House Price Prediction with Multivariate Linear Regression Model

Maryam Ghaffari

Outline

- Business Understanding
- Methodology
  - Obtain Data
  - Scrub Data
  - Explore Data
  - Model Data
  - Interpret Dodle
- Conclusion

Business Understanding

o Realtor.com is a real estate listing website operated by the News Corp subsidiary Move, Inc.

o They want to offer valuations of houses using machine learning techniques to their customers.

o The company asked to design a model that be used to the real estate to provide recommendations for home buyers or sellers.
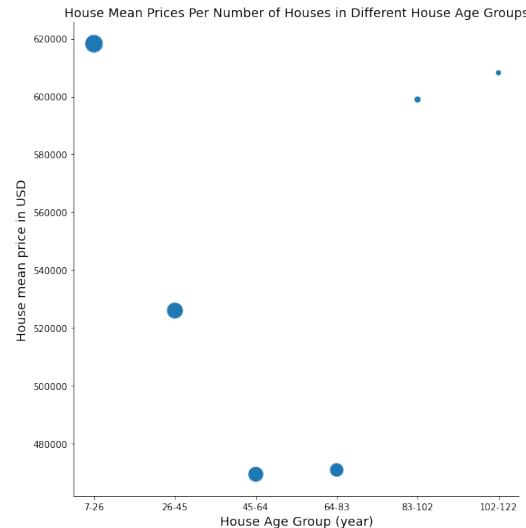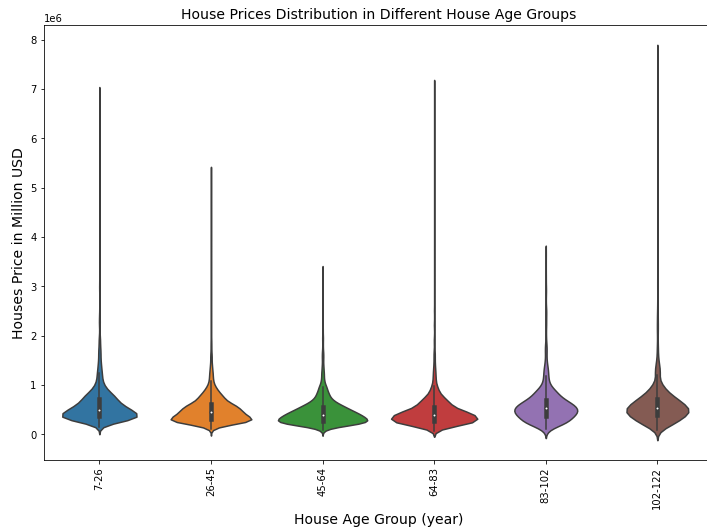
Methadology: Optain & Scrub Data

o   The dataset consisted of historic data of houses sold between May 2014 to May 2015.
o   'kc_house_data.csv' file is source of the data

Name	Description
price	Price the house was sold for
bedrooms	Number of bedrooms in the house
bathrooms	Number of bathrooms in the house
sqft_living	Square footage of the house
sqft_lot	Square footage of the entire lot
floors	Number of floors (levels) in house
waterfront	If a house has a view of a waterfront
condition	A rating of the overall condition of the house
grade	Overall grade given to the housing unit, based on King County grading system
yr_built	Year the house was built

Methodology: Explore Data

Question 1; How does the age of the houses impact the house price?

House Prices Distribution in Different House Age Groups

House Mean Prices Per Number of Houses in Different House Age Groups
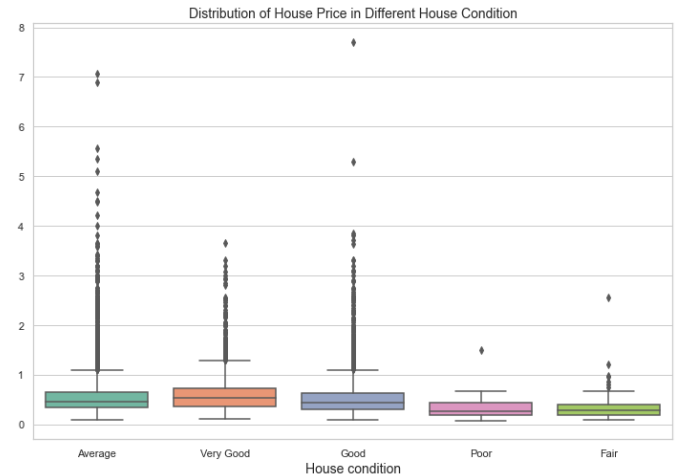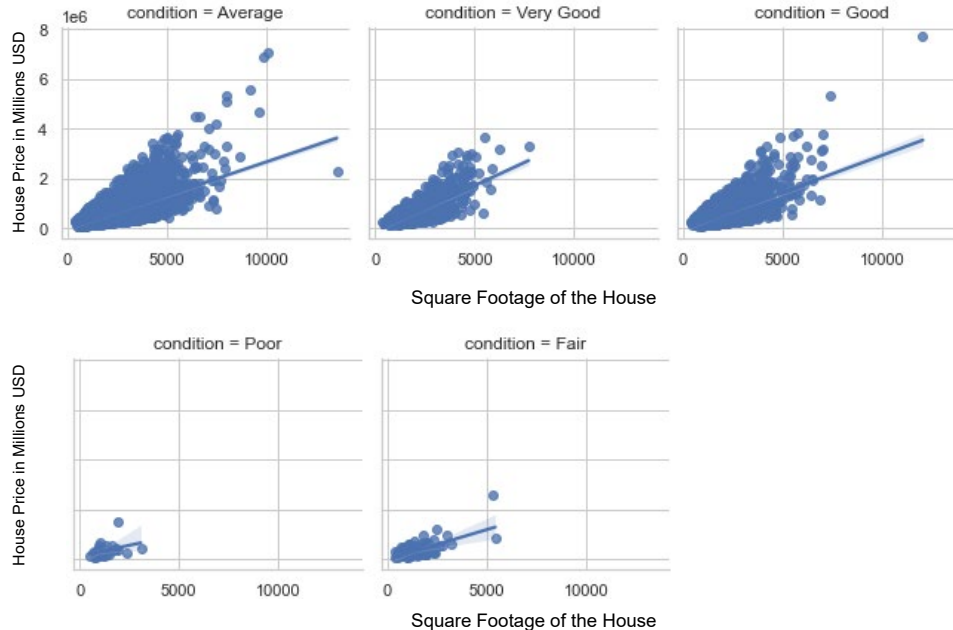
Methadology: Explore Data

Question 2; How does house condition impact the house price?

Relation of House Price with Square Footage of the House Different Based on House Condition

Methadology: Model Data

The first step was to build a "baseline model" from the dataset. Any decisions on how to change features were compared against a baseline model to see if the changes have improved the model or not. Below techniques that were tried to improve the regression model was listed:

- o   Transform categorical variables
- o   Check outliers and remove them
- o   Check the Multicollinearity of features and remove them
- o   Feature scaling and normalization
- o   Check interactions between features

Methadology: Interpret   Model

Adjusted R squared:
o The baseline model Adj R is 0.675. And the final model Adj R is 0.691.
o 69.1% of the variations in price y are explained by the features in our model.
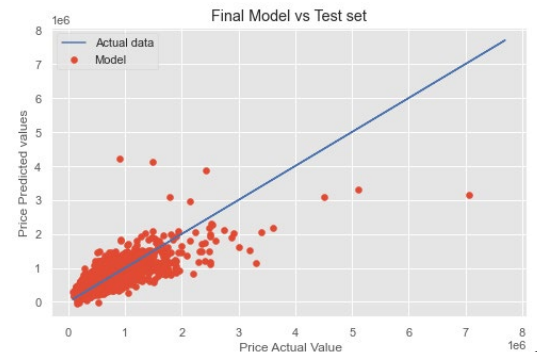
Coefficients:
o House price determiners are waterfront views, floors, and bathrooms.

P-values:
o There is sufficient evidence in the sample to conclude that a correlation exists.

Mean Squared Error (MSE):
o The model underestimated 21.93% of house prices by over 100,000 and overestimated 26.32% of house prices by over 100,000.

Conclusions

o   The linear regression model is not a successful to predict the house price accurately.

o   Other Model should be tried

o   Adding other features that deleted from this dataset might increase the accuracy of this model.

Thank You!

Email: mar.ghaffari@gmail.com
GitHub: https://github.com/MarGhaf/House - Price - Prediction - with - Multivariate - Linear - Regression - Model