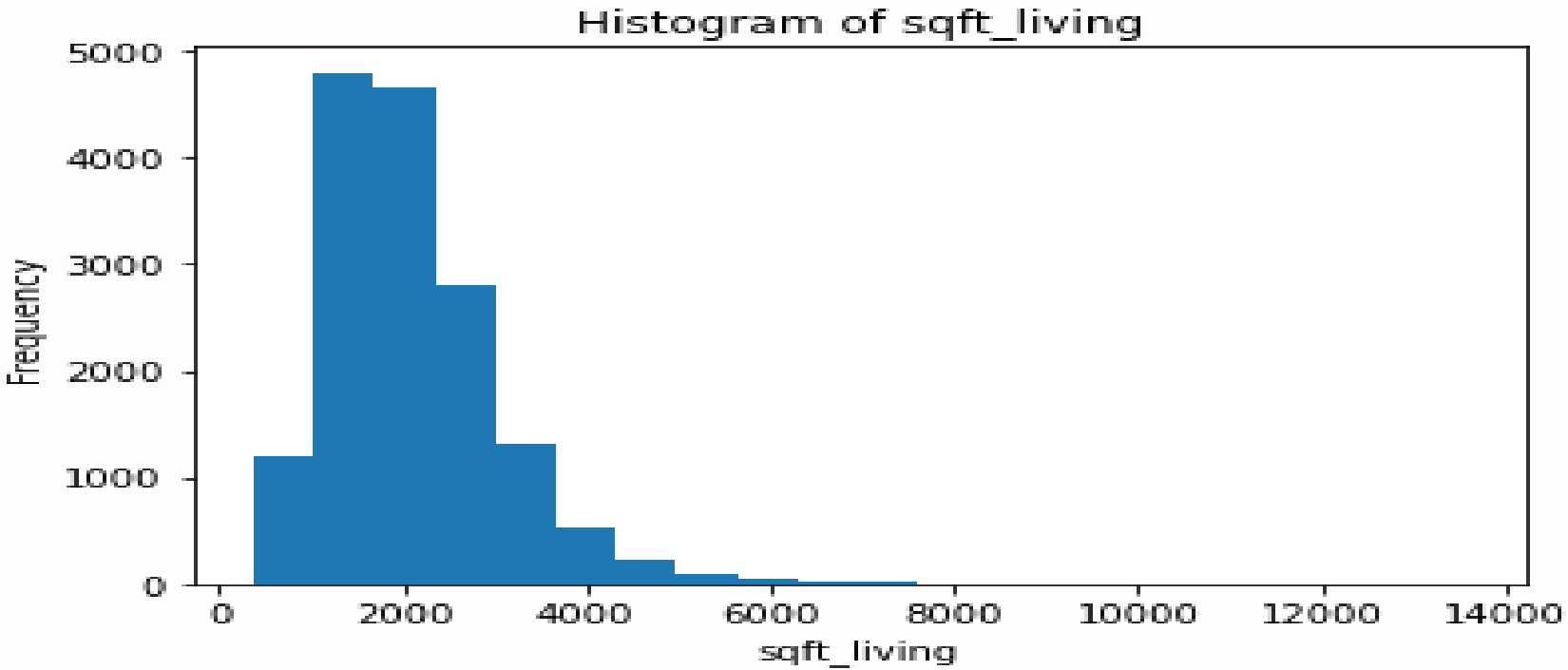
Data Understanding

The data columns used were:

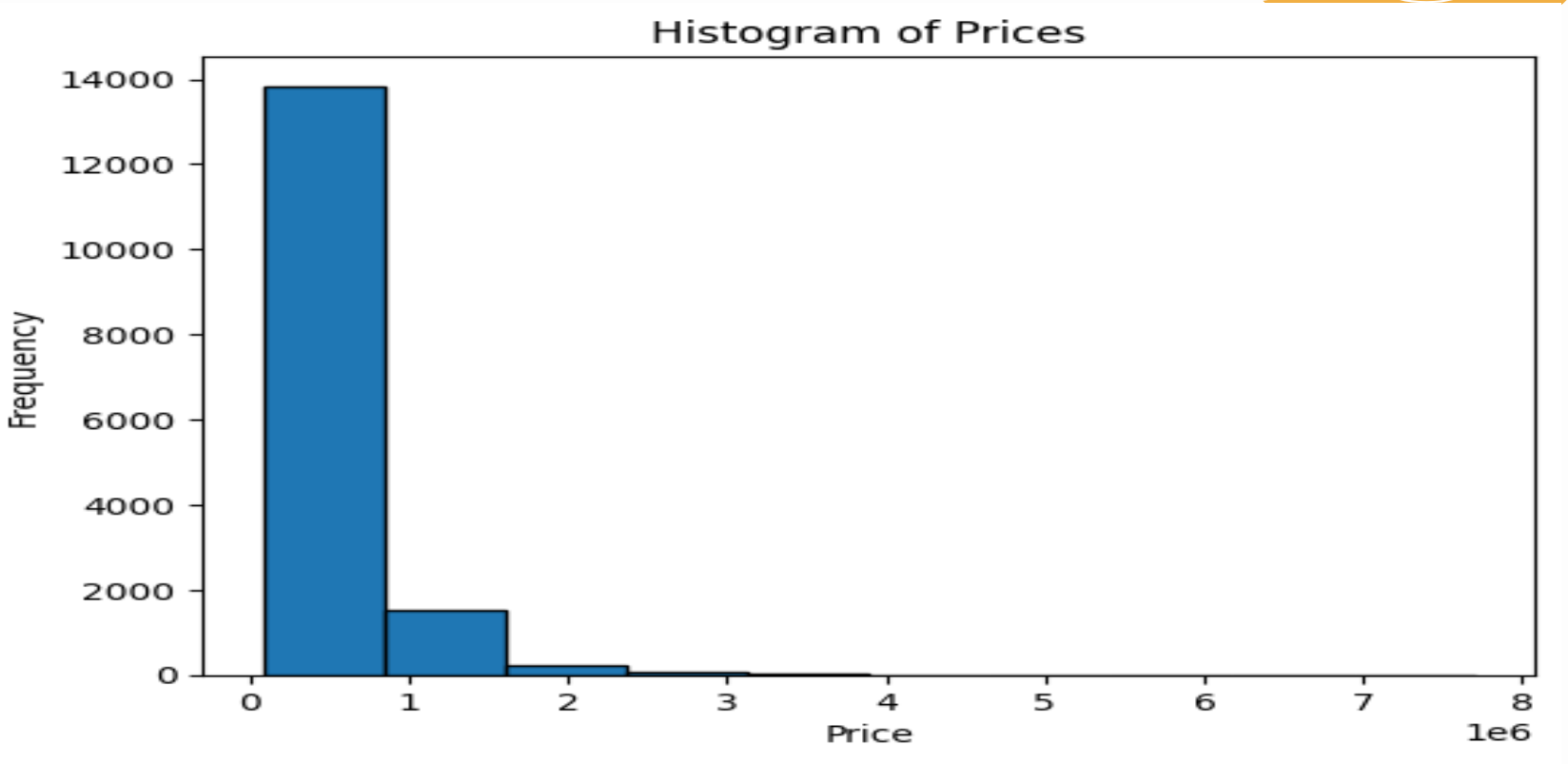
1. Bedrooms - Number of bedrooms.
2. Bathrooms - Number of bathrooms.
3. Condition - How good the overall condition of the house is.
4. Grade - Overall grade of the house. Related to the construction and design of the house.
5. Price - This is the selling price of a house.
6. Sqft_lot - This is the square footage of the lot where a house is built on.
7. Sqft_living - Square footage of living space in the home.
8. Sqft_above - Square footage of house apart from basement.

Exploratory Data Analysis

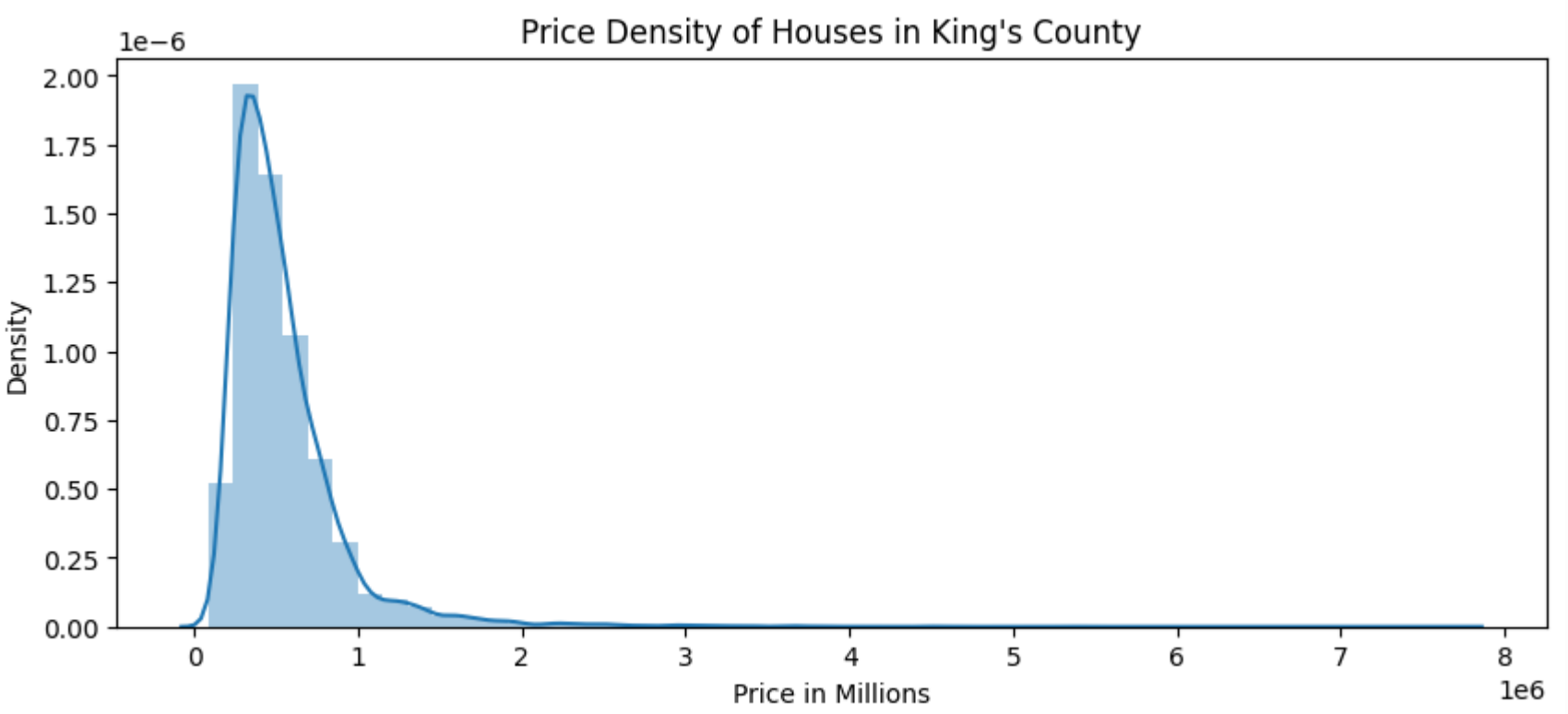


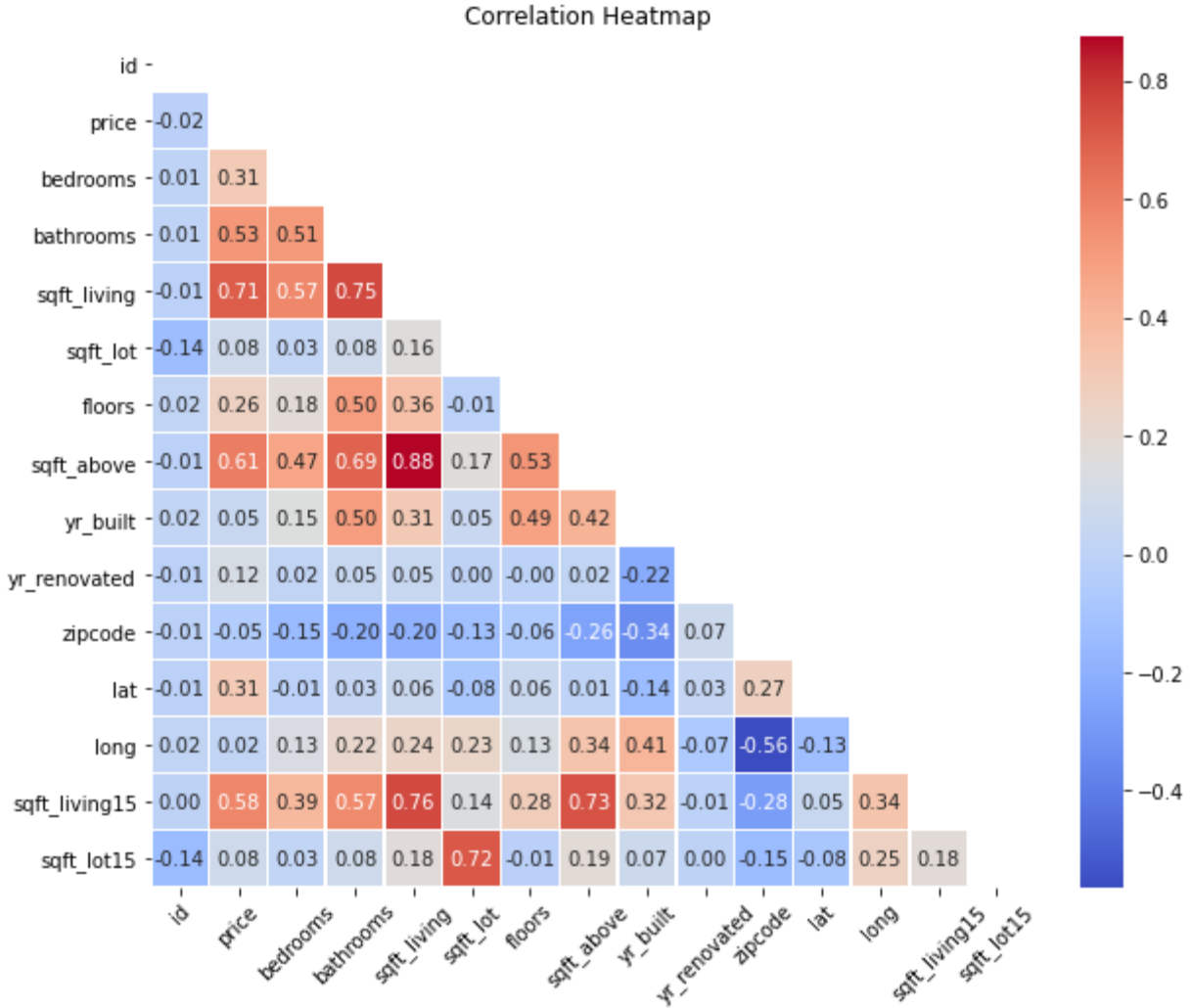


Most of the houses' Square footage of living space lies between 2000 to 4000 square feet.


Histogram of Prices
Frequency
14000
12000
10000
8000
6000
4000
2000
0
0
1
2
3
4
5
6
7
8
Price
1e6

The Distribution Of House prices Is Right-Skewed




Correlation Heatmap
id
price
bedrooms
bathrooms
sqft_living
sqft_lot
floors
sqft_above
yr_built
yr_renovated
zipcode
lat
long
sqft_living15
sqft_lot15
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

DATA

MODELING

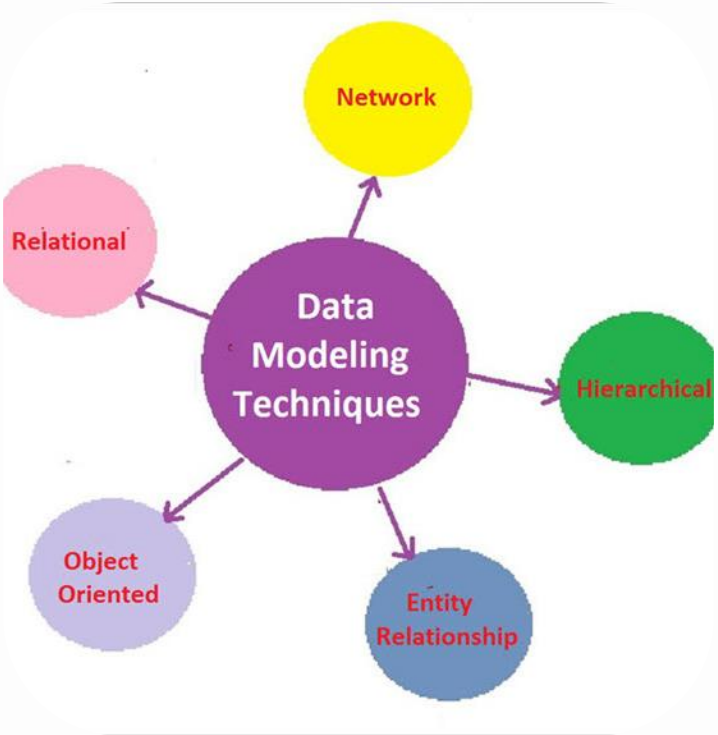
1. We started with a Simple Linear Regression as a baseline model to predict price using Square foot of living.

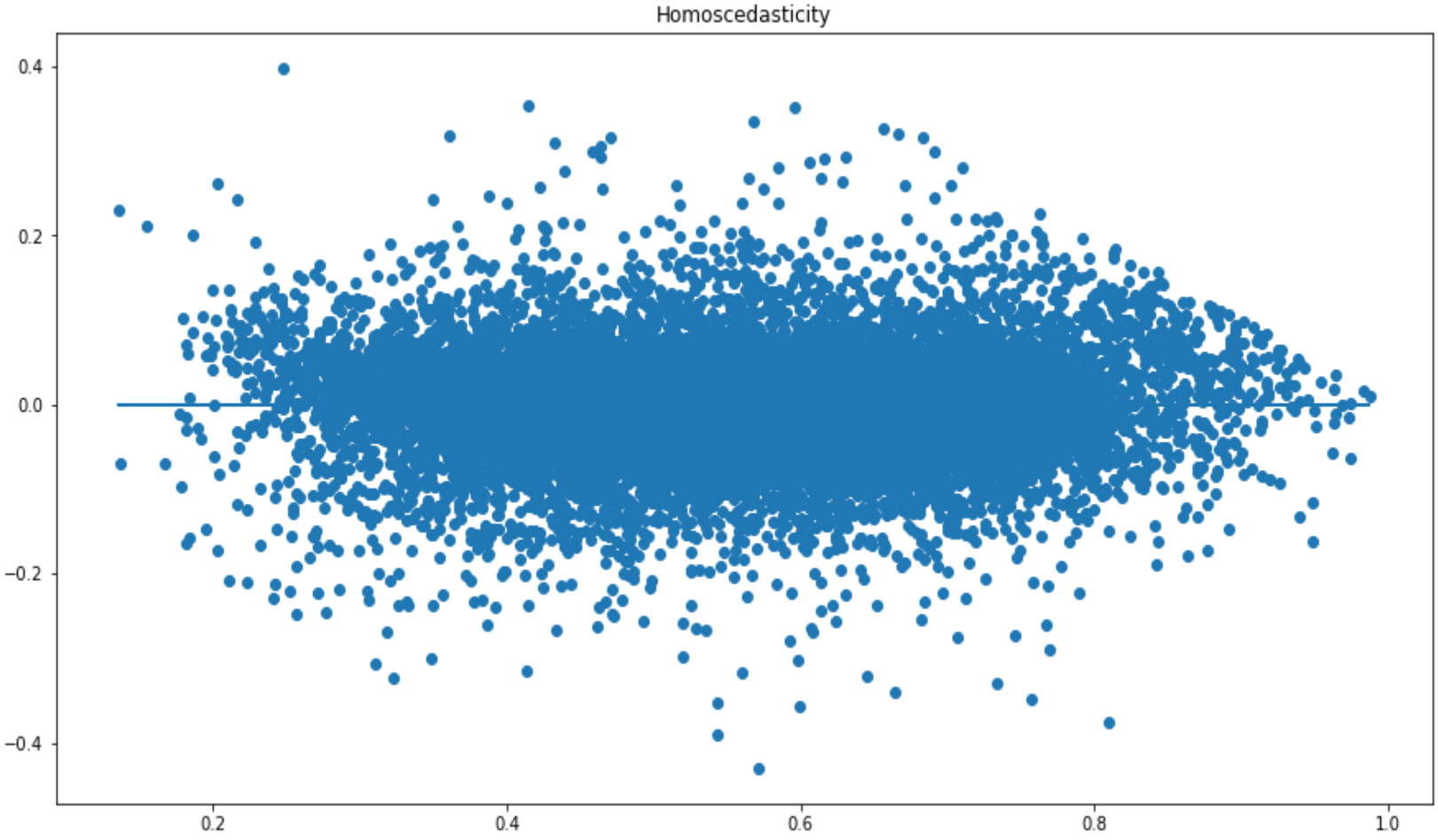
2. We then created a Multiple Linear Regression model to predict price using Sqft_living 15, bedrooms, bathrooms, sq ft lot15, floors, sq_ft above, grade number and basement.

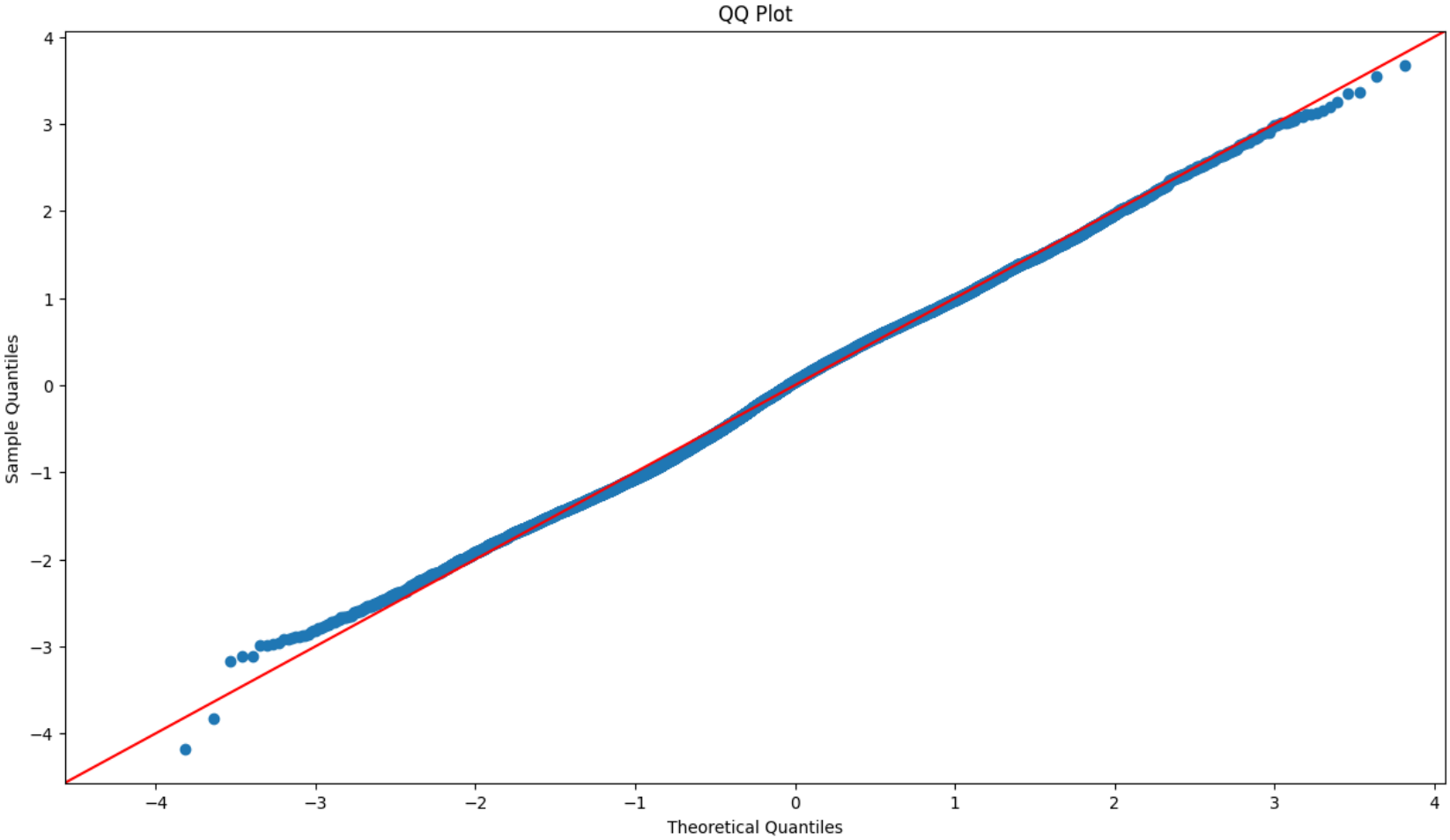
3. The next model was built with outliers removed from features.

4. The next model was created with features log-transformed.

5. Then the final model was created with one-hot-encoding of zip-codes.




Homoscedasticity
0.4
0.2
0.0
−0.2
−0.4
0.2
0.4
0.6
0.8
1.0


QQ Plot
Sample Quantiles
Theoretical Quantiles
4
3
2
1
0
−1
−2
−3
−4
−4
−3
−2
−1
0
1
2
3
4

CONCLUSIONS

1. The model is generally statistically significant with an F-statistic p_value of 0.0 at a significance level of 0.05.

2. The R-squared value is 0.833, indicating that approximately 83.3% of the variation in the price can be explained by the model. This value indicates a great improvement from the previous model.

3. Also, of great importance to note is that the mean RMSE is approximately 0.06465. Then the RMSE in original scale is 0.1135. This means that our model is off by about 0.1135 when making an average prediction, indicating that it is a good model.

CONCLUSIONS

4. The coefficients represent the expected change in the price for a one-unit change in the corresponding predictor variable, assuming other variables are held constant.
For example, Coefficient for `sqft_living` is $123487.74
That means, for a one-unit increase in square-foot living area, we see an associated increase in around $123487.74 in selling price of the houses.

5. The plot to test for homoscedasticity reveals that the residuals are now homoscedastic because they are converging and appear to be having an equal variance. So this assumption remains satisfied.

6. The QQ-plot is used to test for normality of residuals. In this case, the residuals appear to be almost normal as they are following along the line almost neatly, except for the ends where it indicates there could be some skewness in the data.

RECOMMENDATIONS

1. The real estate agency should explore properties that occupy a large square foot of the lot area since, for a one-unit increase in square-foot of the lot area, we see an associated increase in around $ 20891.42 in selling price of the houses.

2. The real estate agency should explore properties that have more bathrooms since, for a one-unit increase in the number of bathrooms, we see an associated increase in around $ 46540.35 in selling price of the houses.

3. The real estate agency should explore properties that occupy a large square foot of living area since, for a one-unit increase in square-foot living area, we see an associated increase in around $ 123487.74 in selling price of the houses.

4. The real estate agency should explore properties with more floors since, for a one-unit increase in number of floors of the house, we see an associated increase in around $ 42072.21 in selling price of the houses.

NEXT STEPS

1. More research is required to have a more integrated and informative dataset for finding more factors that influence the price. Also, use of more complex and robust regression models that will help to deal with the outliers.

2. Using datasets from other counties to be able to better advice our customers from comparing the dataset results.

3. It is also important for the agency to continuously evaluate the effectiveness of the strategies they implement and make adjustments as necessary. This could involve tracking metrics like, this model, social media engagement/reviews, and lead generation to assess the impact of their efforts and identify areas for improvement.

KING COUNTY HOUSES

Thanks!

Github:
https://github.com/BrysonShitsukane77

LinkedIn:
https://www. linkedin.com/in/bryson-shitsukane-584bb7161