
A decorative graphic on the left side of the slide consisting of overlapping geometric shapes. It includes a blue parallelogram, a light green parallelogram, and a dark grey parallelogram, all with sharp, angular edges.

Predicting King County, WA, Housing Prices using Multiple Linear Regression

**Presented by,
Collins Kanyiri**



Stark Realtors Agency is in the business of helping homeowners buy and/or sell houses in King County WA in the US.

Using the provided dataset of house sales from the agency, the task here is to model the real estate housing prices and use the model to accurately predict the housing prices based on a number of features provided within the dataset.

The expectation is that, once completed, the model can be used as a tool in selecting properties for investment in King County.



Data Used

- price - Price is prediction target
- sqft_living - square footage of the home
- view - Does the property have a view?
- grade - overall grade given to the housing unit, based on King County grading system
- sqft_living15 - The square footage of interior housing living space for the nearest 15 neighbors
- sqft_above - square footage of house apart from basement
- bedrooms - Number of Bedrooms/House



Modeling

Why Statistical Modeling?

Statistical Modeling is the use of data along with statistics to provide a framework for understanding data relationships. In statistical modeling, statistical analysis is used rather than the basic data analysis. The reason why this is the case is because data analysis focuses on exploring the data to get useful information out of it while statistical analysis infers what is beyond data analysis by predicting future and hidden trends using the same data. The business problem for this project is predicting housing prices based on provided historical data and this is why statistical modeling will be adopted.



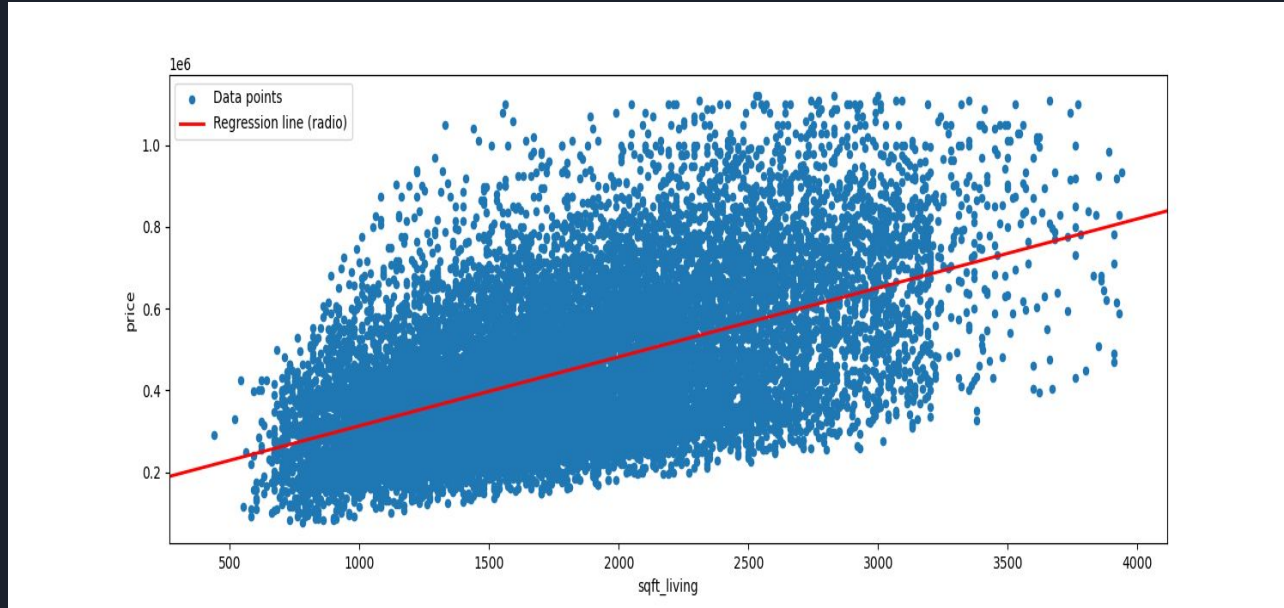
Which Statistical Model should we use?

We are interested in estimating the relationship between a dependant variable (price) vs independent variables (all the other relevant features in the dataset) thus a regression analysis will be used, in particular Linear Regression. Linear Regression analysis is a powerful statistical method that allows one to examine the relationship between two or more variables of interest. A simple linear regression model has only one dependent variable while a multiple linear regression model has two or more independent variables hence it gives a better prediction of the dependent variable. Both linear regression models will be adopted.

Results

Simple Linear Regression Model

Model fit



This shows that the price of property in King County WA is directly proportional to size of the property(sqft_living)



Multiple Regression Modeling

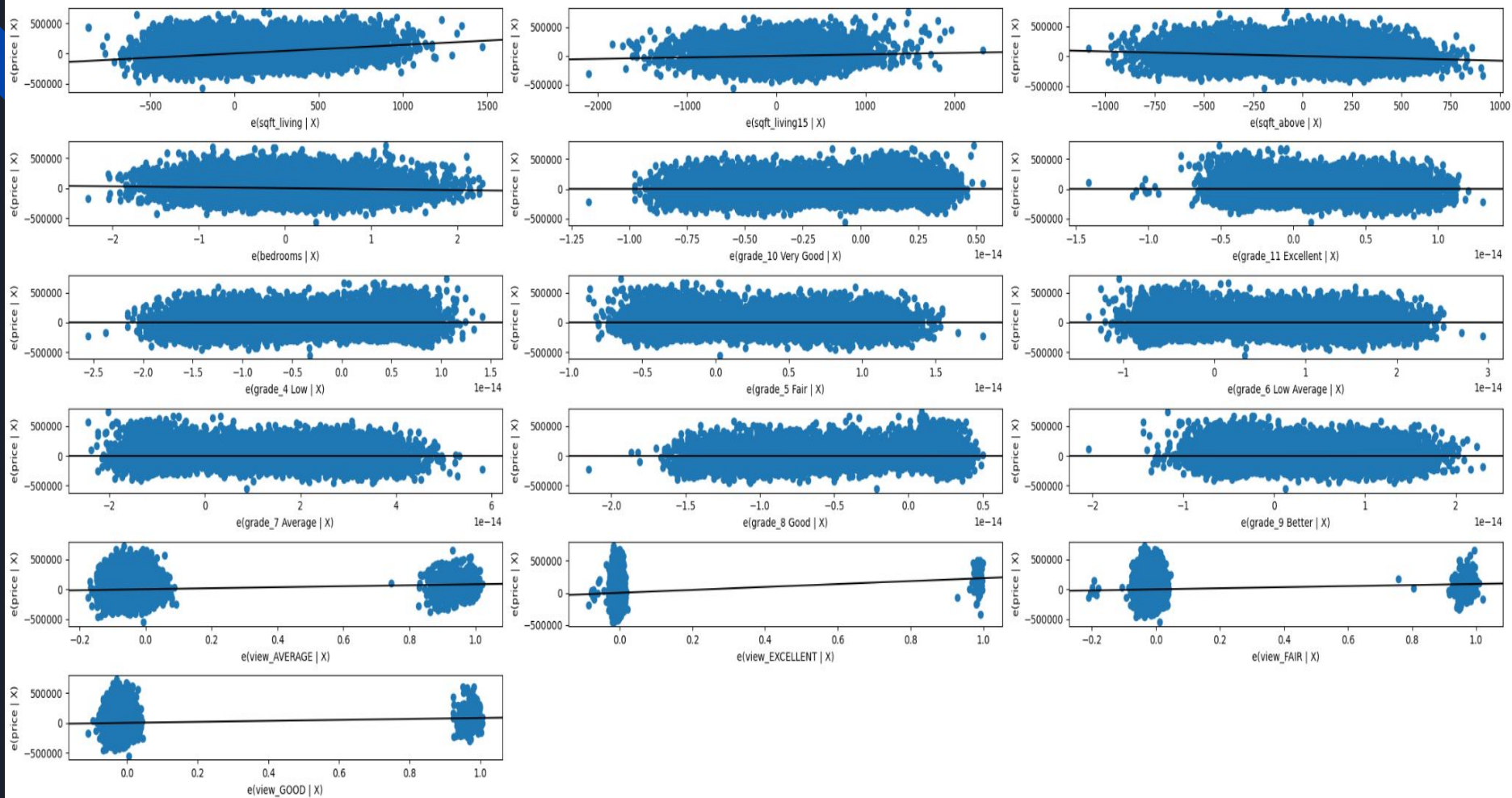
Using variables from the dataset provided,


I ran Multiple Regression Modeling Using Price of the property as the target variable

Against:

- sqft_living
- view
- grade
- sqft_living15
- sqft_above
- Bedrooms

Partial Regression Plot





The partial regression plots don't show a clean linear relationship other than sqft_living. Most the plots show cluster of dots while view_EXCELLENT, view_FAIR and view_GOOD have two clusters on the extreme ends.



Conclusions

Square footage, grade and view are the best predictors of a house's price in King County.

Homeowners who are interested in selling their homes at a higher price should focus on expanding square footage of the living space. This is projected to increase the cost of the house by USD 143 for every increment in square footage.



Recommendations

Future analysis should explore the best predictors of the prices of homes outside of King County, any new data used with the model would have to undergo similar preprocessing.