BUYING A HOUSE FOR DUMMIES®

HOW TO NOT GET RIPPED OFF

A Reference for the Rest of Us!®

Have you ever

Have you ever

Have you ever

Background

- With the ease of Technology, house buyers are making purchases using online platforms, instead of their realtor.
  - *According to the 2020 National Association of Realtors Profile of Home Buyers and Sellers, 51% of buyers found the home they purchased on the internet.*
  - Examples of platforms
    - *Zillow, Realtor.com, Trulia, Homes for Heroes etc*

Background

- Without a professional realtor to help, it is hard to get relative sense of what is a reasonable price.
    - Even with data points of average sale price in the same neighbourhood, each specific house would also be very different in features - especially with renovations.
- How can we equip buyers with more specific information that could help them compare properties?

Problem Statement

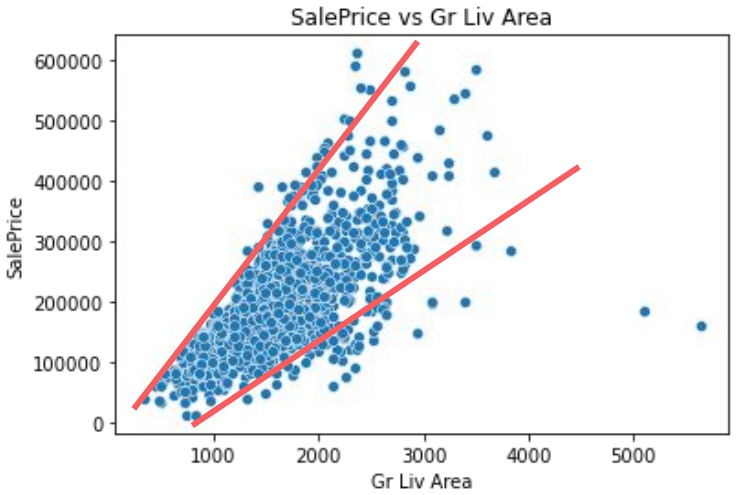
- To create a simplified framework of significant factors that affect house prices.
- The simplified model aims to be usable by human without a computer, so that guesstimate of house prices can be made on-site.
- Allows for quicker evaluation of property prices. From a economic point of view, allows for more efficient and quicker pricing of goods and thereby enabling a smoother economy.

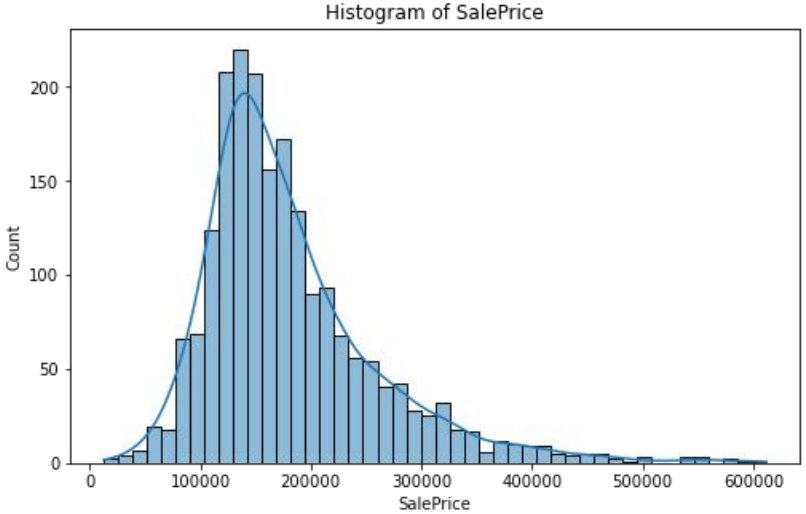
A trip back to high school mathematics

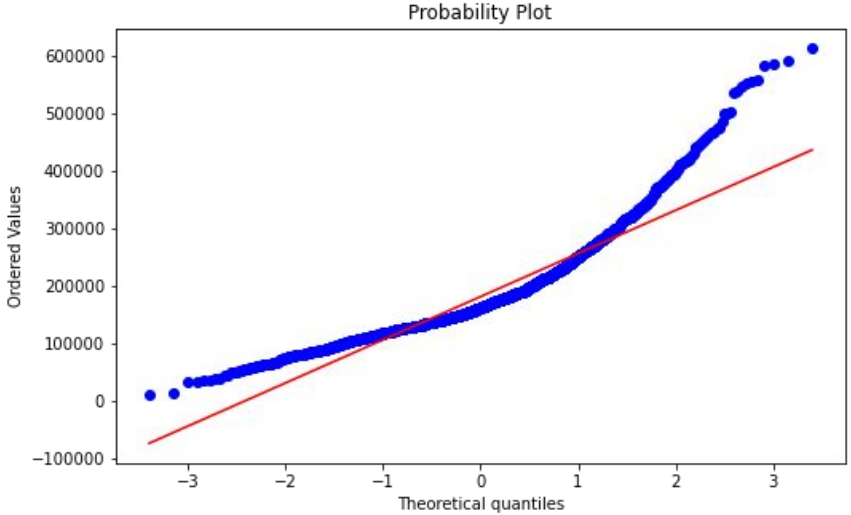
Logarithmic treatment

Raw ‘Sale Price’ is not statistically normal

Exhibits heteroscedasticity (varying variance)

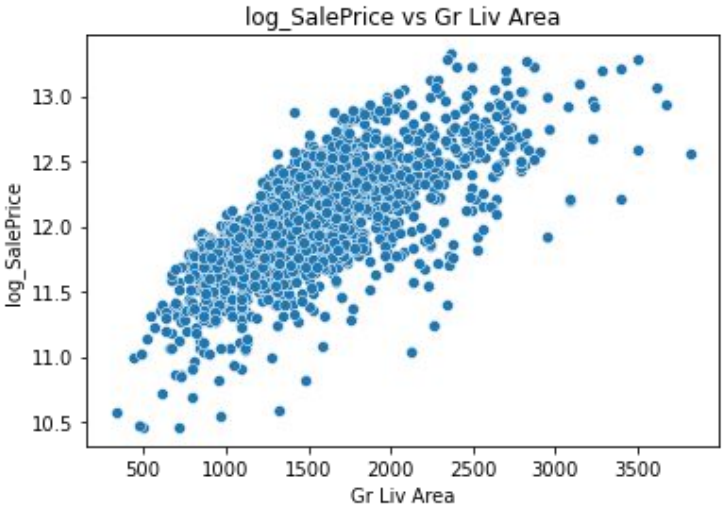


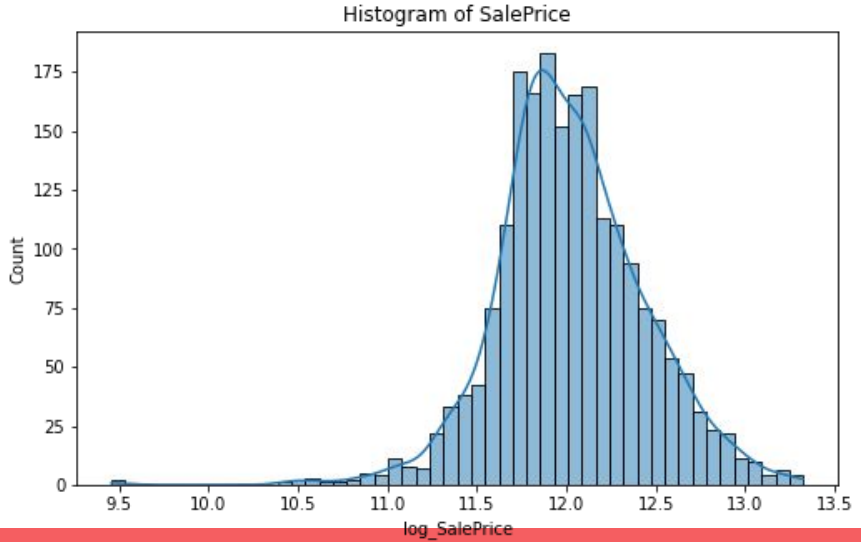


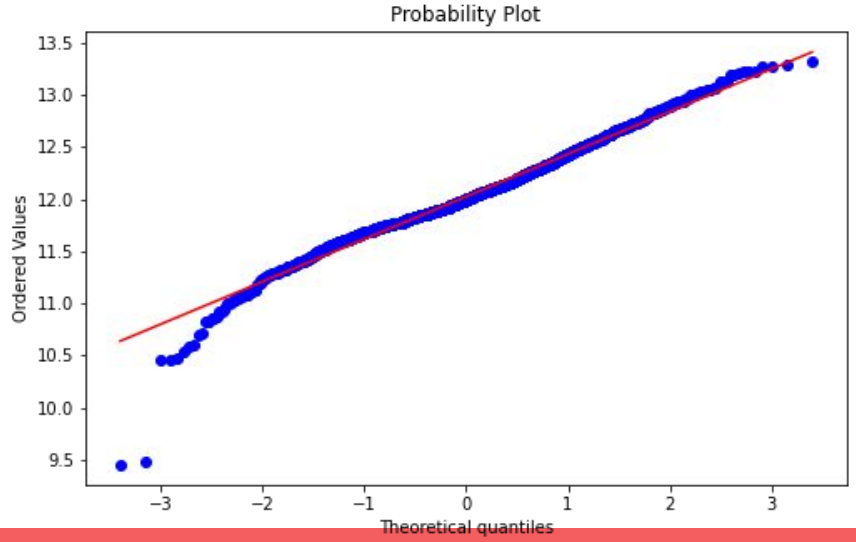


Logarithmic treatment

After logarithmic treatment

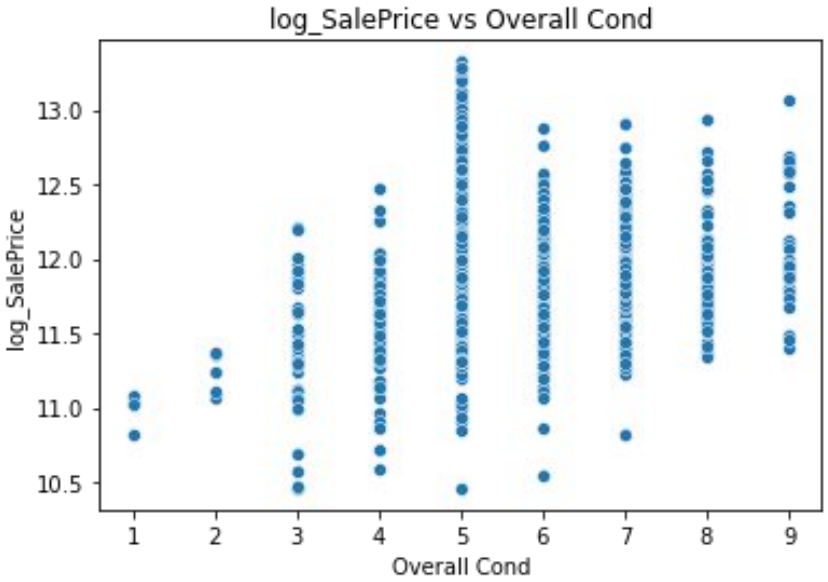


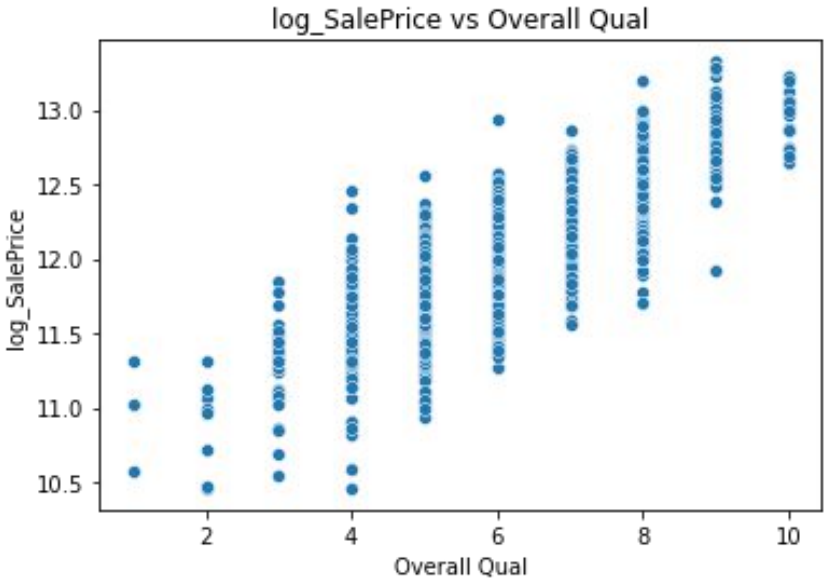




Preliminary guess

Overall Condition seems to be a poor predictor.
However, **Overall Quality** seems to have a strong correlation with the Sale Price.





Top numerical features (using correlation)

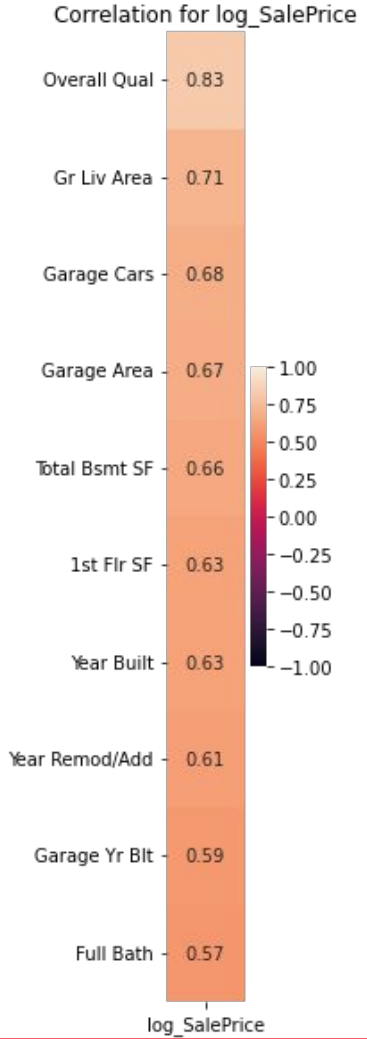
Looking only at |correlations| above 0.5,

We have

Overall Qual	Gr Live Area	Garage Cars	~~Garage Area~~	Total Bsmt SF
~~1st Flr SF~~	Year Built	Year Remod/Add	~~Garage Yr Blt~~	Full Bath

Using only 7 numerical features,

We get a Test RMSE of 27000.



All numerical features

(using coefficient)

We pick the top X features with biggest absolute coefficient

Kitchen AbvGr	Bedroom AbvGr
Overall Qual	Full Bath
Overall Cond	Year Remod/Add
Garage Cars	Lot Frontage
Fireplaces	Gr Liv Area
Bsmt Full Bath	Screen Porch
Yr Sold	Low Qual Fin SF
Half Bath	3Ssn Porch
Bsmt Half Bath	Enclosed Porch
Year Built	Total Bsmt SF
	BsmtFin SF 1

Including ordinal features
(using coefficient)

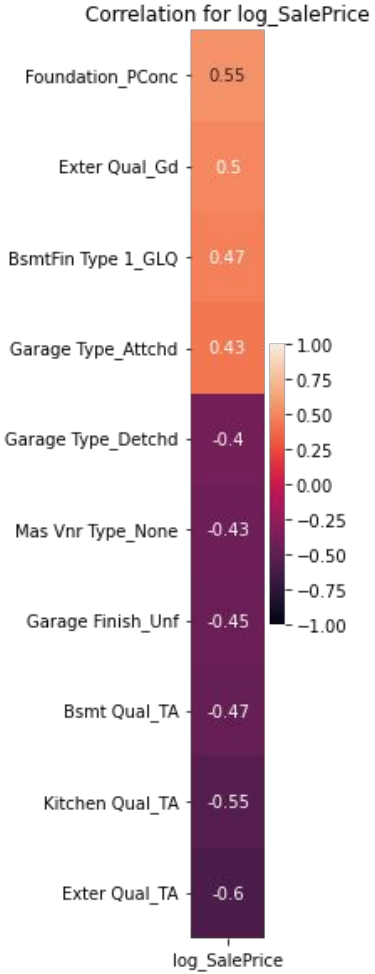
Convert ordinal features to numerical ranking,
Then run through top X coefficient picking again.

Gr Liv Area	Functional	Bsmt Cond
Overall Qual	Heating QC	Land Slope
Year Built	Central Air	Full Bath
Utilities	Exeter Qual	Lot Area
Pool QC	Paved Drive	Fireplaces
Total Bsmt SF	Garage Cars	Bsmt Qual
Kitchen AbvGr	Kitchen Qual	Yr Sold
BsmtFin SF 1	Bsmt Full Bath	Bsmt Exposure
Overall Cond	Garage Cond	
Fireplace Qu	Lot Frontage	
Year Remod/Add	2nd Floor SF	

Including categorical features
(using coefficient)

Filter for features wher abs(correlation) > 0.4

Combine the list of features from previous and pick top X again



Including categorical features
(using coefficient)

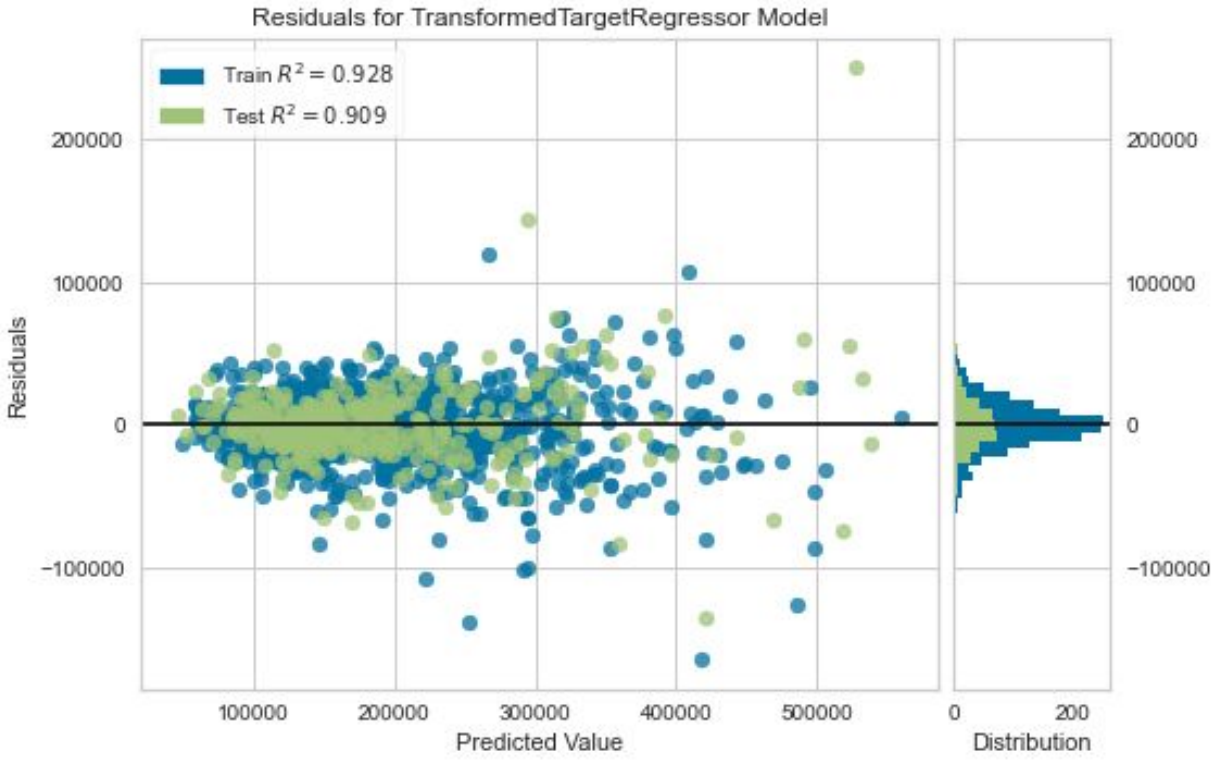
Filter for features wher abs(correlation) > 0.4

Combine the list of features from previous and pick top X again

Gr Liv Area	Paved Drive	Bsmt Cond
Overall Qual	Exeter Qual	Land Slope
Year Built	Foundation_PConc	2nd Flr SF
Kitchen AbvGr	Heating QC	Lot Area
Utilities	Year Remod/Add	Garage Cond
BsmtFin SF 1	Garage Cars	Full Bath
Total Bsmt SF	Kitchen Qual	Yr Sold
Overall Cond	Bsmt Full Bath	Bsmt Exposure
Central Air	Pool QC	
Fireplace Qu	Lot Frontage	
Functional	Fireplaces	

Model performance - Residual Plot

Errors are normally distributed



Model performance - Predicted vs Actual

Prediction matches actual closely

How trustable is the model?

How well are variations in the price described by the 30 features? (R2 score)

92.6%

Test RMSE: 25000, better than the initial guess by ~2000

Final 30 features

'Gr Liv Area', 'Overall Qual', 'Year Built', 'Kitchen AbvGr', 'Utilities', 'BsmtFin SF 1', 'Total Bsmt SF', 'Overall Cond', 'Central Air', 'Fireplace Qu', 'Functional', 'Paved Drive', 'Exter Qual', 'Foundation_PConc', 'Heating QC', 'Year Remod/Add', 'Garage Cars', 'Kitchen Qual', 'Bsmt Full Bath', 'Pool QC', 'Lot Frontage', 'Fireplaces', 'Bsmt Cond', 'Land Slope', '2nd Flr SF', 'Lot Area', 'Garage Cond', 'Full Bath', 'Yr Sold', 'Bsmt Exposure']

Coefficients of the 30 features

Feature	Coefficient
Gr Liv Area	0.178345
Overall Qual	0.125930
Year Built	0.097279
Total Bsmt SF	0.060508
Utilities	0.060137
BsmtFin SF 1	0.057972
Overall Cond	0.042846
Central Air	0.038421
Fireplace Qu	0.038063
Functional	0.036115
Paved Drive	0.033875
Foundation_PConc	0.033565
Year Remod/Add	0.032083
Exter Qual	0.031893
Garage Cars	0.029414
Heating QC	0.029269
Kitchen Qual	0.027057
Bsmt Full Bath	0.026926
Pool QC	0.025572
Lot Frontage	0.020502
Fireplaces	0.019728
Garage Cond	0.015419
Lot Area	0.014572
Full Bath	0.011714
Bsmt Exposure	0.011686
Yr Sold	-0.011441
Bsmt Cond	-0.013280
2nd Flr SF	-0.013830
Land Slope	-0.018550
Kitchen AbvGr	-0.061696

Simplified framework

We can break down the features into 2 main categories for practicality purpose:

Major features:

Gr Liv Area: Above ground living area square feet,

Overall Qual: Rates the overall material and finish of the house (1-10),

Year Built: Original construction date

Minor features:

Total Bsmt SF: Total square feet of basement area,

Kitchen AbvGr: Number of kitchens above ground level ***(negative coefficient)***,

Utilities: Type of utilities available (Elec, Elec+Gas, Elec+Gas+Water, Elec+Gas+Water+Sewage),

BsmtFin SF 1: Basement finished area of type 1 in square feet,

Overall Cond: Rates the overall condition of the house (1-10)

Recommendation and conclusion

Based on what we gathered, we picked the features that has a high coefficient to determine the importance of it.

The other features did not significantly impact the sale prices therefore its not efficient and not worth looking into.

The simplified framework will make it easier for buyers to narrow down by looking at the 8 specific features for information that will affect the price of the sale.