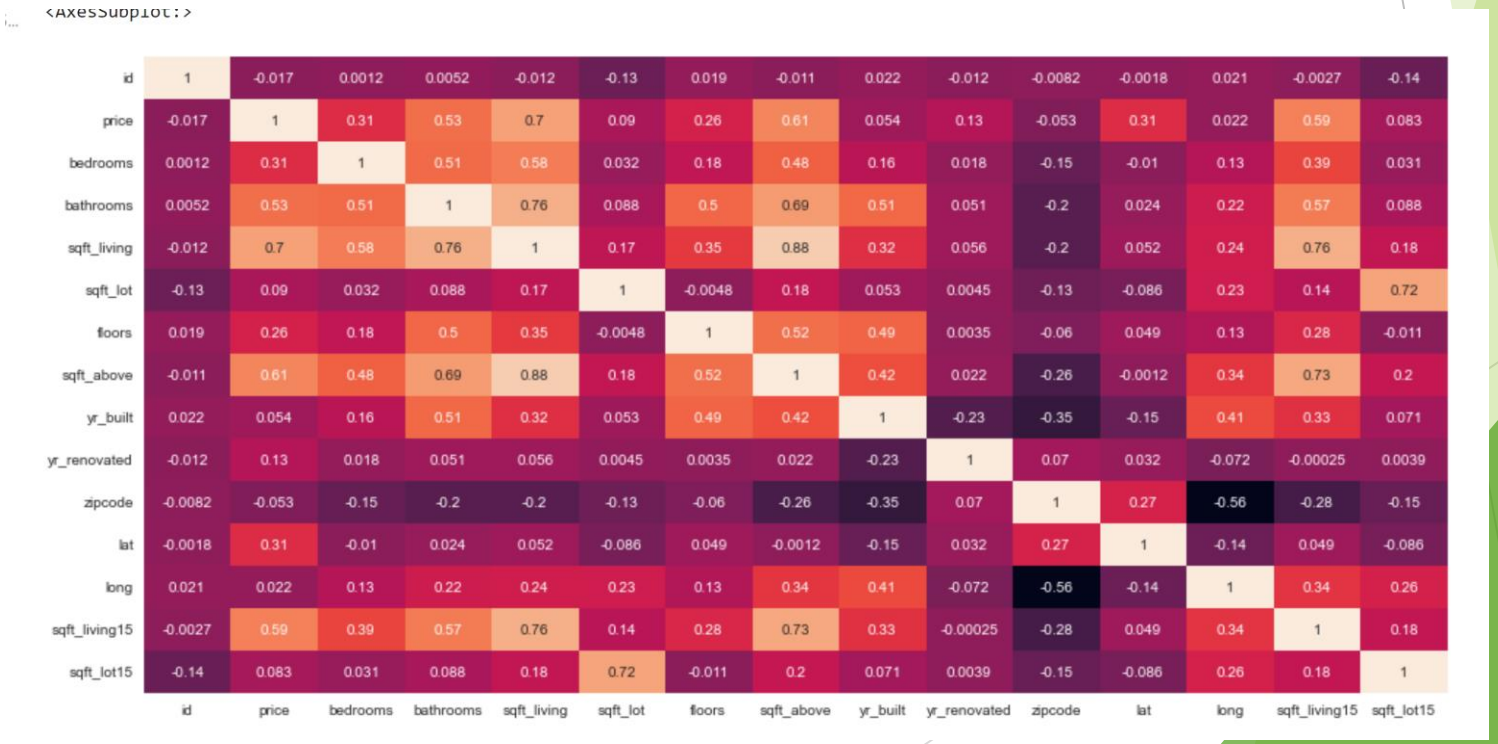
FlatIron Phase 2 Power point presentation

Henry Chung

Business problems

Data-kc_house_data.csv

21597 of home sale between 2014-2015 in King County is used.



Data Source – and processing

21597 of home sale between 2014-2015 in King County is used.

21 columns of columns is included in the dataset.

0	id
1	date
2	price
3	bedrooms
4	bathrooms
5	sqft_living
6	sqft_lot
7	floors
8	waterfront
9	view
10	condition
11	grade
12	sqft_above
13	sqft_basement
14	yr_built
15	yr_renovated
16	zipcode
17	lat
18	long
19	sqft_living15
20	sqft_lot15

Processing Road Map

Processing for this analysis

a) Download the data
b) Split the data into train and test set
c) Data cleaning
d) Set up baseline modeling and fine tune
e) Recommendation

Methodology

3 Linear Regression models were made

1) Every factors from the data set were used as the predictors

2) VLS were used to eliminate some non influential factors.

3) Outliers and conditions were taken from model-2

Model 1 (base-Line)

Method: Linear regression

1) 87 inputs were used

2) R-score is .87 and test score is .85. It means 87% of the variance can be explained by the predictors.

Model 2 (Fine tune BaseLine w WIF)

Method: Linear regression

1) 35 inputs were used with feature selections via Varince_inflation_factor method to reduce multicollinearity issue

2) R-score is .69 and test score is .66.

3) Model is not over fitting or under fitting

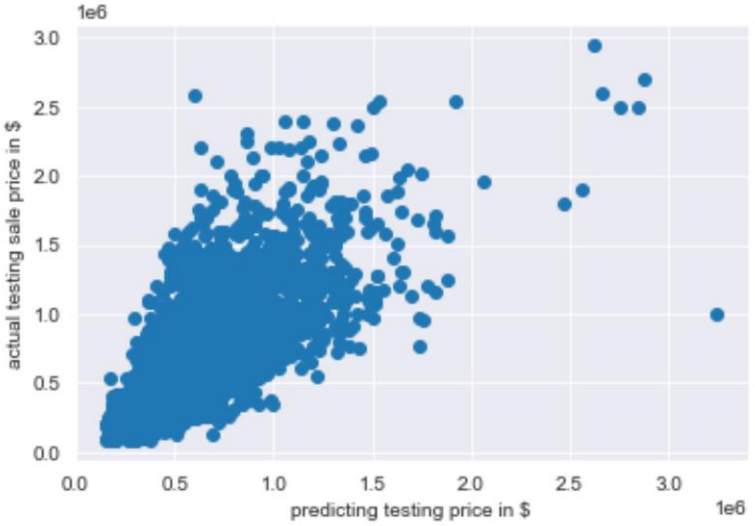
Model 3 (model 2 eliminate outliers, and subset several condition)

Method: Linear regression

1) Several category outliers are removed, and subset of several prediction conitions are used.

2) R-score is .65 and test score is .63.

I would rather take a slight drop off in R2 score and remove the outlier



Compare Pricing Prediction between Seattle, Kent, **Bellevue**

Seattle

	seattle	kent	bellevue	actual	predicted
count	1613.0	1613.0	1613.0	1.613000e+03	1.613000e+03
mean	1.0	0.0	0.0	5.318317e+05	5.209246e+05
std	0.0	0.0	0.0	2.752936e+05	2.253124e+05
min	1.0	0.0	0.0	1.000000e+05	2.456741e+05
25%	1.0	0.0	0.0	3.550000e+05	3.907775e+05
50%	1.0	0.0	0.0	4.650000e+05	4.638699e+05

Ke
nt

Kent- prediction vs actual

	seattle	kent	bellevue	actual	predicted	condition	sqft_lot
count	298.0	298.0	298.0	298.000000	298.000000	298.000000	298.000000
mean	0.0	1.0	0.0	290961.201342	309413.130097	2.503356	0.215284
std	0.0	0.0	0.0	73592.489142	85804.384744	0.626294	0.165722
min	0.0	1.0	0.0	85000.000000	150422.792754	1.000000	0.049571
25%	0.0	1.0	0.0	245250.000000	245002.240748	2.000000	0.129399
50%	0.0	1.0	0.0	278950.000000	306141.613256	2.000000	0.165099

Recommdation

- Focus on interior housing. As they are more influential for the regression data.

- Also look for housing in Kent as their mean of price and condition are cheaper and nicer than Seattle

Next Step

- Further investigate the relationship between individual predictor and outcome
- Adjust house sale for inflation. We may get a more accurate analysis if house sale is adjusted for inflation
- Collect data for house sale during covid.

Thank You