CHOOSING THE BEST PRICED HOMES USING LINEAR REGRESSION MODELING TO MAKE A PROFIT

BY: ALEC HING

OUTLINE

- Business Problem
- The Data
- Methods/ Modeling
- Regression Results
- Conclusion

BUSINESS PROBLEM

- It is sometimes difficult for first time homeowners and beginner investors to choose the right home to flip for a profit
- It can be overwhelming to find the right house among so many, all at different ranges of price
- Deciding what characteristic of a home can be appealing to homeowners is also hard without further research
- Investing into a home is costly and a big risk

DATA

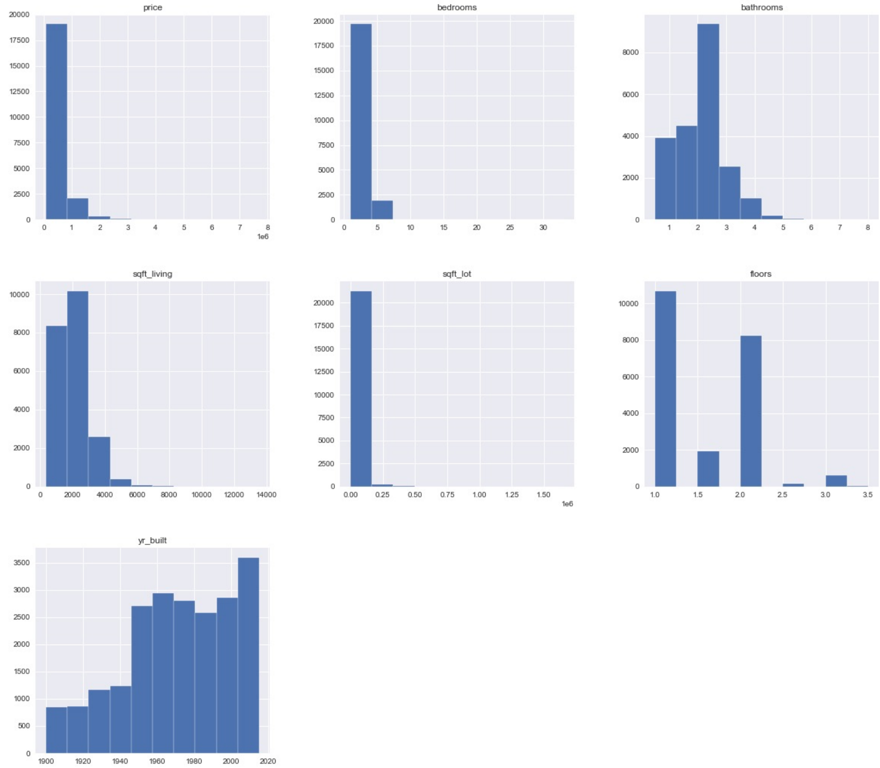
- Data analyzed in this modeling came from the King County Sales dataset
- Variables considered in this analysis include: price, bedrooms, bathrooms, sqft living, sqft lot, floors, waterfront, view, condition, grade, year built, and year renovated
- The dataset consisted of 21,597 rows
- Factors not needed for this analysis were date, ID, latitude, and longitude to name a few

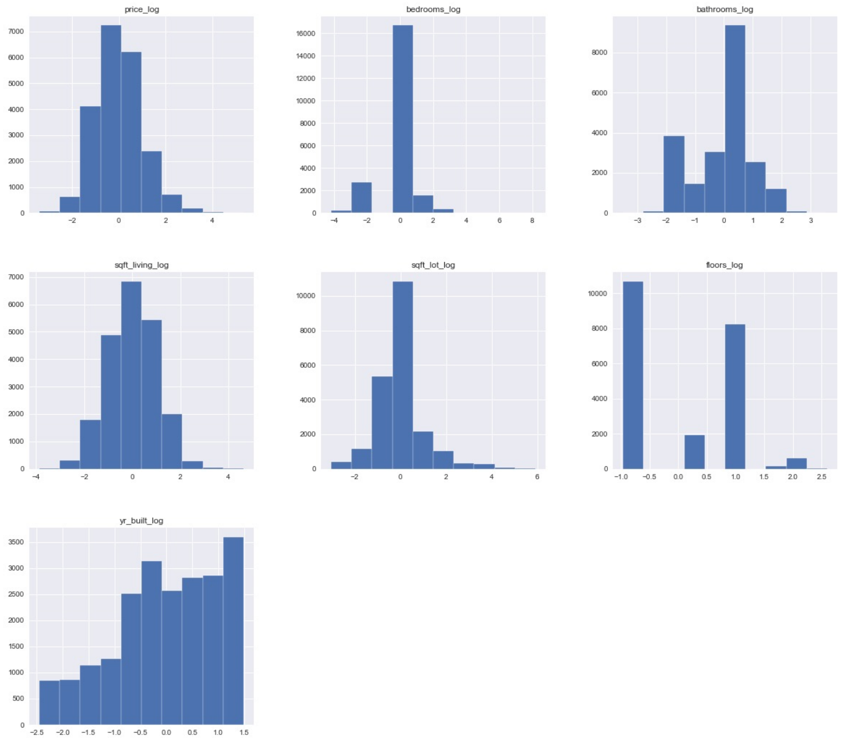
METHODS & MODELING

- Only variable that seemed to directly effect a home's price were chosen
- NaN and 0 values were evaluated to see if they had true purpose to being in the data
- After scrubbing and cleaning data, a baseline model was created with graphs to see linearity and OLS results
- Other graphs created were scatterplots, heatmaps, histograms, and Q-Q Plots
- Manipulations of baseline include: removal of high correlated variables, removal of variables with P-values higher than 0.05, removing outliers outside of 3 STD's, and logging/normalizing data

BASELINE CATEGORICAL HISTOGRAMS

Not much of a normal distribution among these graphs except slightly in ‘sqft_living’ and ‘bathrooms’



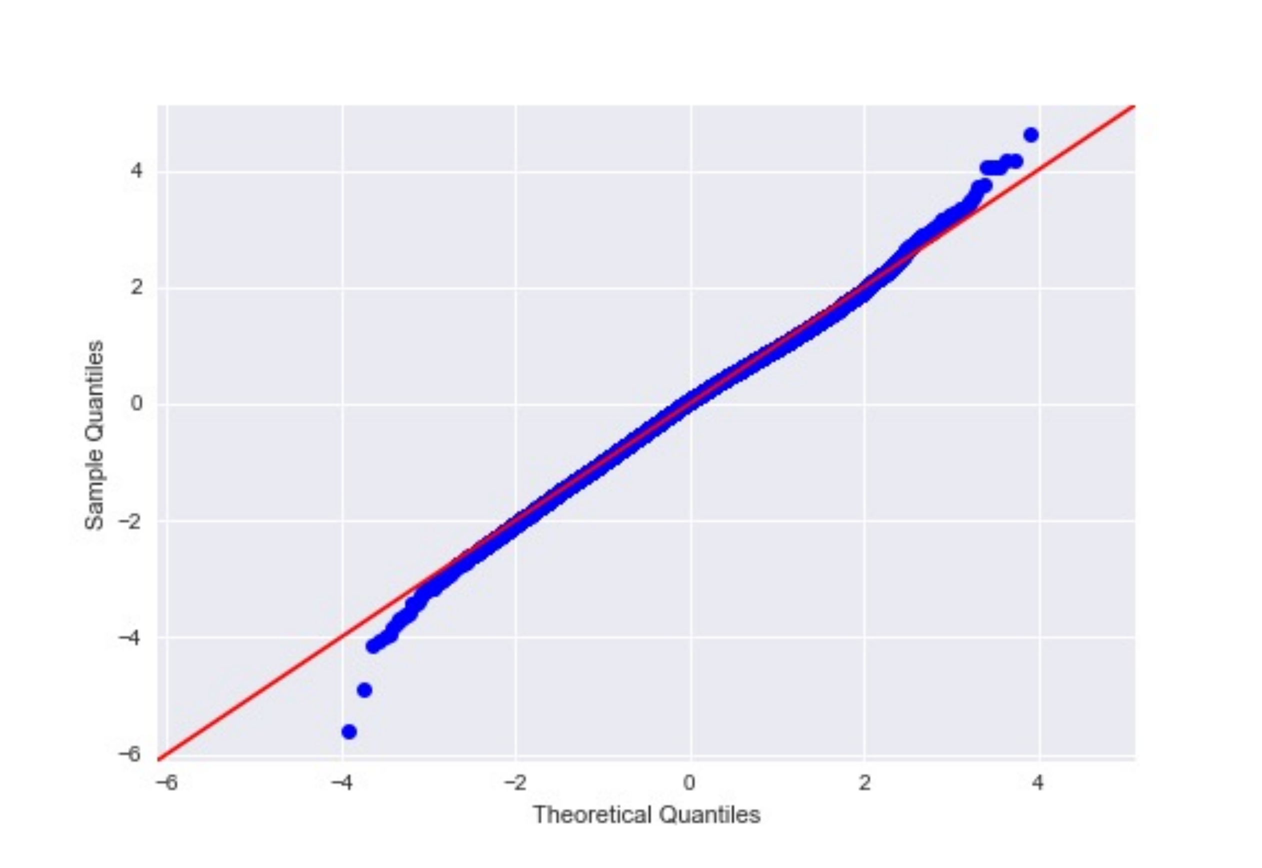


LOGGED & NORMALIZED CATEGORICAL DATA

- More histograms follow a normal distribution now
- Price_log, bathrooms_log, and sqft_living log show the most normal

NORMALIZED Q-Q PLOT

- Among all the other Q-Q plots created the logged and normalized plot was the most linear
- This indicates both sets of quantiles come from the same distribution



CONCLUSIONS

- After performing 4 different manipulations of the baseline dataset, it turns out the baseline data had the best R-squared value of 0.681
- The R-squared represents a statistical measure of fit that indicates how much variance of a dependent variable is explained by the independent variable
- The coefficients chosen from this dataset that would yield the best price for an investor are homes with a grade: low average, grade: fair, grade: poor, and view: none

THANK YOU!

Email: ahing619@gmail.com

GitHub: @ahing

LinkedIn: https://www.linkedin.com/in/alec-hing