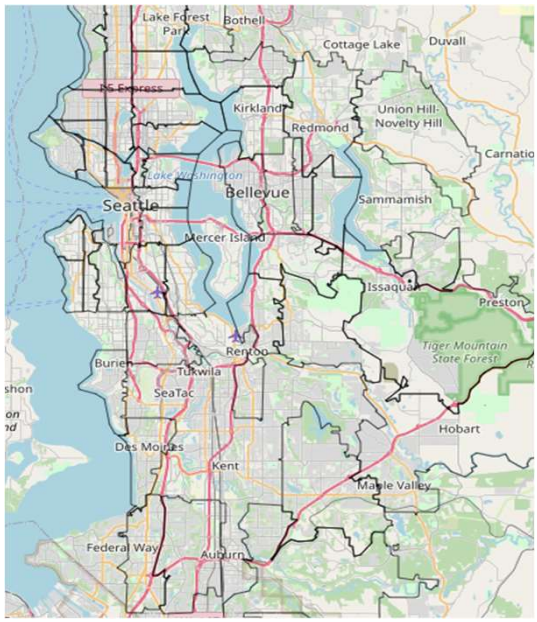
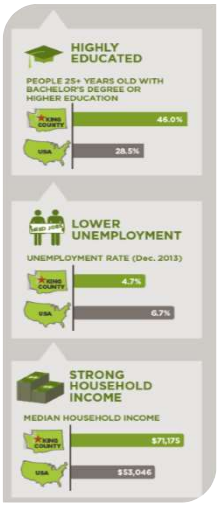
King County, Washington
A great place to buy your first home!



King County, Washington
Zipcodes within King County Outlined

King County, Washington

Reasons to purchase your first home in KC

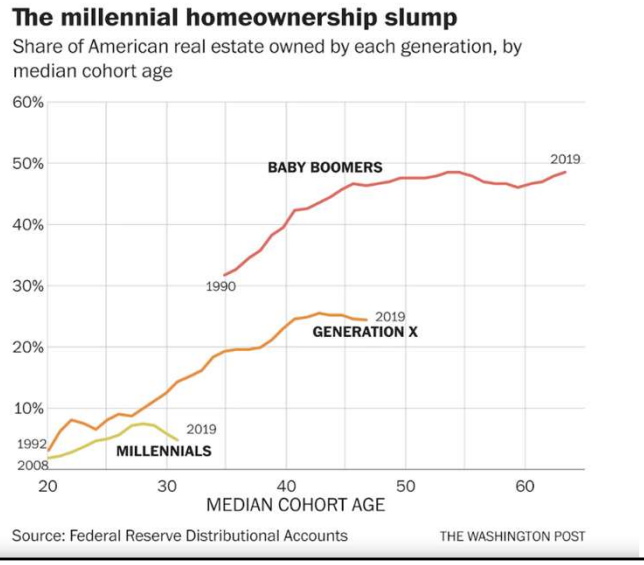


https://www.kingcounty.gov/depts/executive/performance-strategy-budget/regional-planning/Demographics.aspx

We see the location of King County encompasses Seattle and the surrounding area. Several zipcodes make up this area, although diverse, the overall demographics and socio economic indicators are appealing to young families or single people seeking to buy a first home.

A generation of renters?

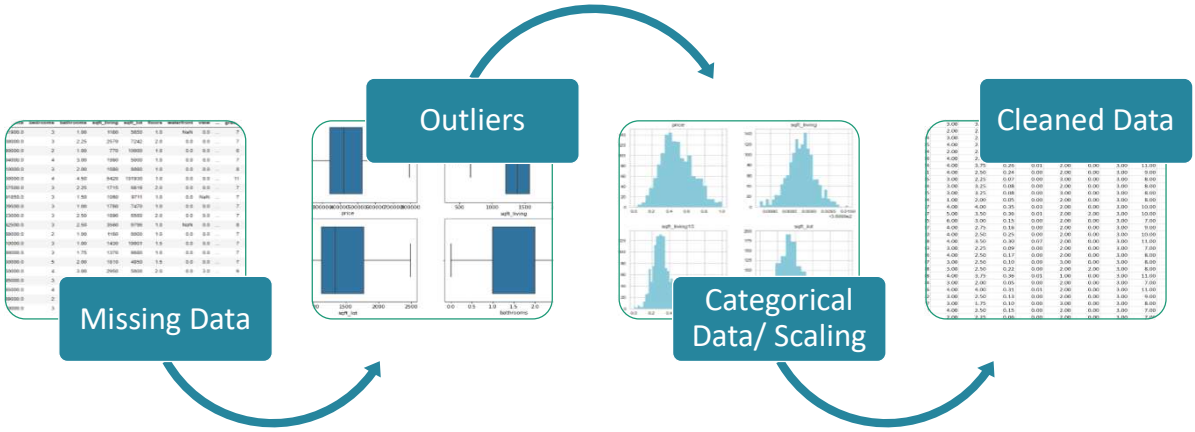
Forming realistic goals can alter that reality



"A data point from the Federal Reserve, highlighted recently in a special report on housing by the Economist, underscores the differences between the financial trajectory of millennials and those of earlier generations: When baby boomers hit a median age of 35 in 1990, they owned nearly one-third of American real estate by value. In 2019, the millennial generation, with a median age of 31, owned just 4 percent." –Christopher Ingraham, Washington Post.

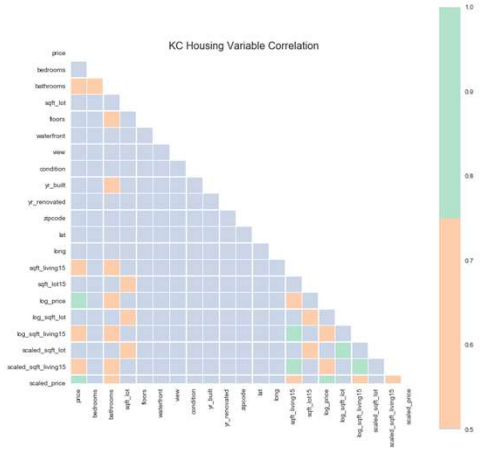
Millenials are reaching an age where they are looking to buy a home but several factors including student loans or sky-high home prices or both are holding them back. In this project the focus is on "affordable" homes between 100 and 800 thousand dollars to create a model capable of predicting a home that is attainable by the vast majority of first time homeowners.

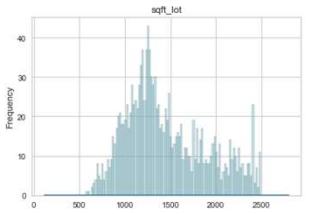
Data Cleaning

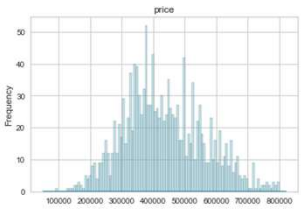


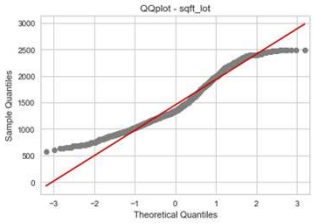
The data cleaning process for the King County House data was approached in a stepwise manner. First

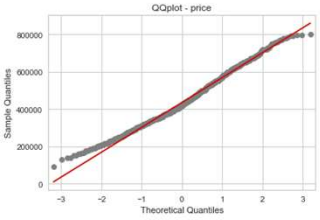
Statistical assumptions







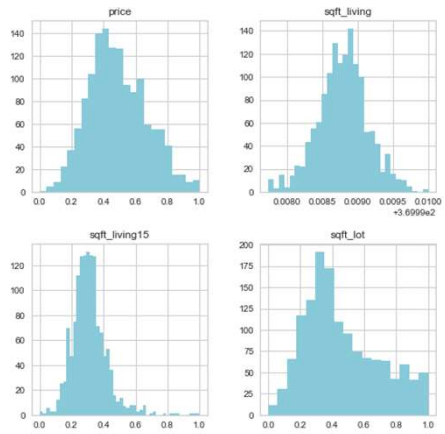




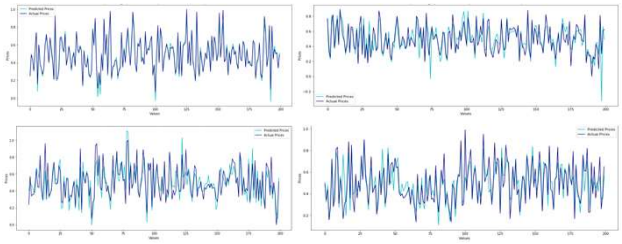
We need to check for basic statistical assumptions for multiple regression. The two primary are normality and collinearity. To check for our data we can use heatmaps and QQ –Plots. The Q-Q plot, or quantile-quantile plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal or exponential. Q-Q plots take your sample data, sort it in ascending order, and then plot them versus quantiles calculated from a theoretical distribution. The number of quantiles is selected to match the size of your sample data. While Normal Q-Q Plots are the ones most often used in practice due to so many statistical methods assuming normality, Q-Q Plots can actually be created for any distribution (UVA Statistical Library).These plots are subjective but are useful as a quick look check. If our data is not normal, we can apply a log transform to help normalize our data, which is what was used on a few of the variables in this dataset. The heatmap shows that some variables are highly correlated, following this check the variable with the highest correlation was dropped.

Modeling the Data

Clean and Engineered Data



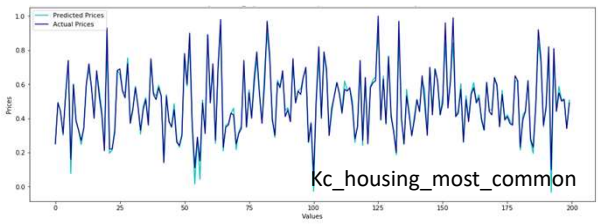
Models Based on Four Methods of Data Cleaning



Once the data has been cleaned and tested to make sure it meets the statistical assumptions of linear and multiple regression, it's time to create our model using all four dataframes. The next step is to split the data into test and train data sets and finally cross validation. We see that each dataframe performed differently when compared to the testing data set likely due to how we treated missing data. Some of the data performs well against the training set while some of the data set is OK. While having a perfectly fit model is great it will likely underperform on real world data. We will discuss this on the next slide.

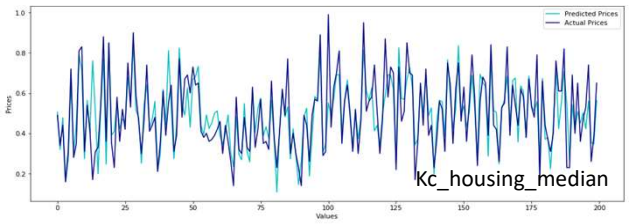
Model Selection: Finding A Balance

Kc_housing_median



R^2 Score: 0.79

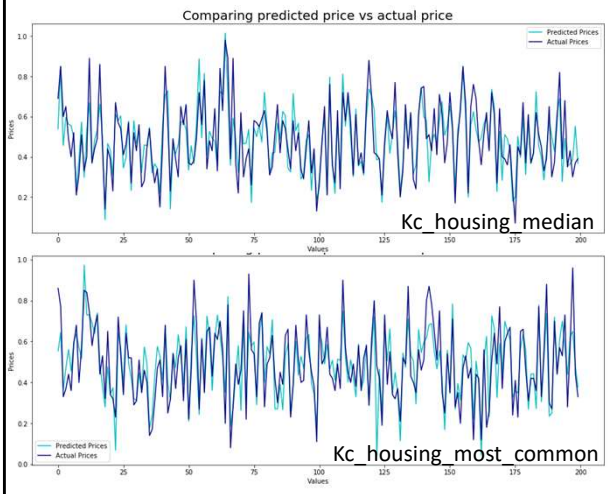
Kc_housing_most_common

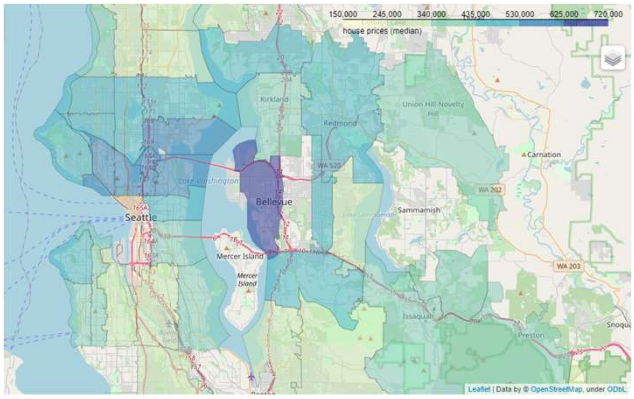


R^2 Score: 0.78

Following our modeling exercise we saw four different models. The best performing model included all of the data and returned an R^2 value of .98. That means we can predict what a house will cost with a 98% accuracy! While that is fantastic this model likely overfit the data and as was mentioned on the last slide will likely underperform on real world data. To solve this, the two models of choice are those that are well correlated but not overfit which is why the kc_housing_median and kc_housing_most_common were chosen in the end and are the two models that predict house price with a 79 -78% accuracy. Next we will evaluate for the factors that impact home price the most in King County on our next slide.

What influences the price of a home in KC?
Location, location, location!





Kc_housing_median

	Features
zipcode_98004	0.753989
zipcode_98053	0.677368
zipcode_98112	0.669149

Kc_housing_most_common

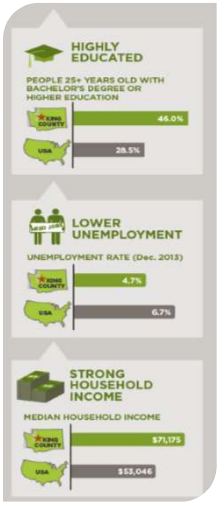
	Features
zipcode_98075	0.765415
zipcode_98004	0.740147
zipcode_98105	0.705379

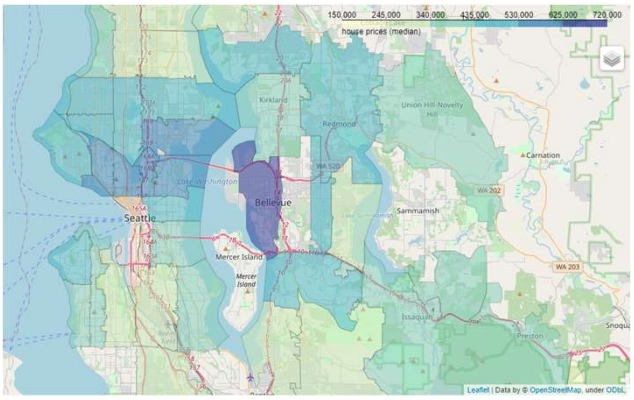
Here we look at the most important features when attempting to predict house price in King County. Between the two models zipcode/location is the key influencer. But why are these locations so influential? Look at the map next to the models, the highest priced homes are also located to large corporations and spectacular views!

Top areas are affluent communities:
Sammamish median income : $183,038
Bellevue hosts a number of satellite offices for large technology companies such as eBay, Oracle, Salesforce, Google, median income $104,839
Redmond median income: $78,430
Seattle median income: $93,500

Conclusions and Recommendations

Build equity in King County!





https://www.kingcounty.gov/depts/executive/performance-strategy-budget/regional-planning/Demographics.aspx

At this point we have been able to deduce from a large amount of data the accuracy in which we can predict a home's price in King County with a fair amount of accuracy. What we've found is that the most important feature is zipcode and therefore location. These zipcodes likely govern price in King County because of certain economic factors, largely the fact that major corporations are within or near these zipcodes. So, while you may not necessarily work for a large corporation in this area, it is recommended to seek homes in these areas as they have strong fundamentals that are popular and therefore would likely lead to increase in demand over the long term. With all these factors not only would you live in a healthy neighborhood, but your journey to financial health would also be aided by accruing equity!

Thanks and good luck!

Thank you for your time! I hope that this exercise has proven useful to allow you a better understanding of King County and what drives house prices in that area. Whether you are in the market or just starting to look, the King County market is full of options for those looking for a reasonably priced home!

Future Work

While this exercise provided us with several models, these models focused on a subset of the whole data set. As a reminder, this project focused on homes at a sale price between 100-800 thousand dollars in order to provide a model that is most accurate for the 'average' buyer. A future exercise could be done for even more subsets to tailor to specific price ranges or 'classes' of buyers such as the amount they are pre-qualified for.