Project Documentation: House Price Prediction Model

Introduction:
This project is about building a predictive model to estimate the prices of houses based on various features. The data set used for this project contains information about houses located in King County, Washington, USA, and includes features such as number of bedrooms, bathrooms, square footage of living area, lot size, age of the house, number of floors, and the season when the house was sold.

Data Cleaning and Exploration:
The data set was first explored and cleaned to ensure that it was suitable for building a predictive model. Data cleaning involved checking for missing values, outliers, and duplicates. Exploratory data analysis was also performed to understand the distribution of the features and identify any correlations between the features and the target variable (price).

Feature Engineering:
To improve the performance of the predictive model, feature engineering was performed. This involved creating new features based on the existing features. For example, the age of the house was converted to a categorical variable based on the decade it was built in. The square footage of living area and lot size were also log-transformed to improve the linearity of their relationship with the target variable.

Model Building and Evaluation:
In the project, we began by building a baseline model to compare the performance of other models. We then developed a multiple linear regression model that predicted the price of a product based on several independent variables. The independent variable with the highest correlation with the price was identified and used in the regression model with multiple independent variables. To address the possibility of a non-linear relationship between the variables, we used a non-linear transformation technique called Log transformations. This technique helped us model a linear relationship between the transformed variables and improved the accuracy of our predictions.

The models were evaluated using metrics such as R-squared, mean squared error (MSE), and root mean squared error (RMSE). The best model was chosen based on its R-squared value, which indicates the percentage of variance in the target variable explained by the model. The final model included log-transformed features for number of bedrooms, bathrooms, square footage of living area, lot size, and number of floors.

Recommendations:

Bathrooms: The number of bathrooms has a positive effect on the price of the house, meaning that houses with more bathrooms tend to be priced higher. The agency may want to consider this factor when pricing and marketing homes with more bathrooms.

Living Area and Lot Size: The size of the living area has a positive effect on the price of the house, while the size of the lot has a negative effect. The agency may want to consider emphasizing the living area in their marketing efforts, while also being mindful of the lot size.

Floors: Houses with more floors tend to be priced higher. The agency may want to consider this factor when pricing and marketing multi-story homes.

Condition and Grade: Houses with higher condition and grade ratings tend to be priced higher. The agency may want to emphasize these ratings in their marketing efforts and pricing strategy.

Age and Renovated: The age of the house and whether or not it has been renovated both have significant effects on the price of the house. The agency may want to consider these factors when pricing and marketing homes, particularly when comparing newer, renovated homes to older ones.

Waterfront View: Houses with a waterfront view are priced significantly higher than those without. The agency may want to emphasize this factor in their marketing efforts for waterfront properties.

Season: The season in which a house is sold can also affect the price, with spring selling for higher prices than fall. The agency may want to consider this factor when planning their marketing and pricing strategies throughout the year.

Conclusion:
The house price prediction model can be used to estimate the prices of houses in King County, Washington, USA, based on their features. The model has a high R-squared value and can provide accurate predictions with a low error.

Next Steps:
Further improvements to the model can be made by incorporating additional features or optimizing the existing features.