King County Business Plan:

REST™

Stephen Enke and Austin Murray

Business Problem

We are a new real estate agency, Rest, with a customer-friendly realtors paired with an online site for price prediction. Initially, we are only operating in King County. In order to both attract customers online and to help our realtors tailor specific needs to our home buyers and sellers, we need a prediction model that will accurately predict home prices after features are uploaded.

Our prediction model will both help online users find out the price of their specific home or to look at potential prices of homes with certain features in specific locations.

For our initial roll-out, Rest is focusing on mid-range houses priced $200,000 to $790,000

Data Used

Data comes from King County House Sales Dataset

This includes many factors of the houses sold in king county such as:

- Date sold
- Price
- Zipcode
- Square Footage

```
In [212]: house_data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 21597 entries, 0 to 21596
Data columns (total 21 columns):
 #   Column         Non-Null Count  Dtype  
---  ------         --------------  -----  
 0   id             21597 non-null  int64  
 1   date           21597 non-null  object 
 2   price          21597 non-null  float64
 3   bedrooms       21597 non-null  int64  
 4   bathrooms      21597 non-null  float64
 5   sqft_living    21597 non-null  int64  
 6   sqft_lot       21597 non-null  int64  
 7   floors         21597 non-null  float64
 8   waterfront     19221 non-null  float64
 9   view           21534 non-null  float64
 10  condition      21597 non-null  int64  
 11  grade          21597 non-null  int64  
 12  sqft_above     21597 non-null  int64  
 13  sqft_basement  21597 non-null  object 
 14  yr_built       21597 non-null  int64  
 15  yr_renovated   17755 non-null  float64
 16  zipcode        21597 non-null  int64  
 17  lat            21597 non-null  float64
 18  long           21597 non-null  float64
 19  sqft_living15  21597 non-null  int64  
 20  sqft_lot15     21597 non-null  int64  
dtypes: float64(8), int64(11), object(2)
memory usage: 3.5+ MB
```

EDA

We cleaned the original data set to remove outliers, fill null values, and drop unnecessary data

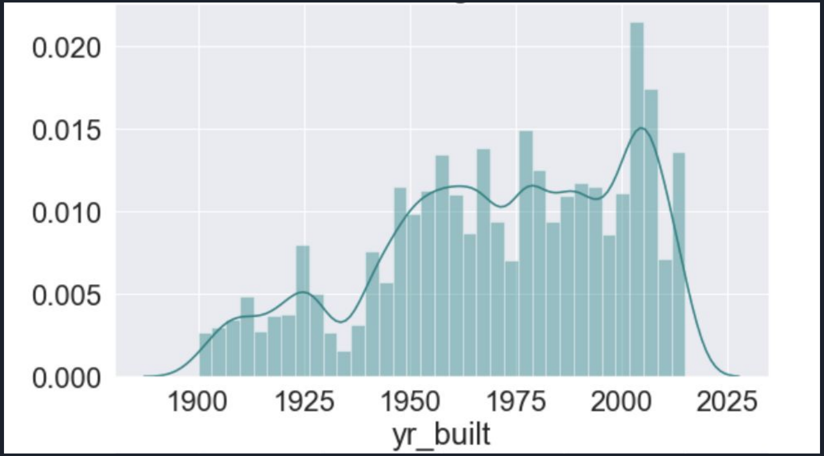
Scatter Matrices and Distplots were used to determine relationships between price and explanatory variables & determine normalcy

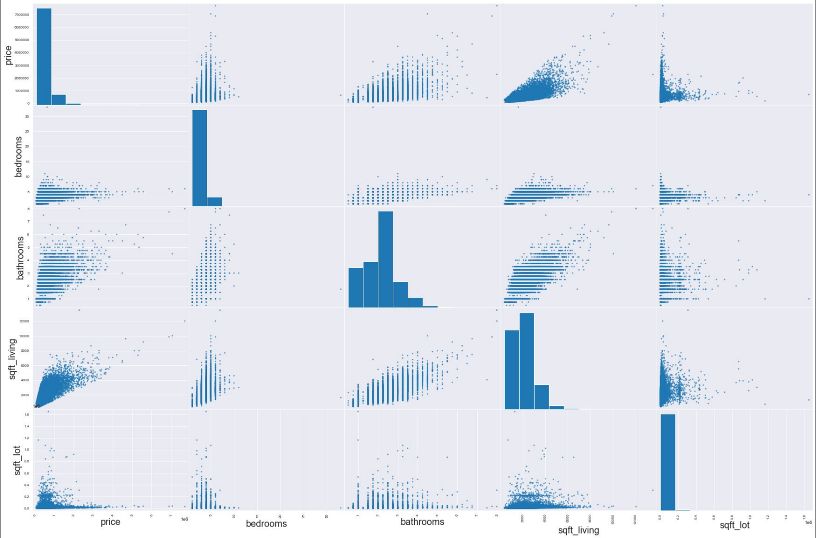
Judged multicollinearity

Feature Engineering:

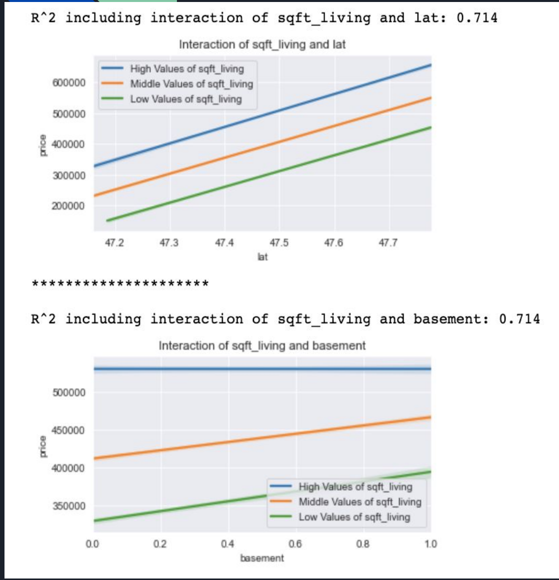
- Cities from zipcodes
- Binary Variables (Renovated)
- Binned Bathrooms
- Dummied Cat Variables

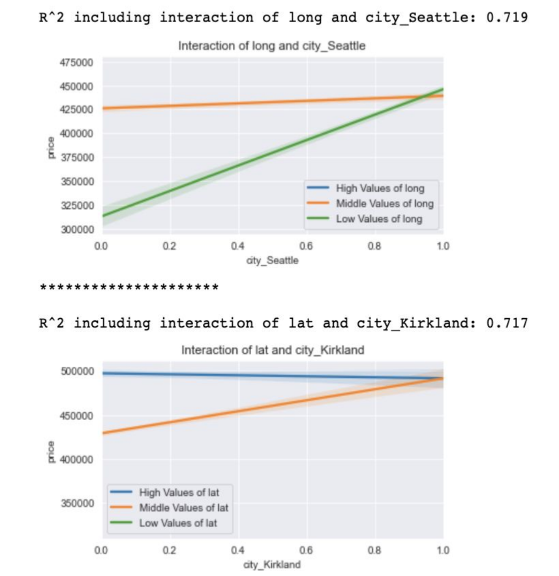
	cc
pairs	
(sqft_above, sqft_living)	0.876448
(sqft_living, grade)	0.762779
(sqft_living, sqft_living15)	0.756402
(sqft_above, grade)	0.756073
(sqft_living, bathrooms)	0.755758
(sqft_above, sqft_living15)	0.731767
(sqft_lot15, sqft_lot)	0.718204
(sqft_living15, grade)	0.713867
(price, sqft_living)	0.701917

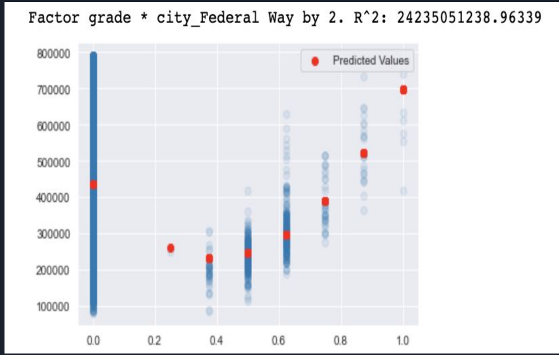




EDA - Interactions & Polynomials







Used functions to determine most relevant interactions and polynomial features to include in model.

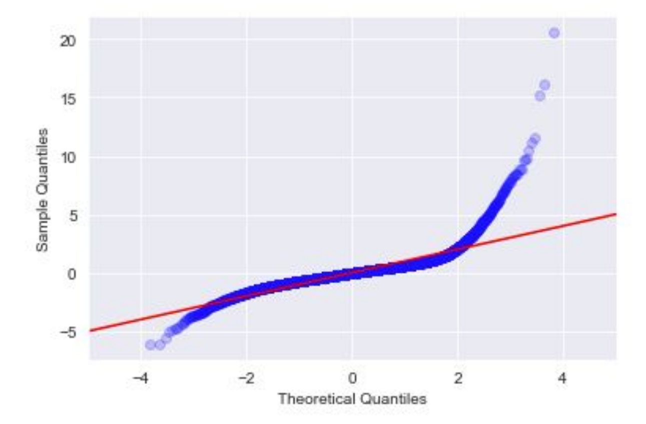
Original Modeling

We may have a high R-squared but our RMSE shows that this is most likely from spurious correlation

QQ plot for the vanilla model shows a heavy tail on the upper end and is definitely not normal

Dep. Variable:	price	**R-squared:**	0.760
Model:	OLS	**Adj. R-squared:**	0.759
Method:	Least Squares	**F-statistic:**	1135.
Date:	Sun, 29 Nov 2020	**Prob (F-statistic):**	0.00
Time:	12:35:11	**Log-Likelihood:**	-2.0418e+05
No. Observations:	15117	**AIC:**	4.085e+05
Df Residuals:	15074	**BIC:**	4.088e+05
Df Model:	42		
Covariance Type:	nonrobust		

```
Train RMSE: 177703.18334308316
 Test RMSE: 188904.6781261696
Percent change:  6.303
Percent change (Base Model vs. Updated Model):  0.0
```



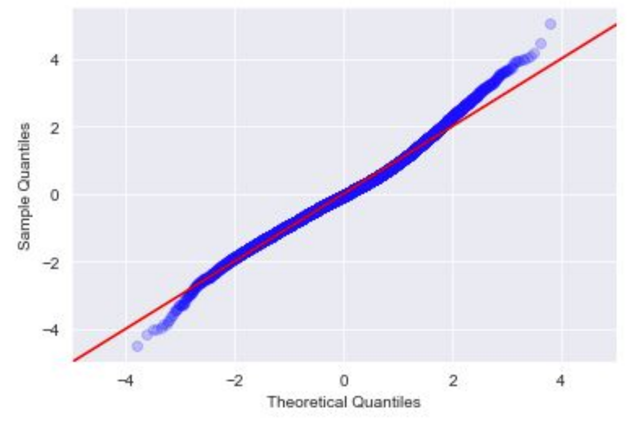
Pricing outliers removed Model

The R-squared dropped for the correlation but the accuracy improved by over 50 percent

QQ plot shows again that our regression line is much more accurate

Dep. Variable:	price	**R-squared:**	0.712
Model:	OLS	**Adj. R-squared:**	0.711
Method:	Least Squares	**F-statistic:**	758.9
Date:	Sun, 29 Nov 2020	**Prob (F-statistic):**	0.00
Time:	12:35:20	**Log-Likelihood:**	-1.6555e+05
No. Observations:	12958	**AIC:**	3.312e+05
Df Residuals:	12915	**BIC:**	3.315e+05
Df Model:	42		
Covariance Type:	nonrobust		

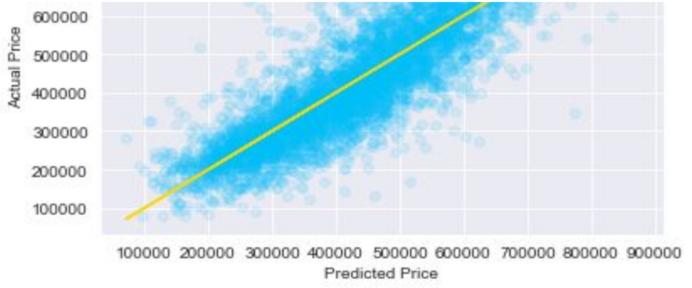
```
Train RMSE: 85556.9037960636
 Test RMSE: 87245.36622062251
Percent change:  1.973
Percent change (Base Model vs. Updated Model):  -51.854
```



Final Model: Stepwise Selection

Dep. Variable:	price	**R-squared:**	0.762
Model:	OLS	**Adj. R-squared:**	0.761
Method:	Least Squares	**F-statistic:**	664.9
Date:	Sun, 29 Nov 2020	**Prob (F-statistic):**	0.00
Time:	12:45:36	**Log-Likelihood:**	-1.6431e+05
No. Observations:	12958	**AIC:**	3.288e+05
Df Residuals:	12895	**BIC:**	3.292e+05
Df Model:	62		
Covariance Type:	nonrobust		

```
Train RMSE: 77772.56378704724
 Test RMSE: 78721.96709791309
Percent change:  1.221
Percent change (Base Model vs. Updated Model):  -56.235
```

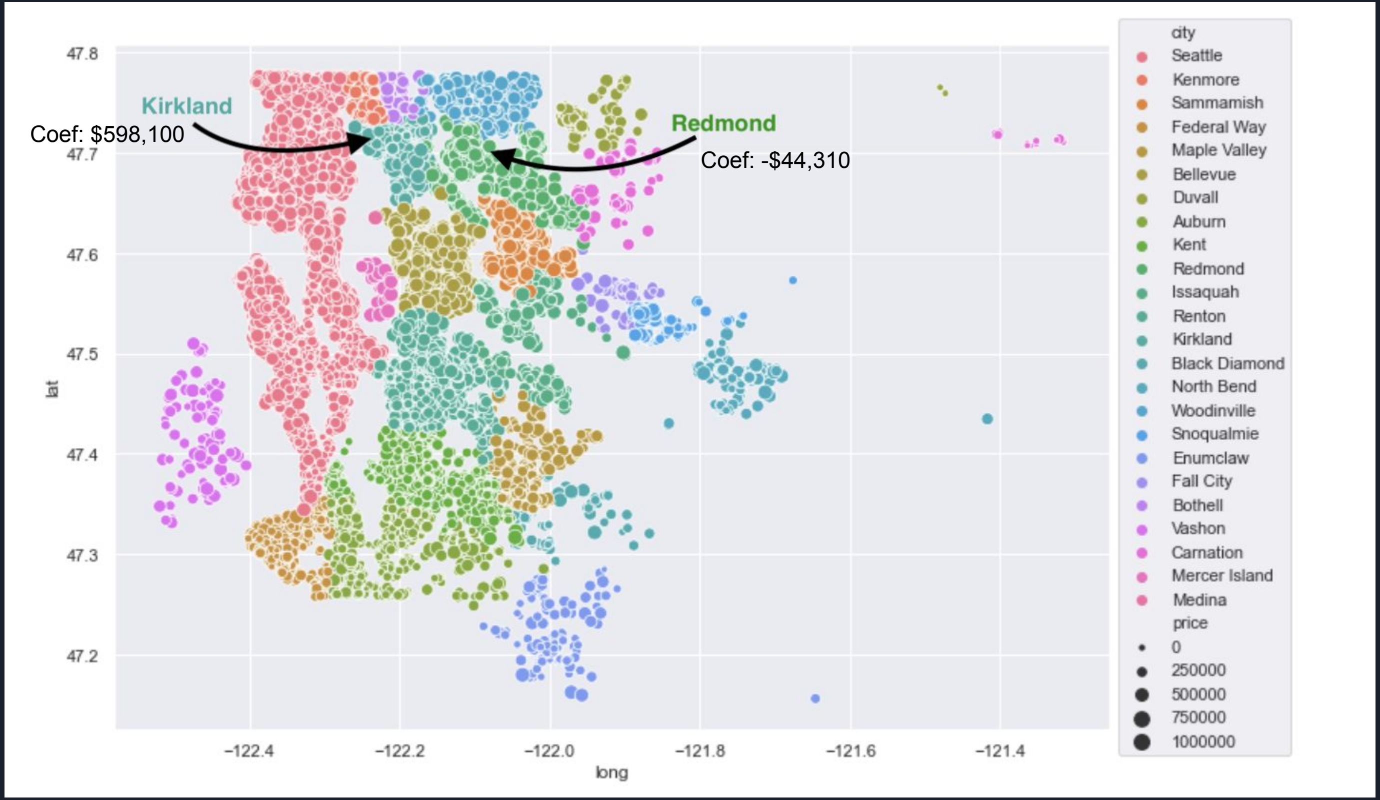



```
************
High Impact Variables:

Variable: grade * lat
Coefficient: 804107.0554313979

Variable: yr_built * lat
Coefficient: -799156.7071081145

Variable: city_Bellevue^2
Coefficient: 1.1066880182318767e+18

Variable: city_Bellevue
Coefficient: -1.1066880182318497e+18

Variable: lat * city_Kirkland^2
Coefficient: -808448.0

Variable: lat
Coefficient: 1149824.0

Variable: long * city_Kent^2
Coefficient: 1018880.0

Variable: sqft_living * floors^2
Coefficient: -1230682.0

Train R^2: 0.7614718257178578
CrossValidated R^2: 0.758080765270026
Test R^2: 0.7546748573809261
```

Unique Opportunity:

Bargain hunting with Young Families

- Better school metrics correlate with more expensive cities
- Certain cities share school districts
- Opportunity for affordable housing at good school districts via neighboring cities



Conclusion

We are focusing on homes in the $200,000 - $790,000 price range as it is what our model is best suited for and aligns with opportunities we see in the data. Using the most impactful variables such as the interaction between grade & latitude, we can find homes catered to clients based on their priorities and price range.

We feel especially equipped for helping young families with well-paying jobs find their first home in Seattle and its suburbs. Using our model, we can find opportunities for customers who are looking for housing in a great school district for younger children that is still affordable.

```
************
High Impact Variables:

Variable: grade * lat
Coefficient: 804107.0554313979

Variable: yr_built * lat
Coefficient: -799156.7071081145

Variable: city_Bellevue^2
Coefficient: 1.1066880182318767e+18

Variable: city_Bellevue
Coefficient: -1.1066880182318497e+18

Variable: lat * city_Kirkland^2
Coefficient: -808448.0

Variable: lat
Coefficient: 1149824.0

Variable: long * city_Kent^2
Coefficient: 1018880.0

Variable: sqft_living * floors^2
Coefficient: -1230682.0

Train R^2: 0.7614718257178578
CrossValidated R^2: 0.758080765270026
Test R^2: 0.7546748573809261
```

Thank You & Questions

Next Steps:

- Drill more into specific data of Redmond to improve our model for the specific suburb we will be focusing on
- Train model on other specific suburbs with negative impacts to go after affordable houses that may be able to be renovated.
- Include school data to improve model's accuracy

Thank You:

- Former DS Students (especially Shawn Sobieski) for GitHub inspo
- Thanksgiving Turkey for fueling the procrastinated last dash
- Yish and her fellow Data Science instructors (Amber, Lindsey, and Abhineet)