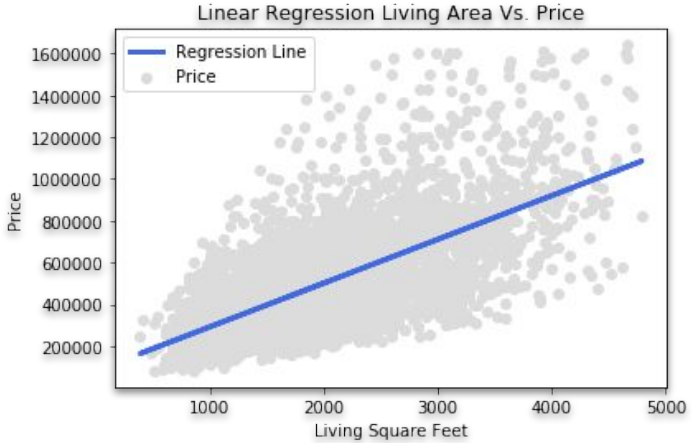
Linear Regression of King County Real Estate Data

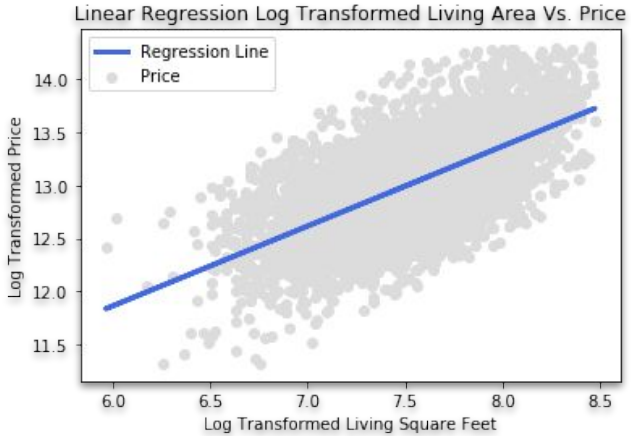
By: Blake and Alex

Purpose

- Use Case Scenario:
  - You are on Zillow looking to buy a house and see a house listed for sale that you like. You don’t have any real estate experience and are not familiar with King county at all. How do you know if the listed sales price is a good or bad price?
- Goals:
  - Foolproof UX
  - Simple, but accurate
  - Use information that is very easily found and verifiable

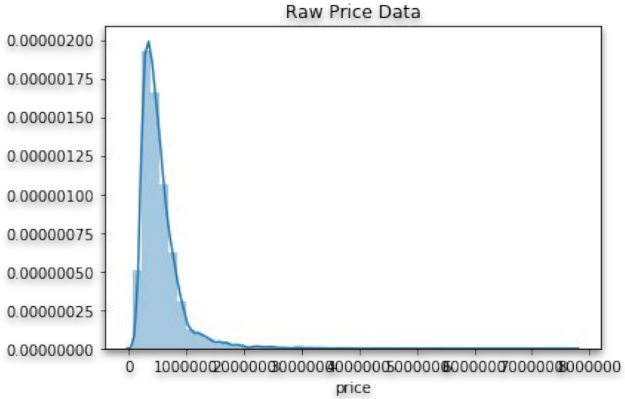
Sqft living vs price

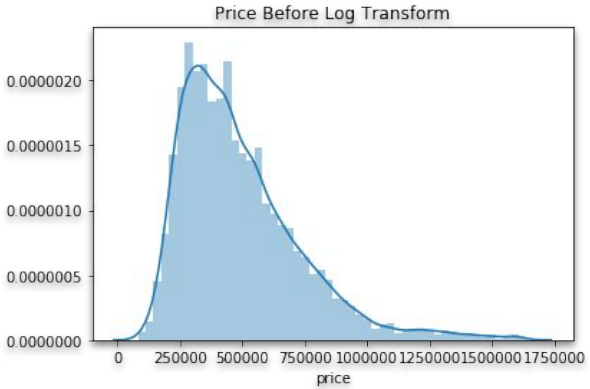


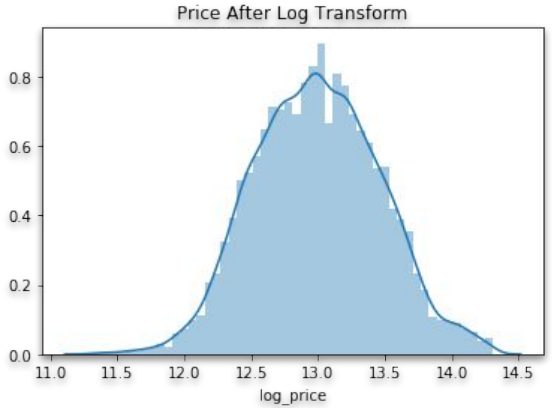


Accuracy did not change in this model due to transformation: .417

Price Before and After log transform







Kurtosis

- Any distribution with **kurtosis ≈3 (excess kurtosis (Fisher) ≈0)** is called mesokurtic. This is a normal distribution
- Any distribution with **kurtosis <3 (excess kurtosis (Fisher) <0)** is called platykurtic. Tails are shorter and thinner, and often its central peak is lower and broader.
- Any distribution with **kurtosis >3 (excess kurtosis (Fisher) >0)** is called leptokurtic. Tails are longer and fatter, and often its central peak is higher and sharper.

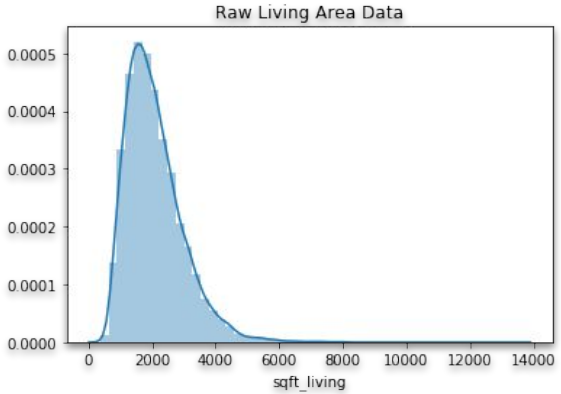
```
Raw Price Data Kurtosis:  2.474
Before Price Log Transform Kurtosis:  2.474
After Price Log Transform Kurtosis:  -0.186
```

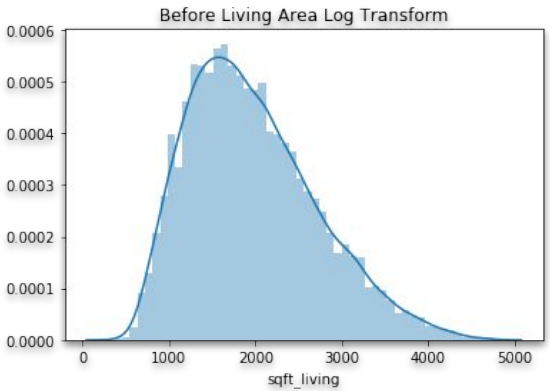
Skewness

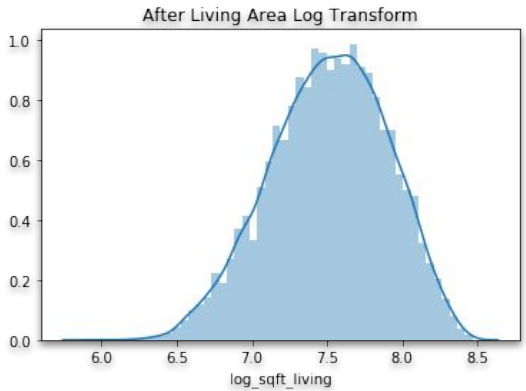
- If skewness is **less than −1 or greater than +1**, the distribution is highly skewed.
- If skewness is **between −1 and −½ or between +½ and +1**, the distribution is moderately skewed.
- If skewness is **between −½ and +½**, the distribution is approximately symmetric.

```
Raw Price Data Skewness:  1.365
Before Price Log Transform Skewness:  1.365
After Price Log Transform Skewness:  0.042
```

Sqft_Living Before and after log transform





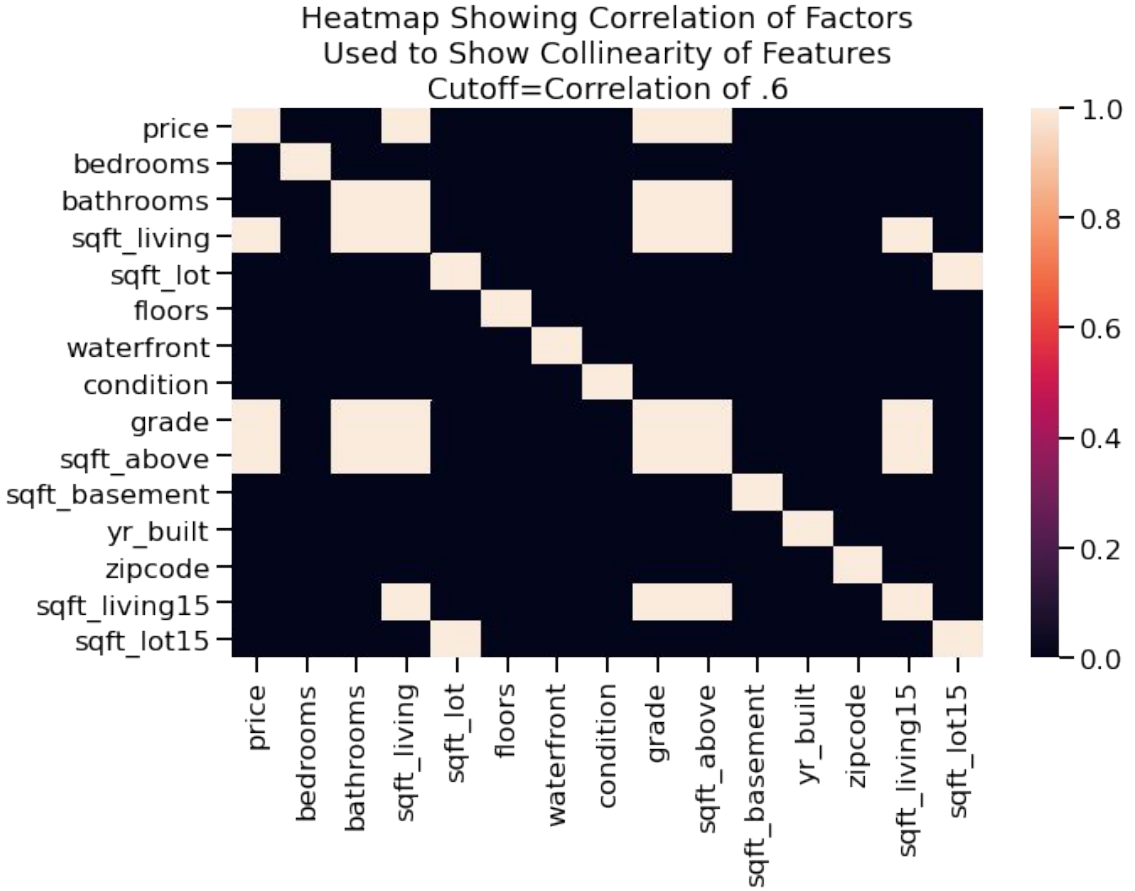



```
Raw Living Area Data Skewness:  0.642
Before Living Area Log Transform Skewness:  0.642
After Living Area Log Transform Skewness:  -0.232
```

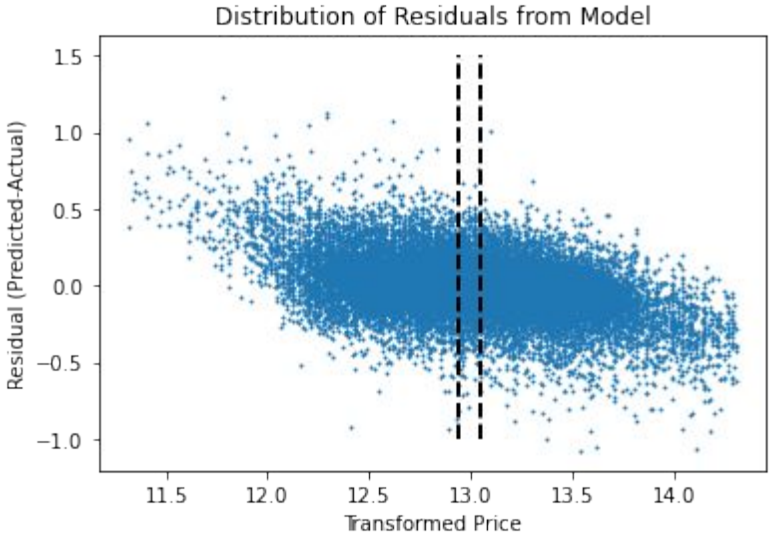
```
Raw Living Area Data Kurtosis:  2.474
Before Living Area Log Transform Kurtosis:  0.027
After Living Area Log Transform Kurtosis:  -0.354
```

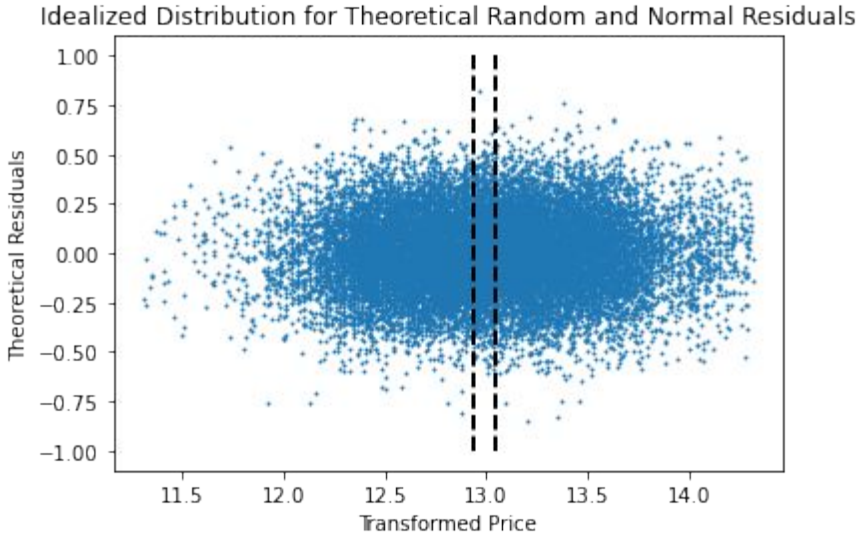
Multicollinearity

- Map of correlation values
- If the absolute value of correlation is greater than 0.6, set to true, otherwise false
- Plot is of these true/false values-easy to visually interpret



Validity of model

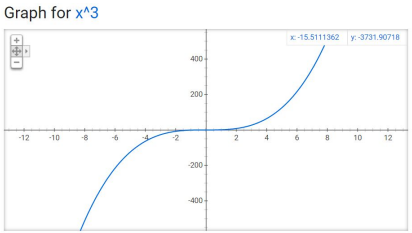


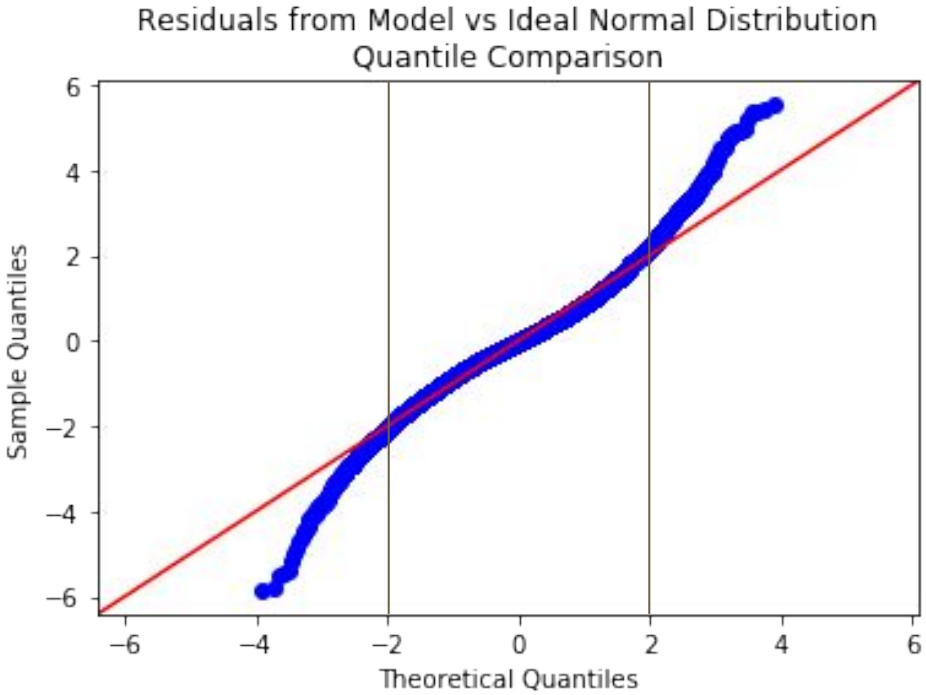


- Plotting of residual between predicted and actual vs transformed price
- Shows that model over-predicts at low prices and under-predicts at high prices

Residuals vs Ideal Normalized Distributions

- Residuals normally distributed within first 2 standard deviations
- Trend in the residuals looks polynomial





Model Output

- R-squared
  - Our model can explain about 82% of the variance in the data
- Cross Validation Score
  - .817 with stdev of .03
- MAE = $73,244

OLS Regression Results

Dep. Variable:	log_price	R-squared:	0.816
Model:	OLS	Adj. R-squared:	0.815
Method:	Least Squares	F-statistic:	1325.
Date:	Thu, 04 Jun 2020	Prob (F-statistic):	0.00
Time:	15:07:32	Log-Likelihood:	3665.5
No. Observations:	20110	AIC:	-7195.
Df Residuals:	20042	BIC:	-6657.
Df Model:	67		
Covariance Type:	nonrobust		

Limitations of Model

- Data is heteroscedastic
  - Underestimates large homes greater than 3,916 sf
    - Larger the house larger the possible error
  - Over Predicts small homes under 244 sf
    - Unrealistic so this error is null
- Model training data is old
  - Real Estate prices fluctuate over time
  - Training data is from 2014-2015 so not valid today
- Model is trained only for KC county
  - Location is very important in real estate
  - Can easily be trained on other counties though!

Interactive Demo

Link

Future Work

- Our Model failed the OLS homoscedastic assumption
  - Re-evaluate using some other model such as GLM,GLS, WLS
- Residuals have a pattern
  - This may need to be a polynomial regression rather than linear
- Have interactive spit out a range instead of singular value
- Use RFECV to see if removing some zipcodes would increase fit
- Update data and repeat for different areas

Thank You!

1. Git Hub Project Links:
    a. Alex:
        i. https://github.com/anbillinger/dsc-mod-2-project-v2-1-onl01-dtsc-ft-041320
    b. Blake
        i. https://github.com/bsamaha/dsc-mod-2-project-v2-1-onl01-dtsc-ft-041320
2. Follow our Blogs:
    a. Alex:
        i. https://anbillinger.github.io/
    b. Blake
        i. https://medium.com/@blake.samaha16