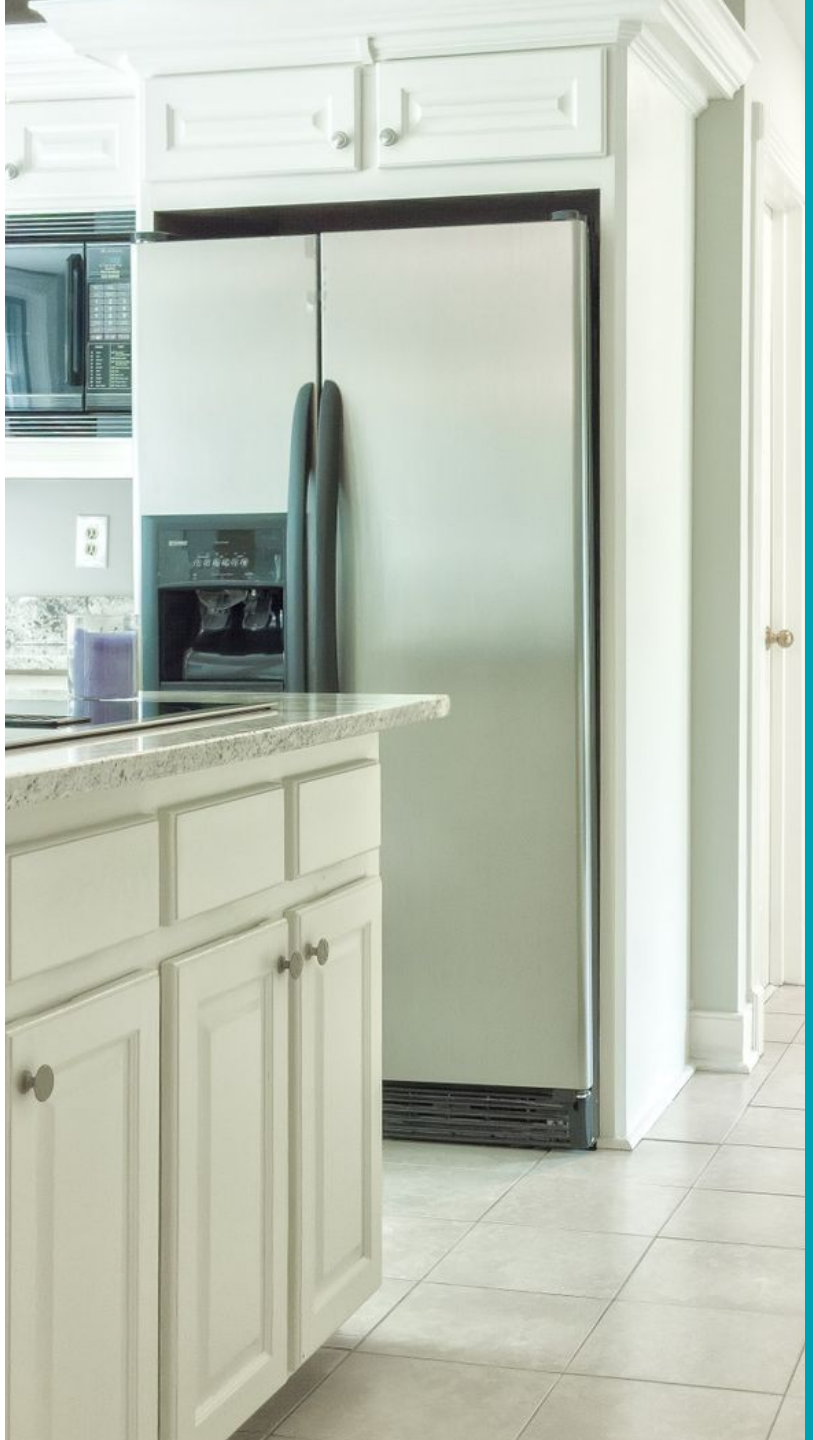


King County, WA

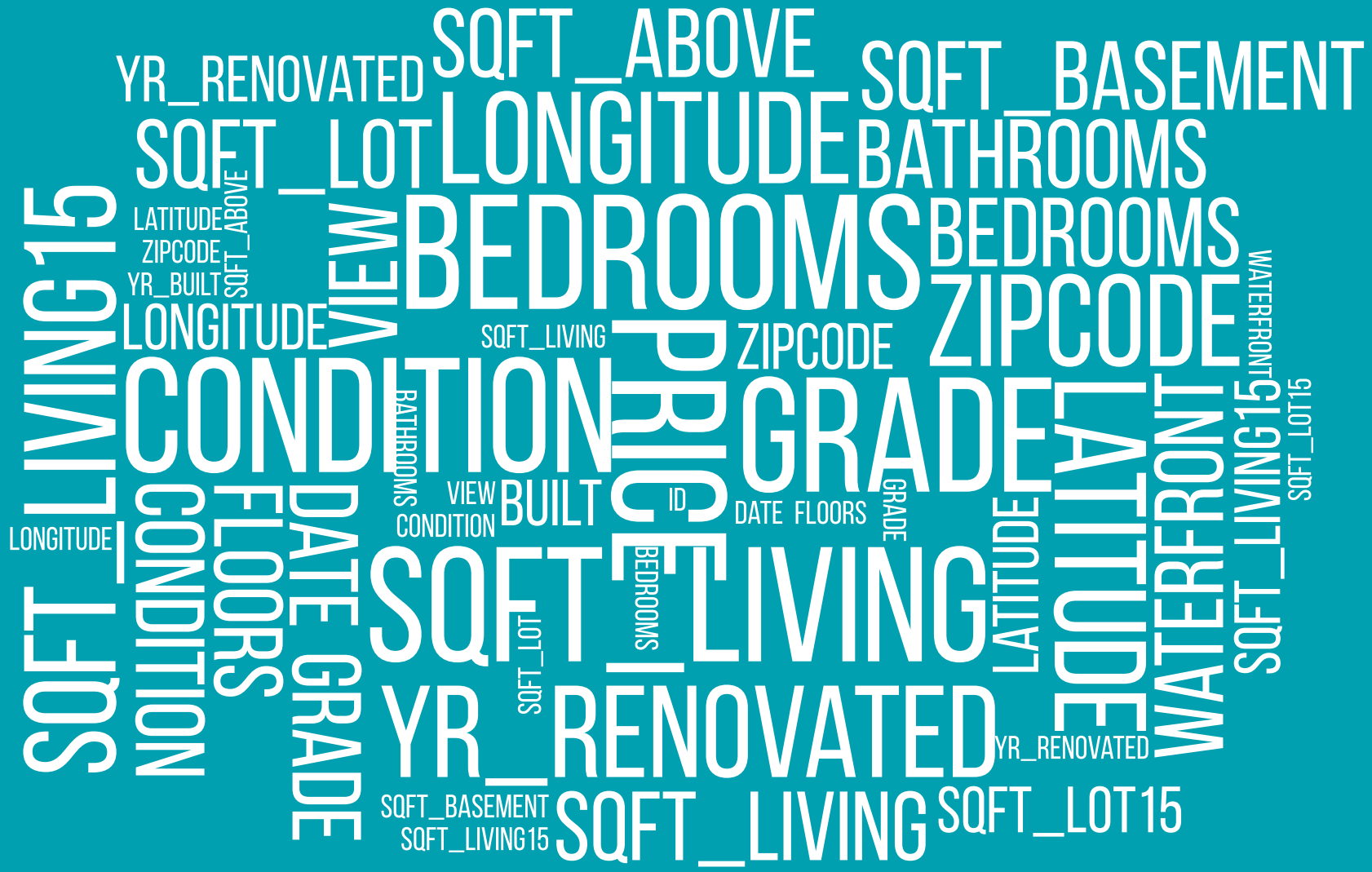
Statistical Analysis of House Sales Prices

Joan Leonard-Short

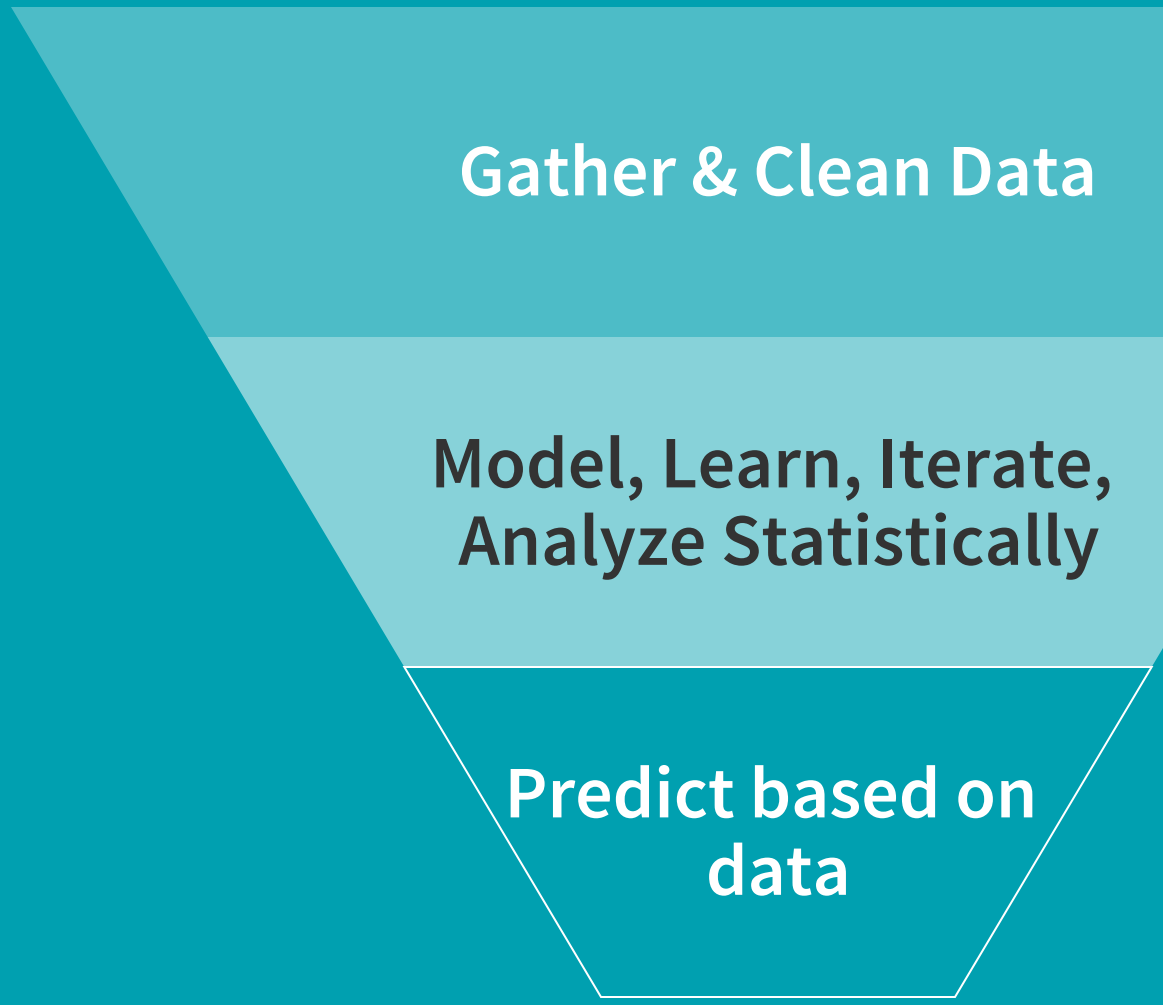
Agenda

- Price Drivers Examined
- Methodolgy
- Results based on data analytics
- Next steps

Initial Price Drivers Analzyed



Methodology



King County real estate database

Multiple regression analysis, multicolinearity, categorical vs. number data, log transformation, normality, train_test, X-validate

Iterate models and determine best drivers of sales prices

Regression Results

(Features of the House)

Items not statiscally relevant 13 out of 19 features

Multicolinearity (7 items)

Sqft_above, sqft_lot, sqft_basement, sqft_lot15, sqft_living 15

Determined to be mulicolinear with sqft_living

Baths _to_beds and grade were also colinear

Not relevant (2 items)

id, date - did not add reasonable data to the analysis

yr_renovated - not enough data to be relevant

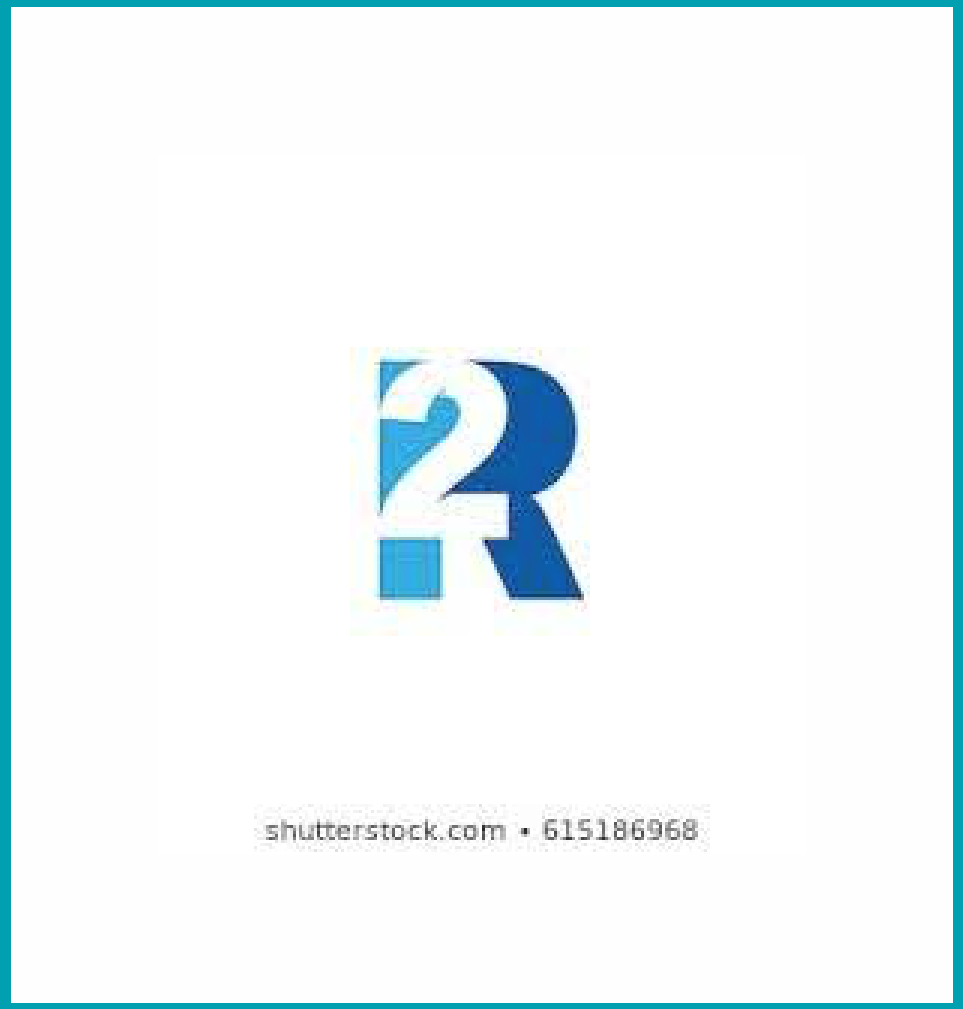
Geo-Spacial (3 items)

latitude, longitude and zipcode were excluded and analyzed using other tools

Other

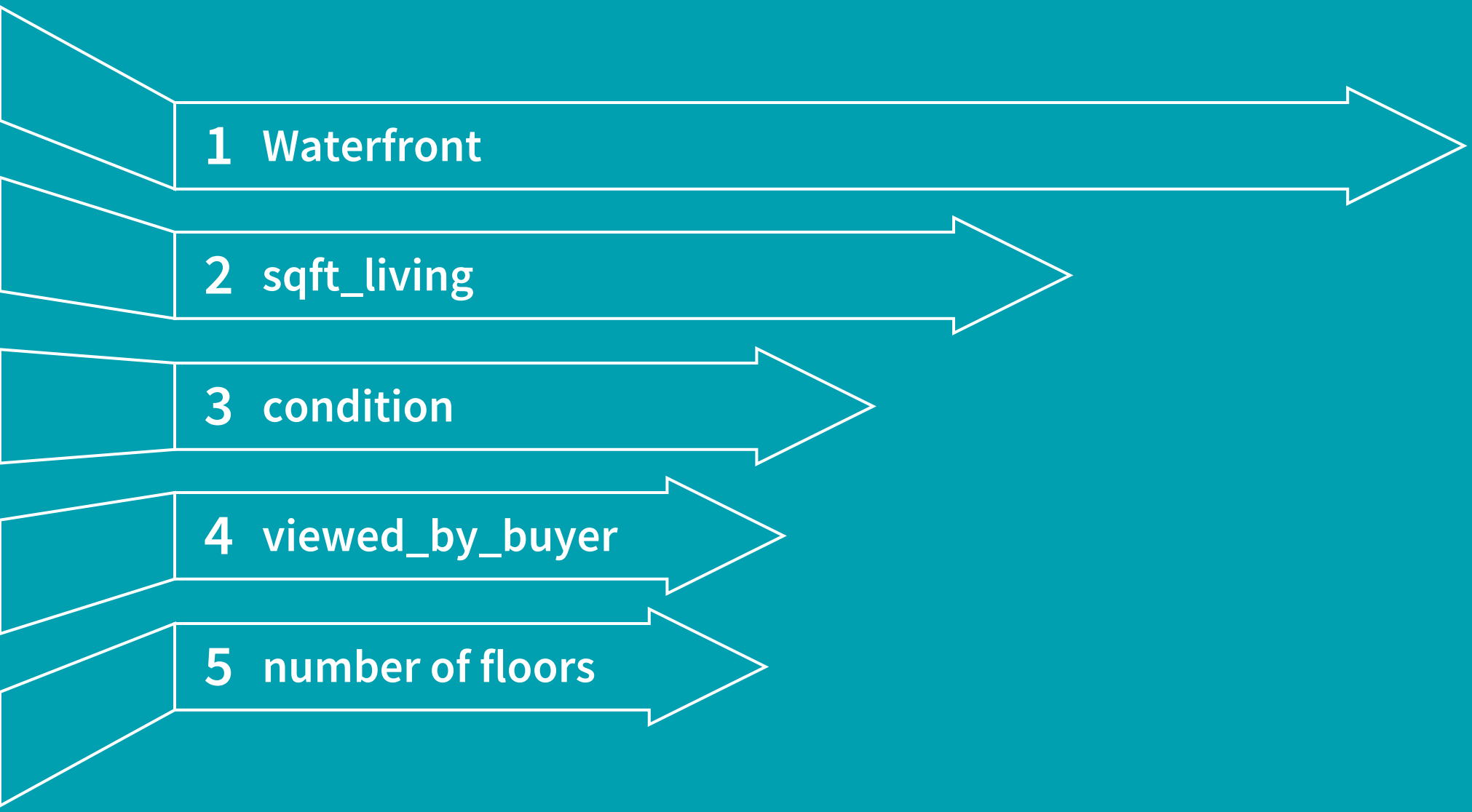
bedrooms

Regression - The R2 Factor



- R2 is the overall score of the remaining data. It tells us if the remaining data is significant and reliable enough to make detailed predictions
- Our testing showed a 40.2 R2 which means that the values generated from the model are 40.2 percent reliablie. The data may be used for directionality as it relates to price but probably not actual dollar amounts

Features that tend to drive house prices higher

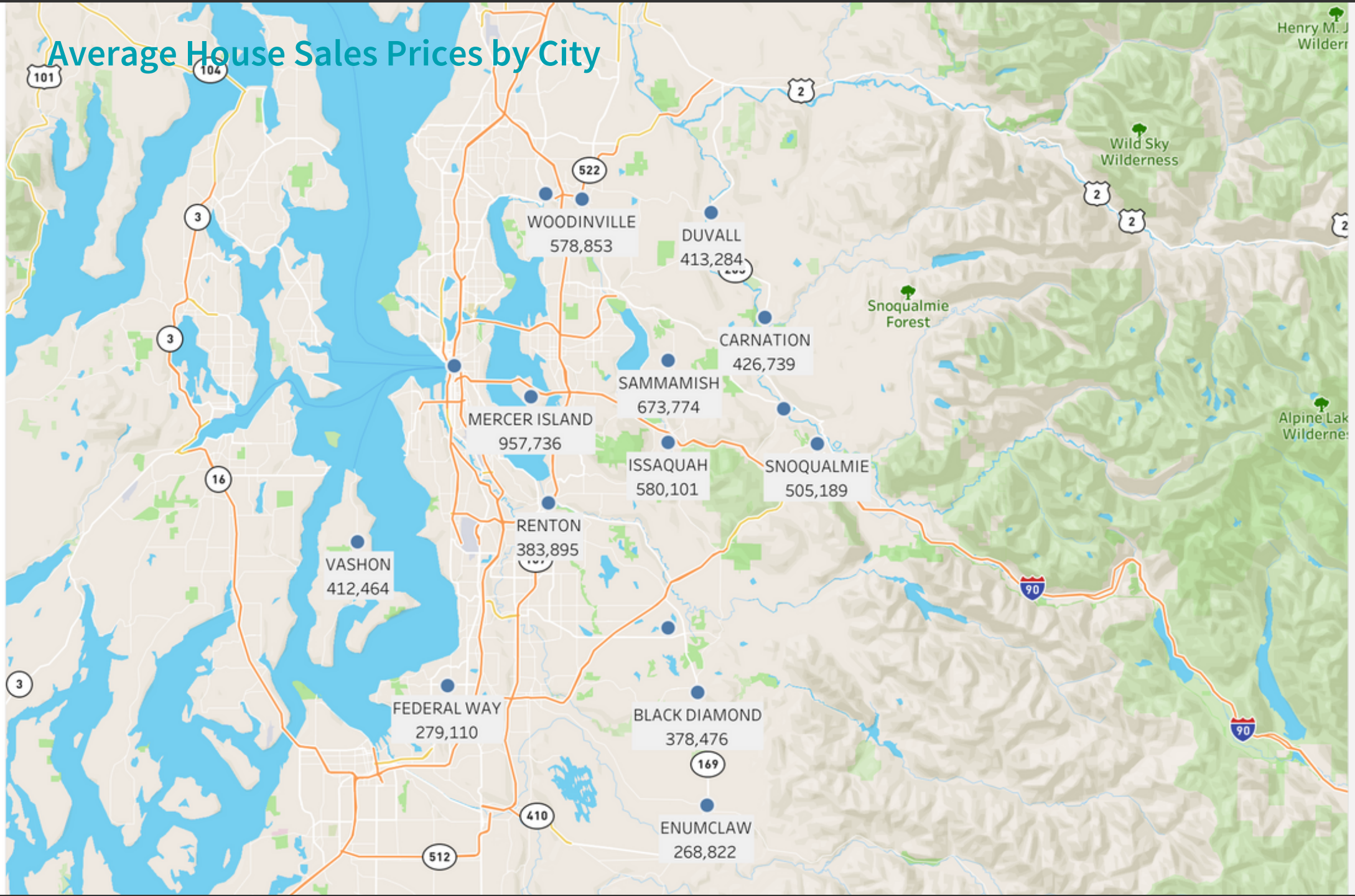


Feature that tends to reduce house prices

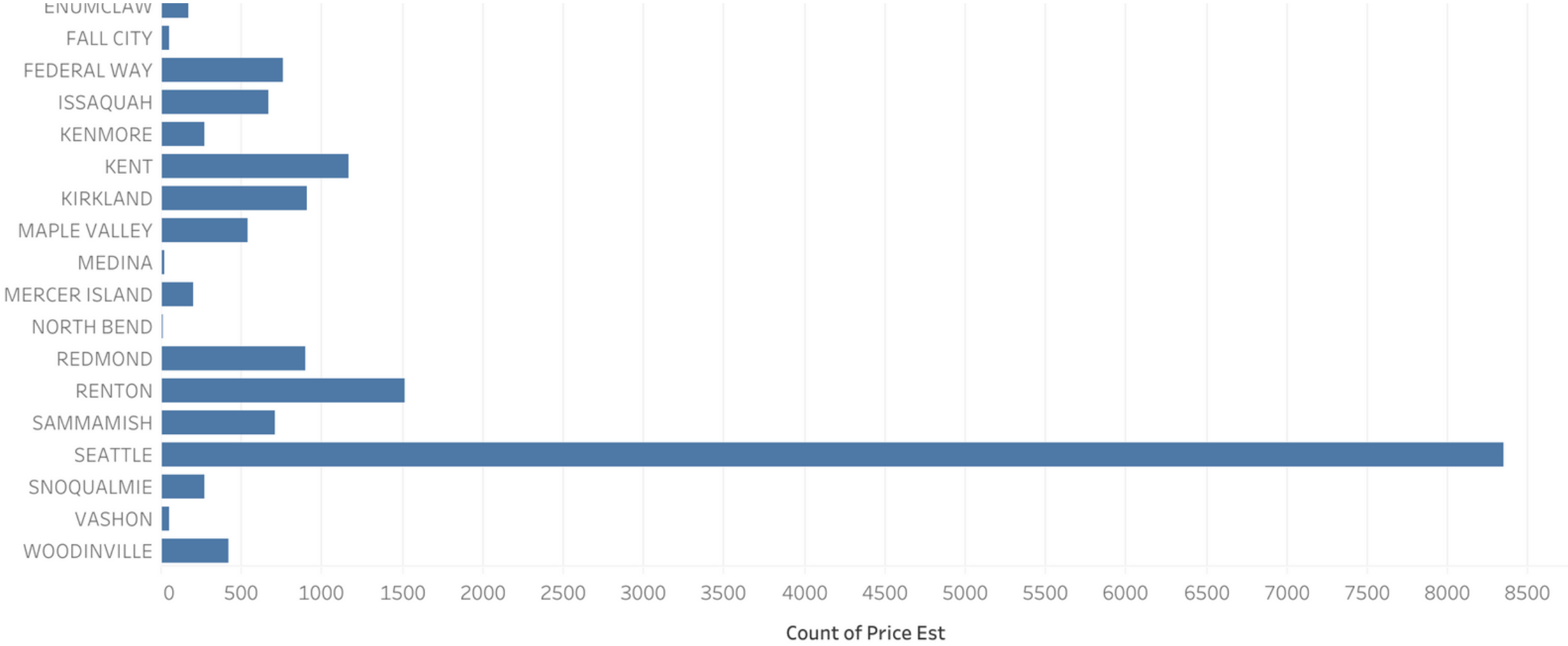
Geo-Spacial

(Where you live)

Average House Sales Prices by City



Number of House Sales by City



Next Steps

- Further analysis of overfitting
- Gather more of the current data to improve the model
- Split the regression into 2 sections - 'Seattle' and 'All Other'

Thanks!

Any Questions?

Appendix

Apendix - Contributors

1. Thanks to the flatiron schools staff especially James Irving for helping where needed
2. Photos courtesy of Beautiful.ai
3. My classmates were helpful while working through problems
4. Graphs from Tableau