PROJECT 2: LINEAR REGRESSION

KING COUNTY HOUSING DATA

ERIC HANSEN

FLATIRON DS; SELF PACED

9/7/2021

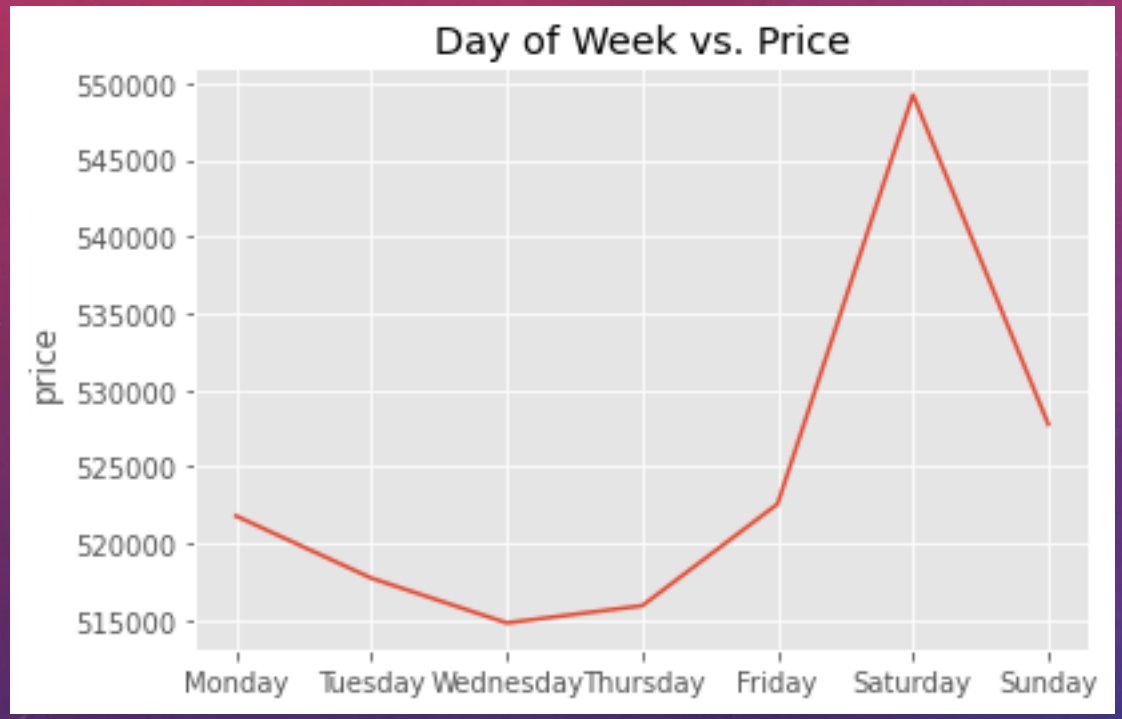
BACKGROUND

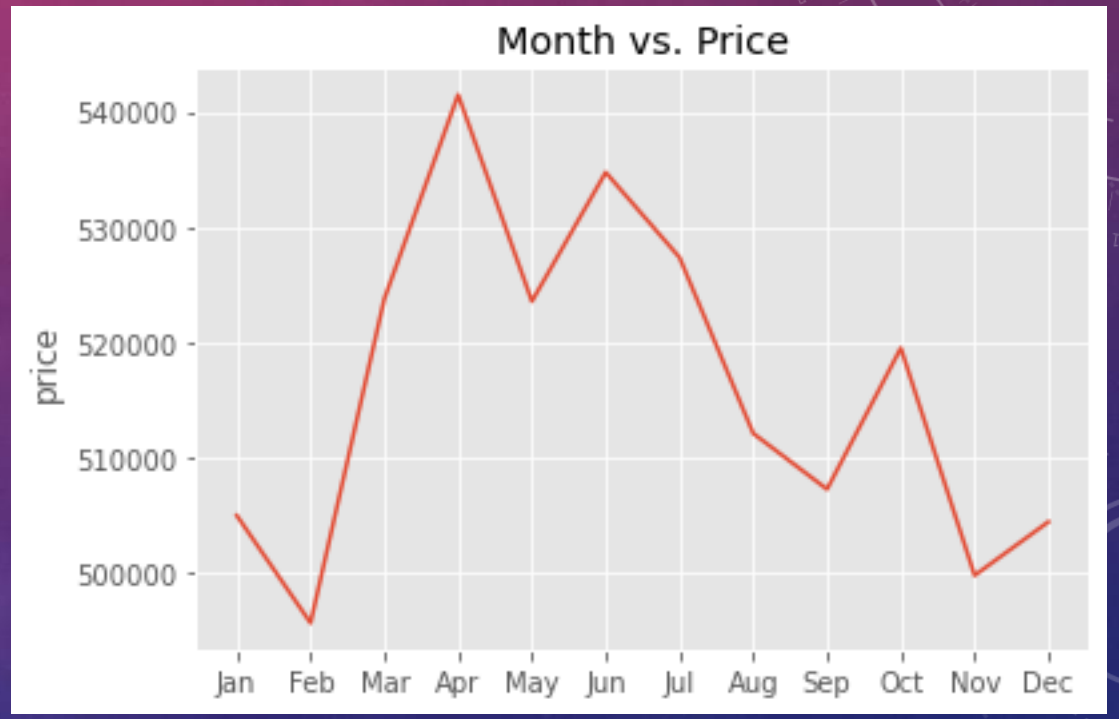
- Problem – Provide insight into most influential factors when buying/selling a home
- Stakeholder – family selling and potentially buying home in Seattle area
- Data – Kings County, WA Home Sale data; 2014-2015
- Model – Multiple Regression
- Goals
  - Identify areas for improvement of current home/ areas of savings for future home
  - Identify strategic timing for sale/purchase

METHODOLOGY

- 1) I performed Exploratory Data Analysis to identify characteristics of the data
- 2) I created several new features including grade squared, day of the week, month.
- 3) I performed multiple linear regression (with one polynomial regression term) under several different permutations (e.g. logging numericals, scaling numericals, and different ways of classifying categoricals/numeric values
- 4) I identified a model that had a good Score as well as minimal multicollinearity.
- 5) I examined the necessary conditions of linear regression
- 6) The business takeaways: sqft_living, grade_squared, zipcode_avg, waterfront, View, day of week (namely Saturday vs Sunday) all have a (normalized) significant impact on price. The top 6 coefficients in descending order are: zipcode_avg, view_3.0, view_4.0, waterfront_1.0, sqft_living, and grade_squared.
- Furthermore, the model is ready to accept any new home predictors for comparison to current prices, in order to establish if a home is over- or under-valued.

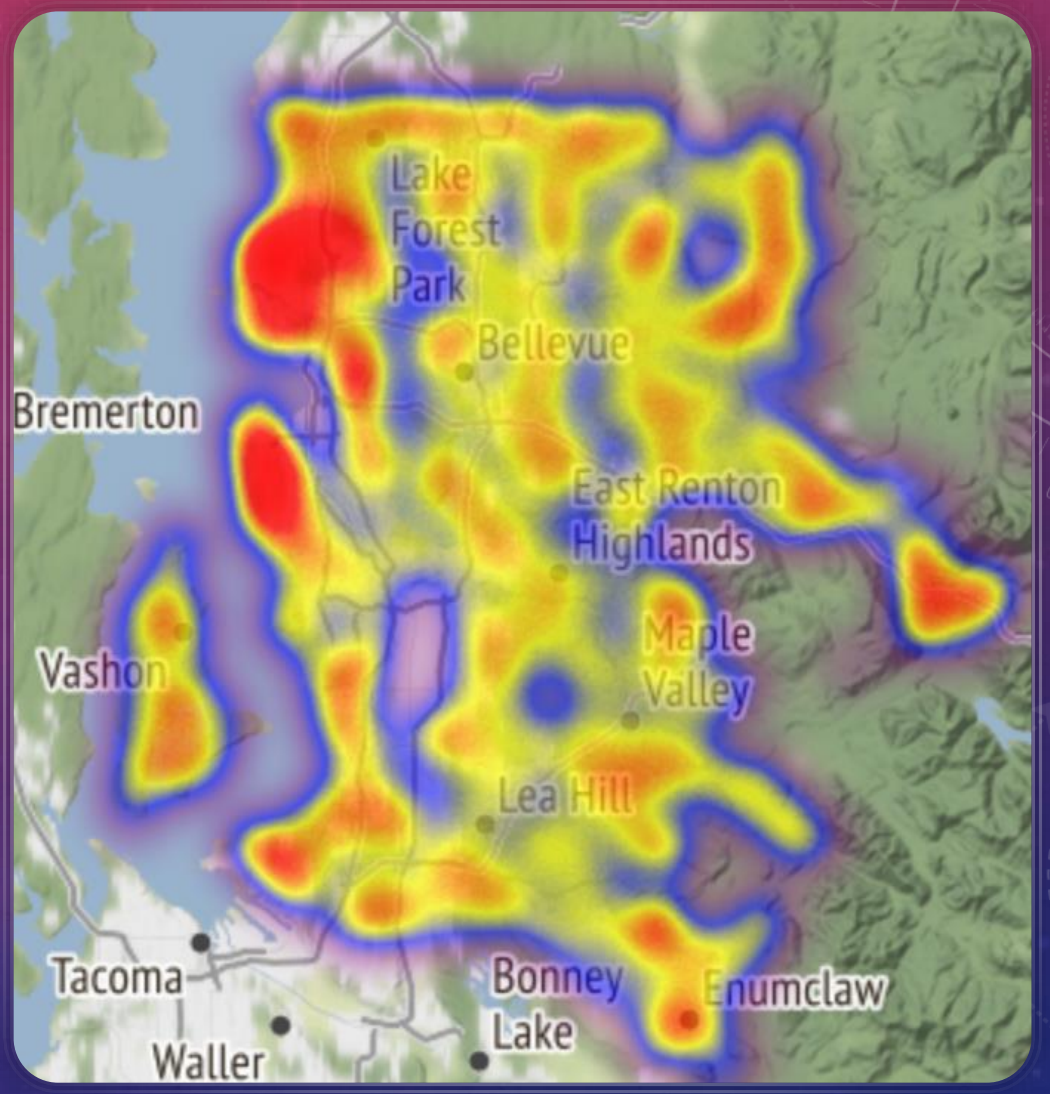
RESULTS 1 – AVERAGE PRICE VS TIME



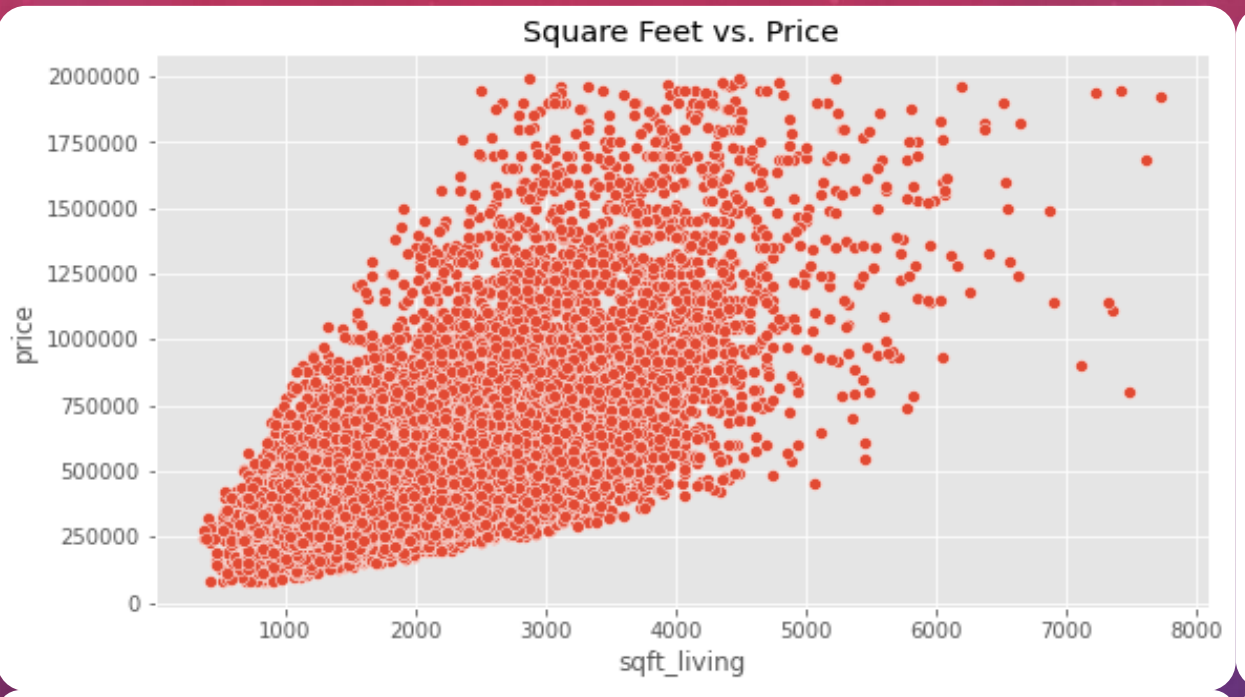


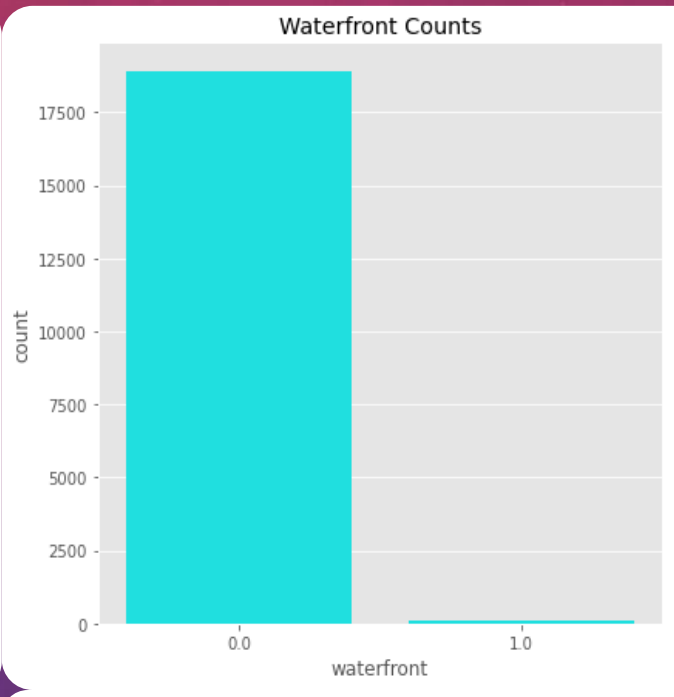
RESULTS 2

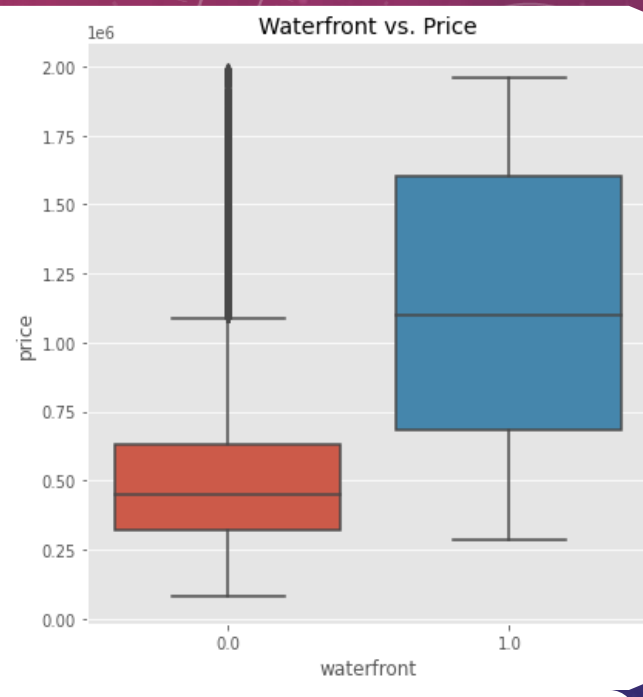
- This map illustrates the price differentials by location. Zipcode_avg_price served as a representation of this in the model.
- Future work could include looking at pricing trends since 2015 to identify what neighborhoods have upwardly trending prices.

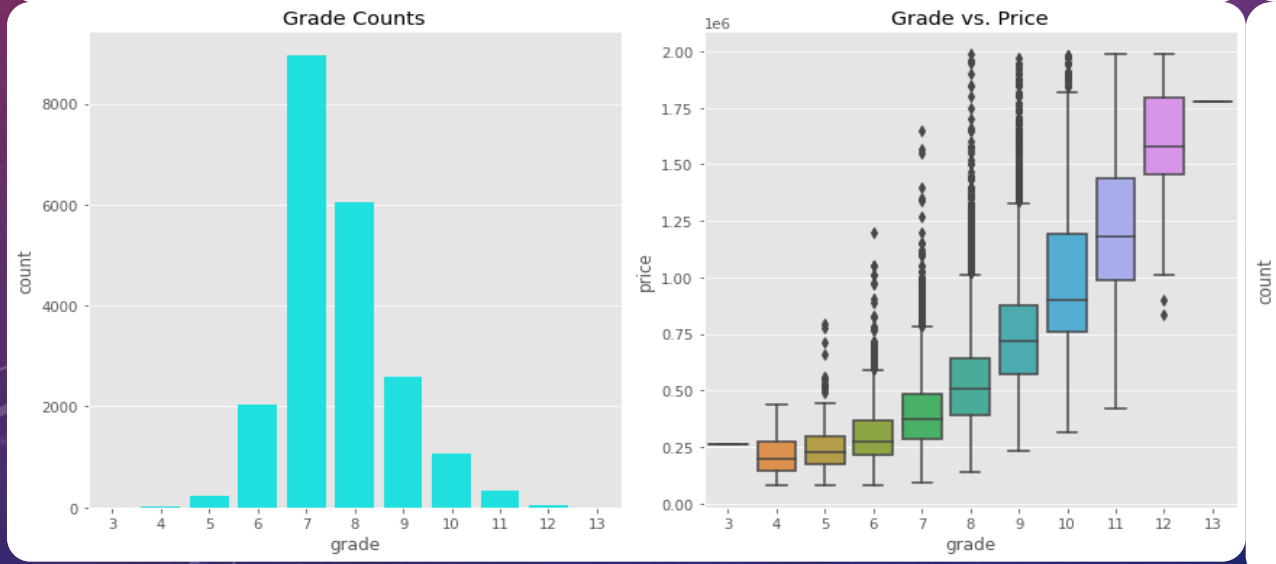


RESULTS 3

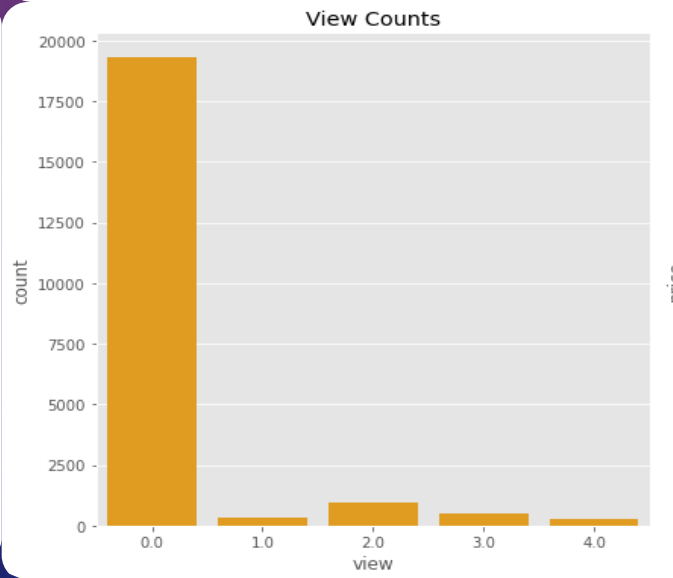

Square Feet vs. Price
price
sqft_living
2000000
1750000
1500000
1250000
1000000
750000
500000
250000
0
1000
2000
3000
4000
5000
6000
7000
8000

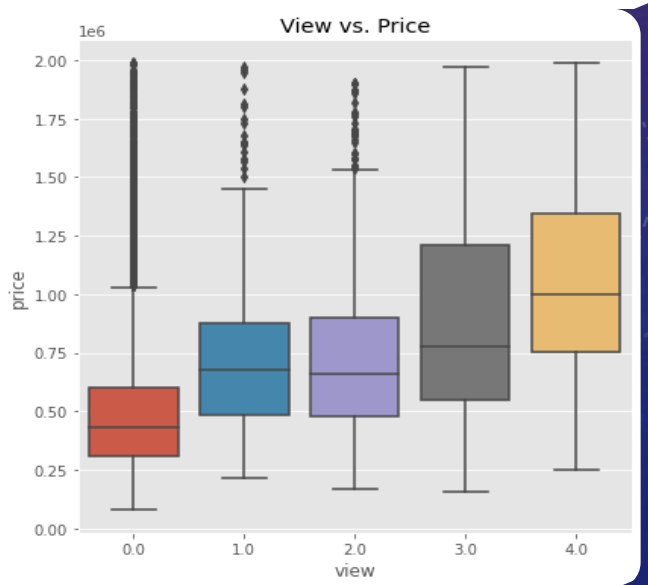

Waterfront Counts
count
waterfront
17500
15000
12500
10000
7500
5000
2500
0
0.0
1.0


Waterfront vs. Price
1e6
price
waterfront
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0.0
1.0


Grade Counts
count
grade
8000
6000
4000
2000
0
3
4
5
6
7
8
9
10
11
12
13

Grade vs. Price
1e6
price
grade
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
3
4
5
6
7
8
9
10
11
12
13


View Counts
count
view
20000
17500
15000
12500
10000
7500
5000
2500
0
0.0
1.0
2.0
3.0
4.0


View vs. Price
1e6
price
view
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0.0
1.0
2.0
3.0
4.0

CONCLUSIONS/QUESTIONS

- Linear regression model has identified some characteristics to dial in on.
  - As stated earlier, the top 6 coefficients of linear regression, in descending order, are: zipcode_avg, view_3.0, view_4.0, waterfront_1.0, sqft_living, and grade_squared.
  - This model had an R^2 value of 0.8241. Compared to a ceiling of 1.0, and maintaining good interpretability of predictors, this is a strong result.
- Actionable items:
  - Strategically time home sale/purchase during the year and even within the week
  - Examine locations/zipcodes carefully. In the future, could project zipcodes with desirable growth
  - Improve grade/square footage of current home to increase its value
  - Identify future homes for purchase where grade/square footage are not yet maximized
  - Use linear regression or other model to investigate current / prospective home price points.
- Questions?