HOW TO PREDICT PRICES IN AMES, IOWA

By Shanell Jones

Data Scientist

PROBLEM/WHY WE ARE HERE

- Problem that we are here to solve
  - What variables can we use to predict housing prices in Ames, Iowa
- Agenda
  - How I cleaned data
  - How I chose features
  - How I chose to add additional variables
  - How I chose interaction variables
  - How I chose my final model
- Data Sources
  - Ames Iowa Sales Data for 2014

INTRODUCTION TO MODELING SELECTION

- Worst Scoring model
  - Based on the mean of the training set sale price sale price
- Slightly Better Scoring Model
  - Based on one categorical variable
- Slight Improvement on Models
  - Inclusion of more independent variables based on numeric categorical variables to give context to other
- Good Model
  - Combination of variables to change the dummy variables type of the variable
- Great Model
  - Based on the dummy and interaction variables to change orientation of other variables
- Best Model
  - Change to the y that allows for a log transformation
- Variable Types
  - Categorical Variables-Ratings, Descriptions
  - Numeric Variables-Numeric Space, Number of Rooms

HOW WE CLEANED DATA

- Null Values
  - Null values were originally assigned to a blank value
  - This caused a problem due to lack of measurement
  - These were zeroed out die to limited information throughout the category
- Incorrect data types
  - Numerous columns that were numeric registered as object
  - Functions were used to change these to strings
- Outliers
  - Due low data these were kept in for the essence of time
- Missing Columns
  - Columns were not present in the test data set that were present in training set
  - These were added as Nan values

INSIGHTS AND HOW WE CHOSE VARIABLE

- How I found correlated values
  - Used a combination of a correlation matrix and a heat map
  - Chose values with strong correlation
  - Later used this information to combine like
- How I engineered my first feature
  - Choose to Add context with a ratings matrix
  - Used this to add context to actual features
- What I found when I did this
  - There is more explanatory power in the quality of how a house is built and or maintained then the presence of additional feature
    - Example is fireplace quality had a .2 and kitchen quality had a .4 while the number present was insignificant

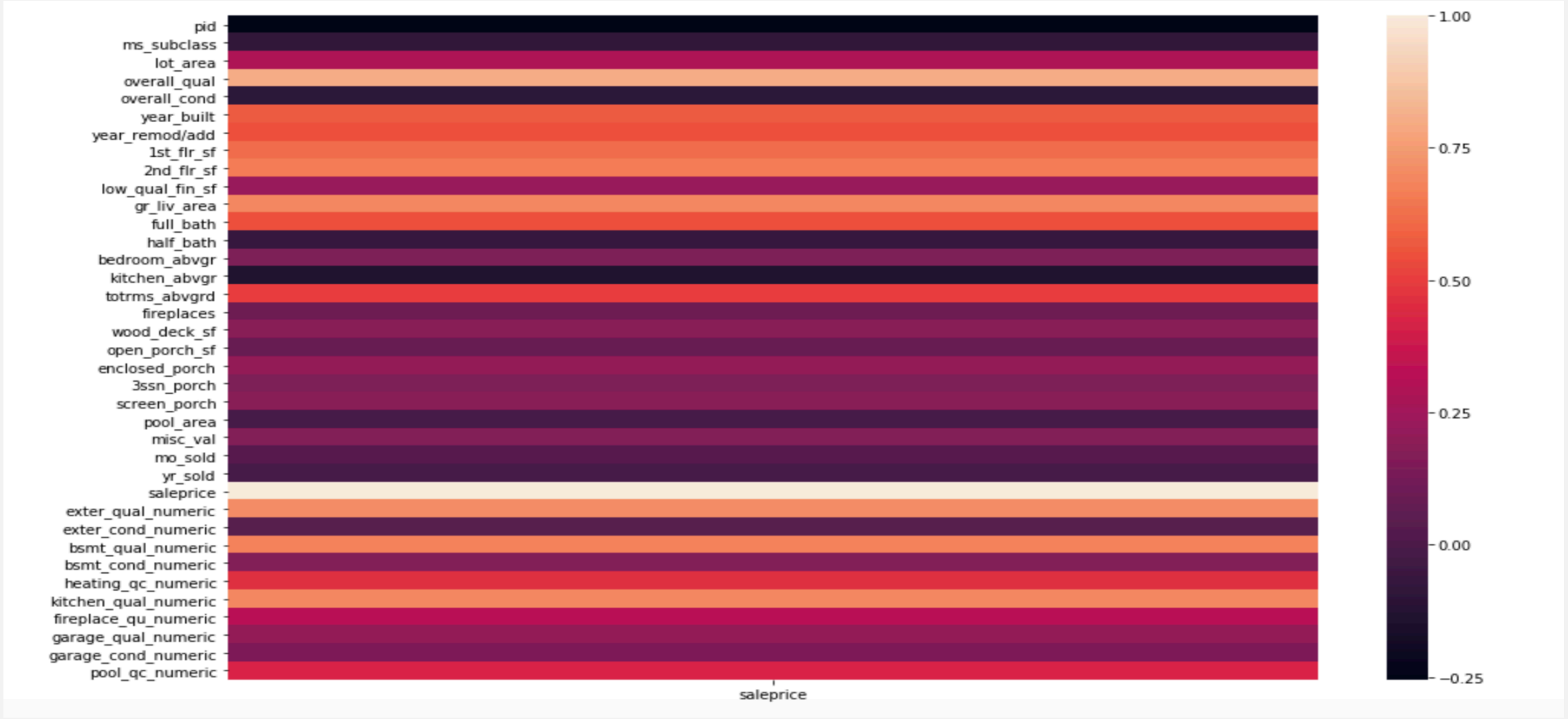
RATINGS MATRIX

Rating Written Info	Numeric Conversion
NA/ Not Included	0
Po/ Poor Quality	1
Fa/ Fair Quality	2
Ta/ Typical or Normal Quality	3
Gd/ Good Quality	4
Ex/ Excellent Quality	5

WHAT WE FOUND WITH OUR FIRST MODEL

- First model info
  - Linear Regression based on beds, baths, sqft, misc values like porches and numeric ratings
- How variables were chosen
  - Variables were chosen via there correlation to the list price
  - Chosen for a strong positive or negative correlation above the absolute value of 0.20
- Findings
  - Certain variables were shown

HEAT MAP USED TO CHOOSE VARIABLES IN FIRST TWO MODELS



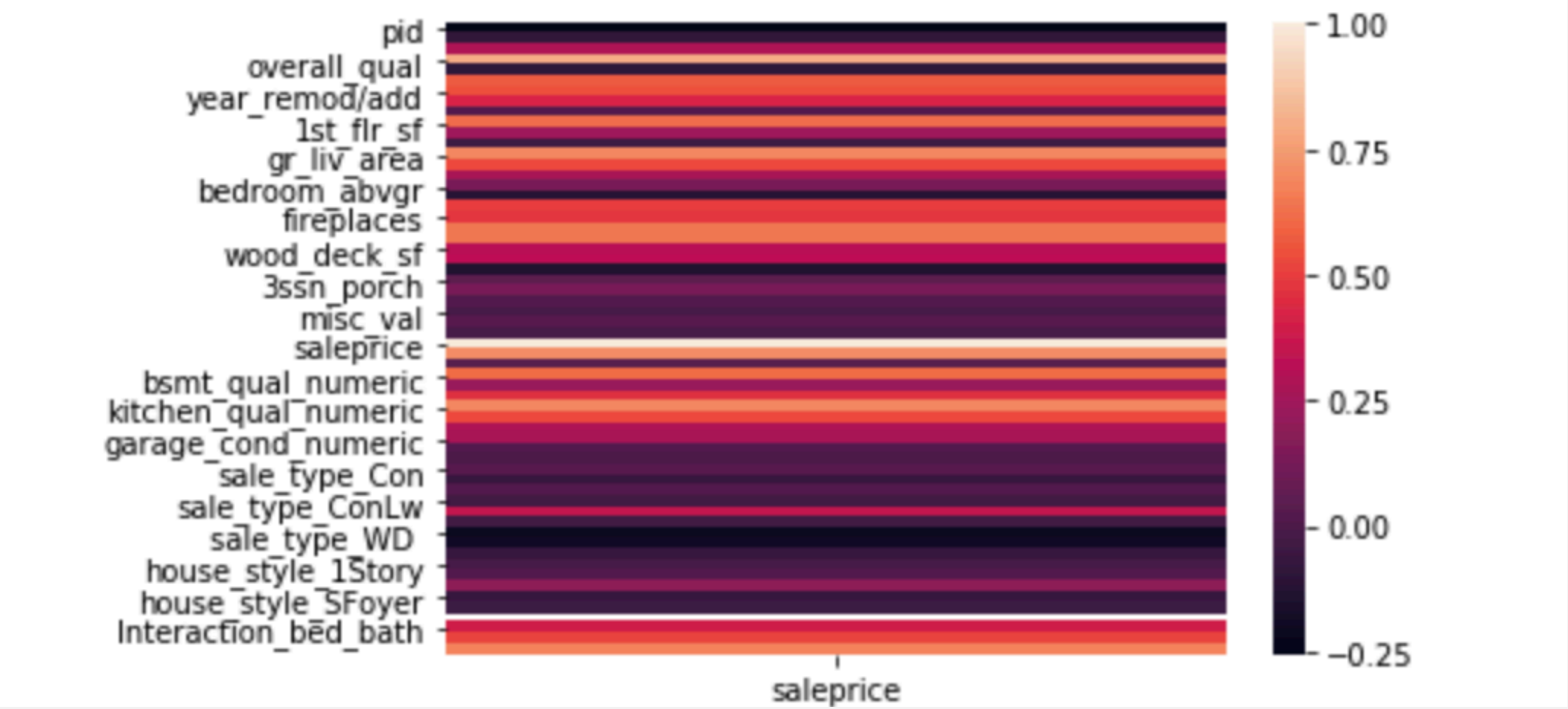
WHAT WE FOUND WITH OUR SECOND MODEL

- How features have been dummied
  - Features attempted were attempted
    - Neighborhood, basement type, zoning, building type
  - Features that were ultimately chosen
    - House type, Sale type
- Which Columns were drop for reference columns
  - Both variables were dropped in comparison to how little values there were
  - This were dropped due to the lowest number of occurrence within the dataset.
- Which features were found to be the most important
  - These were combined into the model to show how the presence of house sale and house
  - Largest impact is shown when looking at New constructed building sales and two story homes

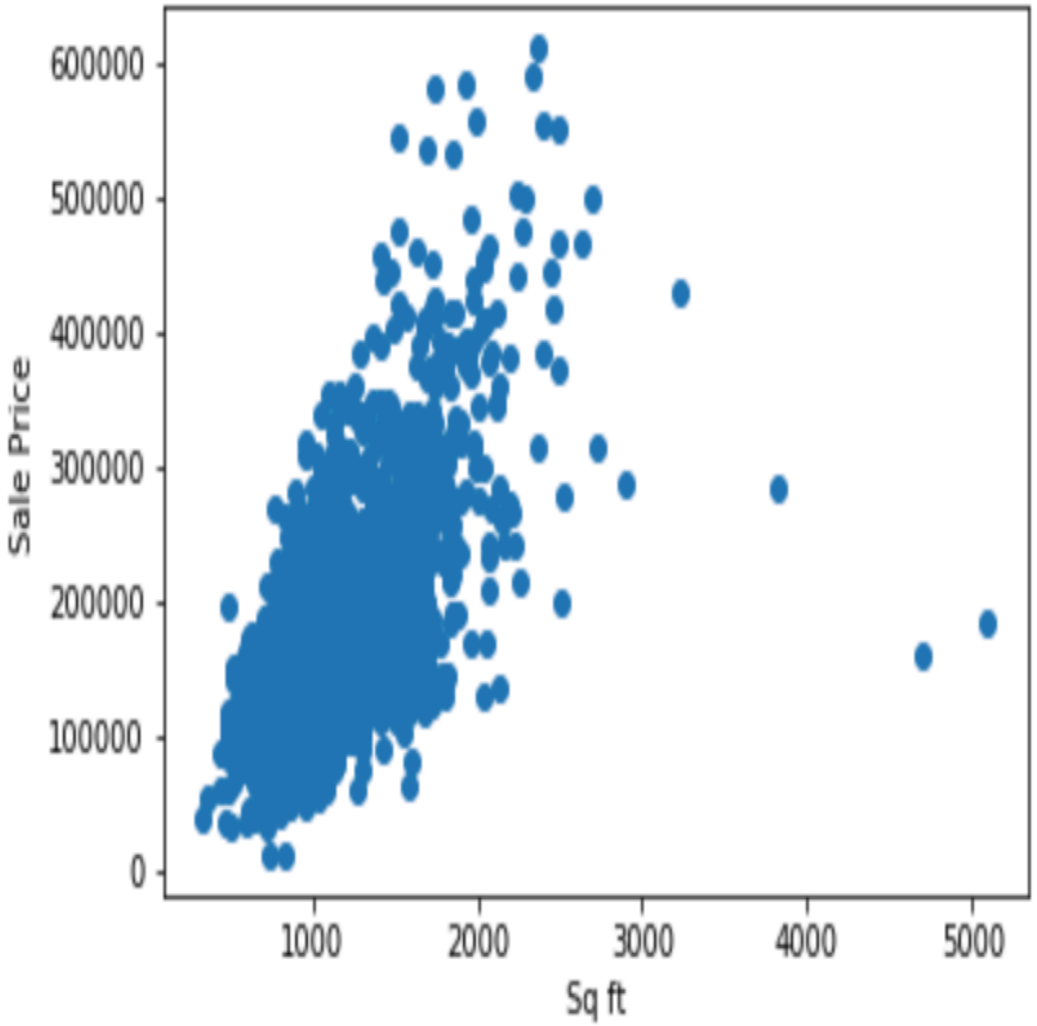
WHAT WE FOUND WITH OUR THIRD MODEL

- How Features were created
- What Transformations were done on the variables
  - Both variables were dropped in comparison to how little values there were
- How did this Change Variable behaviors
  - These numeric variables to be affected by categorical
  - Example is to allow bed, bathrooms, house ratings to affect sq ft
  - Variables showed a significant impact on correlation when combined together
- How did this affect the model
  - Found to have greater explanatory power in order to explain most pricing
  - Much less homodestacity than originally inferred

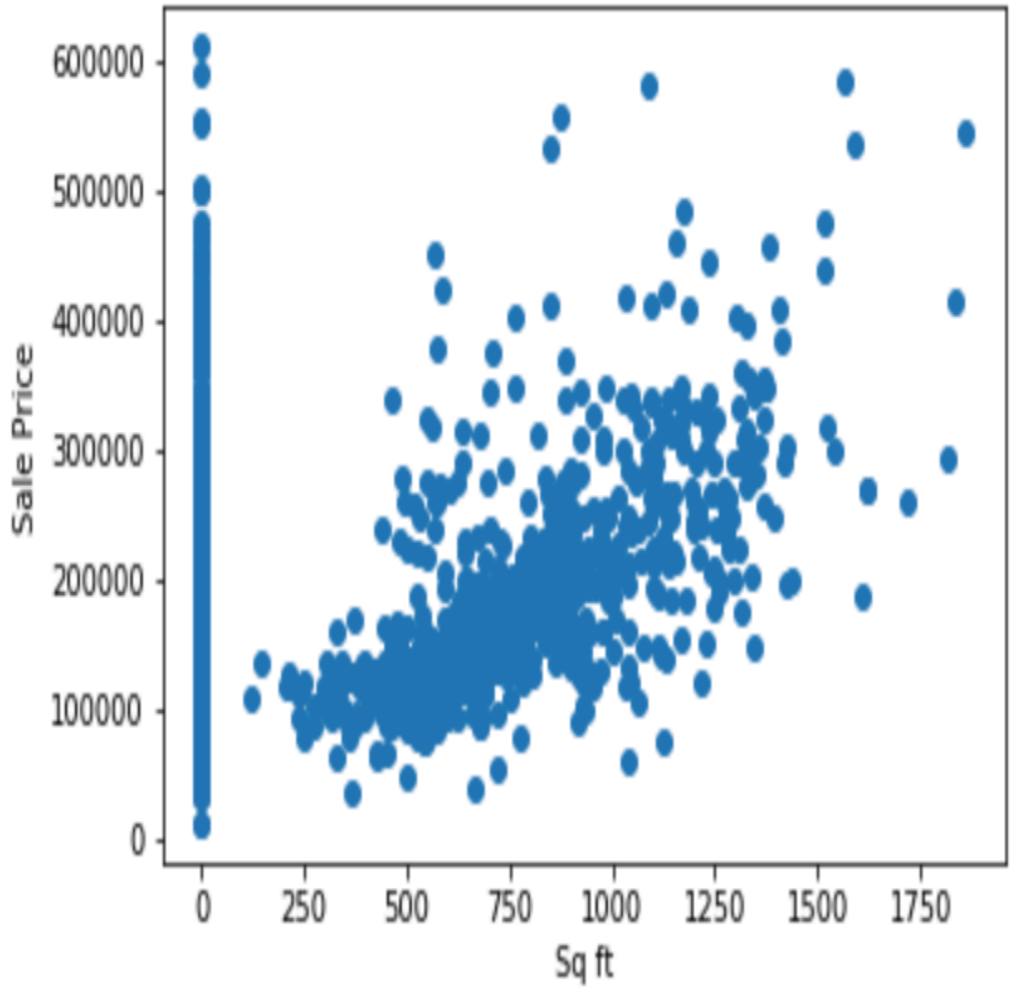
HEAT MAP USED TO CHOOSE VARIABLES IN LAST TWO MODELS



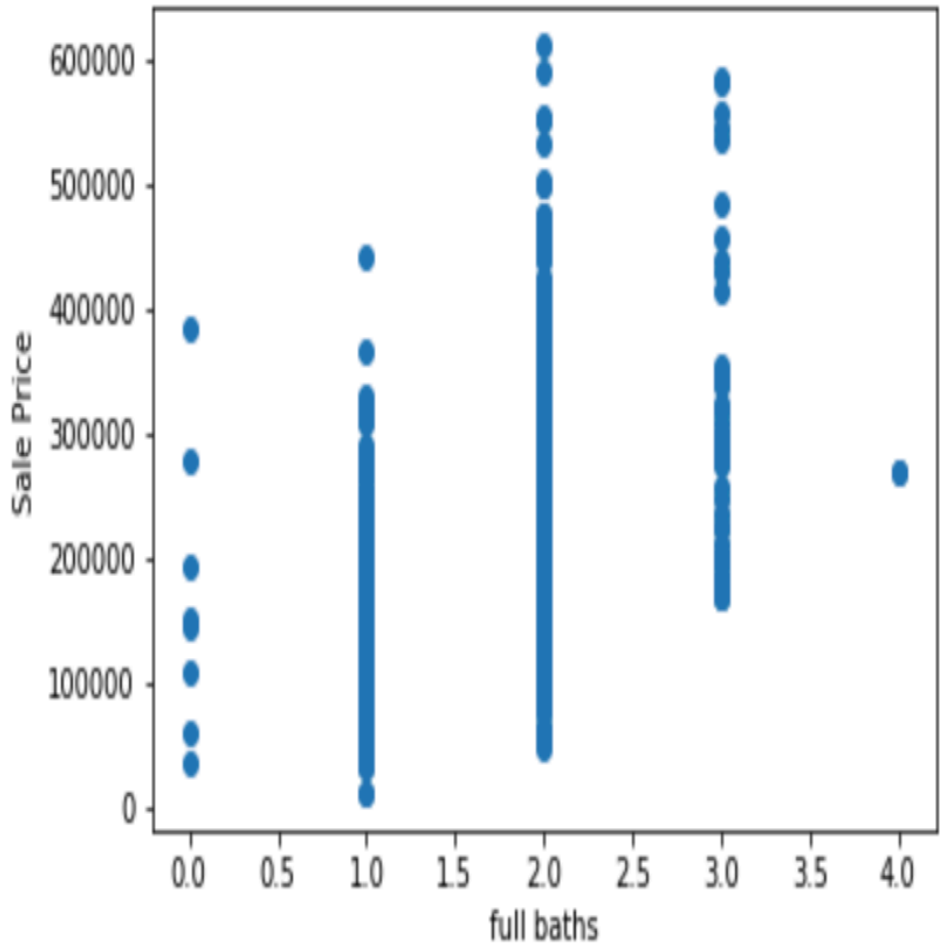
1ST FLOOR SCATTER PLOT



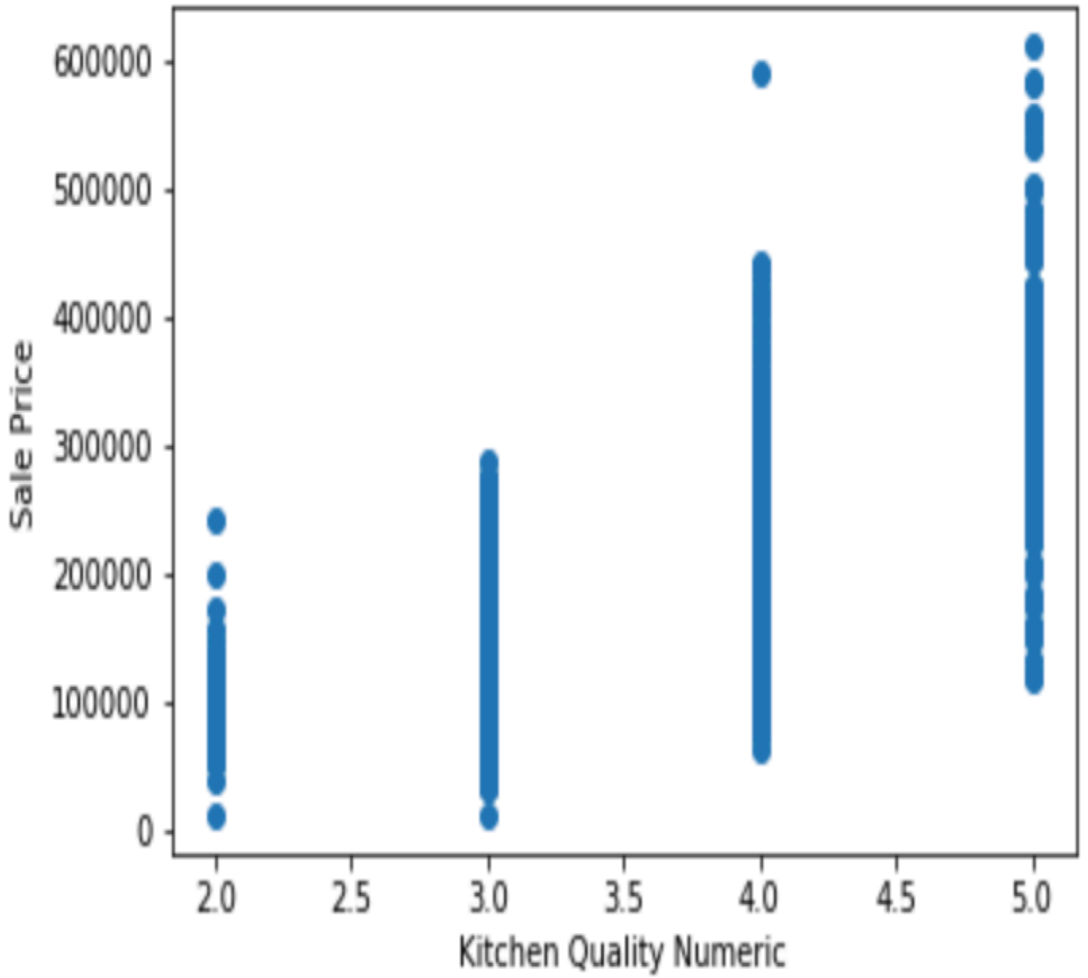
2ND FLOOR SCATTER PLOT



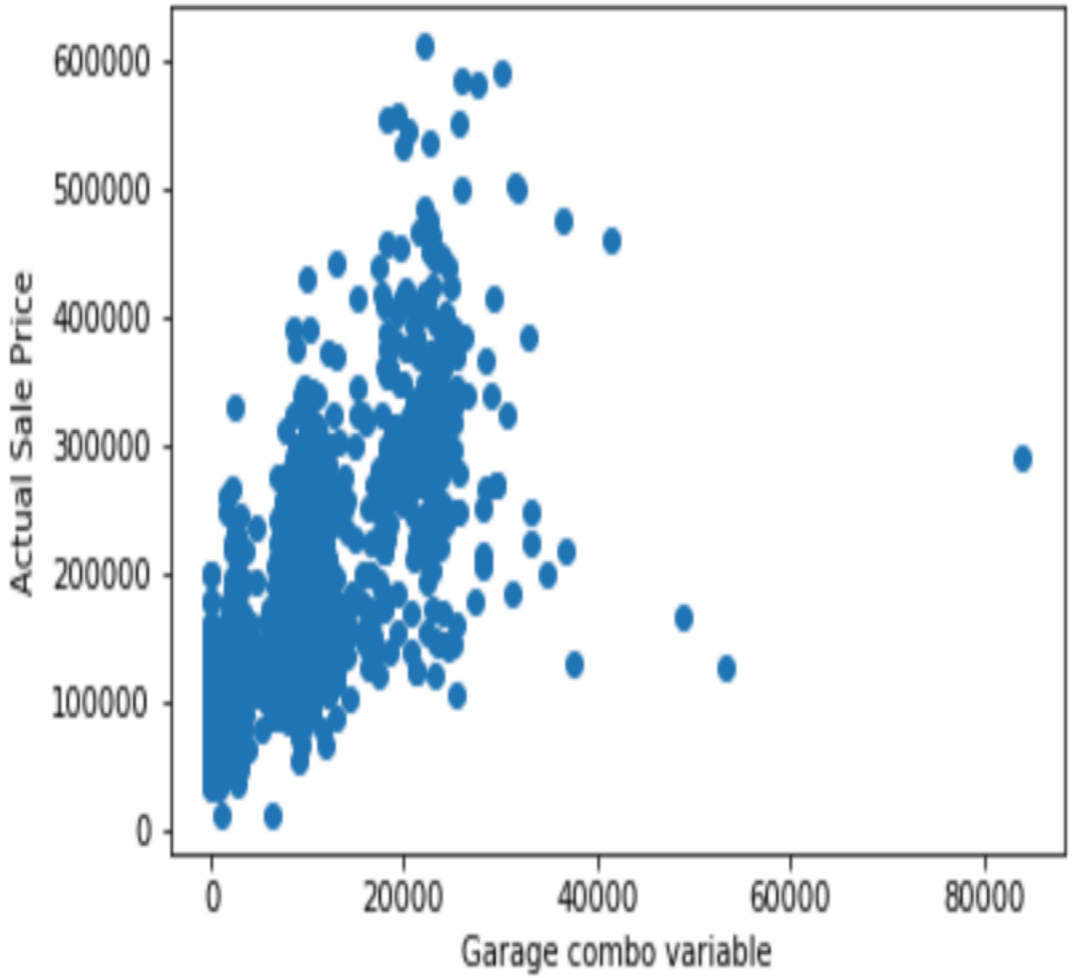
NUMBER OF FULL BATHROOMS IN HOUSE



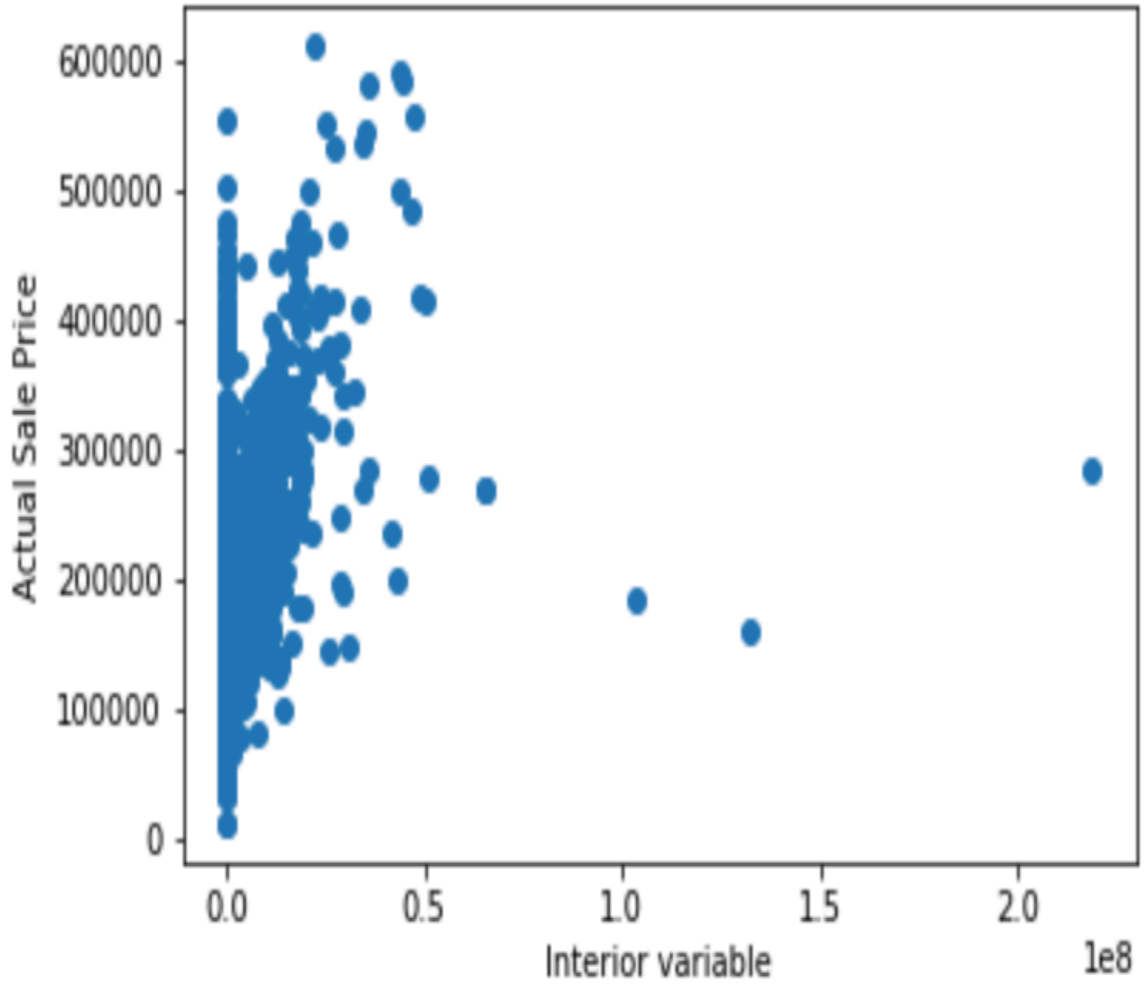
KITCHEN RATING



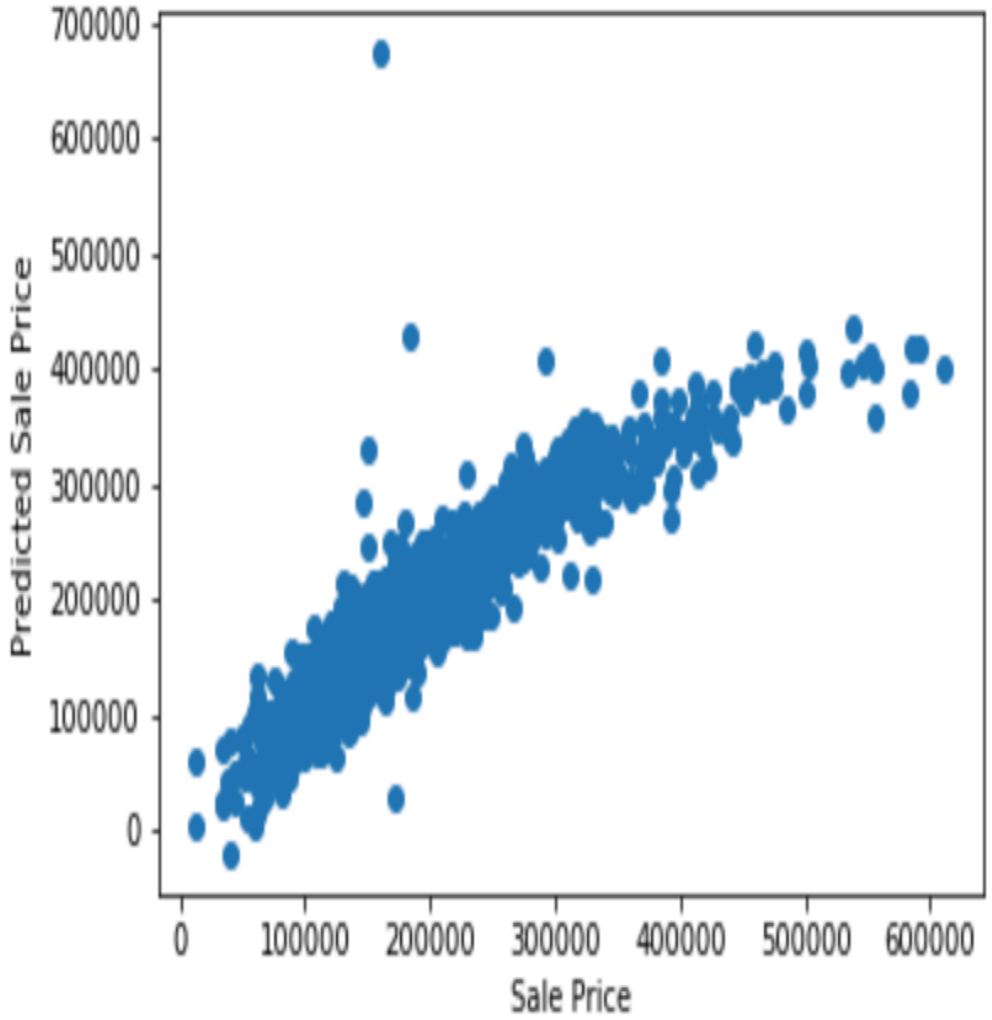
GARAGE INTERACTION



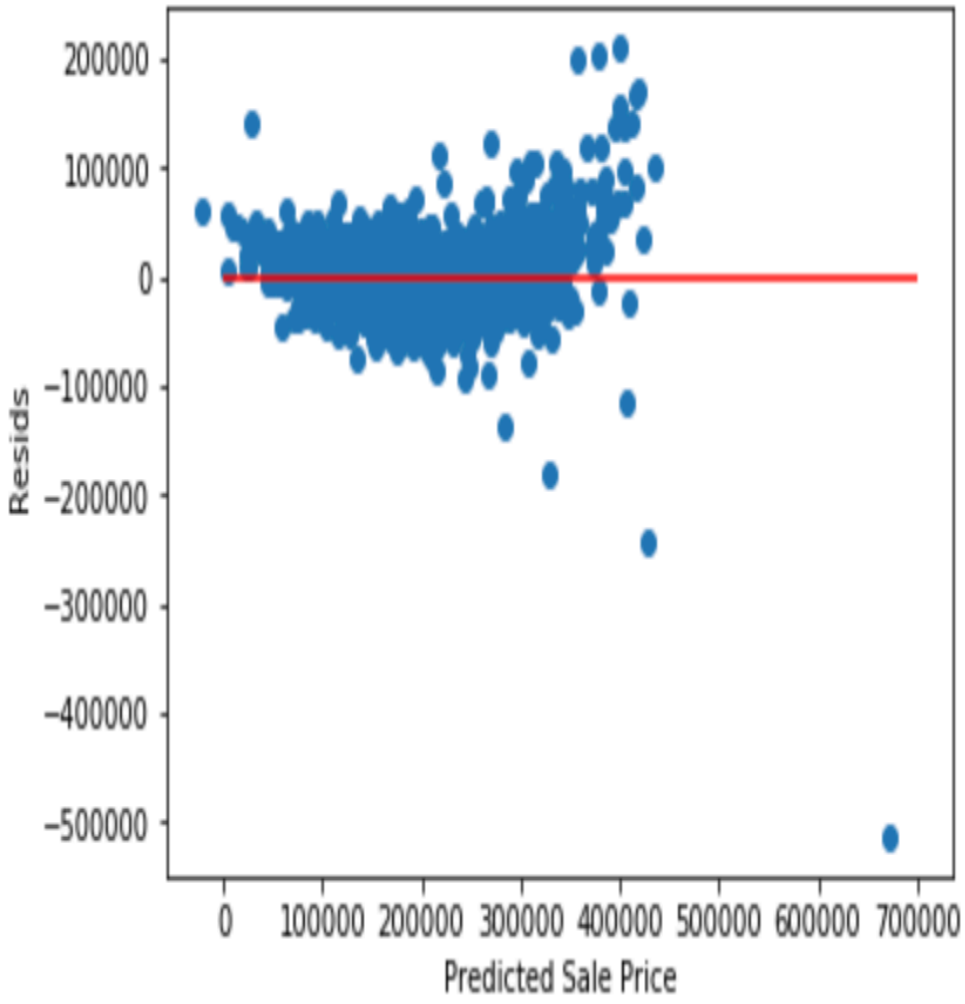
INTERIOR VALUES



LINEAR SALE PRICE



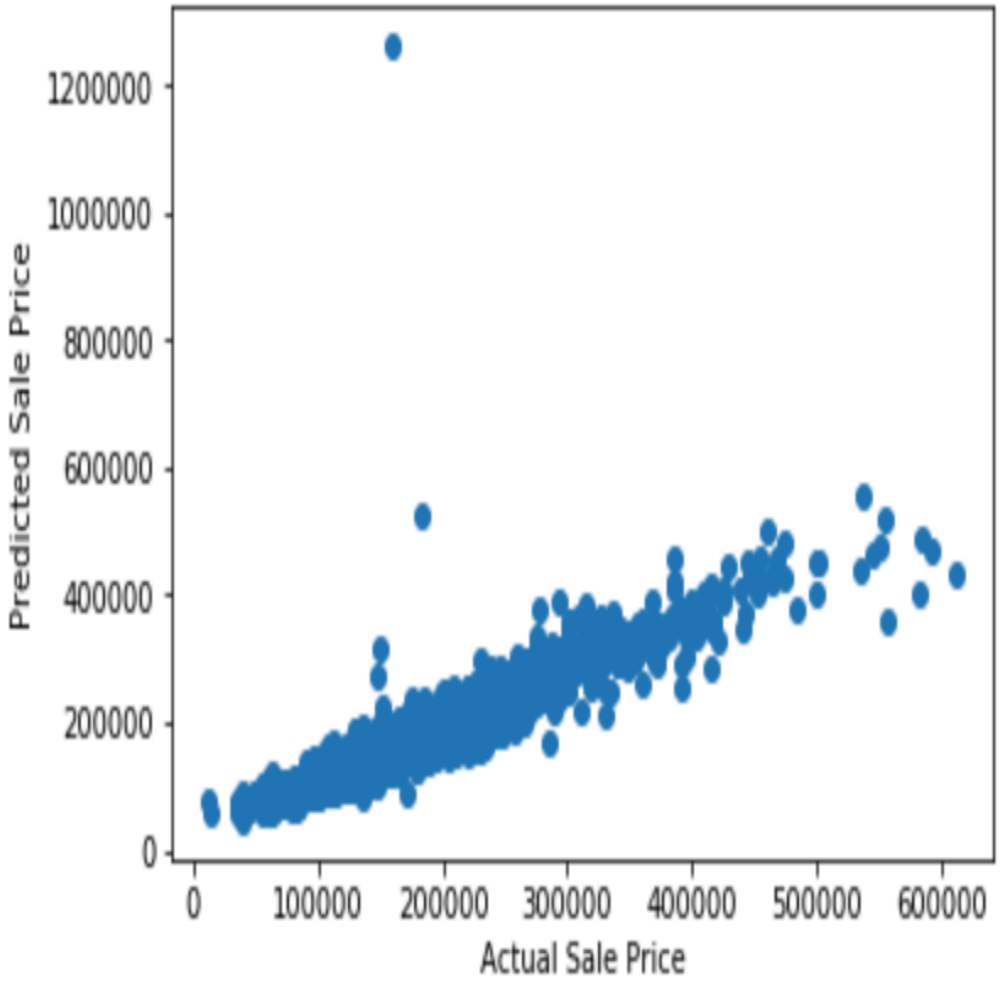
LINEAR RESIDUAL COMPARISON



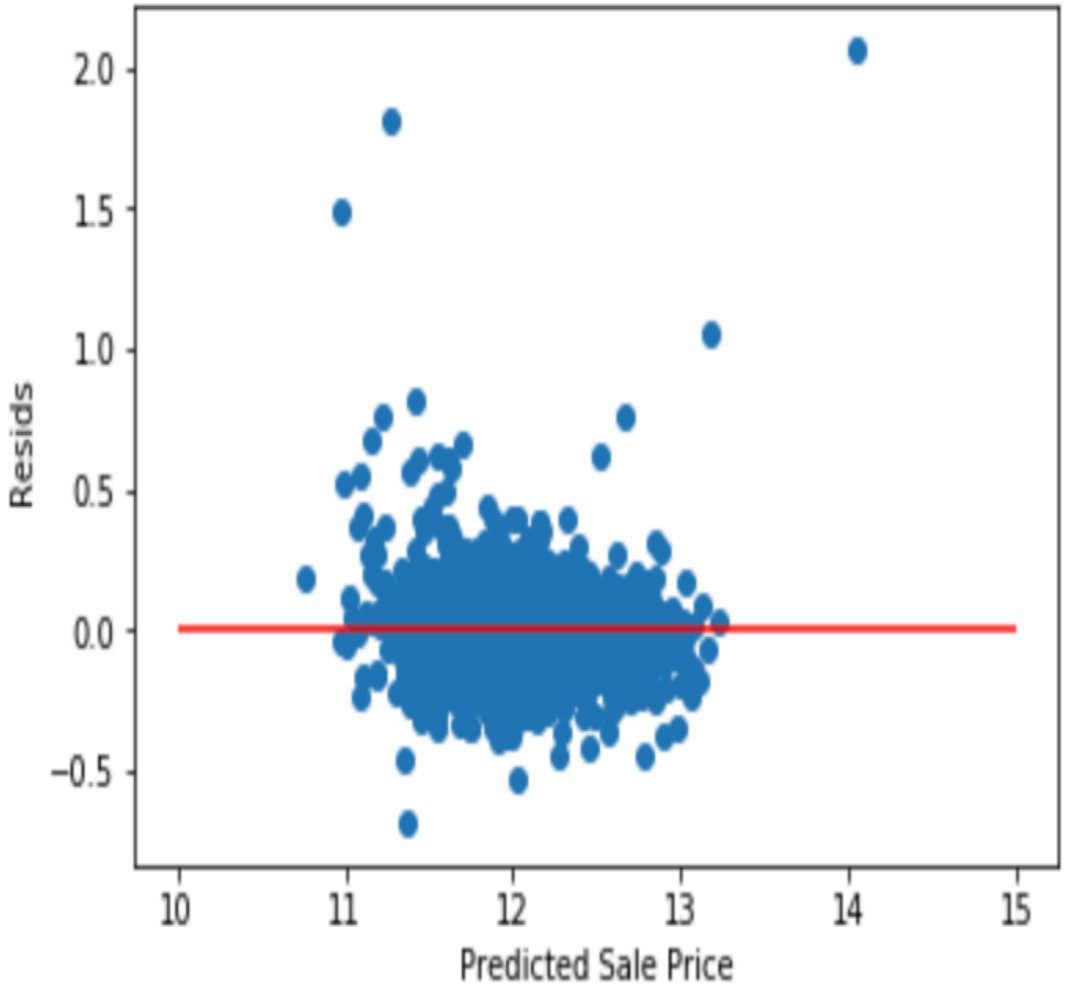
WHAT WE FOUND WITH OUR FORTH AND FINAL MODEL

- Changes were to the model
  - Model appears to have a trouble predicting higher sales values
  - Model appears to have a y=(x)^1/2 behavior
  - Fits a more natural log behavior
- Evaluations that were made to the scaling data
  - Y is transformed into a natural log to ensure behavior of the model
  - Model was rescored with both in mind
- Final submissions were based on the natural log
- Model behavior changed to a more traditional linear distribution
- Behavior between error and predicted values shows understandable homoscedastic behavior
- Final model is scored with a r^2 of 0.87 to 0.88
- When Lassoed r2 was within 5% of that score showing that the model has a good tradeoff between bias and variance
- When checking coefficients all, but two had value to the model

LINEAR SALE PRICE



LINEAR RESIDUAL COMPARISON



CONCLUSIONS AND

- Conclusions
  - Price is based a log equation of ratings, important room info, quality, neighborhood and house type info
- Further Research and Recommendations
  - Further Research is needed in the following categories:
    - Neighborhood info like historic neighborhood and schools
    - Transactional information about previous sale
    - Town economic information-typically income, age demographic information, etc.