FINAL PROJECT SUBMISSION PHASE 2 FINAL PROJECT SUBMISSION

Students name: Matilda Odalo, Wyclife Orimba, Mark Bundi and Charles Kagwanja

Business Problem

- The business problem in this scenario involves providing homeowners with advice on how to increase the estimated value of their homes through renovation projects.

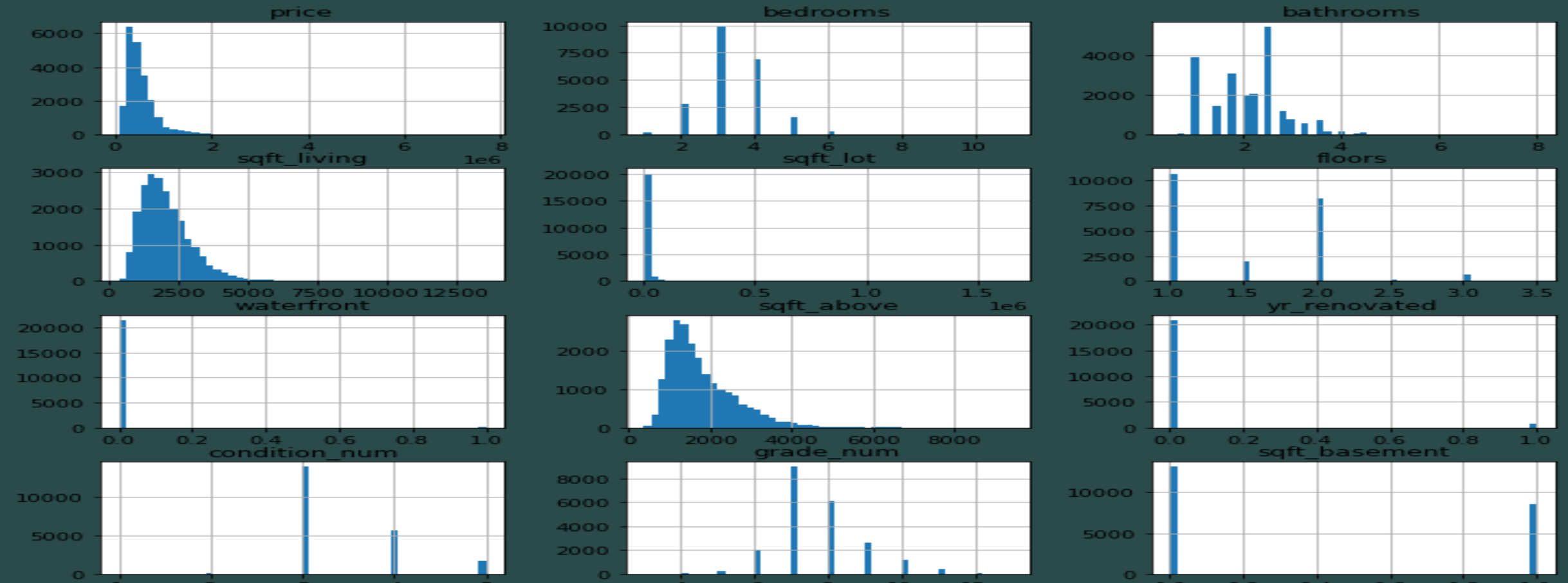
Data Understanding

- This project uses the King County House Sales dataset, which can be found in kc_house_data.csv.
- The data is used to create regression models that predict the trend of house prices in relation to specific variables. The below are the libraries that we are going to use for our analysis.

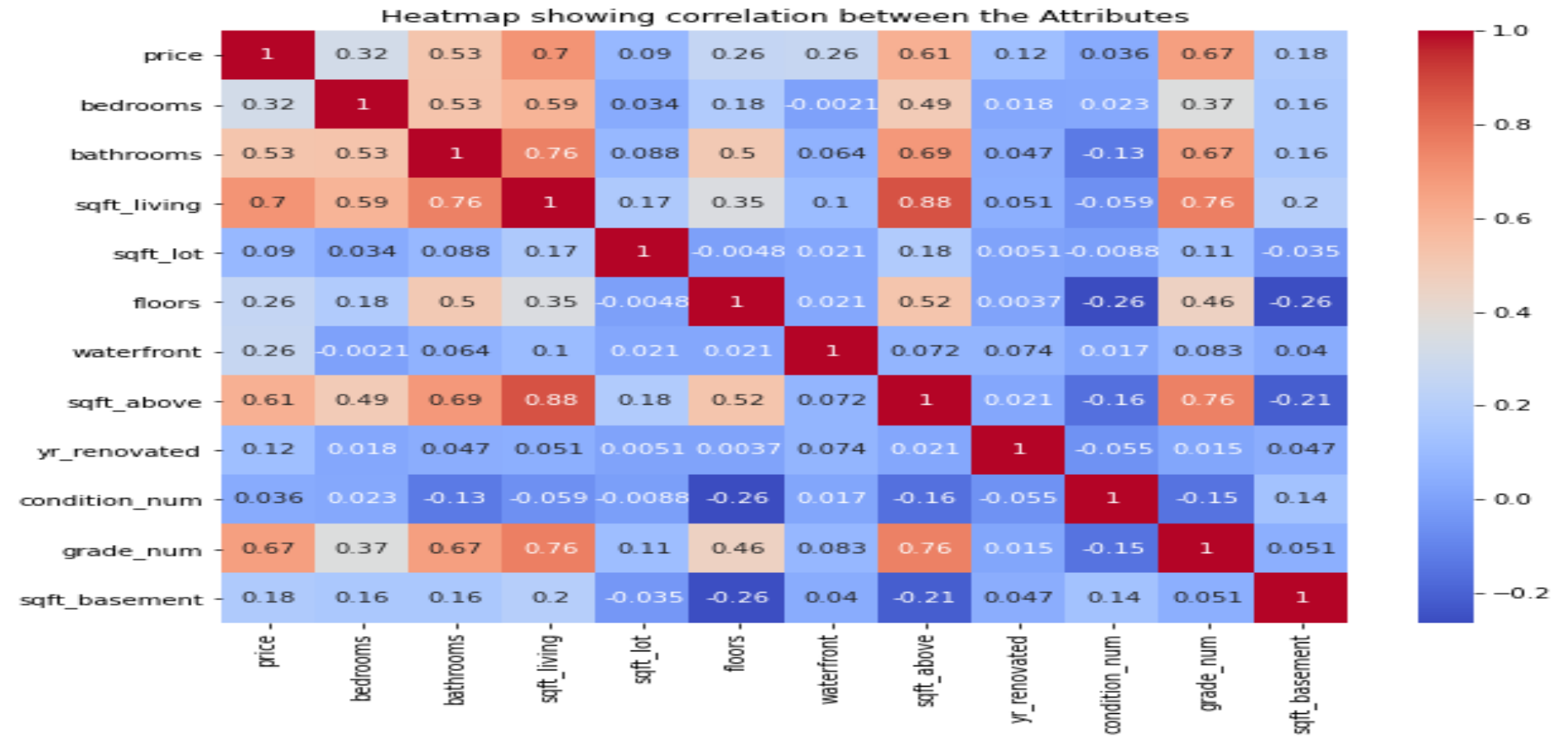
Data Preparation

- We prepared out data to be ready for analysis by handling missing values, outliers and changing of categorical data to numerical data to enable us use them in regression analysis easily.

We then plotted a histogram to check for the distribution in each column including the price column as well as the correlation between the variables.



We also plotted a Heatmap showing correlation between the Attributes



Observations

- The above Heatmap is just to show us how the independent variables are correlated with the dependent variable(price). sqft_living has a high positive correlation with price being 0.7 while condition_num has a low positive correlation with the price.

Modelling

Baseline Model

- The baseline model is a simple model used to contextualize the results of trained models. We create the baseline model to provide a reference point for measuring the performance of other models. We start with a simpler model as our base and work through it to make a much better base. In this instance we used y = price as the dependent variable and square foot living independent variable

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:	Sun, 10 Sep 2023	Prob (F-statistic):	0.00
Time:	15:13:05	Log-Likelihood:	-3.0005e+05
No. Observations:	21596	AIC:	6.001e+05
Df Residuals:	21594	BIC:	6.001e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-4.401e+04	4410.123	-9.980	0.000	-5.27e+04	-3.54e+04
sqft_living	280.8688	1.939	144.820	0.000	277.067	284.670

Omnibus:	14801.492	Durbin-Watson:	1.982
Prob(Omnibus):	0.000	Jarque-Bera (JB):	542642.481
Skew:	2.820	Prob(JB):	0.00
Kurtosis:	26.901	Cond. No.	5.63e+03

Model:	OLS	Adj. R-squared:	0.601
Method:	Least Squares	F-statistic:	2600.
Date:	Sun, 10 Sep 2023	Prob (F-statistic):	0.00
Time:	15:13:31	Log-Likelihood:	-2.3802e+05
No. Observations:	17276	AIC:	4.761e+05
Df Residuals:	17265	BIC:	4.761e+05
Df Model:	10		
Covariance Type:	nonrobust		

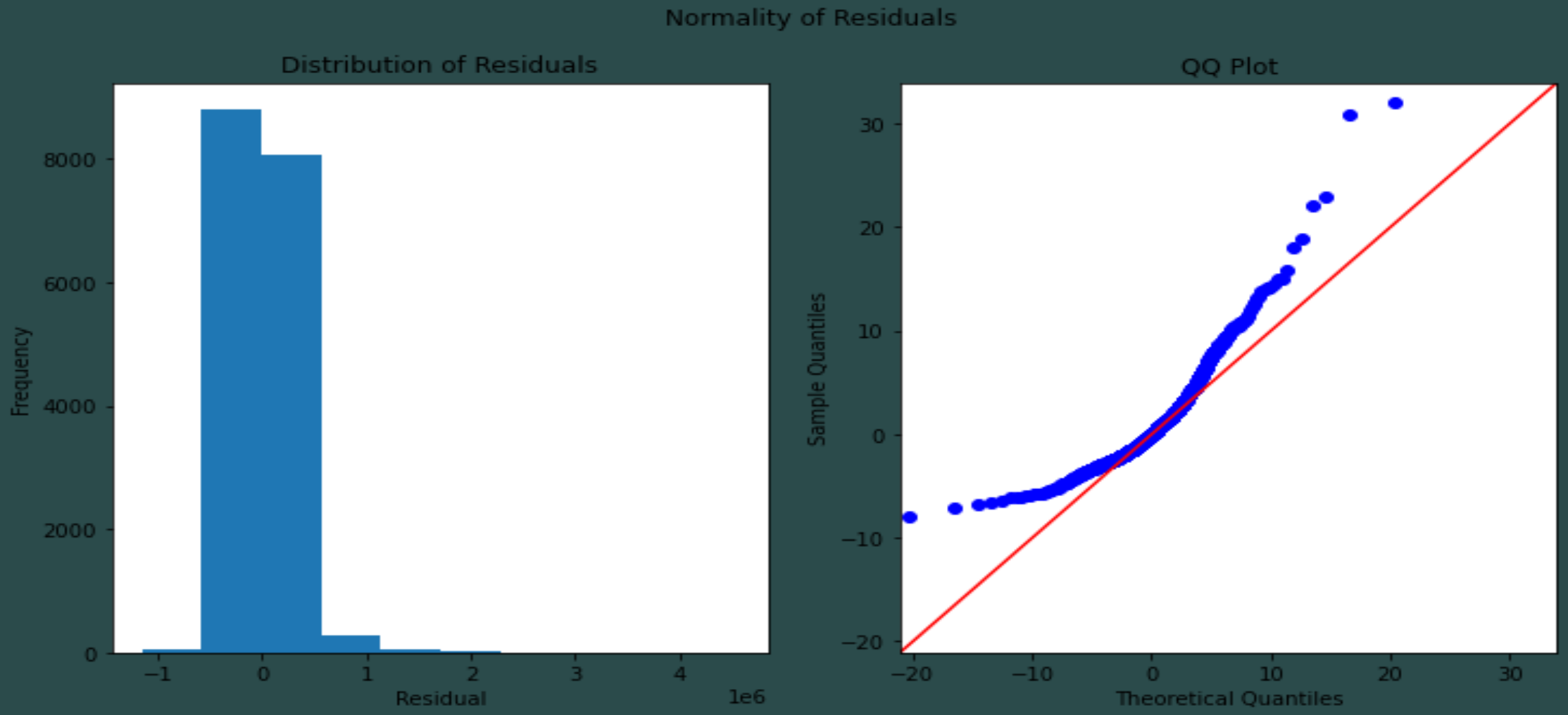
	coef	std err	t	P>|t|	[0.025	0.975]
const	-7.814e+05	1.99e+04	-39.336	0.000	-8.2e+05	-7.43e+05
bedrooms	-4.091e+04	2533.182	-16.148	0.000	-4.59e+04	-3.59e+04
bathrooms	-1.924e+04	4000.505	-4.809	0.000	-2.71e+04	-1.14e+04
sqft_living	202.0745	3.961	51.019	0.000	194.311	209.838
sqft_lot	-0.3158	0.044	-7.182	0.000	-0.402	-0.230
floors	-3494.6175	4318.995	-0.809	0.418	-1.2e+04	4971.050
waterfront	7.626e+05	2.17e+04	35.146	0.000	7.2e+05	8.05e+05
yr_renovated	1.587e+05	9840.279	16.131	0.000	1.39e+05	1.78e+05
condition_num	6.264e+04	2843.796	22.025	0.000	5.71e+04	6.82e+04
grade_num	1.11e+05	2536.920	43.739	0.000	1.06e+05	1.16e+05
sqft_basement	4.276e+04	4094.512	10.443	0.000	3.47e+04	5.08e+04

Omnibus:	12599.031	Durbin-Watson:	1.999
Prob(Omnibus):	0.000	Jarque-Bera (JB):	718345.414
Skew:	2.942	Prob(JB):	0.00
Kurtosis:	34.037	Cond. No.	5.40e+05

Observations

- On our baseline model, we have an R-squared value of 0.601 meaning that 60.1% of variation in price is due to the independent variables. The RMSE of the train set is 232918.7 and the RMSE of test set is 229101.3.The baseline model also has a skew of 2.942 which means that it is positively skewed. This means the dataset has a high percentage of outliers. It has a high Kurtosis of 34.037 indicating the dataset has high outliers. We also note that Floor has a p-value of 0.418 which is greater than alpha(0.05) indicating that it is an insignificant feature.

We also plotted a q-q plot to show the distribution of residuals



We note that the residuals are not normally distributed as per the below plots

In our second model, we removed the outliers in order to improve our model. The summary of the model is as below.

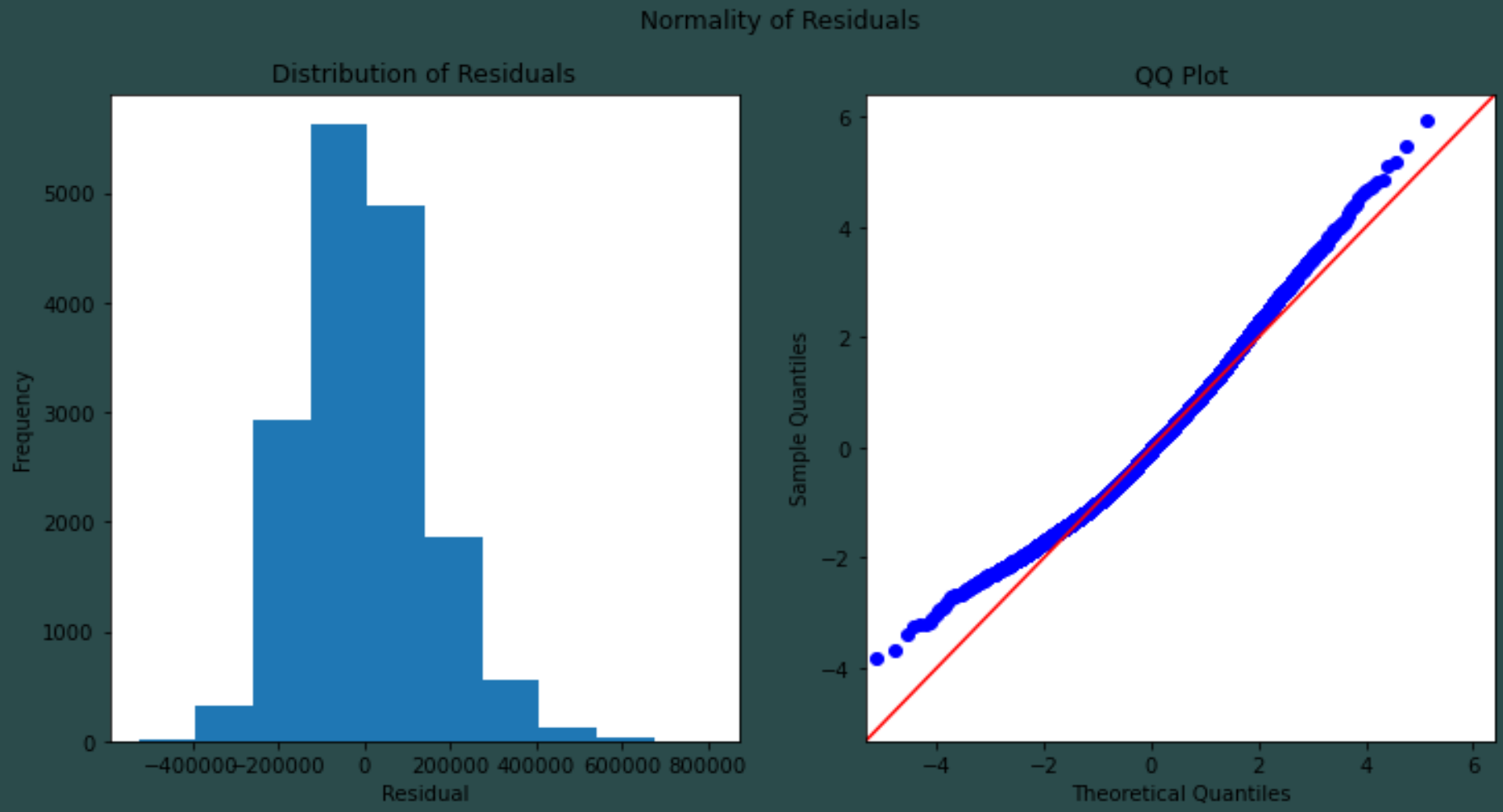
OLS Regression Results

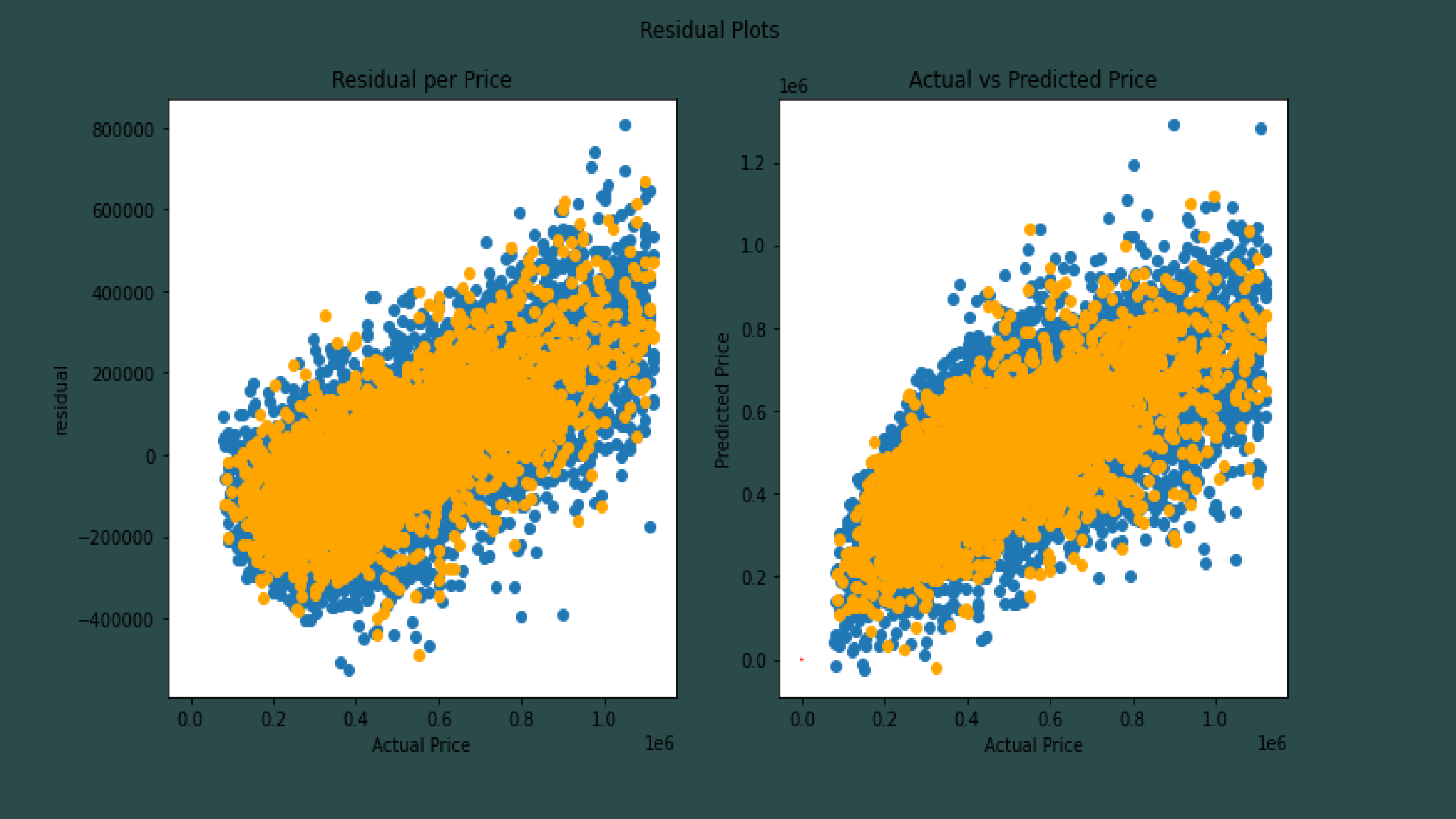
Dep. Variable:	price	R-squared:	0.503
Model:	OLS	Adj. R-squared:	0.502
Method:	Least Squares	F-statistic:	1651.
Date:	Sun, 10 Sep 2023	Prob (F-statistic):	0.00
Time:	15:14:29	Log-Likelihood:	-2.1768e+05
No. Observations:	16348	AIC:	4.354e+05
Df Residuals:	16337	BIC:	4.355e+05
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-4.831e+05	1.32e+04	-36.500	0.000	-5.09e+05	-4.57e+05
bedrooms	-1.367e+04	1694.721	-8.063	0.000	-1.7e+04	-1.03e+04
bathrooms	-2.286e+04	2632.429	-8.686	0.000	-2.8e+04	-1.77e+04
sqft_living	101.6064	2.780	36.544	0.000	96.157	107.056
sqft_lot	-0.0029	0.031	-0.092	0.927	-0.064	0.058
floors	2.492e+04	2833.736	8.794	0.000	1.94e+04	3.05e+04
waterfront	2.022e+05	2.2e+04	9.204	0.000	1.59e+05	2.45e+05
yr_renovated	1.009e+05	6744.622	14.962	0.000	8.77e+04	1.14e+05
condition_num	4.368e+04	1868.331	23.379	0.000	4e+04	4.73e+04
grade_num	8.553e+04	1697.098	50.397	0.000	8.22e+04	8.89e+04
sqft_basement	4.708e+04	2661.184	17.693	0.000	4.19e+04	5.23e+04

Omnibus:	805.029	Durbin-Watson:	2.012
Prob(Omnibus):	0.000	Jarque-Bera (JB):	980.049

qq_plot after removing the outliers. We note that there is a modrate normal distribution of the data

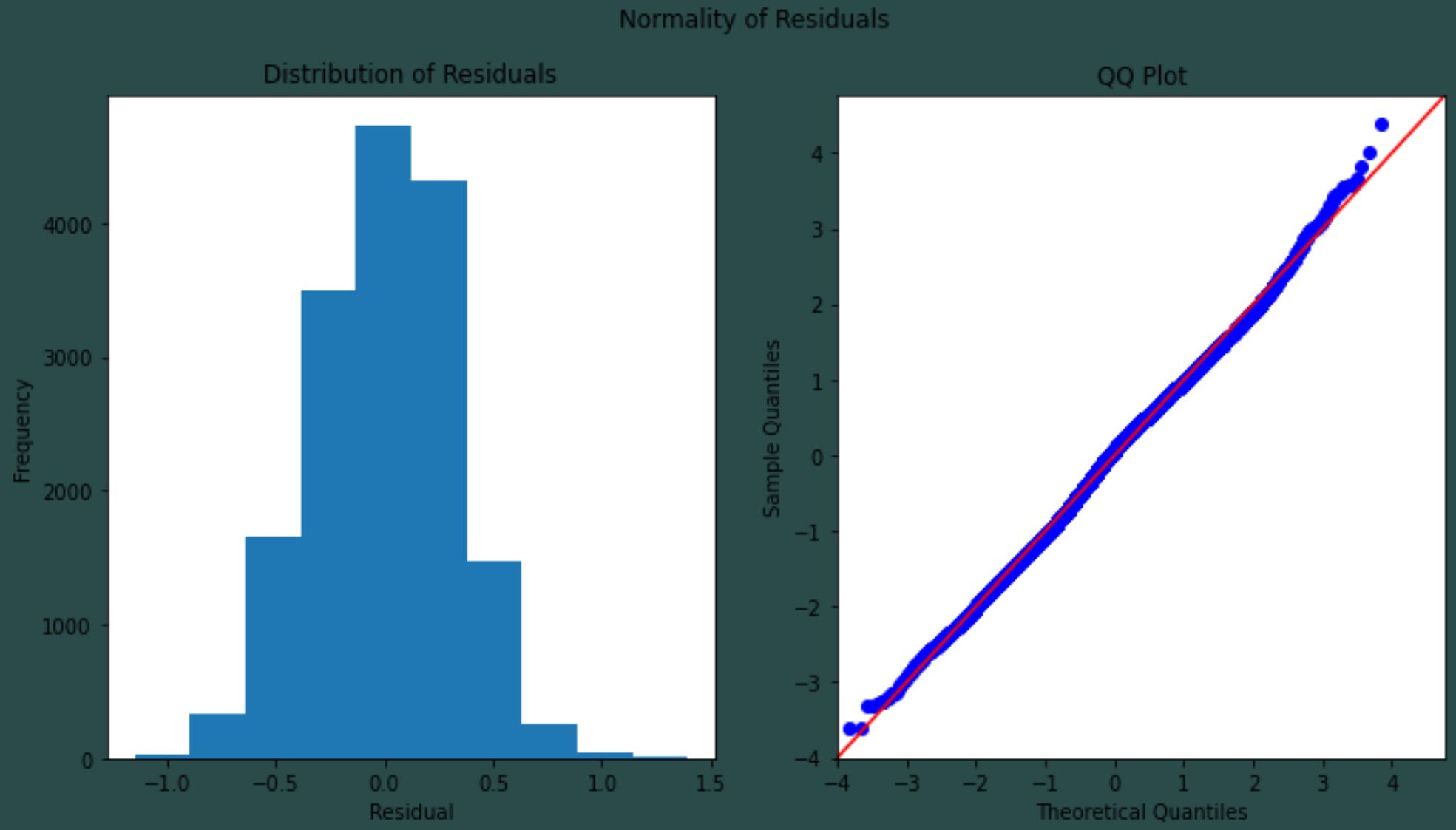



Residual Plots
Residual per Price
residual
800000
600000
400000
200000
0
−200000
−400000
0.0
0.2
0.4
0.6
0.8
1.0
1e6
Actual Price
Actual vs Predicted Price
1e6
Predicted Price
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
1e6
Actual Price

In our final model, We will try to do a log transformation to normalize the data further and the model summary results are as show below.

```
[ ] Train RMSE: 0.3183210543467515
    Test RMSE: 0.31652284300899786
                            OLS Regression Results
    ==============================================================================
    Dep. Variable:                  price   R-squared:                       0.491
    Model:                            OLS   Adj. R-squared:                  0.491
    Method:                 Least Squares   F-statistic:                     1578.
    Date:                Sun, 10 Sep 2023   Prob (F-statistic):               0.00
    Time:                        15:15:27   Log-Likelihood:                -4483.3
    No. Observations:               16348   AIC:                             8989.
    Df Residuals:                   16337   BIC:                             9073.
    Df Model:                          10
    Covariance Type:            nonrobust
    ==============================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
    ------------------------------------------------------------------------------
    const          8.3078      0.074    112.295      0.000       8.163       8.453
    bedrooms      -0.1342      0.012    -11.203      0.000      -0.158      -0.111
    bathrooms     -0.1000      0.011     -9.333      0.000      -0.121      -0.079
    sqft_living    0.4583      0.014     33.900      0.000       0.432       0.485
    sqft_lot      -0.0353      0.003    -10.332      0.000      -0.042      -0.029
    floors         0.0727      0.010      6.985      0.000       0.052       0.093
    waterfront     0.4193      0.048      8.779      0.000       0.326       0.513
    yr_renovated   0.1853      0.015     12.662      0.000       0.157       0.214
    condition_num  0.0921      0.004     22.708      0.000       0.084       0.100
    grade_num      0.1836      0.004     50.537      0.000       0.177       0.191
    sqft_basement  0.1017      0.006     16.323      0.000       0.089       0.114
    ==============================================================================
    Omnibus:                       22.098   Durbin-Watson:                   2.015
    Prob(Omnibus):                  0.000   Jarque-Bera (JB):               21.732
```

qq_plot after the log transformation of the data



Observations

- On our model 1, we have an R-squared value of 0.491 meaning that 49.1% of variation in price is due to the independent variables.
- The RMSE of the train set is 0.3183210543467515 and the RMSE of test set is 0.3165228430089978 6.
- Model 1 also has a skew of -0.077 which is not very significant. The data is more normally distributed
- It has a high Kurtosis of 2.910 indicating the dataset has few outliers.

Conclusion

- The r squared of model one is 0.491 compared to 0.493 of our baseline model. Though it's a bit less than the baseline model, however the difference is not substantial.
- "Model one" is considered a better model than the "baseline model" for several reasons, despite a slightly lower R^2: It has significantly lower kurtosis, indicating that its residuals are closer to a normal distribution, which is an assumption of linear regression. It also has lower skewness, suggesting a more symmetric distribution of residuals.

Recommendations

- From the regression model we can interpret that features that improve household value are in descending priority:

✓ Square_foot living

✓ Waterfront

✓ Grade of the house

✓ Renovation

The homeowner should focus on these features when renovating the house for maximum profits