Project-2 Housing Predictions

Juston Dea, DSI 0911

Problem Statement

Buying a house can be exciting but also frustrating when it comes to finding the right price for the right house. When purchasing a home there are many factors to measure besides the price itself such as : How many cars can the garage fit? What is the basement square footage? How good is the quality of the kitchen? For this project I will be using the Ames housing data and applying the possible features of an *initial* open house to help home buyers predict the sale price of a house after just one viewing. Most home buyers don’t possess the technical knowledge of data science but not too worry I am here to help!

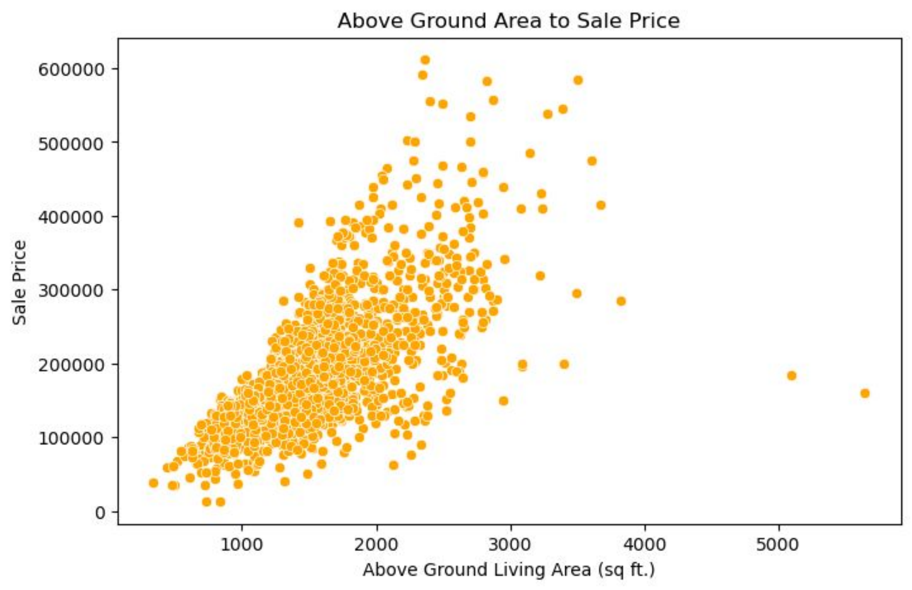
House Features

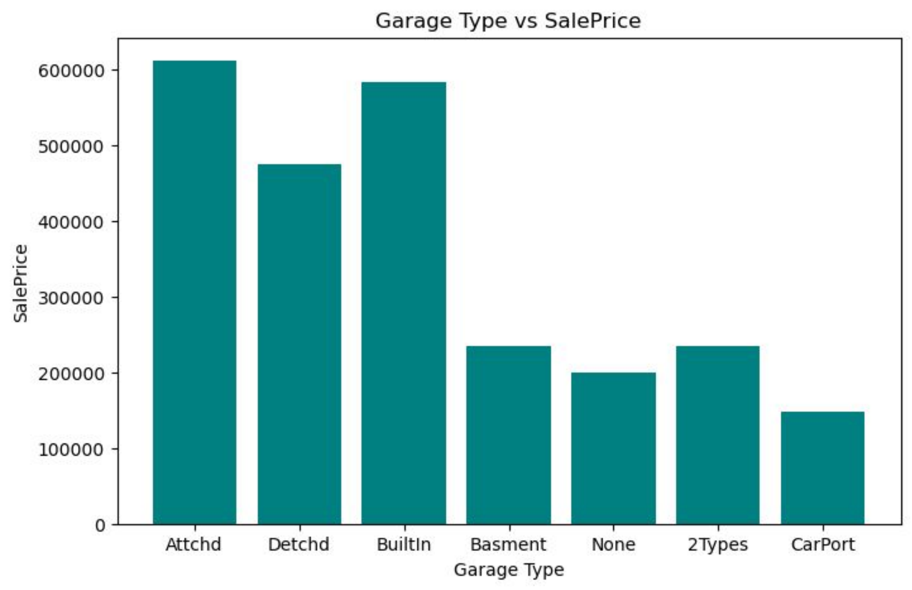
Here are some house features a home buyer may consider on a first viewing:

- Overall quality
- House style
- Total basement square footage
- First floor square footage
- Second floor square footage
- Living space above ground
- Garage square footage
- How many cars can fit in garage
- Garage type (attached, built-in)
- Condition of the garage
- Wall finish of the garage
- General garage quality
- Neighborhood within Ames city limits
- Exterior covering
- Exterior quality
- Central heating system
- Central air conditioning
- Kitchen quality
- Pool quality
- Miscellaneous Features
  - Sheds
  - Elevators
  - Tennis court
  - 2nd garage
  - Anything extra

Why these features?

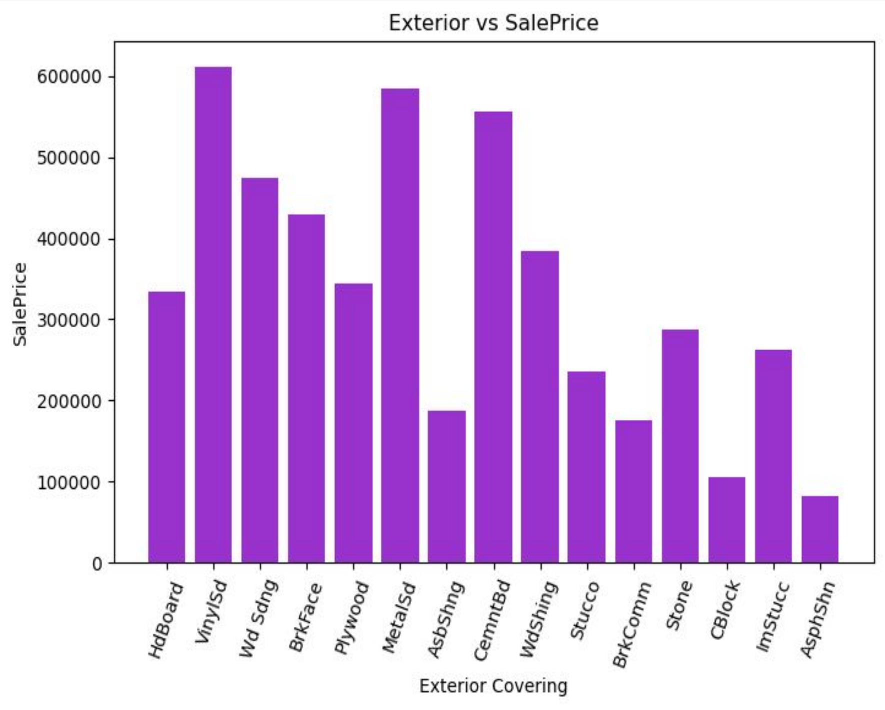
While organizing the data set and performing exploratory data analysis, I went through each features information to diagnose if it would be a good fit for the problem statement. I created several visual plots to look for relevance, data presence, data type and data correlation. For example, below we have the features of “Above Ground Area” and “Garage Type” compared to “Sale Price”

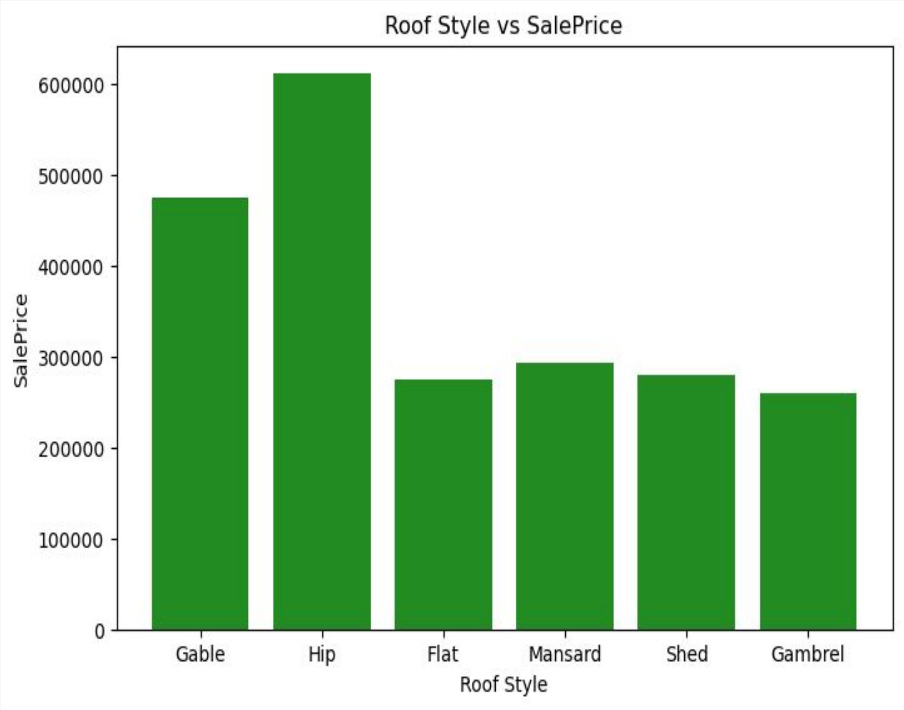




Why these features? (cont.)

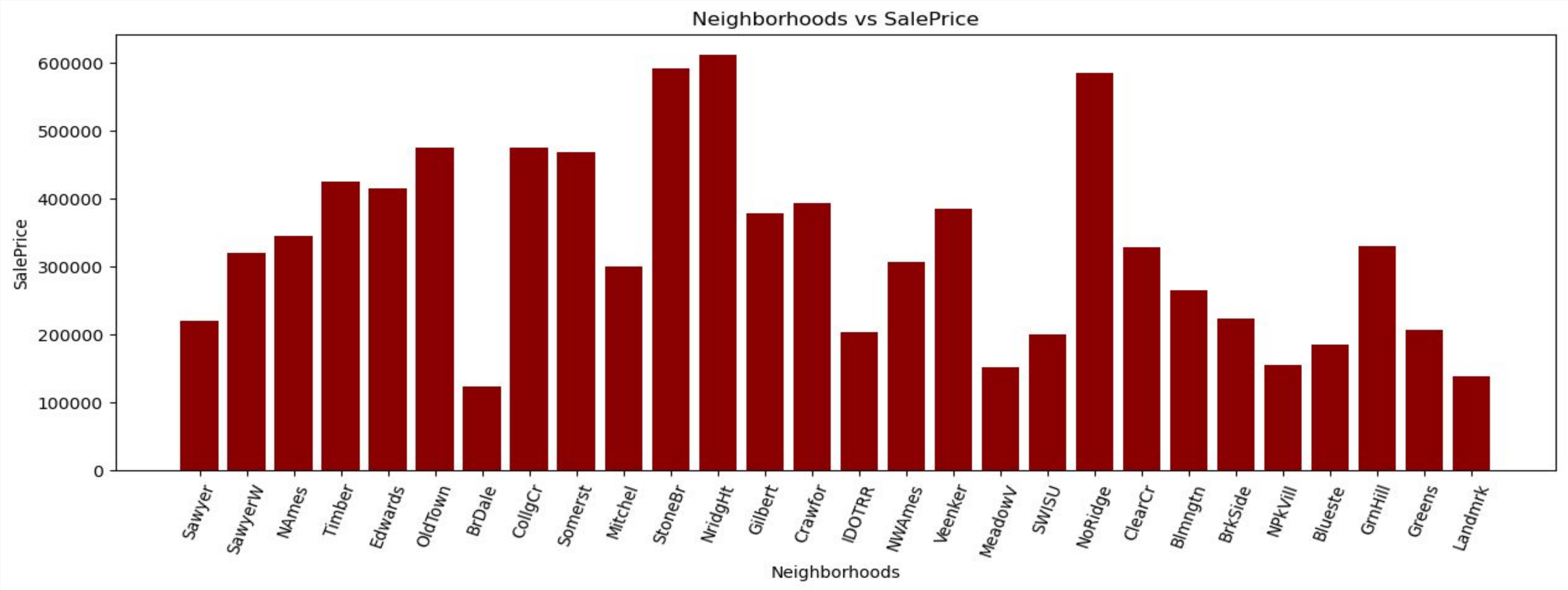
Here we can see the correlation of type of “Exterior Coverings” and “Roof Styles” to “Sale Price”





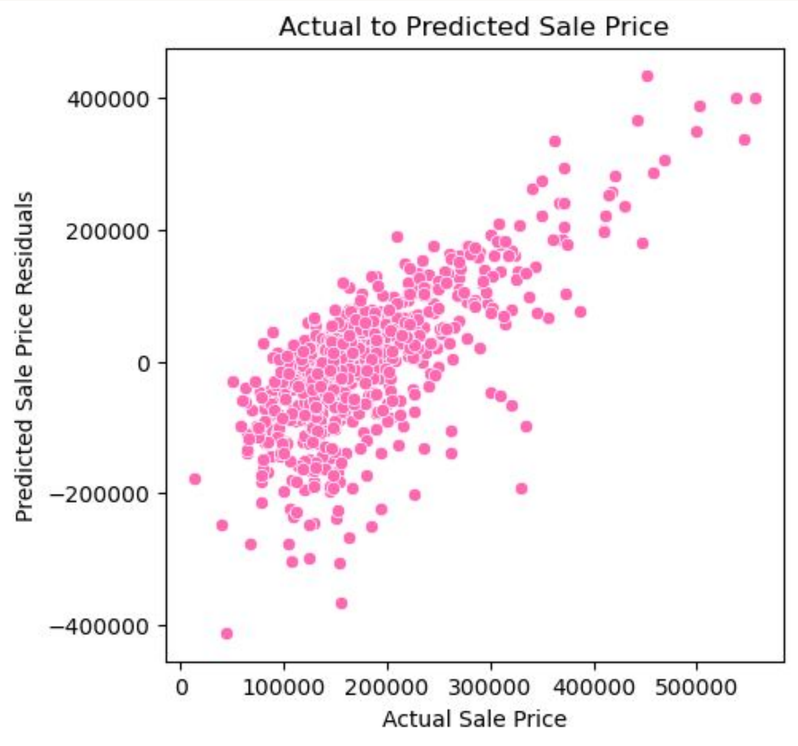
Why these features? (cont.)

Houses in certain “Neighborhoods” in Ames, Iowa compared with their “Sale Price”



Model Process

After organizing, exploring and choosing the features in the data set it was time to create a model to show predicted house prices. My mission is to take the house features I was given, create a linear regression model to the actual sale prices of the houses and then apply that same model to an unseen test data set that has houses without a sale price. I decided to use the **train_test_split method**, followed by the preprocessing technique of **OneHotEncoder** to transform and prepare my chosen data. When preprocessing was complete, it was time to fit my data to a **linear regression model** to see how well my model would perform against unseen data. One metric to gauge how well a model fits is the **R^2 score** which is the percentage of variability in the dependent variable (y) that can be explained by the independent variable (x). R^2 scores range from 0, meaning the model failed, to 1 where the model is a perfect fit. The model I created has an **R^2 score of 85%** which is a solid fit and will properly predict the sale price on the testing data.



Conclusion and Recommendations

It seems my linear regression model fit well and new home buyers can have a good idea on what sale price to expect after visiting a house for the first time. Although my model was strong, was it the best? I can assume there are other house features and modeling methods that could raise the R^2 score and reflect better predictions.

In conclusion I would need to continue to gather more data, preferably data that’s similar to the current housing market, try to add or remove house features, and create other linear regression models using different preprocessing techniques.

Thank you,

Questions???