Capstone Project 1

King County Houses Final Report

Jennyfer Vu

Problem Statement

For most people, becoming a homeowner is one of the greatest accomplishments in life. It is a lifelong dream for many. For potential homebuyers, the search and transaction process can be tricky. It is a delicate game of offers, counter offers, and bidding wars. How does a buyer know if they are offering too little or too much to a seller? The buyer would like to offer the lowest amount possible to win the house. The seller would like to receive the highest offer possible to sell their house. How many bedrooms, bathrooms, and square feet make up the purchase price of a home?

My client will be potential home buyers in King County, WA.

If a potential homebuyer is equipped with housing market knowledge, they will be able to make confident offers knowing the true worth of a property.

I will be using data from King County GIS data portal.

Data Wrangling:

Data preprocessing: After importing necessary libraries and reading the csv dataframe through pandas, we take a closer look at the dataset. Using the .shape[ ] function, we see that this dataframe has 21,613 rows and 21 columns of data--quite a large dataset! With the .info( ) function we can see what type of data we are dealing with. There are integers, floats, and objects. The .head( ) function lets us take a sneak preview at what the dataset looks like row x columns.

Missing values: In dealing with missing values, we replace NaN with 0 or 1.

Outliers: There are a few outliers in this dataset that we need to treat. For example, number of bedrooms range from 0-33. It is unlikely that most houses will have 33 bedrooms. We can remove unwanted columns or rows with .dropna( ) function.

Data cleaning: After inspecting the data and seeing what type of data we are working with, it is time to whittle down our dataframe.

Objectives: During data wrangling, the most important goal is to keep in mind our hypothesis. What can we ask of the data? What story can we weave from this information? Going back to our problem statement, we want to create a prediction model in order to better inform and equip homebuyers in King County, WA.

Looking at the different variables, we will take a deeper dive into the relationship between bedrooms, bathrooms, square feet, waterfront, condition, zipcode, latitude, longitude and the influence they have on price of a property.

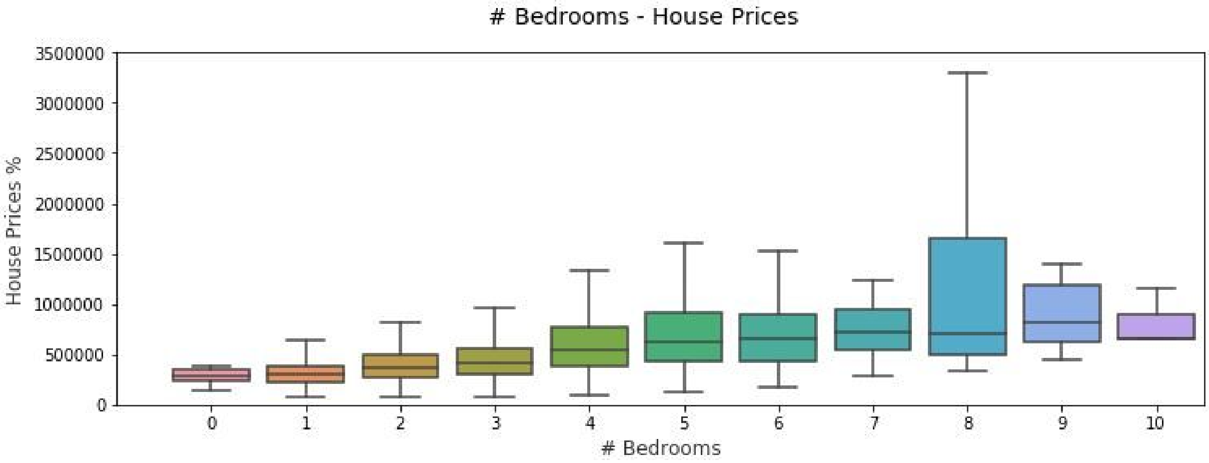
We will also be comparing the distribution of zipcodes (98001-98199) across King County to see how zipcode density differs in areas.

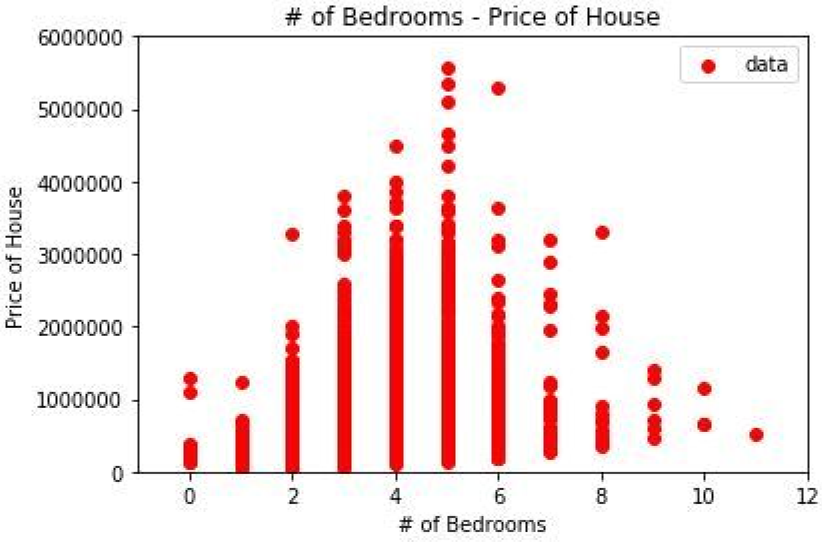
Finally, I will be exploring more in depth the distribution of location (zipcodes) vs. house prices to see just how much location matters when it comes to purchasing a home.

Exploratory Data Analysis:

https://github.com/vujennyfer/capstone1/blob/master/capstone_one_data_wrangling(3).ipynb

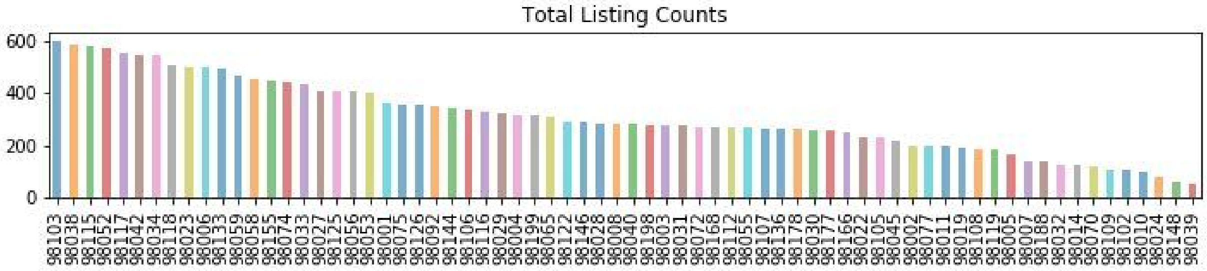
After we have cleaned our data, it is time to explore and visualize the relationship between the variables we are comparing. First, let's explore how number of bedrooms will affect the price of a property.





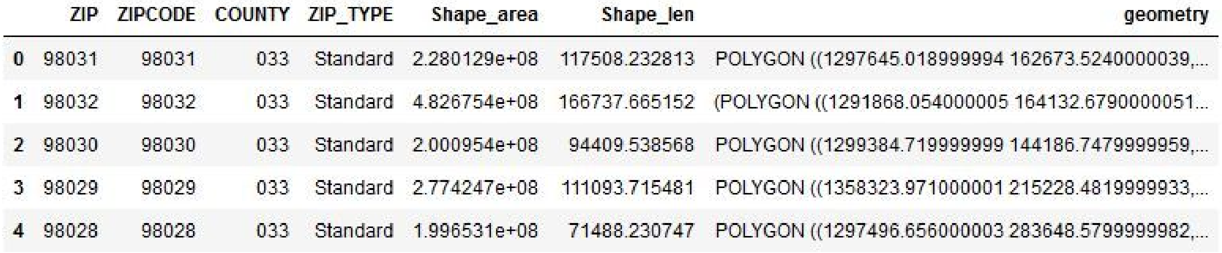
The above boxplot and scatter plot shows that there is a linear correlation between number of bedrooms and house prices until 5 bedrooms. After 5 bedrooms, the results vary widely whether or not bedrooms play a key factor in how expensive a house is. Outliers were removed with "showfliers=False" and axes were adjusted to show main data.

Next, we take a look at the the distribution of properties in King County, WA by zipcode. By using the .value_counts( ) function on the zipcode column, we see that there are max 602 houses located in zipcode 98103 and min 50 houses located in zipcode 98039.



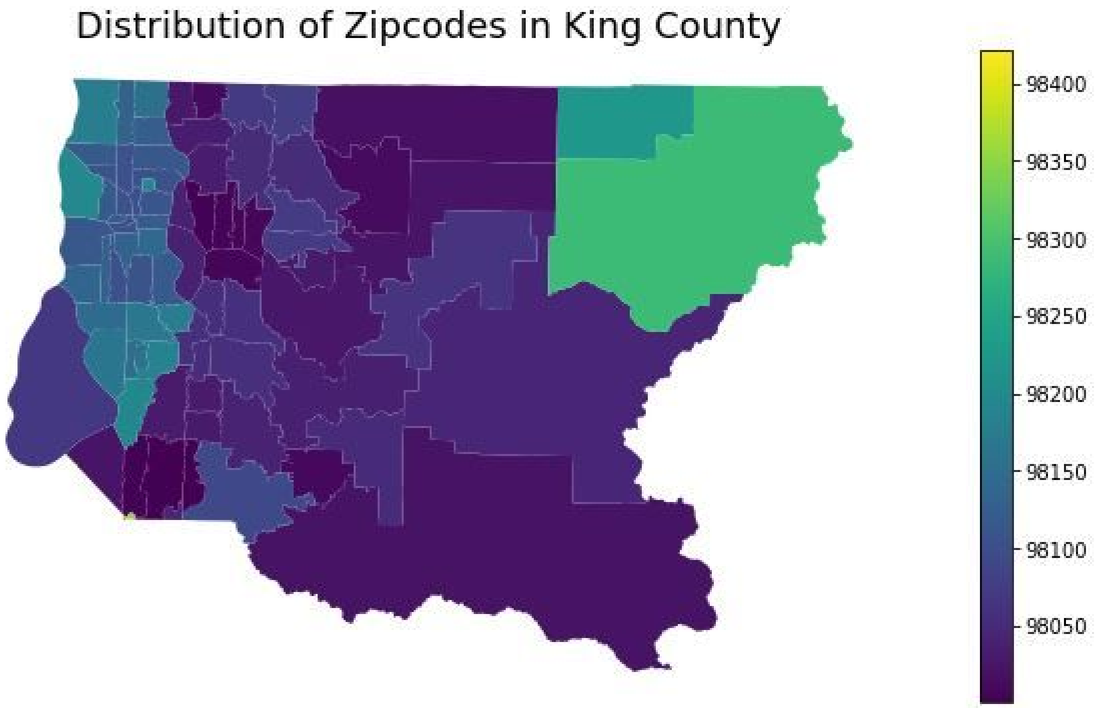
Above is a visual representation of the total listing counts in each zipcode.

After we have explored the csv dataset, we will now import the shapefile dataframe for further investigation and mapping. Using geopandas, we read the shp dataframe and take a look at the data with .head( ) function.

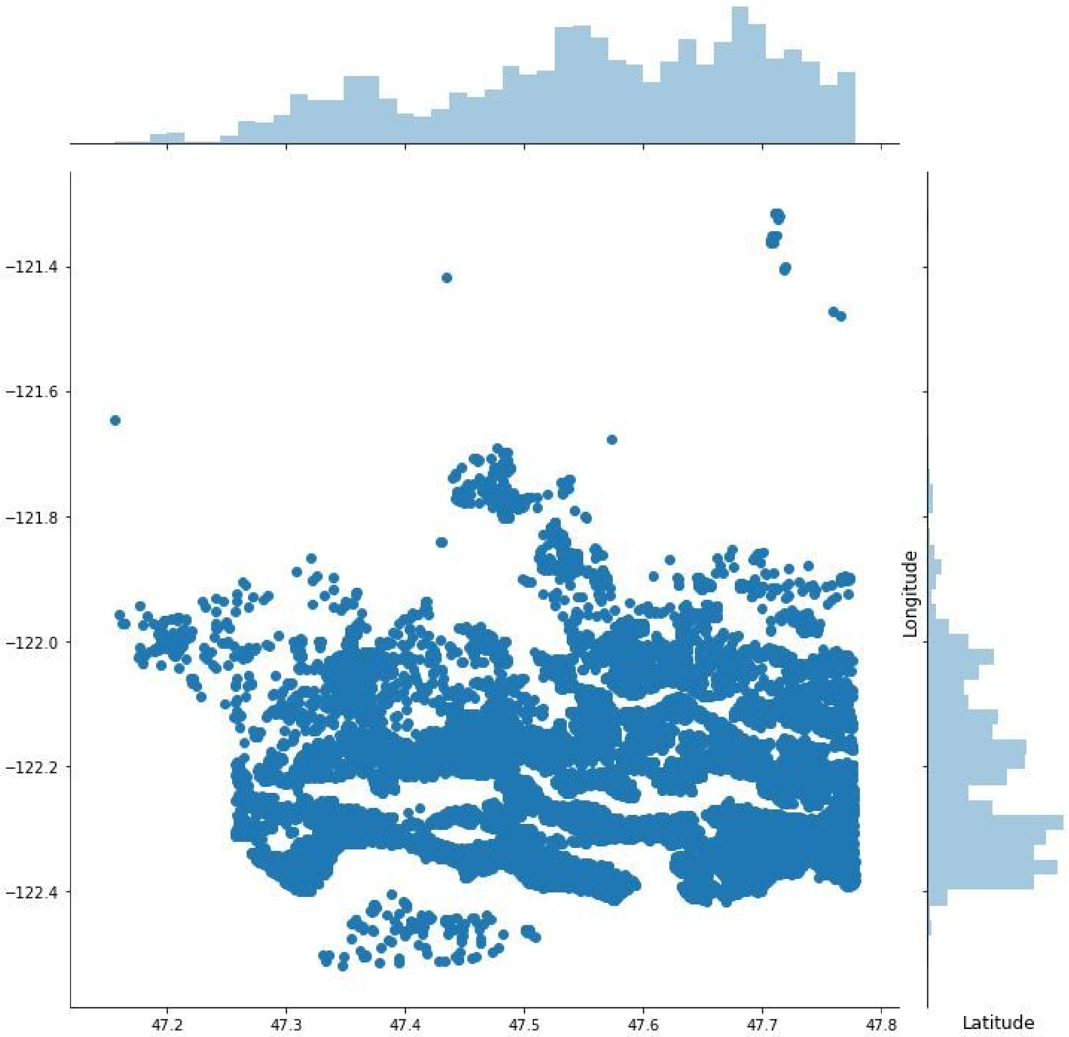


	ZIP	ZIPCODE	COUNTY	ZIP_TYPE	Shape_area	Shape_len	geometry
0	98031	98031	033	Standard	2.280129e+08	117508.232813	POLYGON ((1297645.018999994 162673.5240000039,...
1	98032	98032	033	Standard	4.826754e+08	166737.665152	(POLYGON ((1291868.054000005 164132.6790000051...
2	98030	98030	033	Standard	2.000954e+08	94409.538568	POLYGON ((1299384.719999999 144186.7479999959,...
3	98029	98029	033	Standard	2.774247e+08	111093.715481	POLYGON ((1358323.971000001 215228.4819999933,...
4	98028	98028	033	Standard	1.996531e+08	71488.230747	POLYGON ((1297496.656000003 283648.5799999982,...

Next, we will plot zipcodes in King County.

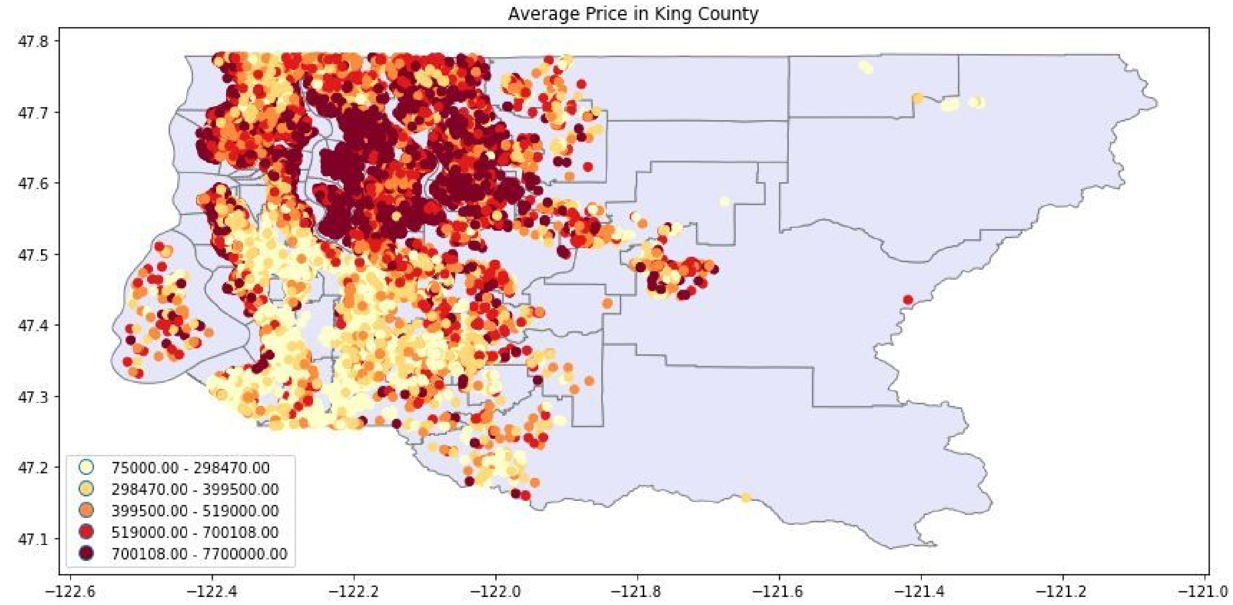


King County map with legend that shows the highest concentration of homes are located in the dark blue region of zipcode 98200 and below. The map above demonstrates that the highest concentration of homes is located closest to metropolitan Seattle.

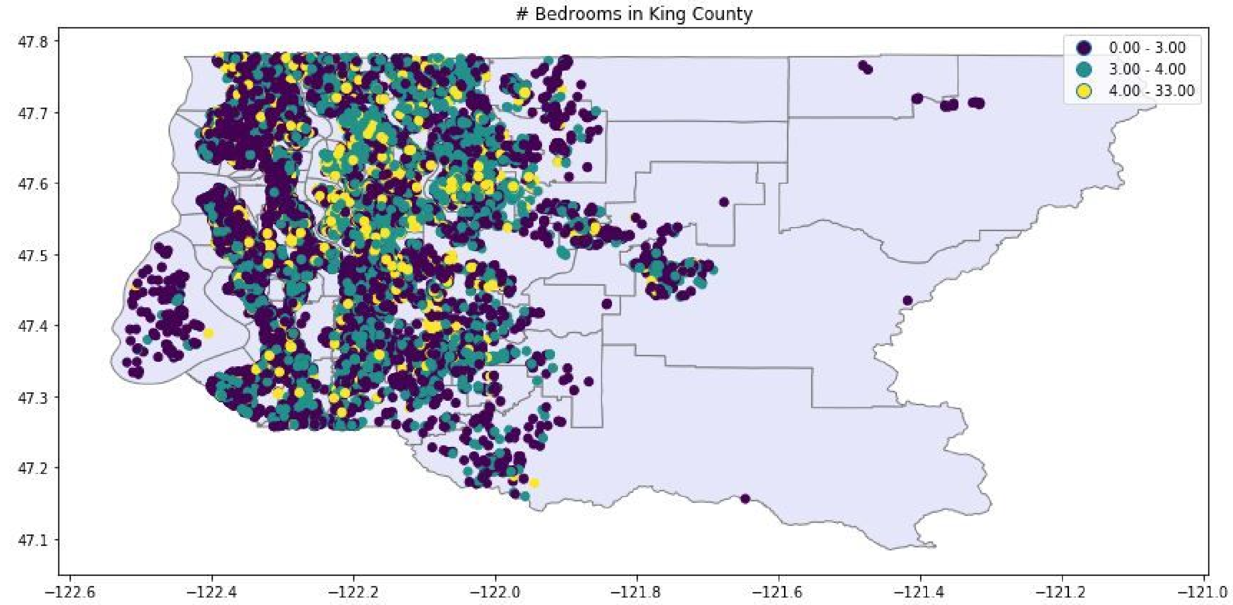


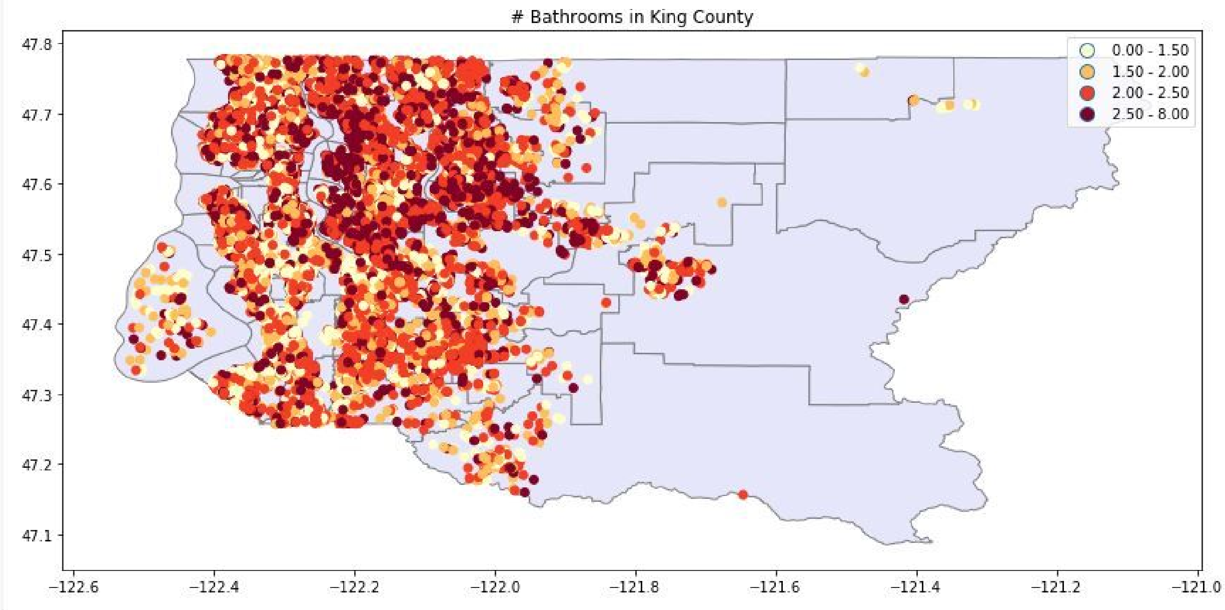
The figure above shows visualization of the location of homes based on longitude and latitude. For latitude between 47.5 and 47.75 and longitude between -122 and -122.4 there is the highest concentration of homes.

In order to further map using shapefiles and geopandas, we need to convert our data to the correct crs which is 'epsg:4326'. After this conversion, we are able to map different features and layer it atop of King County, Wa.

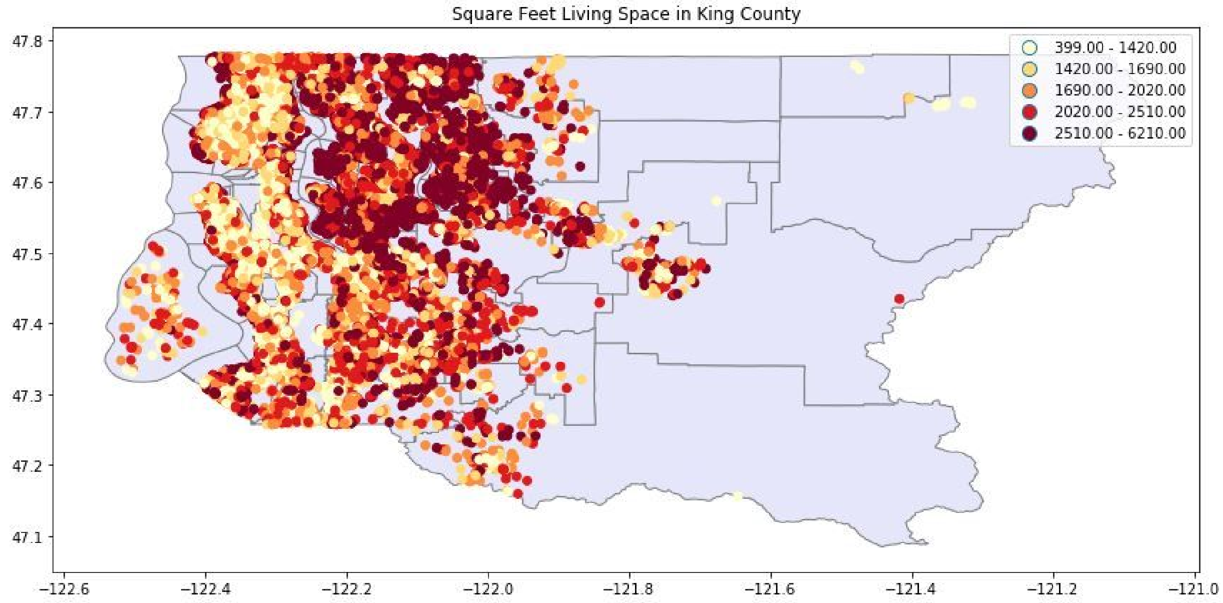


The figure shows the average listing price by latitude/longitude. As you can see, the closer the house is to metropolitan Seattle, the higher the average listing price.

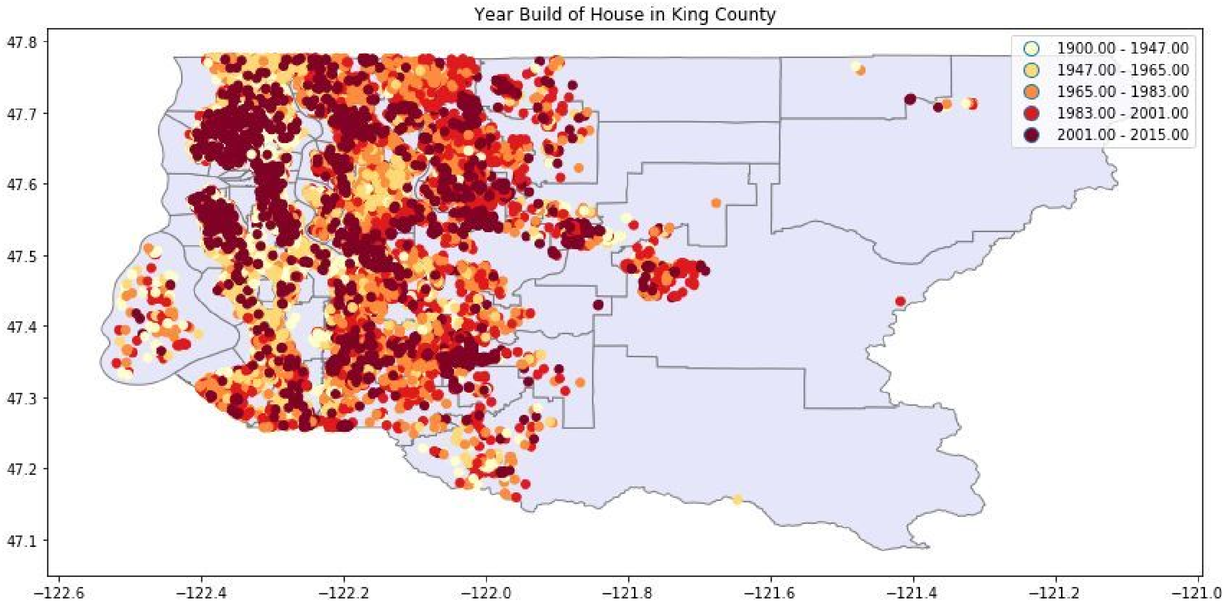




# Bathrooms in King County
0.00 - 1.50
1.50 - 2.00
2.00 - 2.50
2.50 - 8.00
47.8
47.7
47.6
47.5
47.4
47.3
47.2
47.1
−122.6
−122.4
−122.2
−122.0
−121.8
−121.6
−121.4
−121.2
−121.0



Square Feet Living Space in King County
399.00 - 1420.00
1420.00 - 1690.00
1690.00 - 2020.00
2020.00 - 2510.00
2510.00 - 6210.00
47.8
47.7
47.6
47.5
47.4
47.3
47.2
47.1
−122.6
−122.4
−122.2
−122.0
−121.8
−121.6
−121.4
−121.2
−121.0



Inferential Statistics:

https://github.com/vujennyfer/capstone1/blob/master/capstone_one_inferential_statistics.ipynb

After exploratory visual data analysis, we will take a look at what the statistics of the data mean.

Variables such as price of the house, number of bedrooms, bathrooms, square footage, location (zipcode) of a property. All of these factor into the purchase price of a home. By exploring how these variables statistically affect the price of a property, we will be able to better answer the project question to help homebuyers make more confident offers.

Splitting the data into ‘cheap’ (properties below 500k) and ‘expensive’ (properties above 500k), we are able to better understand the pricing of homes. The mean of house prices in King County is $540,182, the mean of ‘cheap’ properties is $338,387, and the mean of ‘expensive’ properties is $817,435.

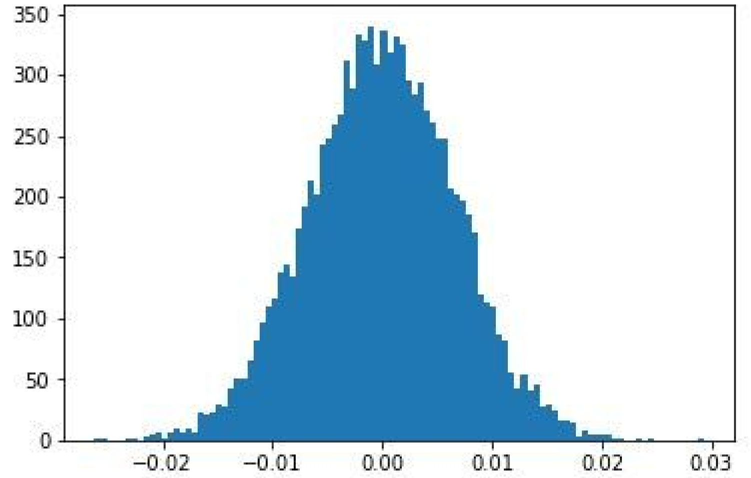
Pearson Correlation:

In order to test correlation between average property price & location We must conduct a hypothesis test, here we will use Pearson Correlation.

Null hypothesis: King county listings have the same mean price in each zip code.

Alternative hypothesis: King county listings do not have the same mean price in each zip code.

The observed correlation is -0.05316819852916175 and the p-value is 0.00. Because the observed value is smaller than the p-value, we reject the null hypothesis and accept the alternative.



For significance level α = 0.01, the 99% confidence interval of the correlation coefficient is (-0.017309107324504743, 0.017257484252865236)

As the testing above shows, the observed correlation is -0.05316819852916175 which is lower than the p-value of 0.00 The correlation coefficient should also be within the confidence interval and it is beyond that of: (-0.017376812656229552, 0.01766896799684194). We can also see this visually from the above histogram.

This means that we reject the null hypothesis and that there is statistical significance between price of a property and location. This coincides with the possibility that properties closer to convenience cities are more expensive.

There appears to be a strong correlation between average price and location of a property. The location of the property is the independent variable and the price of the home is the dependent variable. From the data visualizations that we have done, it appears that the closer the property is to Seattle metropolitan area, the higher the value of the home. This suggests that homebuyers are willing to pay a higher premium for convenience to the larger city for work, tourism, restaurants etc.

Frequentist tests:

We should use the t-test because the sample size is 10 (n < 30). We generally use z-tests when we have a large sample size (n > 30), when we know the standard deviation, when samples are drawn at random, and when the samples are taken from an independent population. Alpha level: 0.05

The calculated t-test p-value is 0.00129 which is lower than the alpha value 0.05. This means it is statistically significant and we reject the null hypothesis. This suggests that the sample size of 10 mean price of homes does not differ significantly from the population mean price of homes.

In-Depth Machine Learning Analysis:

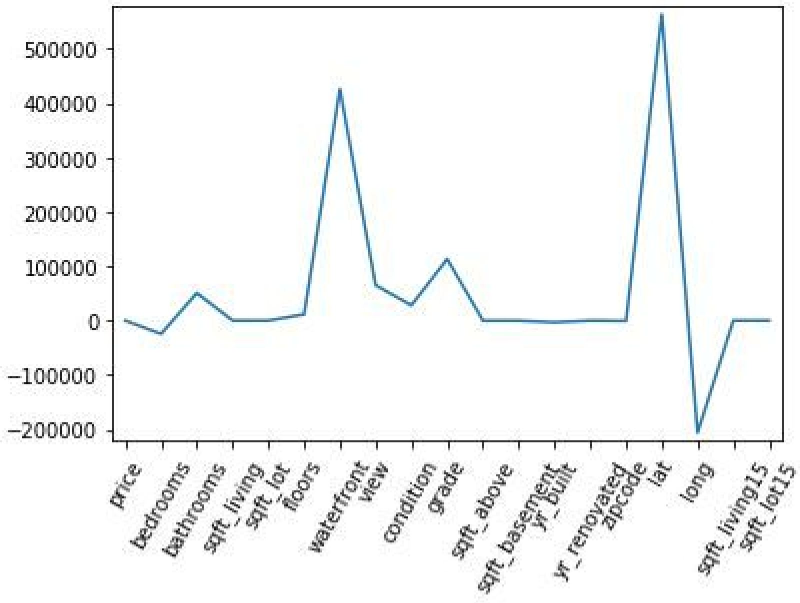
https://github.com/vujennyfer/capstone1/blob/master/capstone_one_machinelearning(1).ipynb

Supervised Learning:
Since this project is looking at how labeled data such as bedrooms, square footage, and area affect price of a house, this is considered supervised learning (labeled or categorized).

1. **Linear Regression**: A simple linear regression model was created to fit the training data and predict house prices. First we create the linear regressor and clean the data. Replacing dates with 1's and 0's as to not influence the data. Using 0 for new houses built after 2014. After splitting the data into 90% train and 10% test set, we are ready to fit our data to the model. We fit the model to the training set and then test the accuracy of prediction on the test set. The linear regression model predicted 73%. This is lower than our aim of 85%.

2. **Gradient Boosting**: In the previous linear regression model, it scored 73% which proved it had room for improvement. Gradient boosting regression is a machine learning model that is constructed from an ensemble of weak prediction models such as decision trees. There are various reasons why linear regression did not score higher. Sometimes data read by a machine can be lost. The model could be a weak learner that needs visualization, such as decision trees. After importing the library and defining the gradient boosting regressor, we fit the training data to the model. Checking for accuracy, we get 91.98% prediction accuracy! It appears that gradient boosting is a brilliant tool for weak prediction models.

3. **Lasso Regression**: Lasso is used to select the most important features in prediction, in this case prediction of house prices. It shrinks the coefficients of features to zero that are not important in prediction. This is useful in this data set in particular because there are various features that could influence price of a property. Lasso chooses them intelligently for you. After instantiating the regressor and fitting it to the data, we print the coefficients and plot them.



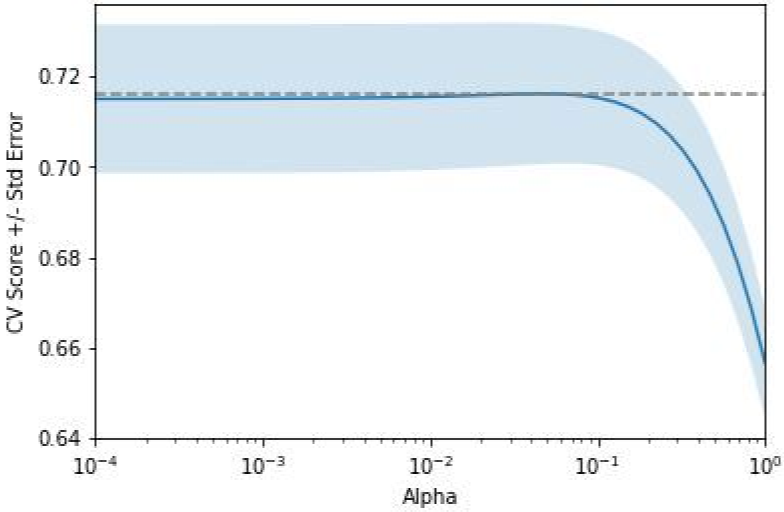
Lasso regression above illustrates that waterfront, lat, long, grade, and number of bathrooms are the most important features selected by lasso regression coefficients for predicting price of a home. It performs regularization by multiplying an alpha value by the absolute value of each function. This is known as L1 regularization.

We then calculate the R^2 for the training set 0.69 and test set 0.74 which shows that the model was able to learn from the training set and predict a higher accuracy on the test set. Whereas R^2 is a relative measure of fit, RMSE is an absolute measure of fit. RMSE was calculated at $195,929 and can be interpreted as the standard deviation of the unexplained variance. Lower values indicate better fit and is extremely useful to see how well the model predicts the response and fit.

4. **GridSearch CV:** After performing lasso regression, we attempted to hyperparameter tune to improve our model. R^2 for training set decreased to 0.70, R^2 for test set decreased to 0.73 and RMSE increased $197,976. The best alpha determined by GridSearchCV() is 0.78

5. **Ridge Regression**: Lasso is great for feature selection, but when building regression models, ridge regression is highly robust and should be your first choice. Ridge regression takes the sum of squared values of the coefficients and multiplies it by an alpha value. This is computing the L2 norm. R^2 for the training set is 0.70 and R^2 for

the test set is 0.73 which is the same as the GridSearchCV. RMSE increased to $197,990 and alpha value is 0.004.



Displayed above, we have fitted ridge regression models over a range of different alphas plotted against cross-validated R^2 scores.

6. **ElasticNet**: This model is a linear regression that is a combination of L1 and L2 regularizations. After importing necessary library ElasticNet, we set up the array and ridge regressor. Fit the model to the data and calculate the R^2 for training data is 0.70, R^2 for test data is 0.73, and RMSE is $197,982. The best hyperparameter for alpha is 1.0

Very similar results for R^2 and RMSE were obtained for Lasso, Ridge Regression, ElasticNet. Linear regression appears to have the best results for model prediction. Lasso was interesting because it visually illustrated what features are most important in predicting house prices. The smallest alpha calculated was 0.004 in Ridge Regression, and the largest alpha calculated was 1.0 in ElasticNet.

Conclusion:

It is possible to use machine learning to predict property prices in King County. These methods are not perfect but they do allow for homebuyers to be more informed and equipped with knowledge when buying a house. Location, condition/grade, waterfront view, and number of bathrooms were all important features of influence on the price of a property. This data and analysis is extremely useful to sellers trying to price and list their home, and it is useful to buyers who want to get the most value out of their offer. This is a business strategy and model that can be applied to anywhere in order to make intelligent data driven decisions.