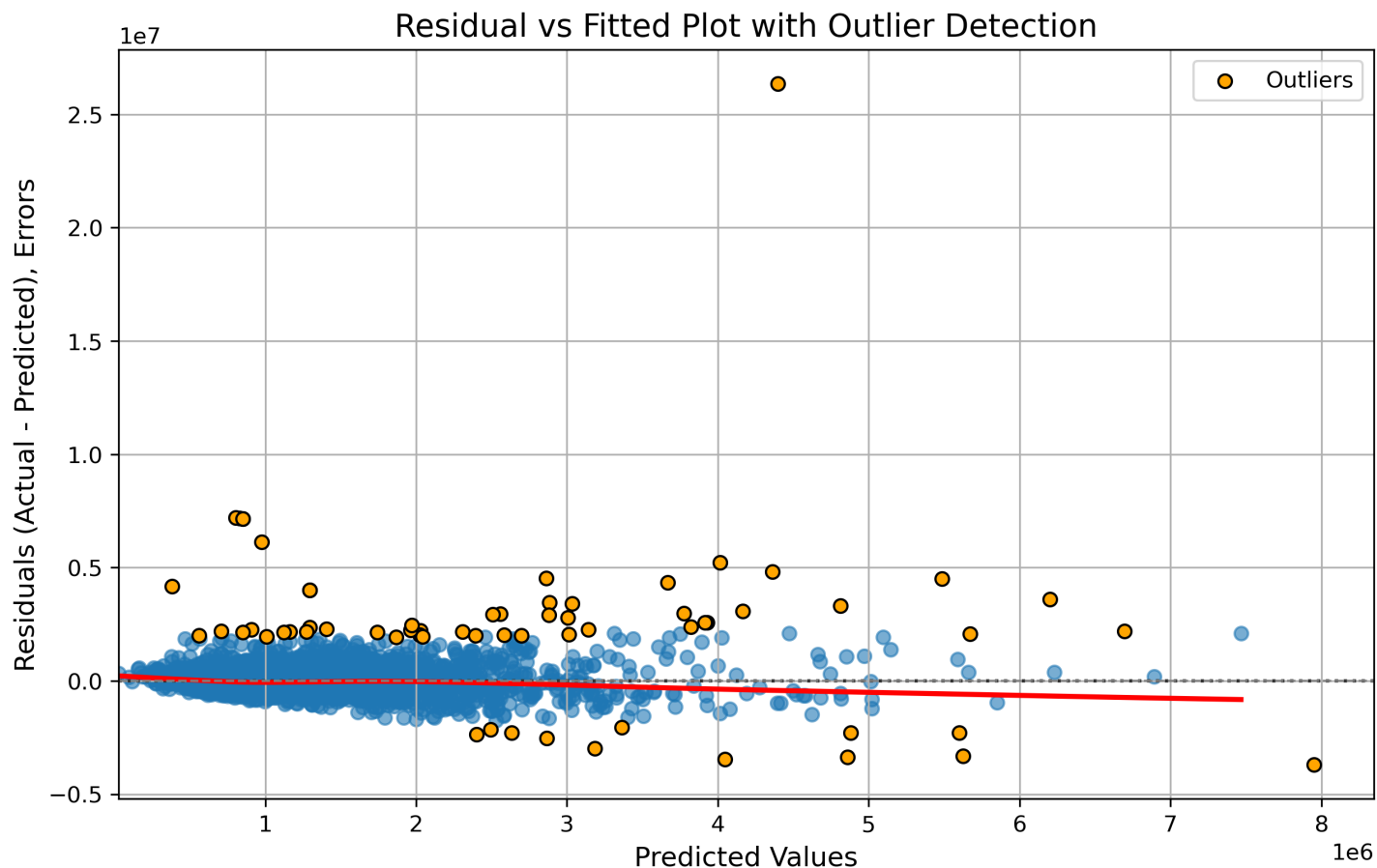


This model and explanation were generated by ModelBot, an agent designed to help non-technical users perform basic machine learning modeling, powered by Llama 3. It is not a replacement for a human data scientist, and there may be discrepancies and inaccuracies within this report.

Linear Regression Results						
Dependent Variable:	price	R-squared:	0.536			
Model:	linear	Adjusted R-squared:	0.536			
No. Observations:	24124	F-statistic:	647.150			
Df Residuals:	24093	Prob (F-statistic):	1.11e-16			
Df Model:	31	AIC:	711,201.890			
RMSE:	609,128.252	BIC:	711,557.893			
	coef	std err	t	P> t	0.025	0.975
waterfront_1	8.1e+05	3.7e+04	22.03	0	7.3e+05	8.8e+05
view_EXCELLENT	5.9e+05	3.7e+04	16.01	0	5.2e+05	6.6e+05
sewer_system_PUBLIC	2e+05	1.3e+04	15.94	0	1.8e+05	2.3e+05
sqft_above	2e+05	1.7e+04	11.69	0	1.6e+05	2.3e+05
bathrooms	8.2e+04	7.3e+03	11.22	0	6.7e+04	9.6e+04
sqft_basement	7.9e+04	8.2e+03	9.61	0	6.3e+04	9.5e+04
sqft_garage	-5.2e+04	5.6e+03	-9.31	0	-6.3e+04	-4.1e+04
bedrooms	-5.3e+04	5.6e+03	-9.38	0	-6.4e+04	-4.2e+04
yr_built	-8.8e+04	6.6e+03	-13.34	0	-1e+05	-7.5e+04
condition_Very Good	9.6e+04	1.4e+04	6.75	1.5e-11	6.8e+04	1.2e+05
sqft_patio	3e+04	4.5e+03	6.57	5.1e-11	2.1e+04	3.9e+04
floors	-3.2e+04	5.8e+03	-5.60	2.2e-08	-4.4e+04	-2.1e+04
sqft_living	9.3e+04	1.8e+04	5.06	4.1e-07	5.7e+04	1.3e+05
yr_renovated	2e+04	4.3e+03	4.66	3.2e-06	1.2e+04	2.8e+04
Intercept	2.8e+06	6.1e+05	4.64	3.4e-06	1.6e+06	4e+06
sqft_lot	2e+04	4.3e+03	4.61	4.1e-06	1.1e+04	2.8e+04
nuisance_YES	4.7e+04	1.1e+04	4.44	9e-06	2.6e+04	6.8e+04
view_FAIR	1.9e+05	4.8e+04	3.94	8.3e-05	9.5e+04	2.8e+05
grade_4 Low	-2.3e+06	6.2e+05	-3.77	0.00016	-3.5e+06	-1.1e+06
grade_3 Poor	-2.3e+06	6.4e+05	-3.67	0.00025	-3.6e+06	-1.1e+06
grade_5 Fair	-2.2e+06	6.1e+05	-3.60	0.00032	-3.4e+06	-1e+06
grade_6 Low Average	-2.2e+06	6.1e+05	-3.60	0.00032	-3.4e+06	-1e+06
grade_7 Average	-2.1e+06	6.1e+05	-3.47	0.00051	-3.3e+06	-9.2e+05
grade_13 Mansion	2.2e+06	6.3e+05	3.44	0.00059	9.3e+05	3.4e+06
condition_Good	3.4e+04	1e+04	3.37	0.00074	1.4e+04	5.4e+04
	coef	std err	t	P> t	0.025	0.975
grade_2 Substandard	-2.5e+06	7.5e+05	-3.35	0.0008	-4e+06	-1e+06
grade_8 Good	-2e+06	6.1e+05	-3.20	0.0014	-3.1e+06	-7.5e+05
grade_9 Better	-1.6e+06	6.1e+05	-2.67	0.0075	-2.8e+06	-4.4e+05
heat_source_Gas/Solar	1.7e+05	7.1e+04	2.36	0.018	2.9e+04	3.1e+05
sewer_system_PRIVATE RESTRICTED	-5.8e+05	2.5e+05	-2.29	0.022	-1.1e+06	-8.4e+04
condition_Fair	-1.1e+05	4.6e+04	-2.29	0.022	-2e+05	-1.5e+04



## ***PREDICTION TARGET***

The model is predicting the price of a property based on various features such as waterfront location, view, sewer system, square footage, number of bedrooms and bathrooms, year built, condition, and more. This prediction is useful for real estate agents, property owners, and potential buyers to estimate the value of a property. Accurate predictions can help them make informed decisions about buying, selling, or investing in properties.

## ***METRIC EXPLANATIONS***

**R<sup>2</sup> (Coefficient of Determination):** This metric measures the proportion of the variance in the target variable (price) that is explained by the independent variables (features). A high R<sup>2</sup> value (close to 1) indicates that the model is able to accurately predict the target variable, while a low value (close to 0) suggests that the model is not able to capture the underlying patterns in the data.

**RMSE (Root Mean Squared Error):** This metric measures the average magnitude of the errors in the model's predictions. A low RMSE value indicates that the model is making accurate predictions, while a high value suggests that the model is making large errors.

**Overfitting:** When a model is overfitting, it is too complex and is able to fit the noise in the training data, resulting in poor generalization performance on new, unseen data. This is often indicated by a high difference between the training and testing R<sup>2</sup> values.

Underfitting: When a model is underfitting, it is too simple and is unable to capture the underlying patterns in the data, resulting in poor performance on both training and testing data. This is often indicated by a low  $R^2$  value for both training and testing data.

## **OVERFITTING OR UNDERFITTING?**

Based on the provided scores, the model is overfitting. The training  $R^2$  value is 0.54, while the testing  $R^2$  value is 0.49, indicating a significant drop in performance when the model is applied to new data. This suggests that the model is too complex and is fitting the noise in the training data, resulting in poor generalization performance.

## **ANOVA TABLE OVERVIEW**

The ANOVA table provides information about the model's ability to explain the variance in the target variable (price). The columns are:

- Degrees of Freedom: The number of independent variables (features) and the number of observations (data points) minus the number of independent variables.
- F-statistic: A measure of the ratio of the variance explained by the model to the variance not explained by the model.
- p-value: The probability of observing the F-statistic or a more extreme value under the null hypothesis that the model does not have any predictive power.

The ANOVA table suggests that the model has statistically significant predictive power, as the p-value is very low ( $1.11e-16$ ).

## **SIGNIFICANT FEATURES ANALYSIS**

The top 5 most significant features are:

- waterfront\_1 (coefficient =  $8.1e+05$ , p-value = 0.0): Properties with a waterfront location have a significantly higher price.
- view\_EXCELLENT (coefficient =  $5.9e+05$ , p-value = 0.0): Properties with an excellent view have a significantly higher price.
- sewer\_system\_PUBLIC (coefficient =  $2e+05$ , p-value = 0.0): Properties with a public sewer system have a significantly higher price.
- sqft\_above (coefficient =  $2e+05$ , p-value = 0.0): Properties with more square footage above ground have a significantly higher price.
- bathrooms (coefficient =  $8.2e+04$ , p-value = 0.0): Properties with more bathrooms have a significantly higher price.

These features are significant, meaning that they have a statistically significant impact on the target variable (price).

## **RESIDUAL PLOT EVALUATION**

A residual plot is used to evaluate the model's performance by examining the distribution of the residuals (the differences between actual and predicted values). The characteristics of a good residual plot are:

- Randomness: Residuals should be randomly scattered around zero.
- Patterns: There should be no systematic shape or pattern in the residuals.

- Homoscedasticity: The variance of the residuals should be constant.
- Outliers: There should be no extreme residuals that suggest poor predictions.
- Normality: Residuals should roughly follow a normal distribution.

The residual plot for this model shows some signs of heteroscedasticity (non-constant variance) and some outliers, indicating that the model may not be performing well.

## ***KEY INSIGHTS FROM RESIDUALS AND PREDICTIONS***

Based on the residuals and predictions, we can conclude that:

- The residuals are not evenly distributed, suggesting that the model is not capturing the underlying patterns in the data.
- There are signs of heteroscedasticity, indicating that the variance of the residuals is not constant.
- The predictions are not accurate, as the residuals show large errors.
- The residuals suggest systematic problems, indicating that the model may not be generalizing well to new data.

## ***MODEL RATING***

Based on the model's performance, I would rate it 6/10. The model has some significant features and is able to explain a significant portion of the variance in the target variable (price). However, the model is overfitting, and the residuals show signs of heteroscedasticity and outliers, indicating that the model may not be generalizing well to new data.