Housing Market Predictions for Prospective Sellers

Ryan McArthur, Flatiron School DS 2020

Understanding Our Goal

In this presentation, we will:

- Construct two models - One prioritizing **accuracy**, another prioritizing **interpretability**

- Give context on the housing market through additional exploratory data analysis

King County Dataset

- Collected from 2014 to 2015
- One county, covering 2,307 sq miles
- Seattle is in King County, making it the most populous county in Washington

Predictors:

- ID
- Date
- Bedrooms
- Bathrooms
- Sqft_Living
- Sqft_Lot
- Floors
- Waterfront
- View
- Condition
- Grade
- Sqft_above
- Sqft_basement
- Yr_built
- Yr_renovated
- Zip code
- Latitude
- Longitude
- Sqft_living15
- Sqft_lot15

Data Cleaning and Preparation

Yr_renovated Predictor

- Yr_renovated can be recorded 0.0
- Converted this column to yrs_since_construction column
- Yrs_since_construction gives a better predictor than yr_renovated

Categorical Predictors

The following columns were classified as categorical predictors

- Waterfront
- View
- Condition
- Grade
- Zipcode

Multicollinearity Checks

The following columns display high multicollinearity, and must be dropped

- Yrs_since_construction
- Sqft_above
- Sqft_living15
- Sqft_lot15

Raw Modeling

Our data has been altered as little as possible. We have completed:

- Data Cleaning
- Categorical Classification
- Multicollinearity Checks

We have yet to:

- Remove Outliers
- Transform Data for normality

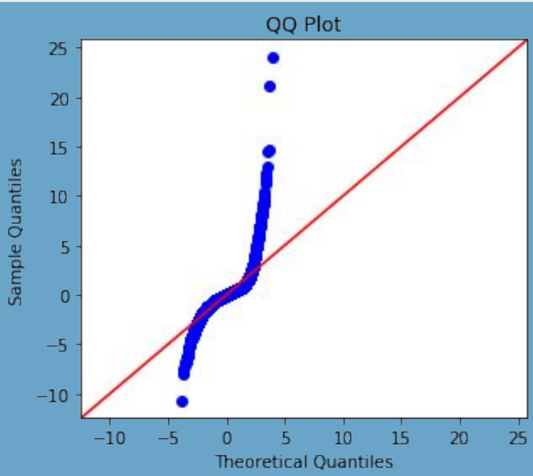
$R^2 = 0.834$

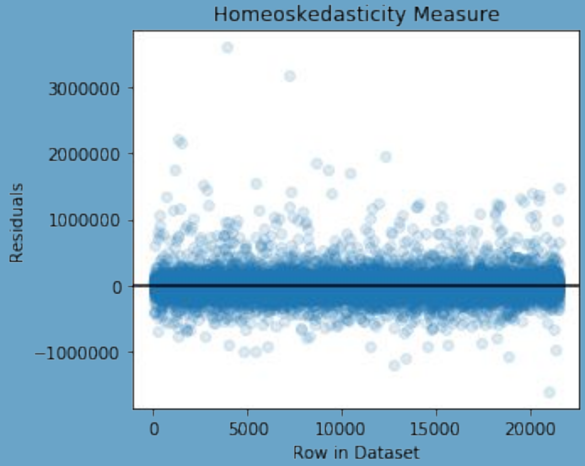
Raw Modeling

- Relatively homoskedastic.

- Residuals are not normally distributed, suggesting our data is not normally distributed.

- P-Value for yrs_since_construction is 0.232. To address this, we will drop yrs_since_construction and bring Yr_built back into our predictors.





Refined Modeling

For our refined model, we attempted to transform our data to match a normal distribution

- Outliers were removed from columns based on distribution Z-scores
    - 1086 rows of data were removed from our dataset, a reduction in sample size of 5%.

- Box-Cox Transformations normalized our predictor columns to approach normality
    - Box-Cox transformations optimize normality applications.

$R^2 = 0.840$

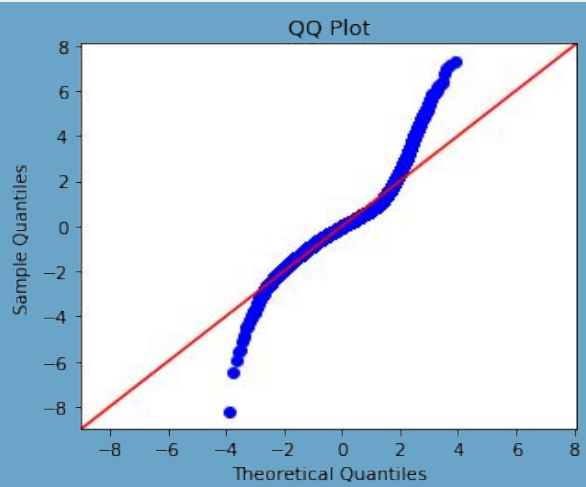
Refined Modeling - Outlier Removal

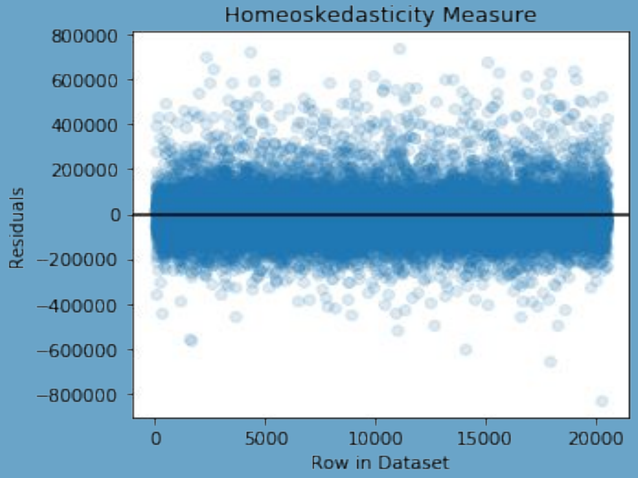
- Homoscedasticity was preserved from earlier mode

  l

- Residuals are approaching normality. Removing outliers vastly improved this aspect of our model.

- P-Values suggest no predictors need to be removed.

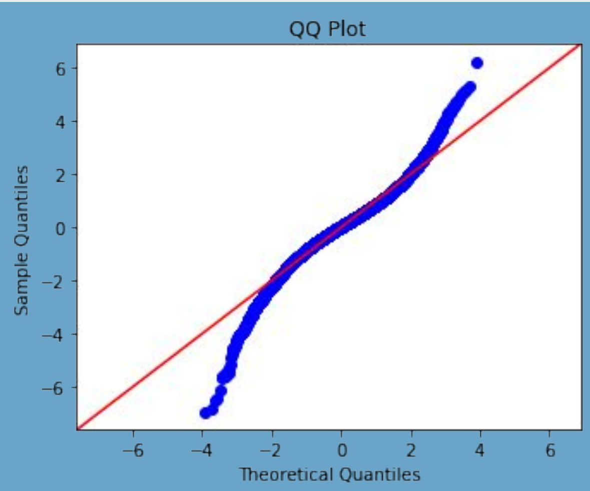


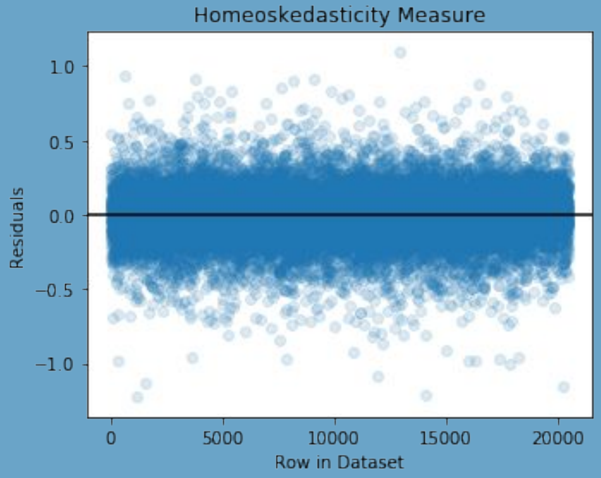


$R^2 = 0.863$

Refined Modeling - Transformed Data

- QQ plot has far improved from our raw model
- R^2 misses less than 15% of data variations.
- A test-train splitting process on our dataset validates this R^2



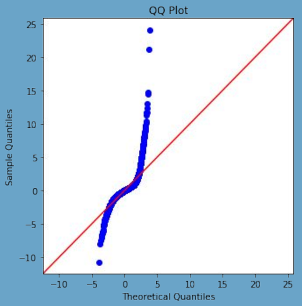


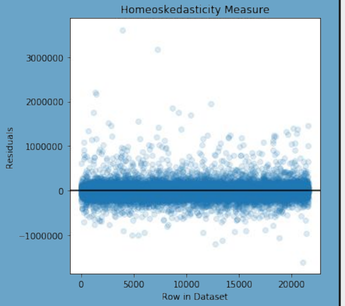
Comparing Models

Raw Model

$R^2 = 0.833$
n = 21,597

- Non-Normal Residuals
- Homoskedastic
- Many Outiers

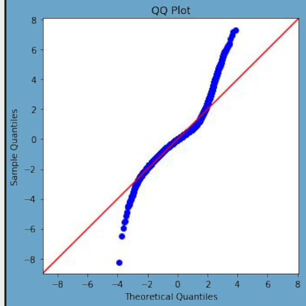


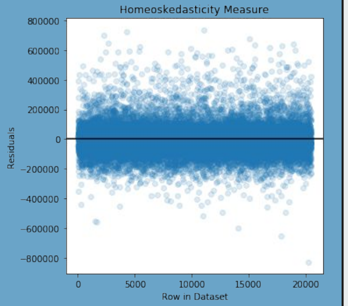


Outlier Removed Model

$R^2 = 0.840$
n = 20, 511

- Somewhat Normal Residuals
- Homoskedastic
- Skewed Residual Distribution

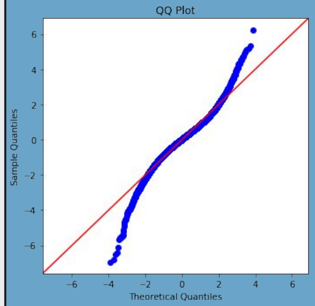


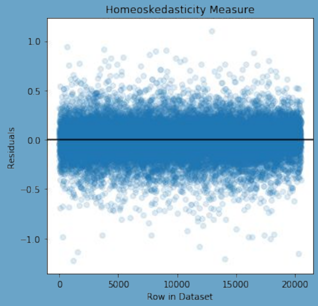


Normal-Transformed Model

$R^2 = 0.863$
n = 20, 511

- Normal Residuals
- High R^2 result
- Strong Testing Result





Modeling Conclusions

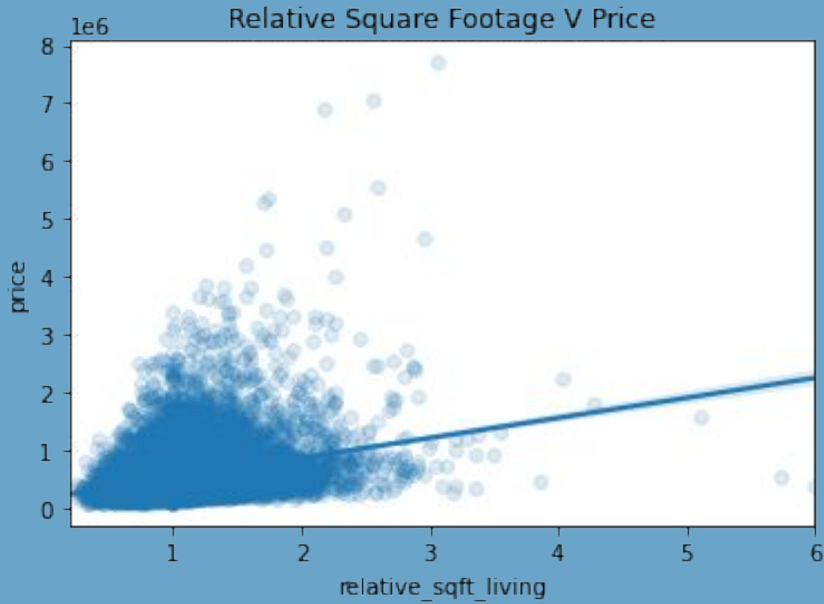
Depending on whether interpretability or accuracy is required, we have provided models that would satisfy either requirement.

- Our refined model, using a Box-Cox transformation to normalize our predictor data, would be preferred by those needing a model with a high probability of accurately predicting home prices.
- Our cleaned model, using only data that had outliers removed, would be preferred by those seeking interpretability and inference from a model. While this model is less accurate than our refined model by 3%, all coefficients are easily interpretable.

Housing Market Irrationality - Anchoring

Anchoring: When consumers rely *too much* on pre-existing information when making purchasing decisions.

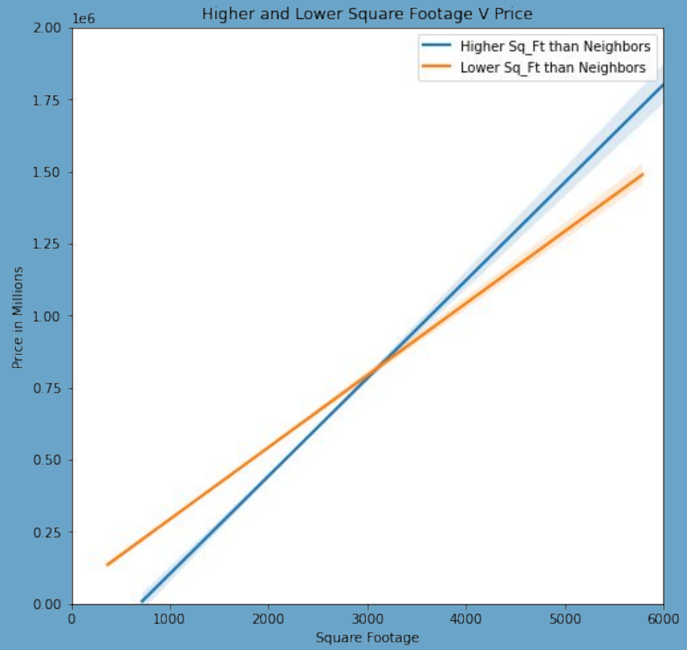
In the housing market, this phenomena materializes through comparisons between square footage in a neighborhood. Neighbors are overly concerned with their relative size.



Housing Market Irrationality - Anchoring

At higher square footages, the difference in size matters more to consumers, since the average price for homes that are larger than their neighbors is so much greater than homes with the same square footage, but are smaller than their neighbors.

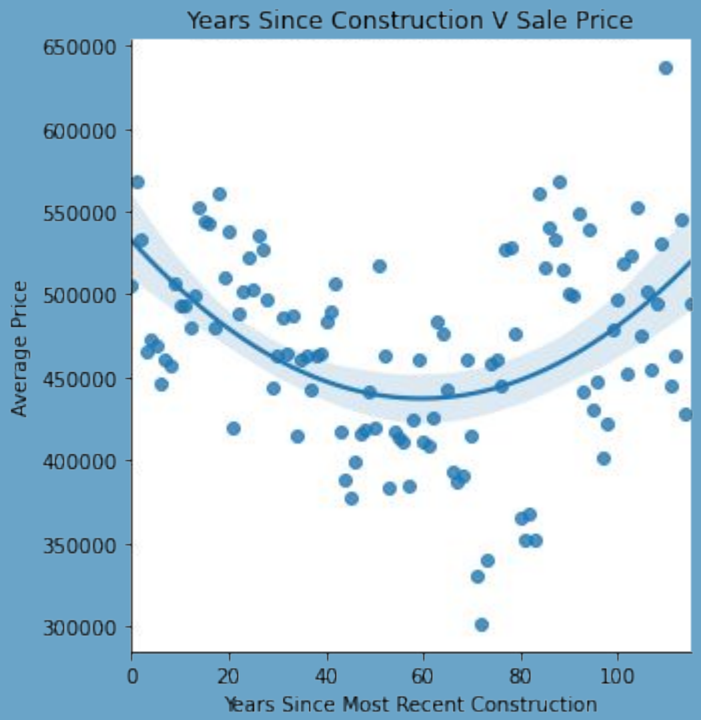
At lower square footages, the opposite is true. A smaller home will experience a value increase due to the larger neighbors, as consumers are anchored to the large home size.



Renovation Choices

A perfect way to add value to your home is through a renovation - but how can you know if it's worth it?

As a home gets older, it becomes an antique, and a renovation may cut into that value.



EDA Conclusions

- If a home is **less** than 3000 square feet and smaller than its neighbors, it will experience a price increase relative to similarly sized houses.
- Iif a home is **more** than 3000 square feet, and larger than its neighbors, it will also experience a price hike.
- Both of these scenarios are materializations of the anchoring fallacy, as they are anchored to the neighbor's size.

- 'A renovation is not always the best choice for adding value to a home. If a home has not had work in 60 years, it will tend to experience value appreciation as the home gains an 'antique' status.

Future Work

- A user interface that allows prospective sellers to predict their own home’s price would be the logical next step for this investigation.

- In order to better represent the housing market as a whole, a larger sample should be taken, one that does not limit the sample to one metropolitan area.

Thank You!