Multiple Linear Regression Analysis of House Prices

Group 4:-
Najma Abdi
Victor Keya
Millicent Kanana
Lewis Ngunjiri’
Amina Saidi
Janet Maluka

Outline

1. Project Overview
2. Objectives
3. Data Understanding
4. Data cleaning and preparation
5. Exploratory data analysis –Bivariate and Univariate analysis
6. Statistical Analysis
7. Modelling
8. Regression Results
9. Results/Findings
10. Recommendations
11. Conclusion

Project Overview

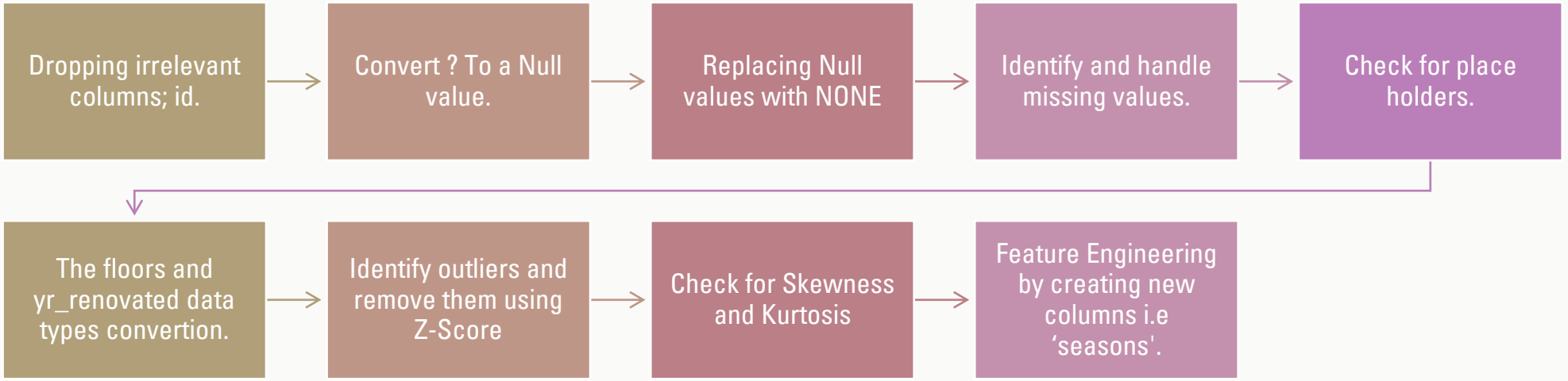
Objectives

- Determine the key factors such as square foot living, the number of bedrooms and bathrooms, the condition of the house and others that significantly influence the house prices.
- Develop a model that can accurately predict house prices based on these factors.
- To determine which seasons have the highest sales.
- To provide valuable insights for real estate agents, property developers, and investors in the company’s portfolio to make informed decisions regarding pricing, renovations, and marketing strategies

Data Understanding

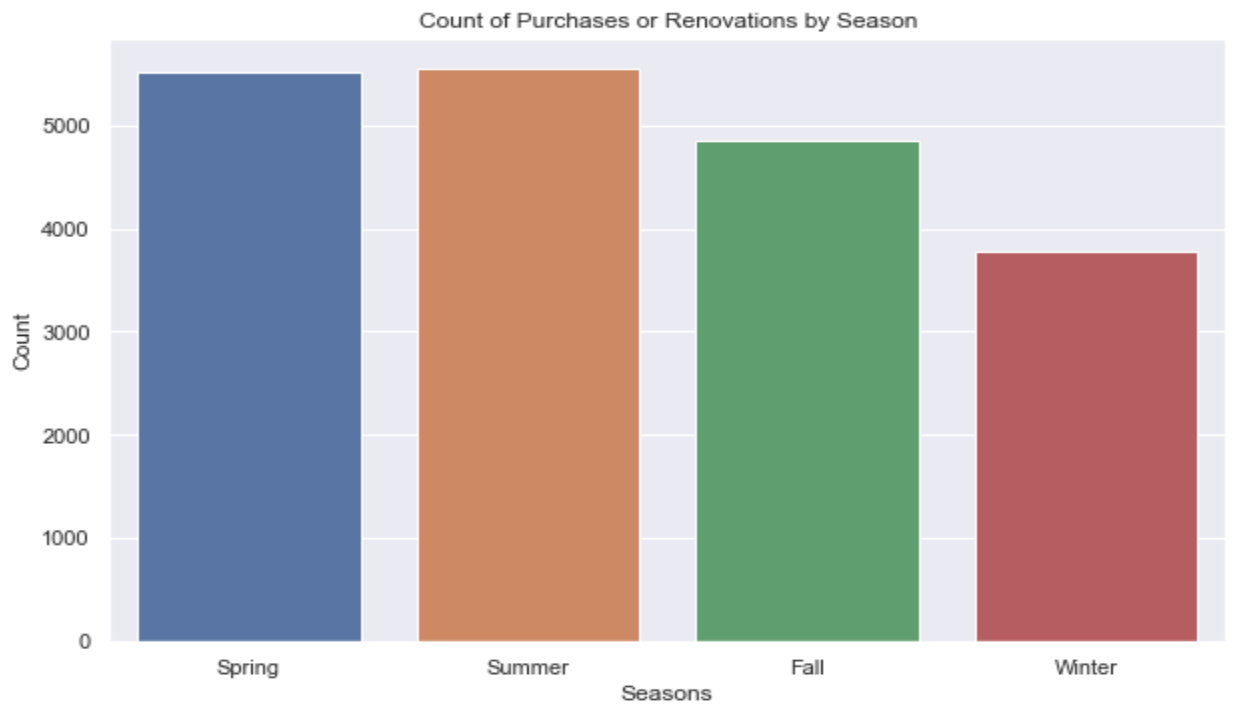
- `id` - Unique identifier for a house
- `date` - Date house was sold
- `price` - Sale price (prediction target)
- `bedrooms` - Number of bedrooms
- `bathrooms` - Number of bathrooms
- `sqft_living` - Square footage of living space in the home
- `sqft_lot` - Square footage of the lot
- `floors` - Number of floors (levels) in house
- `waterfront` - Whether the house is on a waterfront
    - Includes Duwamish, Elliott Bay, Puget Sound, Lake Union, Ship Canal, Lake Washington, Lake Sammamish, other lake, and river/slough waterfronts
- `view` - Quality of view from house
    - Includes views of Mt. Rainier, Olympics, Cascades, Territorial, Seattle Skyline, Puget Sound, Lake Washington, Lake Sammamish, small lake / river / creek, and other
- `condition` - How good the overall condition of the house is. Related to maintenance of house.
- `grade` - Overall grade of the house. Related to the construction and design of the house.
- `sqft_above` - Square footage of house apart from basement
- `sqft_basement` - Square footage of the basement
- `yr_built` - Year when house was built
- `yr_renovated` - Year when house was renovated
- `zipcode` - ZIP Code used by the United States Postal Service
- `lat` - Latitude coordinate
- `long` - Longitude coordinate
- `sqft_living15` - The square footage of interior housing living space for the nearest 15 neighbors
- `sqft_lot15` - The square footage of the land lots of the nearest 15 neighbors

Data Cleaning and Preparation



Exploratory Data Analysis

Univariate Analysis



Most houses were bought and renovated during the summer and spring seasons; less houses were bought during winter season.

Statistical Analysis

Statistical analysis is used to understand relationships within the dataset, identifying patterns, and gaining insights.

- Descriptive Statistics
- Correlation matrix
- Distribution Analysis
- Inferential Statistics using Hypothesis Testing and Analysis of Variance
- Multicollinearity

Descriptive Analysis

Understanding data characteristics. For Mean, Mode, Variance and standard deviation

Price Distribution:

- The prices of houses in the dataset vary widely, with a mean price of approximately 540,296.6 to 367,368.1. The prices range from 78,000 to 7,700,000.

Property Characteristics:

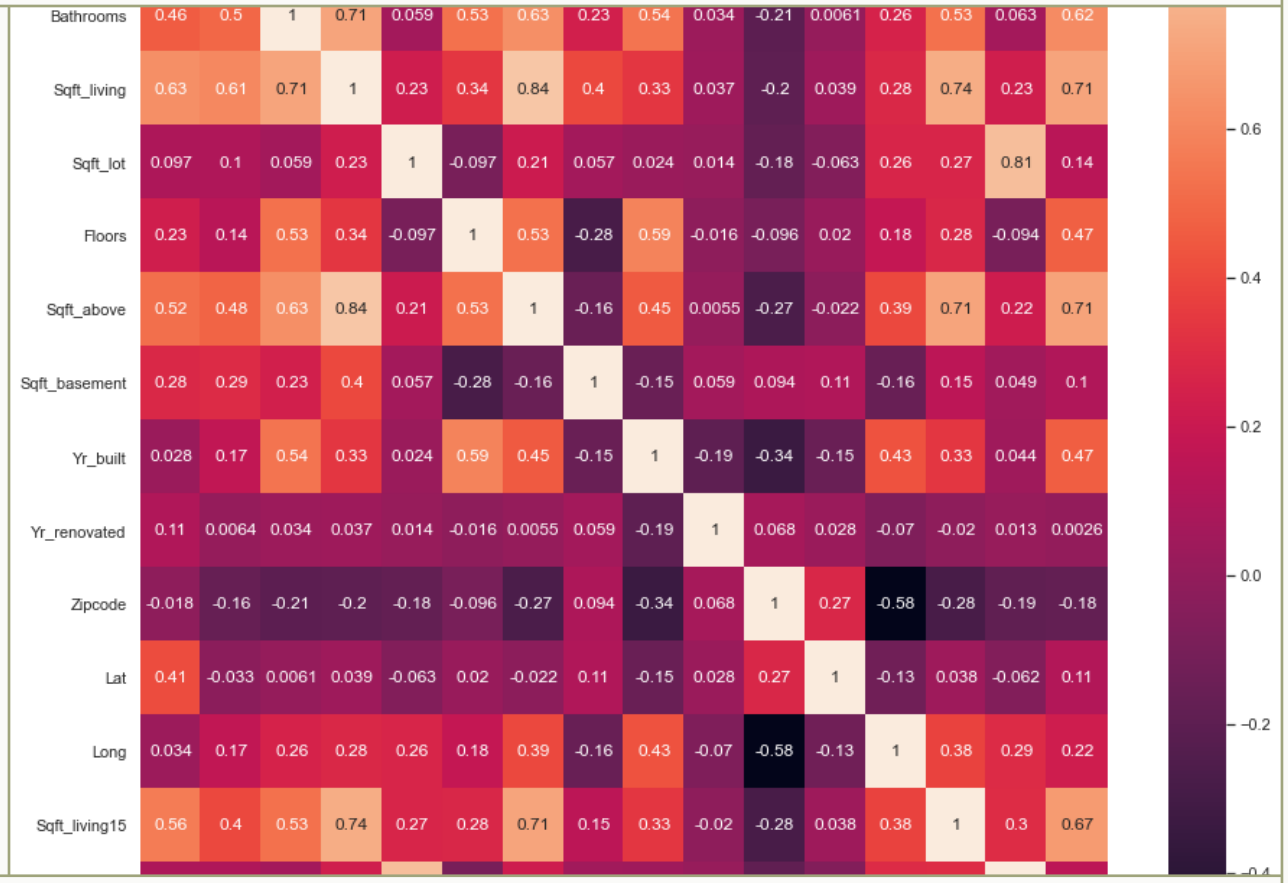
- The dataset contains information on various property characteristics such as the number of bedrooms, bathrooms, square footage of living space, and lot size. For example, the average number of bathrooms is approximately 2.11, with a standard deviation of about 0.77.

Year Built:

- The houses in the dataset were built between 1900 and 2015, with an average year of construction around 1971.

Correlation Matrix

- Price has a moderate positive correlation with sqft living (0.63), sqft above (0.52), sqft living15 (0.56) and Grading(0.64). This means that as the values of these features increase, the price of the house also tends to increase.
- There is a weak positive correlation between price and floors (0.23) and sqft_basement (0.28).
- Price has a weak negative correlation with zipcode (-0.018).



Hypothesis Testing

ANOVA ANALYSIS

Ho: There is no statistically significant interaction effect between sqft_living and grading on the price of a house

H1: There is a statistically significant interaction effect between sqft_living and grading on the price of a house.

	Bathrooms	Bedrooms	Sqft_living	Grading	Sqft_above	\
Bathrooms	1.000000	0.496519	0.712976	0.621718	0.630923	
Bedrooms	0.496519	1.000000	0.605821	0.341800	0.483000	
Sqft_living	0.712976	0.605821	1.000000	0.714065	0.843143	
Grading	0.621718	0.341800	0.714065	1.000000	0.708186	
Sqft_above	0.630923	0.483000	0.843143	0.708186	1.000000	
Sqft_living15	0.529788	0.395662	0.740939	0.673105	0.710484	
Lat	0.006097	-0.032532	0.038606	0.109614	-0.021854	
Floors	0.526388	0.137811	0.337780	0.470591	0.529241	
Yr_renovated	0.034426	0.006353	0.037267	0.002597	0.005483	
Sqft_lot	0.059252	0.102367	0.225068	0.135724	0.209207	
Sqft_lot15	0.063438	0.102743	0.228282	0.151002	0.217345	
Long	0.257063	0.170344	0.277599	0.221835	0.394315	
Yr_built	0.538480	0.168718	0.334372	0.466052	0.450352	
Zipcode	-0.205413	-0.163142	-0.197157	-0.177999	-0.267725	

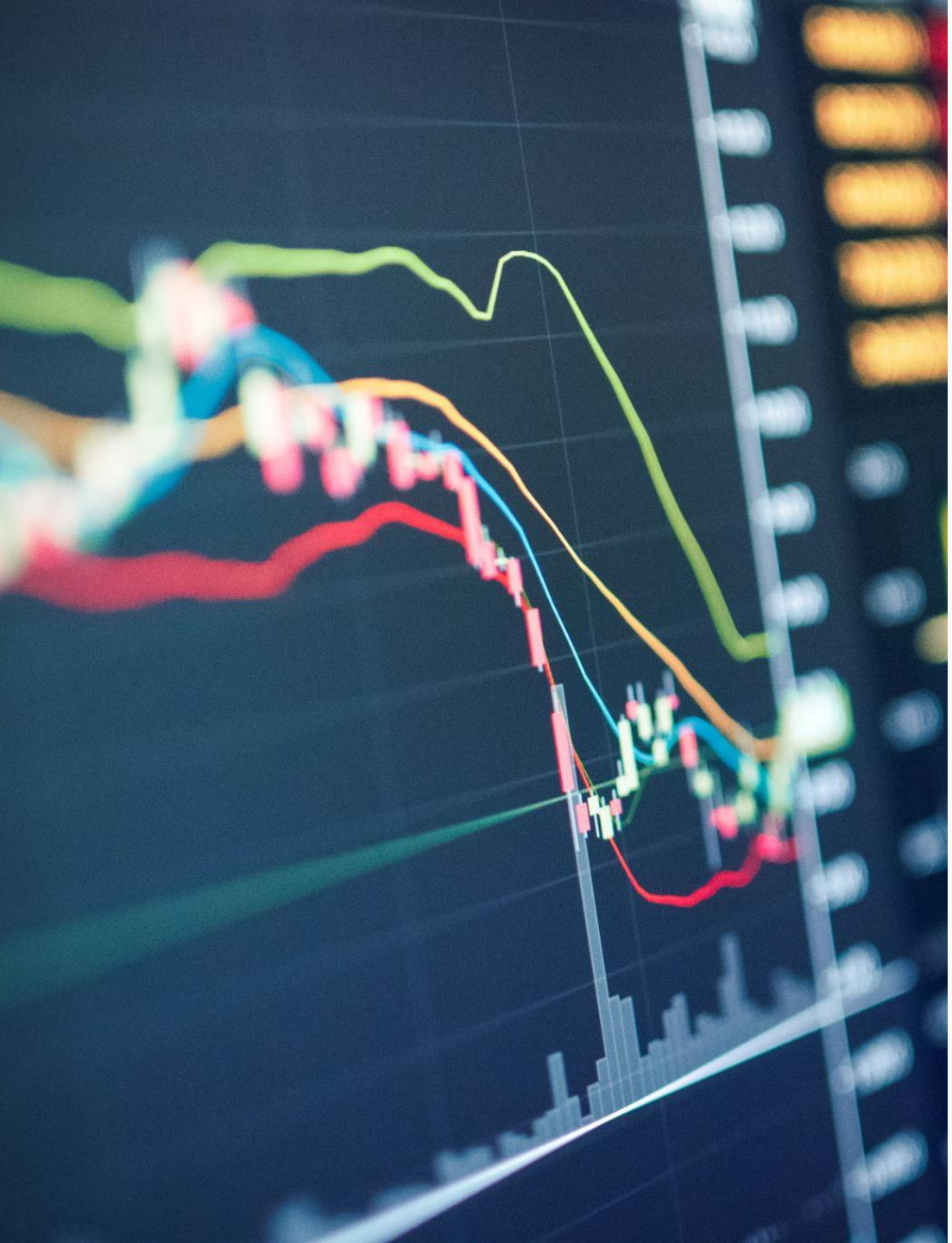
	Sqft_living15	Lat	Floors	Yr_renovated	Sqft_lot
Bathrooms	0.529788	0.006097	0.526388	0.034426	0.059252
Bedrooms	0.395662	-0.032532	0.137811	0.006353	0.102367
Sqft_living	0.740939	0.038606	0.337780	0.037267	0.225068
Grading	0.673105	0.109614	0.470591	0.002597	0.135724
Sqft_above	0.710484	-0.021854	0.529241	0.005483	0.209207
Sqft_living15	1.000000	0.038245	0.276098	-0.019999	0.268220
Lat	0.038245	1.000000	0.019597	0.027616	-0.062716
Floors	0.276098	0.019597	1.000000	-0.015763	-0.097265
Yr_renovated	-0.019999	0.027616	-0.015763	1.000000	0.013946
Sqft_lot	0.268220	-0.062716	-0.097265	0.013946	1.000000
Sqft_lot15	0.295851	-0.062436	-0.094464	0.012742	0.813587
Long	0.381006	-0.130564	0.178326	-0.070311	0.264670
Yr_built	0.333088	-0.152746	0.590668	-0.194751	0.023701
Zipcode	-0.282029	0.271737	-0.096346	0.068337	-0.179550

	Sqft_lot15	Long	Yr_built	Zipcode
Bathrooms	0.063438	0.257063	0.538480	-0.205413
Bedrooms	0.102743	0.170344	0.168718	-0.163142
Sqft_living	0.228282	0.277599	0.334372	-0.197157
Grading	0.151002	0.221835	0.466052	-0.177999
Sqft_above	0.217345	0.394315	0.450352	-0.267725
Sqft_living15	0.295851	0.381006	0.333088	-0.282029
Lat	-0.062436	-0.130564	-0.152746	0.271737
Floors	-0.094464	0.178326	0.590668	-0.096346
Yr_renovated	0.012742	-0.070311	-0.194751	0.068337
Sqft_lot	0.813587	0.264670	0.023701	-0.179550
Sqft_lot15	1.000000	0.294880	0.043846	-0.192665
Long	0.294880	1.000000	0.426155	-0.582936
Yr_built	0.043846	0.426155	1.000000	-0.341316
Zipcode	-0.192665	-0.582936	-0.341316	1.000000

	Feature	VIF
0	Bathrooms	2.924803e+01
1	Bedrooms	2.725006e+01
2	Sqft_living	5.241568e+01
3	Grading	1.507519e+02
4	Sqft_above	3.644921e+01
5	Sqft_living15	3.011027e+01
6	Lat	1.363032e+05
7	Floors	1.760715e+01
8	Yr_renovated	1.119356e+00
9	Sqft_lot	5.194917e+00
10	Sqft_lot15	6.240169e+00
11	Long	1.720040e+06
12	Yr_built	9.401162e+03
13	Zipcode	2.020593e+06

Multicolinearity

- High correlations(eg above 0.7 ot 0.8), between pairs of variables suggest potential multicollinearity
- If the VIF is greater than 5 or 10,it indicates multicolinearity. Higher VIF values signify stronger correlation with other predictors.



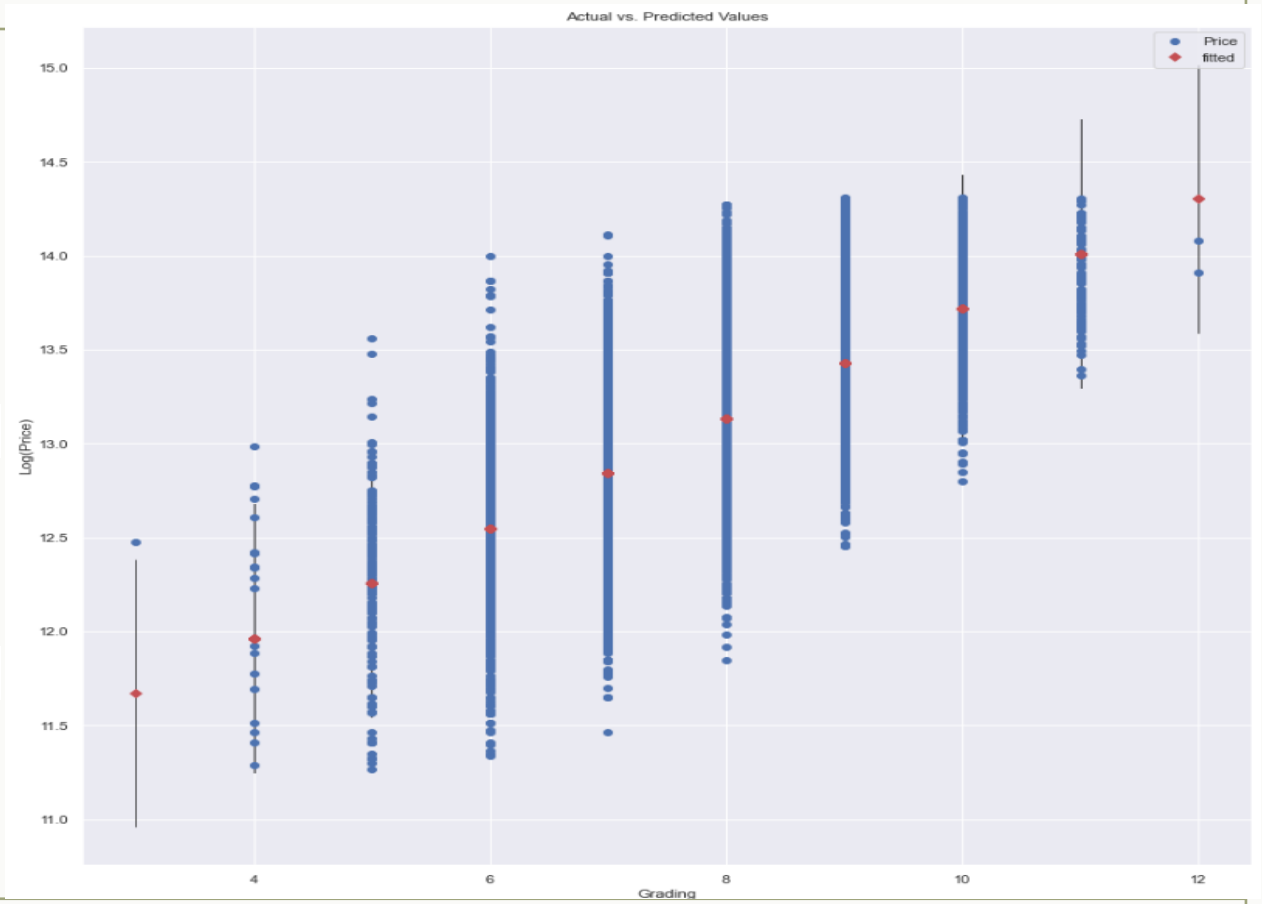
Data Modelling

We demonstrated the three models in the notebook as stated below but we will recommend only polynomial because it is the most suitable regression model to analyze the prices for a more accurate output.

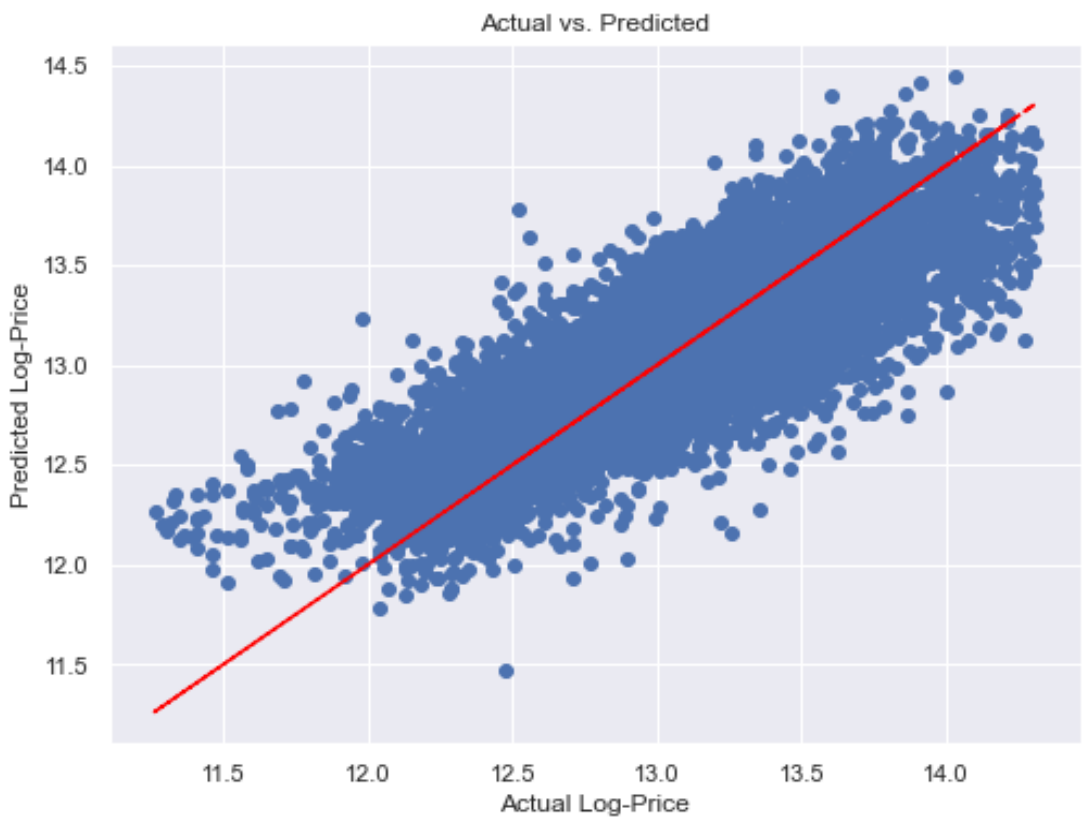
- Baseline Model
- Multilinear Regression
- Polynomial Regression

Baseline model

- This visualization provides insights into how well the model fits the training data and how the predicted values compare to the actual target values. Our model indicates a good fit between the predicted values and the actual values because the fitted line closely follows the diagonal (a 45-degree line from the bottom-left to the top-right of the plot).



Multilinear Regression

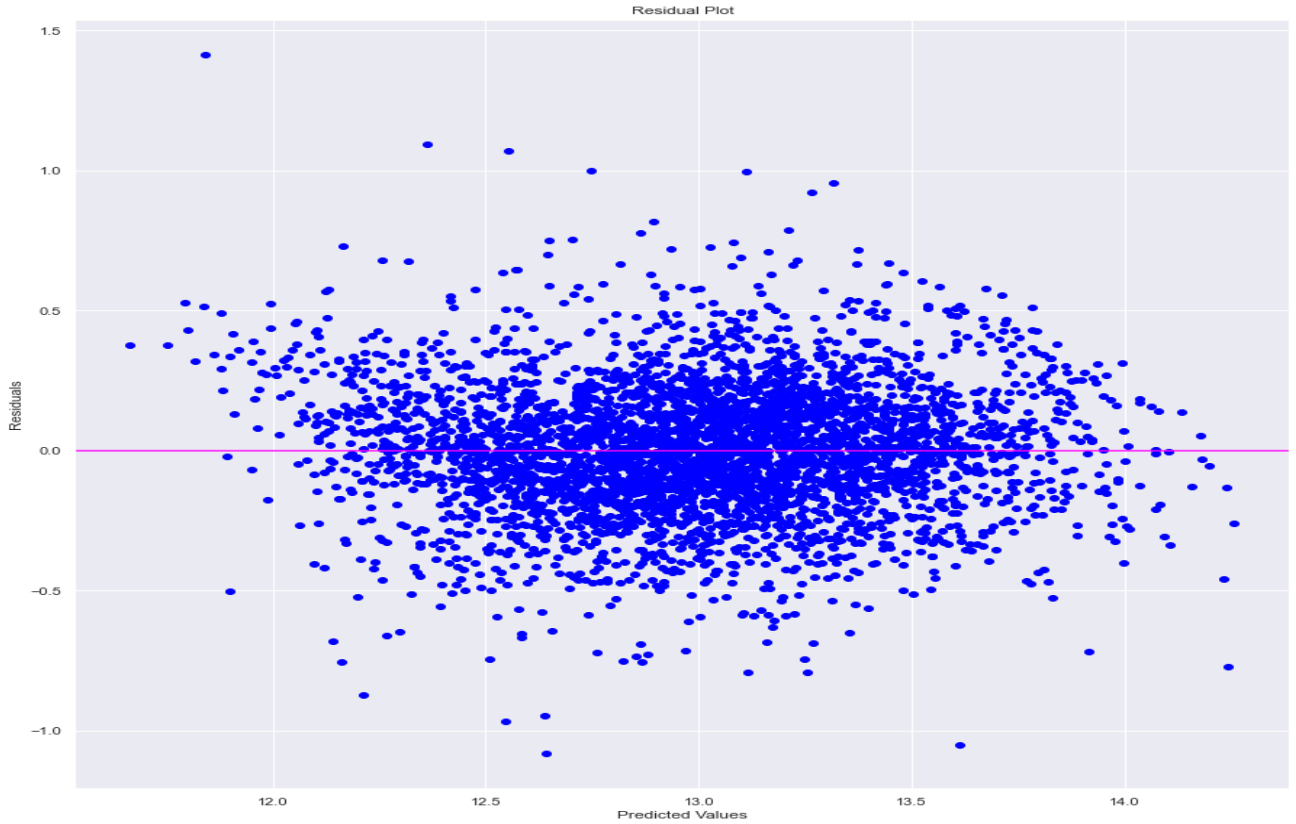


- Adding more predictors to our baseline model can enhance its predictive power by capturing additional complexity and potential relationships within our data. This can lead to improved model accuracy and potentially reducing bias by considering more factors in the prediction process.

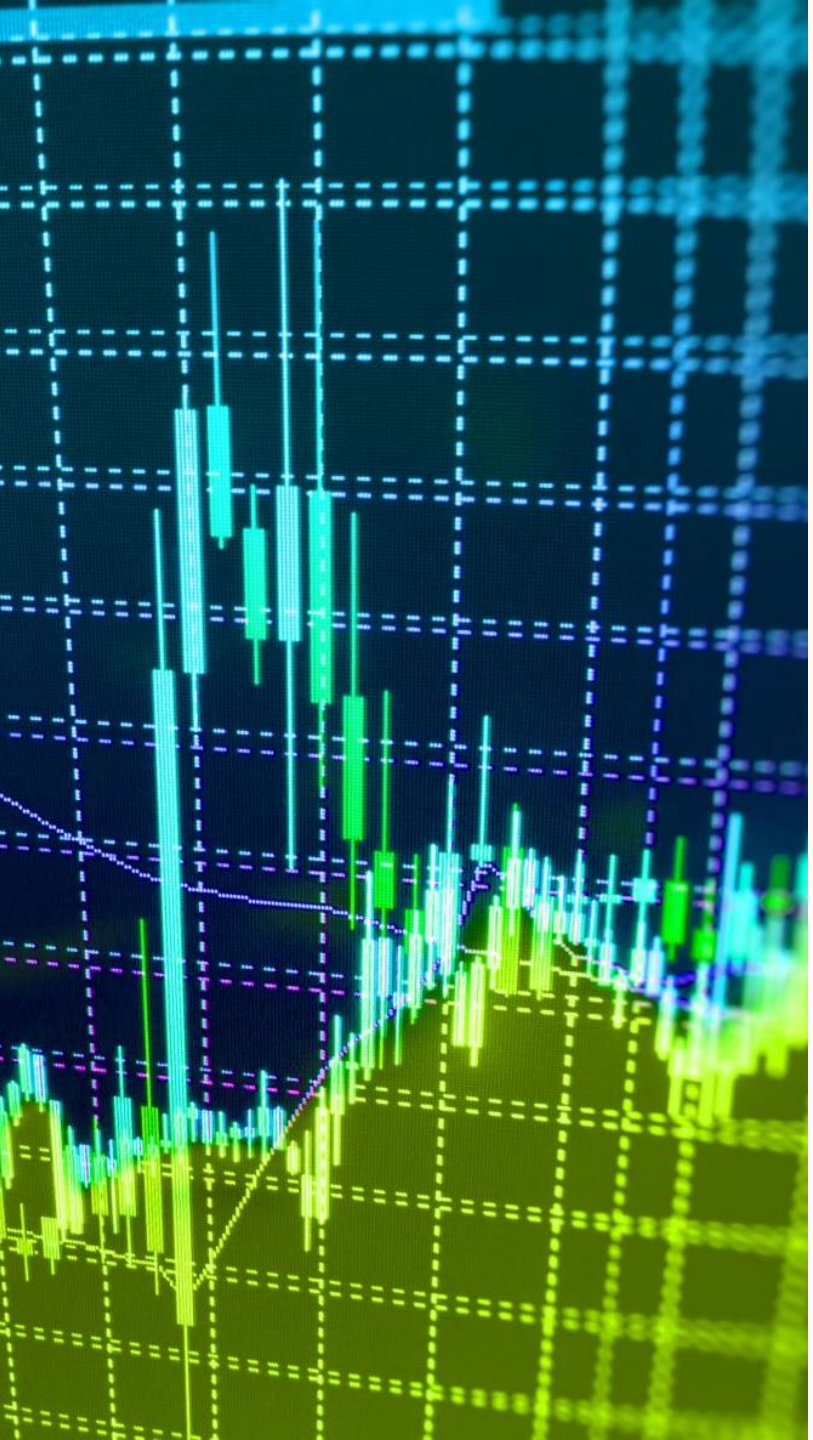
- The model explains 69% of the variance in price as indicated by R squared
- Therefore, the multiple linear regression is better than our baseline regression

Polynomial Regression

Polynomial RMSE: 0.23061511620028652
R_squared: 0.7655407239791847



Given that polynomial regression accounts for roughly 76% of the volatility in house prices—a percentage somewhat greater than that explained by multiple linear regression—polynomial regression is a superior model in this situation. Polynomial regression also has a less Root Squared Mean Error than Multiple Linear Regression Model. This suggests that the polynomial regression yielded superior results. The RMSE of the polynomial regression also suggests the same since it is also lower than the multiple linear regression model.

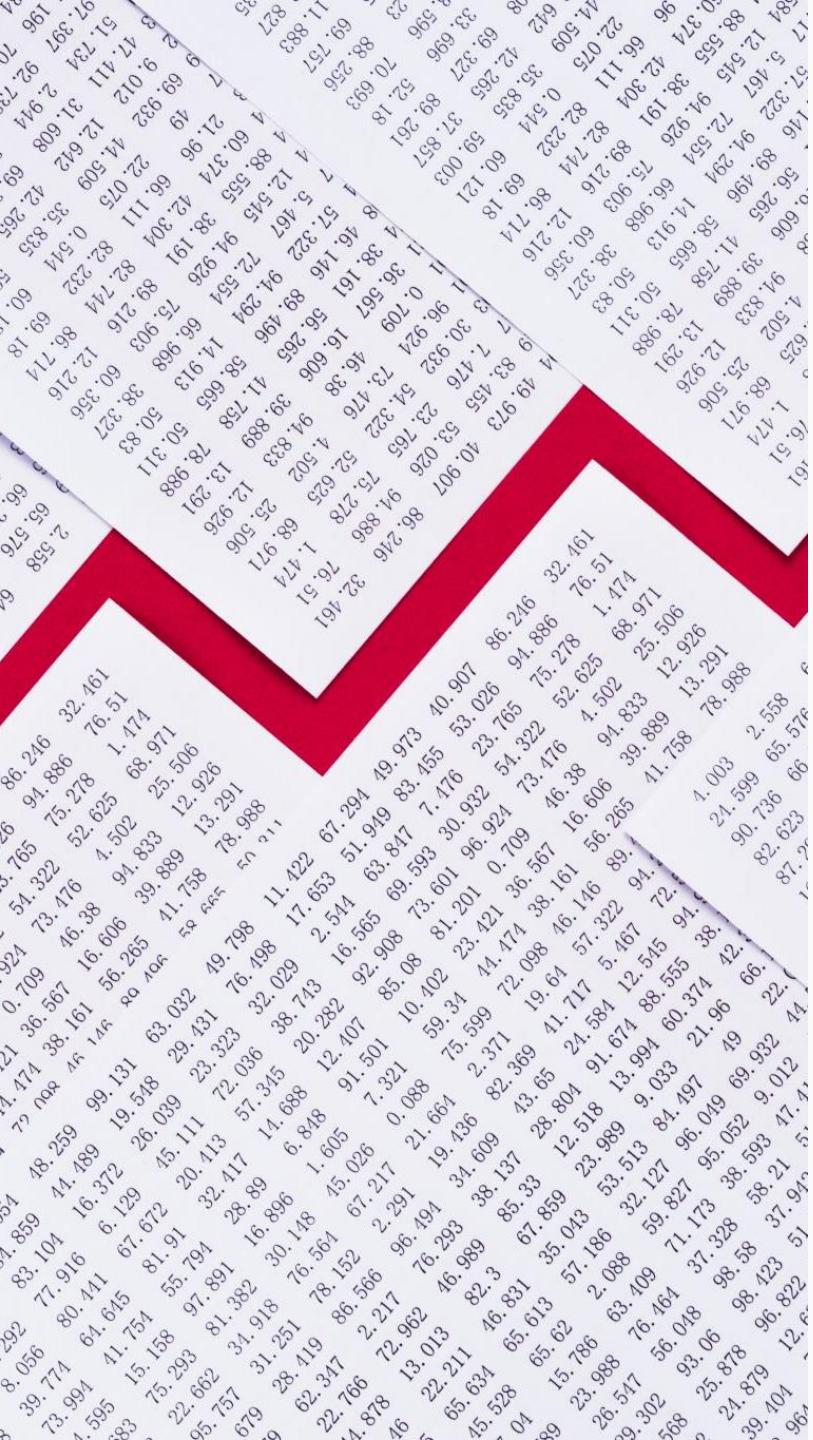


Regression Results

We embarked on an iterative statistical modelling process where we started with a simple linear regression, multilinear regression modelling and finally Polynomial regression where we noted that for each regression, the model became better with each iteration.

1. The baseline model had an R Squared of 0.42
2. The multilinear regression had an R squared of 0.69 and an RMSE of 0.27
3. The polynomial regression model had an Rsquared of 0.77 and an RMSE of 0.23

- For the final polynomial regression RMSE value , our model is off by 0.23 dollars in each prediction.

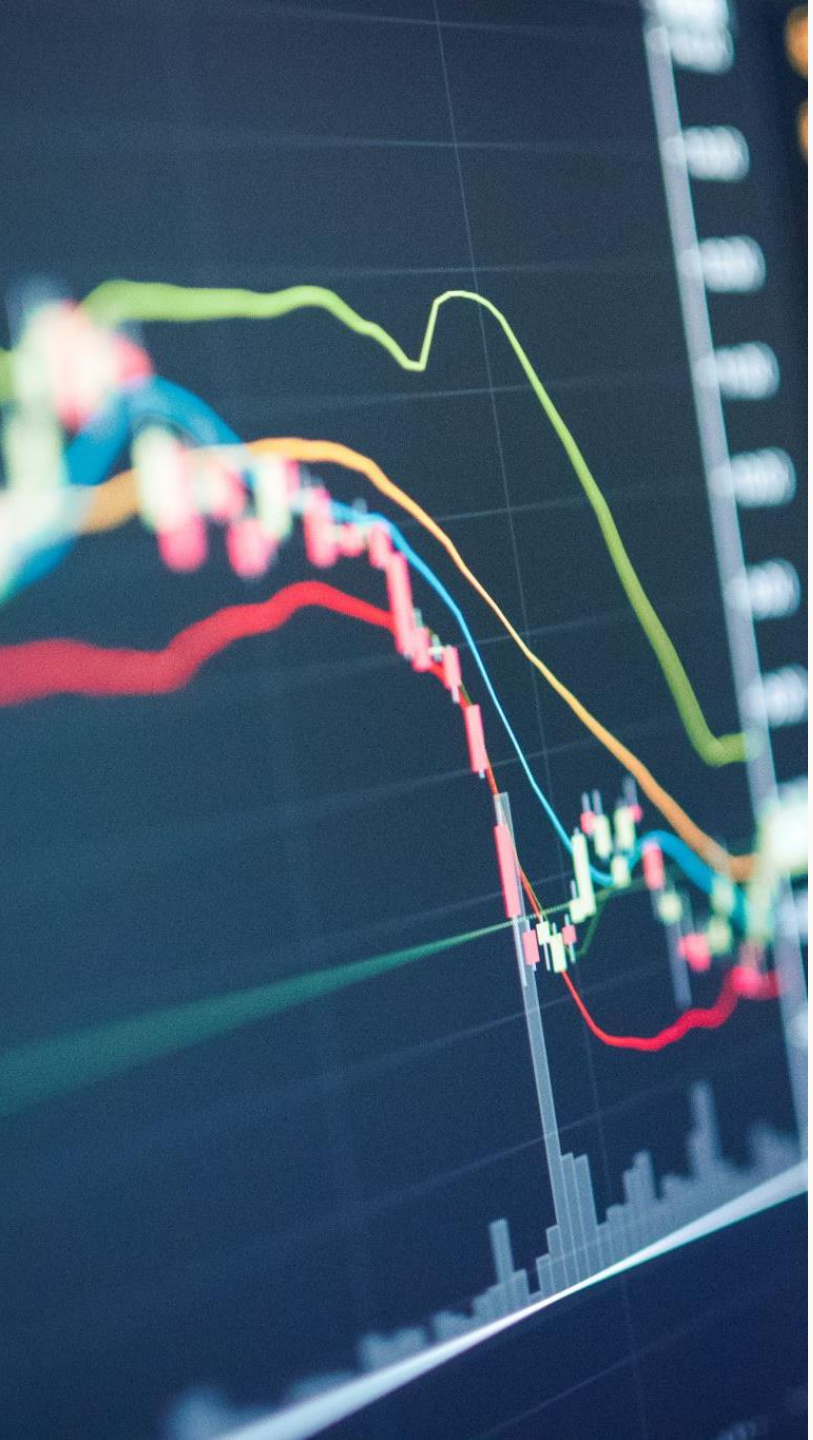


Results/Findings

- Grading: The coefficient of 0.1349 suggests that for each unit increase in the grading, the log-transformed price increases by approximately 0.1349, holding other variables constant. The p-value of 0.000 indicates that this coefficient is statistically significant.
- Yr_renovated: The coefficient is 4.34e-05, indicating a very small effect size. This suggests that year of renovation has a minimal impact on the log-transformed price. Howe
- Sqft_living15: The coefficient is approximately 0.0001, indicating that for each unit increase in square footage of living space (15 nearest neighbors), the log-transformed price increases by approximately 0.0001. The p-value of 0.000 indicates that this coefficient is statistically significant.
- Bedrooms: The coefficient of -0.0164 suggests that an increase in the number of bedrooms leads to a decrease in the log-transformed price by approximately 0.0164, holding other variables constant.

Recommendations

- The square footage of living space has a significant influence on house pricing as well. This information can be used by the Real Estate Agency to support higher listing prices for homes with larger square footage.
- There are several benefits to using a polynomial regression model instead of a typical linear model when predicting property prices. By taking into consideration the non-linear effects of important features like square footage, location, and on-site facilities, this method enables us to capture intricate correlations and fluctuations in home pricing. The polynomial regression's high R-squared value indicates that the chosen features account for a significant amount of the variability in home prices, which translates into more accurate and consistent price projections. Using this cutting-edge modeling technique gives us, as real estate market participants, more decision-making power and improves pricing tactics, investment analyses, and market competitiveness overall.



Conclusion

In this project, we conducted an in-depth analysis of the King County (KC) house dataset to predict house prices using advanced regression techniques. The goal was to provide valuable insights for stakeholders in the real estate industry.

Key Findings:

- Feature Importance: Through regression analysis, we identified several key predictors that significantly influence house prices, including grading, square footage of living space, number of bedrooms, and bathrooms. Effect of Features:
- Effect of Features: Our findings reveal nuanced relationships between house features and pricing. For instance, higher grading and larger living spaces positively impact prices.
- Model Performance: We compared different regression models and found that polynomial regression outperformed multiple linear regression model, capturing non-linear relationships more effectively and yielding higher predictive accuracy.