

King County Houses

By: Jennyfer Vu





How do buyers know what the asking price of a house consists of? Number of bedrooms? View? Location?



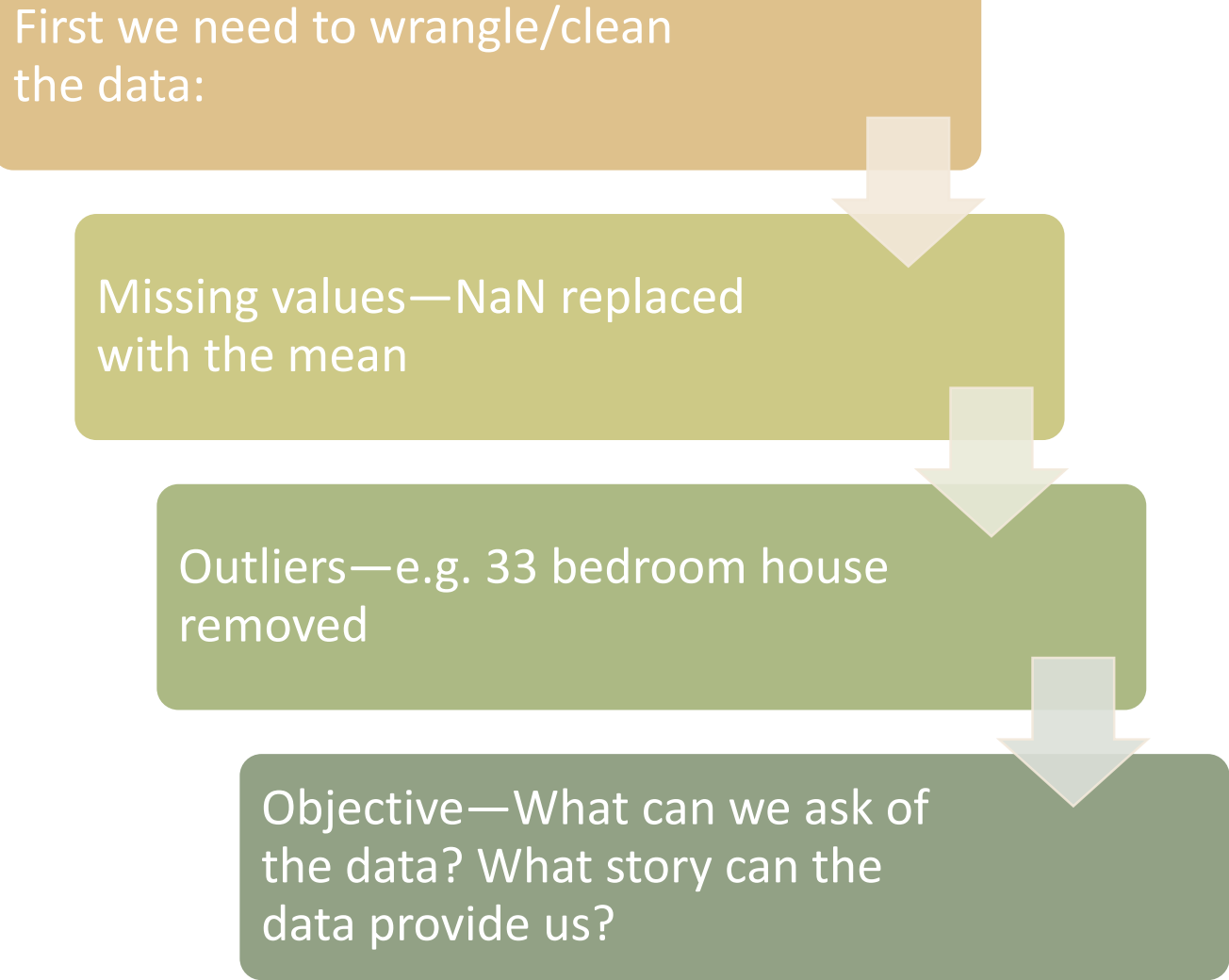
We will be diving into different factors that all contribute to the price of a home



The client for this project will be potential home buyers in King County, WA

Problem Statement

First we need to wrangle/clean
the data:



```
graph TD; A[First we need to wrangle/clean the data:] --> B[Missing values—NaN replaced with the mean]; B --> C[Outliers—e.g. 33 bedroom house removed]; C --> D[Objective—What can we ask of the data? What story can the data provide us?];
```

Missing values—NaN replaced
with the mean

Outliers—e.g. 33 bedroom house
removed

Objective—What can we ask of
the data? What story can the
data provide us?

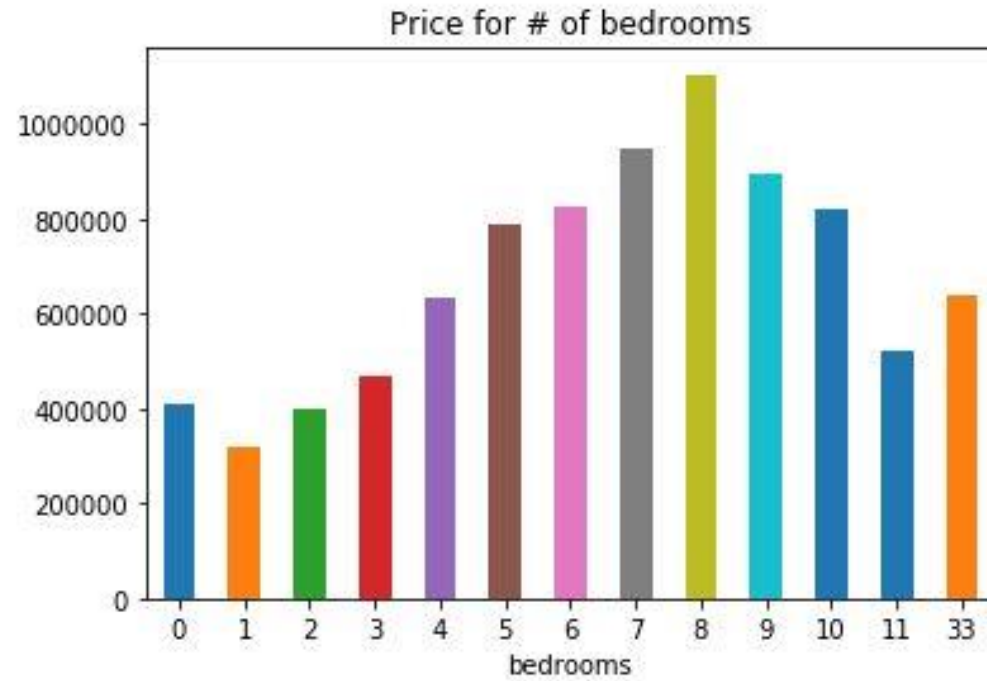
Data Wrangling

Exploratory Data Analysis

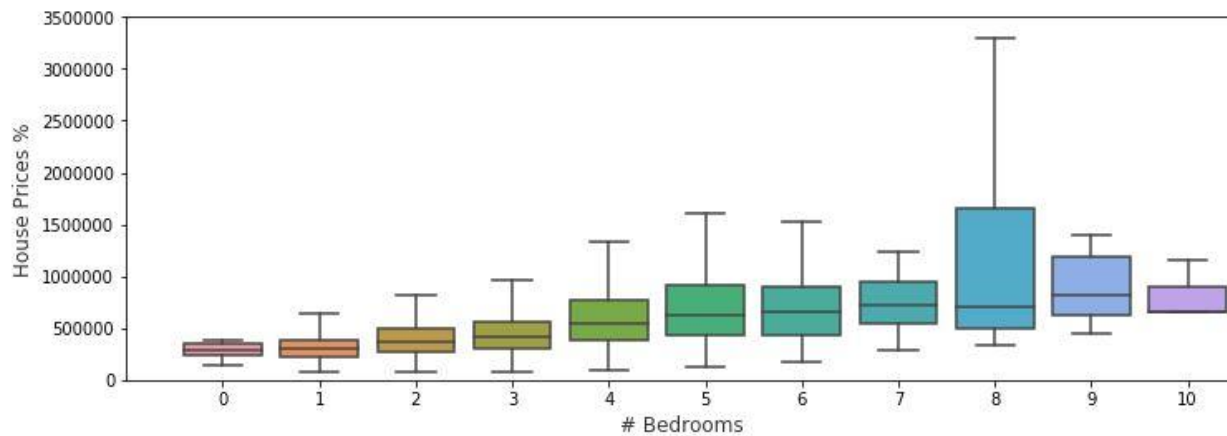


- After cleaning the data, we take a look at the relationship between variables
- First, let's explore how number of bedrooms will affect the price of a property.

Bar chart showing house price vs. # of bedrooms

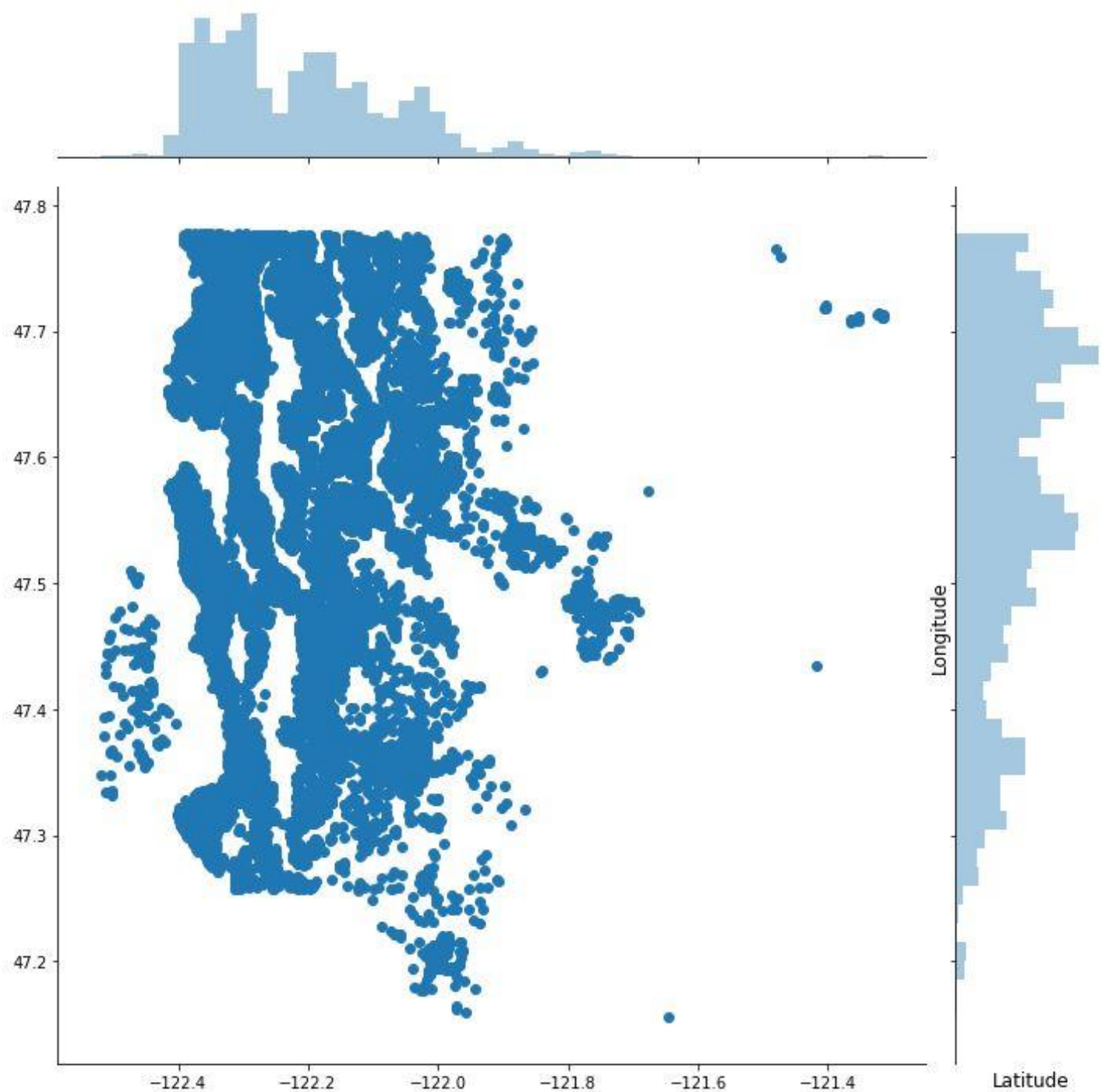


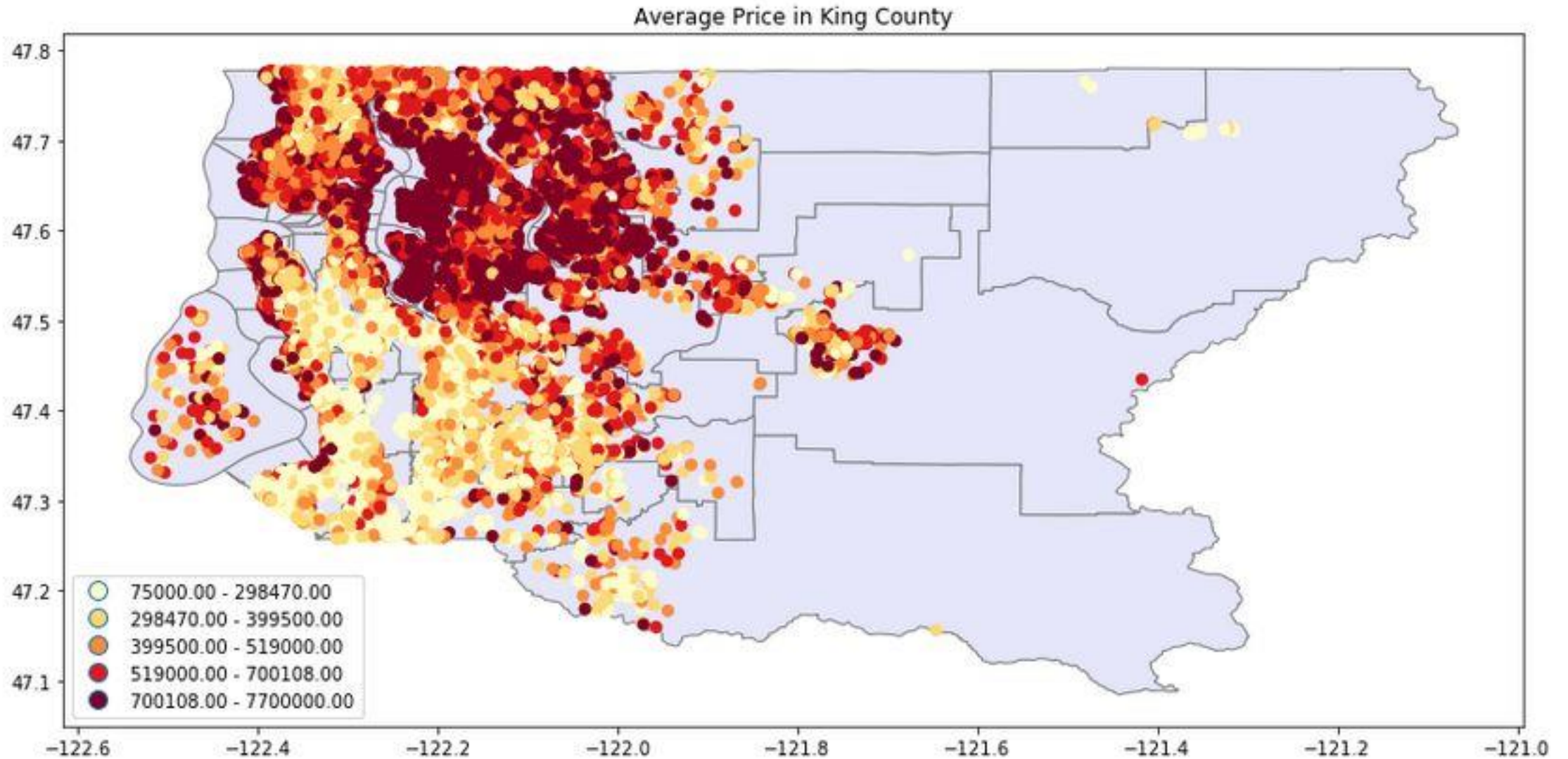
Bedrooms - House Prices



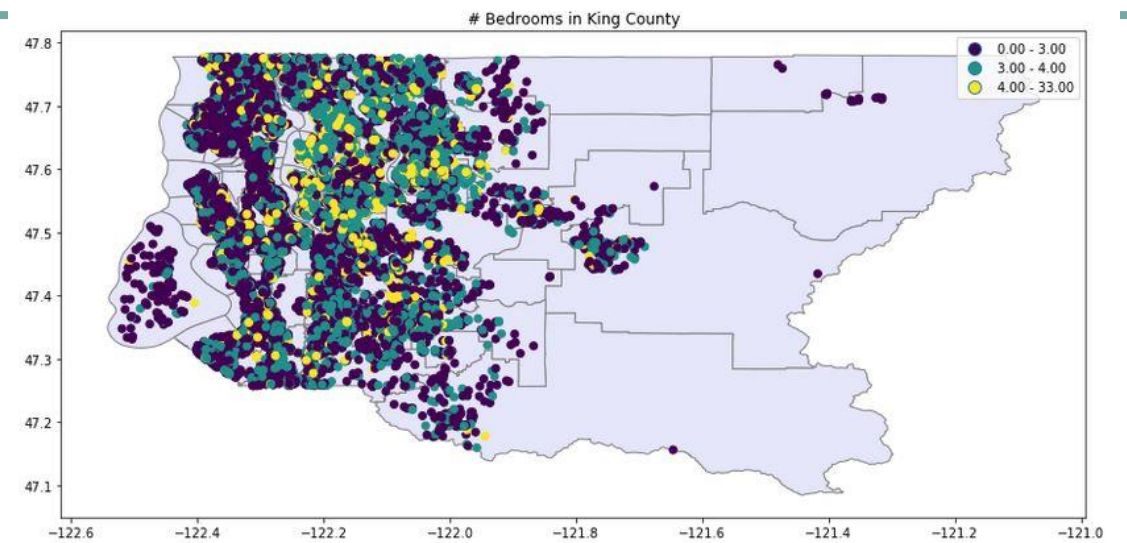
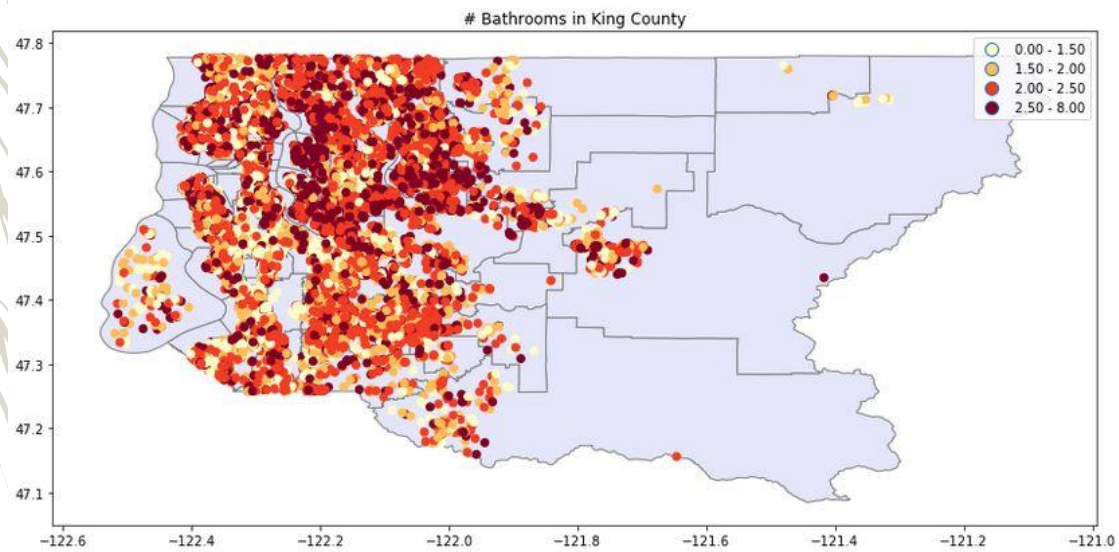
Box plot showing house price vs. # of bedrooms

- Next, we will plot properties by latitude and longitude in King County.
- The highest concentration of homes are between latitude -122.4 and -122 and longitude 47.25 and 47.8
- The coordinates for Seattle, WA are 47.6062° N, 122.3321° W

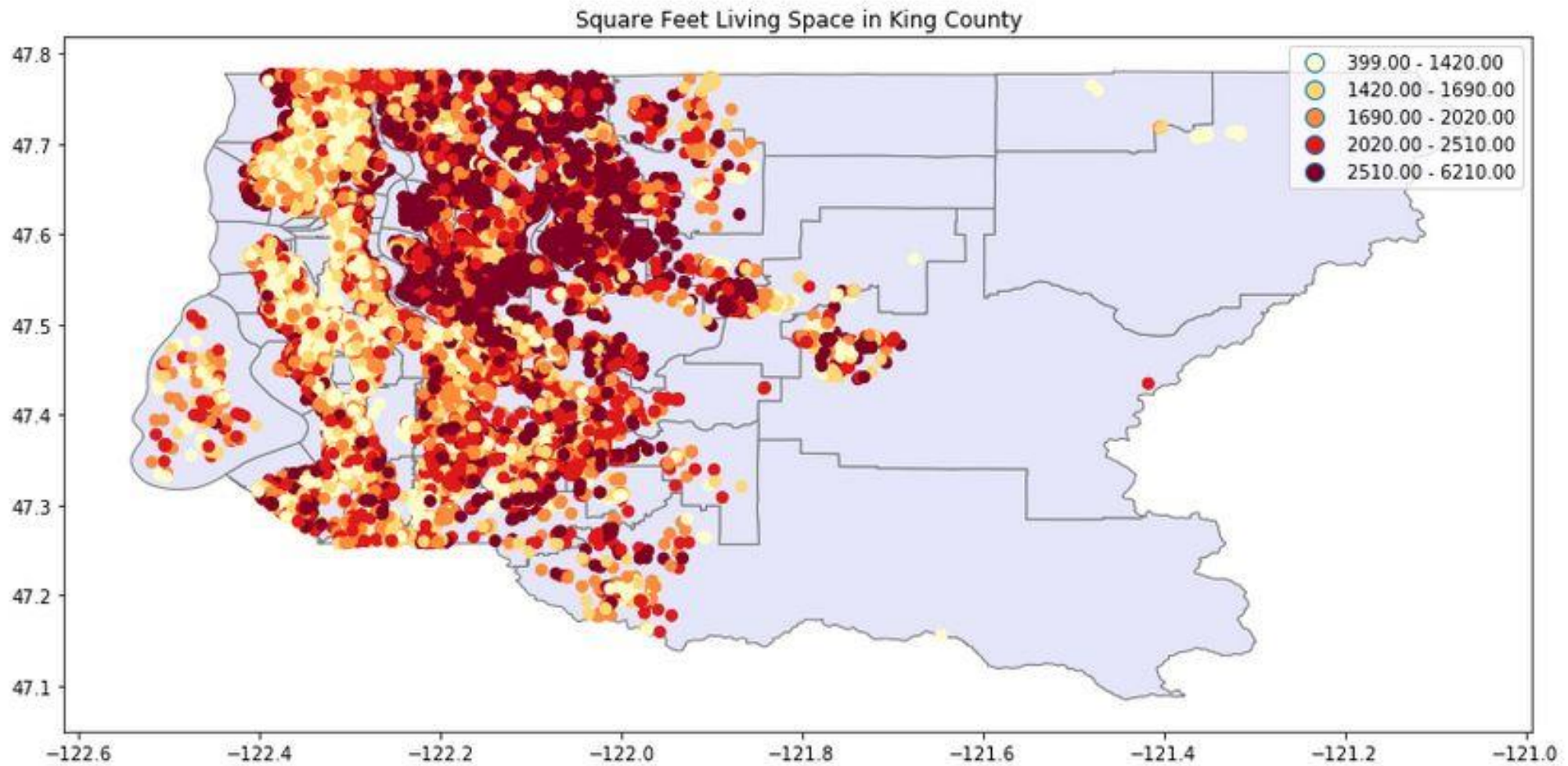




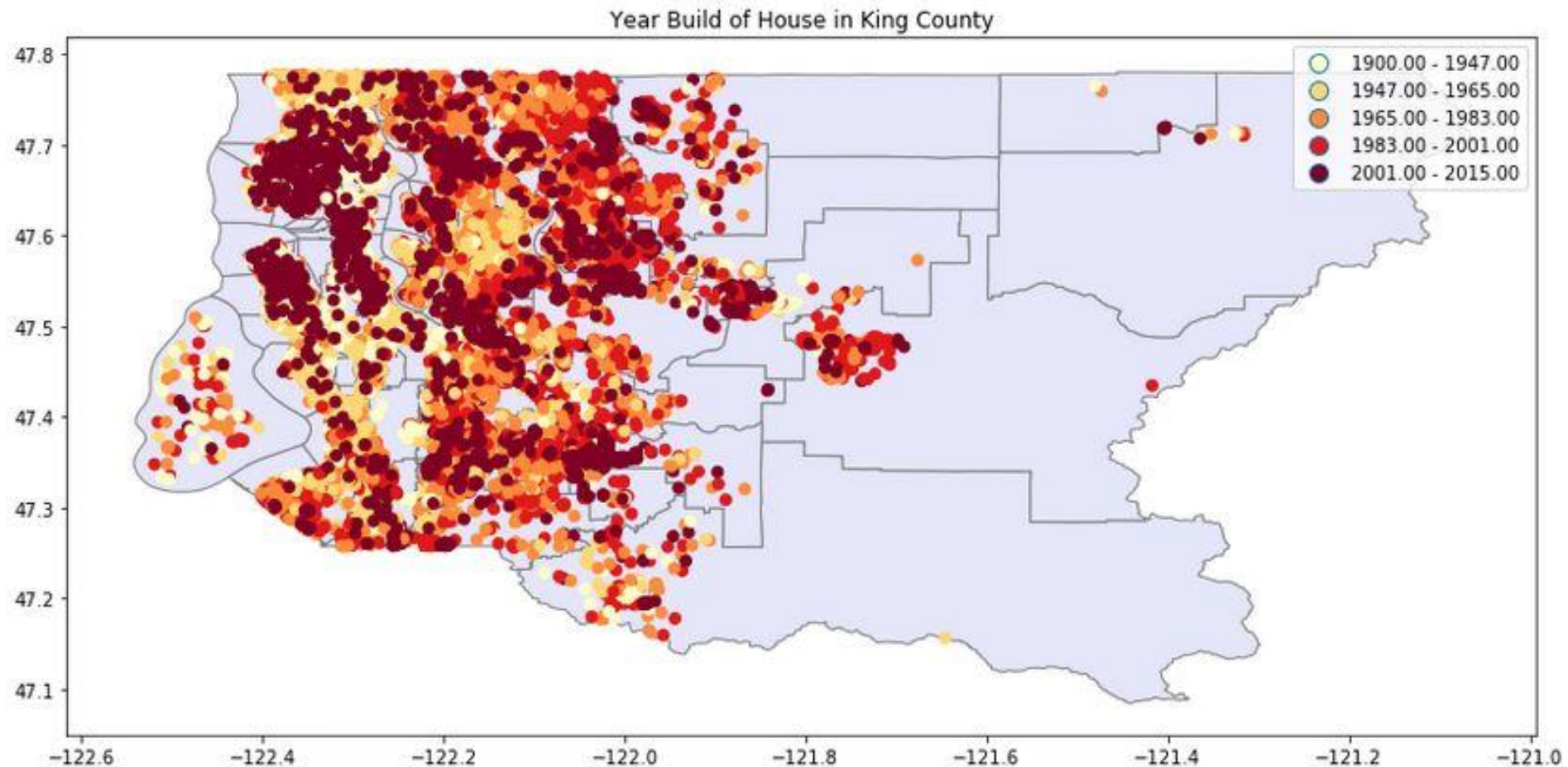
The figure above shows the average listing price by latitude/longitude. As you can see, the closer the house is to metropolitan Seattle, the higher the average listing price.



The two figures above show number of bedrooms and bathrooms. This is to show that there is not a strong correlation between these two variables and the area that they are located. This is important for homebuyers to know that these variables are not location dependent.



This figure shows square feet living space in King County. It appears that there are smaller properties closer to the waterfront. This is explained by there not being as much space and square footage for large properties closer to the water. Houses get larger in square footage further west.



This figure shows that year build of houses in King County is extremely varied. It appears that older properties (white-yellow, 1900-1965) are located towards the waterfront which may be explained by early settlers building there first. There is tremendous new construction over these older areas and all over King County.



Inferential Statistics

- 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.

One-Way ANOVA



- We will be conducting the One-Way ANOVA to compare if the mean price of multiple zipcodes are equal.
- Null hypothesis: The mean price between different zipcodes are the same.
- Alternative hypothesis: The mean price between different zipcodes are different.

OLS Regression Results

```

=====
Dep. Variable:          price      R-squared:                0.407
Model:                  OLS        Adj. R-squared:            0.405
Method:                 Least Squares    F-statistic:            214.6
Date:                   Fri, 26 Apr 2019    Prob (F-statistic):      0.00
Time:                   19:55:01      Log-Likelihood:         -3.0197e+05
No. Observations:       21613          AIC:                    6.041e+05
Df Residuals:           21543          BIC:                    6.046e+05
Df Model:                69
Covariance Type:        nonrobust
=====

```

	coef	std err	t	P> t	[0.025	0.975]
Intercept	2.808e+05	1.49e+04	18.861	0.000	2.52e+05	3.1e+05
cat_zip[T.98002]	-4.652e+04	2.5e+04	-1.861	0.063	-9.55e+04	2476.469
cat_zip[T.98003]	1.331e+04	2.25e+04	0.590	0.555	-3.09e+04	5.75e+04
cat_zip[T.98004]	1.076e+06	2.18e+04	49.369	0.000	1.03e+06	1.12e+06
cat_zip[T.98005]	5.295e+05	2.64e+04	20.023	0.000	4.78e+05	5.81e+05
cat_zip[T.98006]	5.791e+05	1.96e+04	29.601	0.000	5.41e+05	6.17e+05
cat_zip[T.98007]	3.364e+05	2.81e+04	11.965	0.000	2.81e+05	3.92e+05
cat_zip[T.98008]	3.648e+05	2.25e+04	16.231	0.000	3.21e+05	4.09e+05
cat_zip[T.98010]	1.429e+05	3.2e+04	4.464	0.000	8.01e+04	2.06e+05
cat_zip[T.98011]	2.096e+05	2.52e+04	8.329	0.000	1.6e+05	2.59e+05
cat_zip[T.98014]	1.748e+05	2.95e+04	5.931	0.000	1.17e+05	2.33e+05
cat_zip[T.98019]	1.44e+05	2.54e+04	5.675	0.000	9.43e+04	1.94e+05
cat_zip[T.98022]	3.49e+04	2.38e+04	1.469	0.142	-1.17e+04	8.15e+04
cat_zip[T.98023]	5938.1210	1.96e+04	0.304	0.761	-3.24e+04	4.43e+04
cat_zip[T.98024]	2.998e+05	3.48e+04	8.611	0.000	2.32e+05	3.68e+05

- Looking at the p-value under $P > |t|$ we can see that most zipcodes are so small that they are close to 0.000

- This means that they are statistically significant and the means do differ and are unequal. This can be interpreted as the zipcode is a statistically significant predictor of price.

- This means that the difference is not statistically significant and we can't reject the hypothesis that their means are equal.

- The computed F-statistic is 214.6 Typically a high F-value means that your data does not support the null hypothesis well. This means that we reject the null hypothesis and that there is statistical significance between price of a property and location.

OLS Regression Results

=====						
ble:	price	R-squared:	0.407			
	OLS	Adj. R-squared:	0.405			
	Least Squares	F-statistic:	214.6			
	Fri, 26 Apr 2019	Prob (F-statistic):	0.00			
	19:55:01	Log-Likelihood:	-3.0197e+05			
ations:	21613	AIC:	6.041e+05			
ls:	21543	BIC:	6.046e+05			
	69					
Type:	nonrobust					
=====						
	coef	std err	t	P> t	[0.025	0.975]

	2.808e+05	1.49e+04	18.861	0.000	2.52e+05	3.1e+05
98002]	-4.652e+04	2.5e+04	-1.861	0.063	-9.55e+04	2476.469
98003]	1.331e+04	2.25e+04	0.590	0.555	-3.09e+04	5.75e+04
98004]	1.076e+06	2.18e+04	49.369	0.000	1.03e+06	1.12e+06
98005]	5.295e+05	2.64e+04	20.023	0.000	4.78e+05	5.81e+05
98006]	5.791e+05	1.96e+04	29.601	0.000	5.41e+05	6.17e+05
98007]	3.364e+05	2.81e+04	11.965	0.000	2.81e+05	3.92e+05
98008]	3.648e+05	2.25e+04	16.231	0.000	3.21e+05	4.09e+05
98010]	1.429e+05	3.2e+04	4.464	0.000	8.01e+04	2.06e+05
98011]	2.096e+05	2.52e+04	8.329	0.000	1.6e+05	2.59e+05
98014]	1.748e+05	2.95e+04	5.931	0.000	1.17e+05	2.33e+05
98019]	1.44e+05	2.54e+04	5.675	0.000	9.43e+04	1.94e+05
98022]	3.49e+04	2.38e+04	1.469	0.142	-1.17e+04	8.15e+04
98023]	5938.1210	1.96e+04	0.304	0.761	-3.24e+04	4.43e+04
98024]	2.998e+05	3.48e+04	8.611	0.000	2.32e+05	3.68e+05



In-Depth Machine Learning Analysis

- **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).
- **Linear Regression:**
- In general, the closer the coefficient R^2 is to 1.0, the better the model fits the data. Our aim will be to get as close to 1.0 as possible. In the model above, we are returned a score of $R^2 = 70.7\%$



In-Depth Machine Learning Analysis

- **Gradient Boosting:**

- Gradient boosting regression is a machine learning model that is constructed from an ensemble of weak prediction models such as decision trees. Gradient boosting increased model's score to 91.98%!

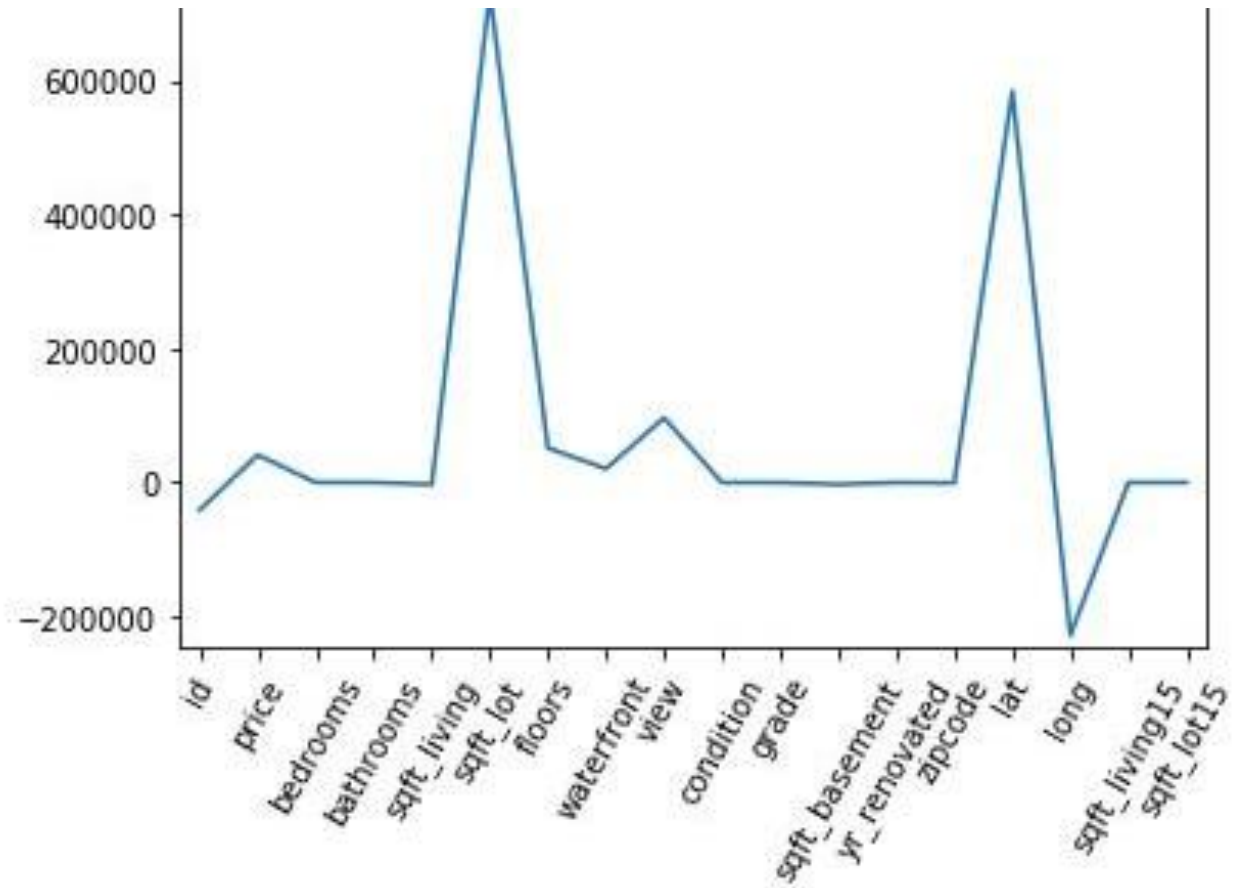
- **Random Forest:**

- Random forest is a more complex model that was able to lower the RMSE from 200,914 to 139,025 dollars. Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction.

- **Fitting Linear Regressions models:**
- Looking at the coefficient in the middle table, let's look at the bathroom coefficient. It's pretty small, essentially zero.
- This means that the number of bathrooms is statistically not significant as a predictor of house price.
- Negative correlation: a negative relationship between two variables (inverse relationship)

In-Depth Machine Learning Analysis

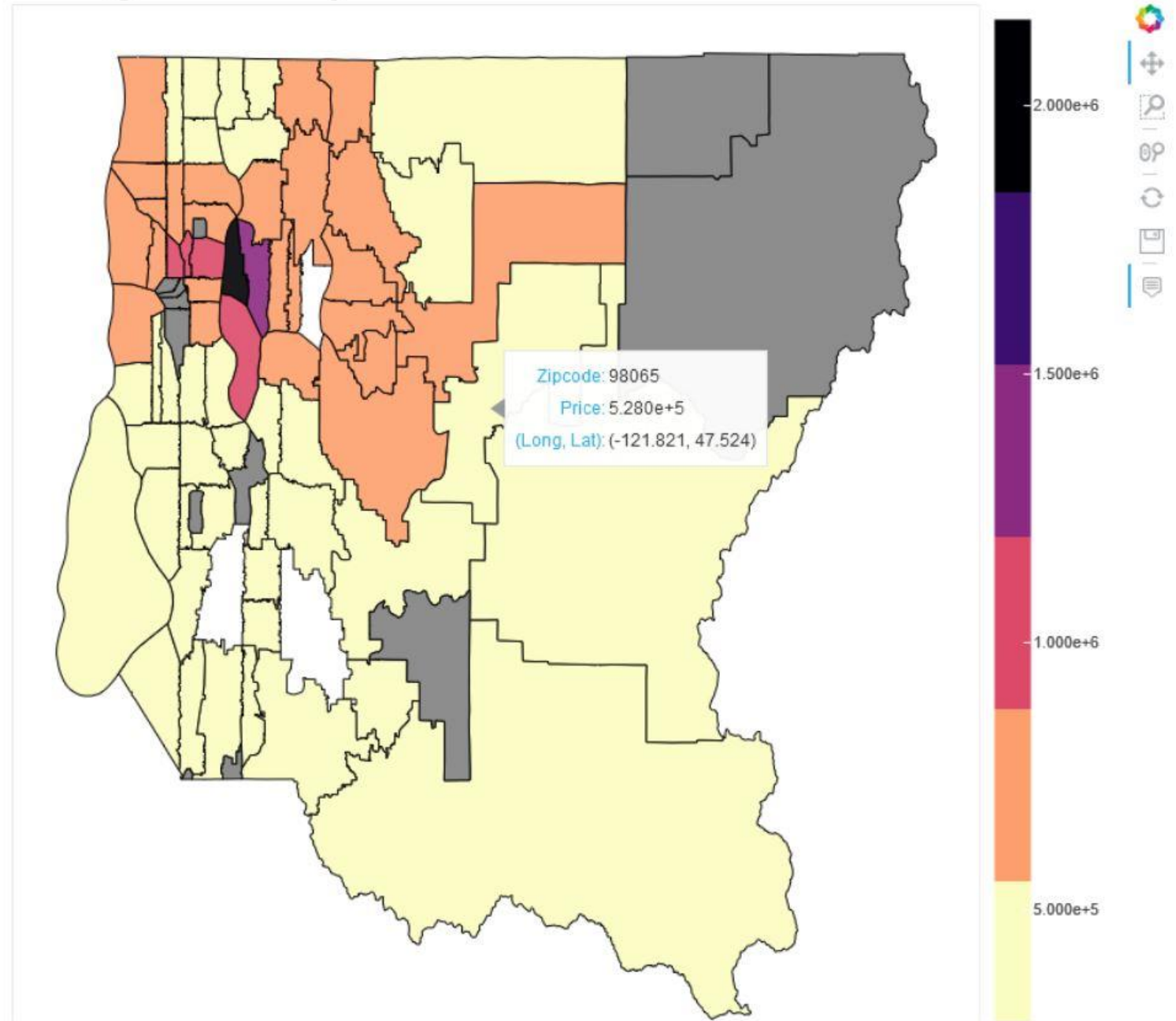
- **Lasso Regression:**
- Lasso is used to select the most important features in prediction, in this case prediction of house prices.
- Lasso regression above illustrates that square footage for living/lot size, waterfront view, zipcode, and lat/long are the most important features selected by lasso regression coefficients for predicting price.



Bokeh

- In addition to EDA, statistics, and machine learning there is a very important tactic to convey your data to an audience.
- Through python, we decided to convey the data visually by using Bokeh which is an impressive library that can handle large datasets.
- As we can see, Bokeh is able to allow users to interact and display different values depending on what variable they are interested in

King County House Prices



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.