**King County House Price Analysis**

City University of Seattle

CS612 Summer 2020

Group Members:

Anil Erturk [erturkanil@cityuniversity.edu](mailto:erturkanil@cityuniversity.edu)

Minh Nguyen [minh.truong@cityuniversity.edu](mailto:erturkanil@cityuniversity.edu)

Faculty Advisor: Professor Brian Maeng

**Abstract**

The paper analyzes relationship between house features and prices. The data was imported, cleaned, and prepared for analysis. The meta data was analyzed. Different types of data required different graphs to be analyzed and visualized. We looked at the distribution of values for each house attribute to see trends and popular values for attributes. We also looked for correlation between house features and house prices to show any relations. Our findings show that size related attributes of a house affect its price positively. Quality related attributes don’t have a significant affect on the price with the exception of waterfront”.

**Introduction**

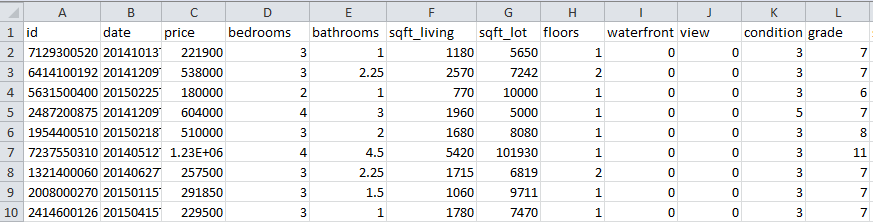
In the past, looking for a house consisted of driving around in neighborhoods to look for “for sale” signs and going to real estate agencies to look at their list of houses. With the advance of technology, today’s real estate listings can be accessed on websites like Zillow. These online real estate services keep houses in their databases with many attributes. This significantly improves the lives of sellers and buyers in the sector. Filtering by different attributes allows sellers to easily reach the types of customers interested in their houses. It also allows buyers to easily find the types of houses they are looking for. In the last 30 years, computers have drastically improved virtually every single industry. The analysis in this paper can also be used by construction companies to figure out what types of houses (features) are in high demand.

Aside from improving the industry, data analysis shows information that was previously unknown. By analyzing different houses based on their attributes and prices, the relationship between each of these attributes and house prices can be seen.

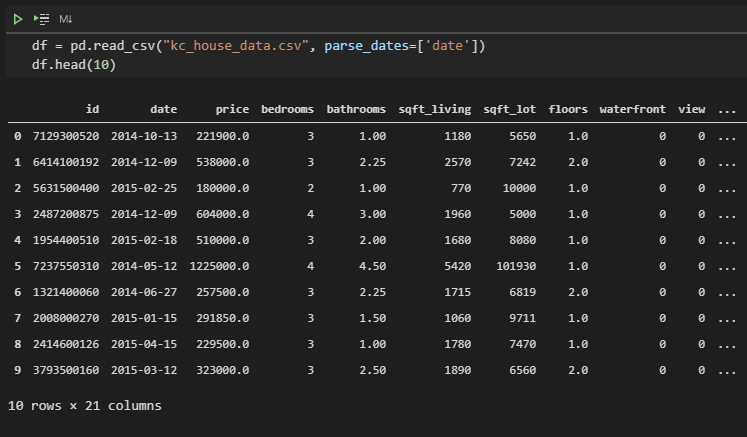
On top of looking for a correlation between different attributes and prices, the distribution of each of these attributes can be used to see the popularity of different values for each attribute. This will show what types of houses are best-selling in an area. This distribution of each attribute will also show how normally distributed it is.

**Data**

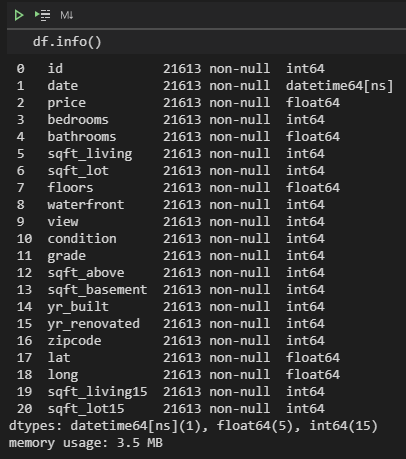
The data is taken from (Kaggle, 2016) in the form of a CSV (Comma Separated Values) file that contains 21614 records. Each record has 21 columns including the id which is the primary key. The primary key is a unique value that allows identification.



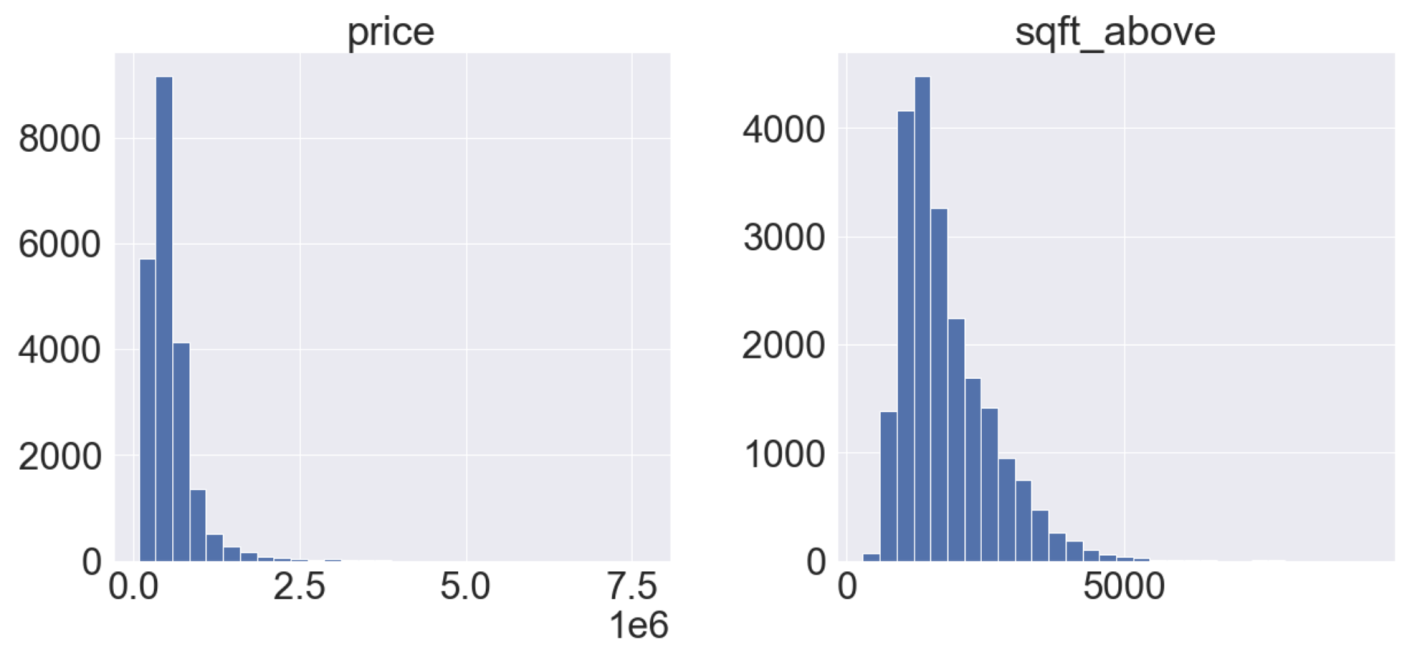
Since the data is two dimensional, a DataFrame is a suitable data type. The CSV file is read as a DataFrame object and then assigned to the variable df. The head function is used to examine the first 10 rows of the data.

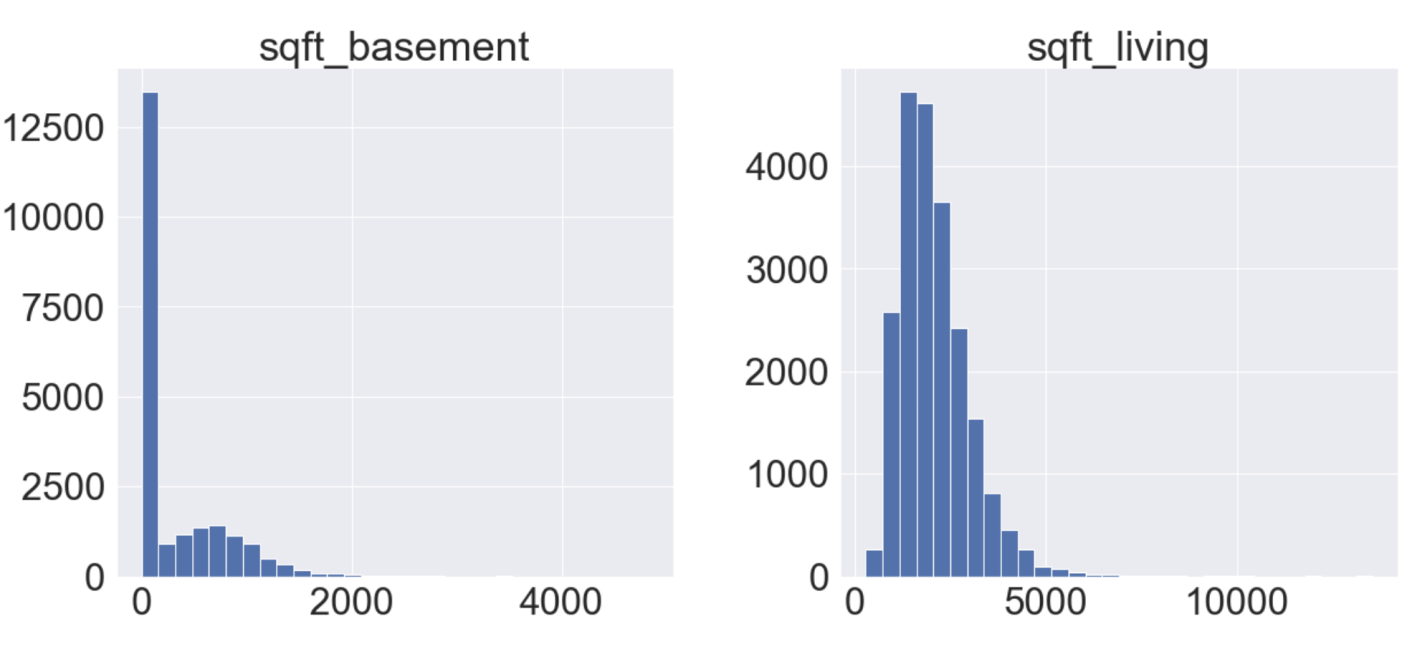


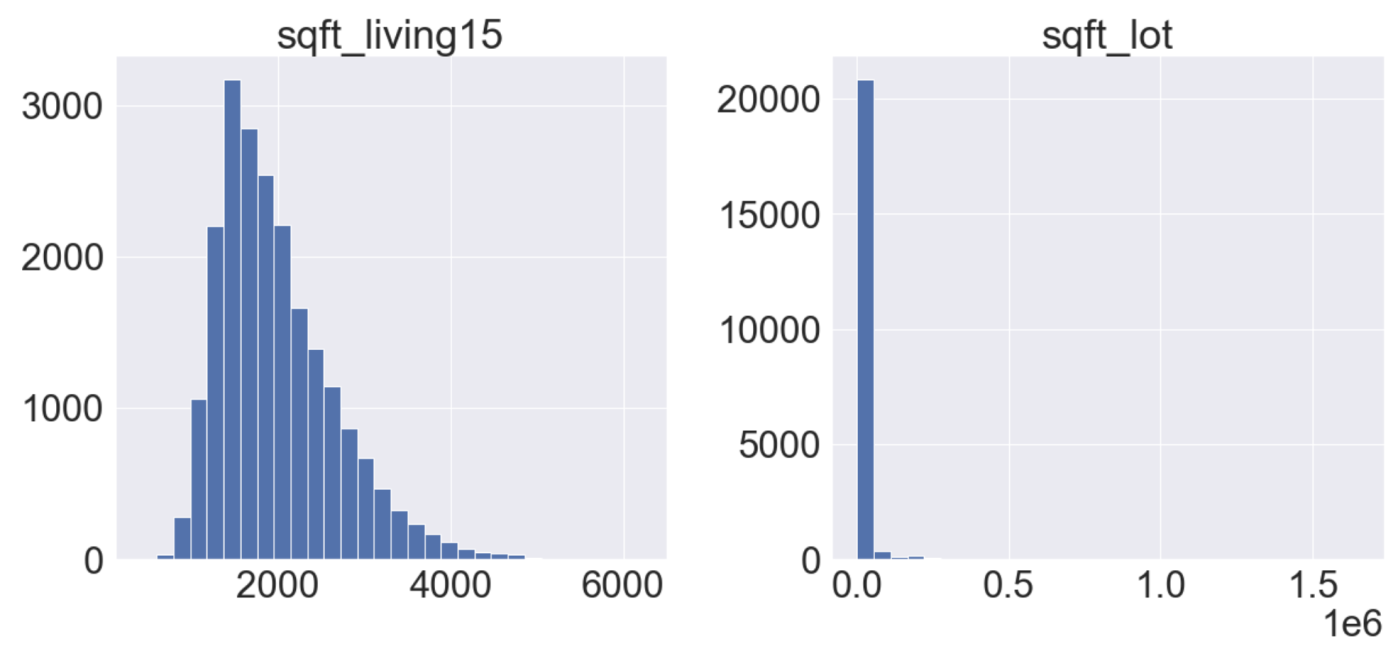
info function gives us an overview of the structure of our data. One interesting observation is how the python DataFrame uses 3.5MB (Megabytes) for the CSV file that is actually 2.39MB and 2.40MB on disk. The size on disk is usually slightly higher than the actual size because of the way files are divided into segments. Higher allocation unit size will cause better performance with higher wasted space.

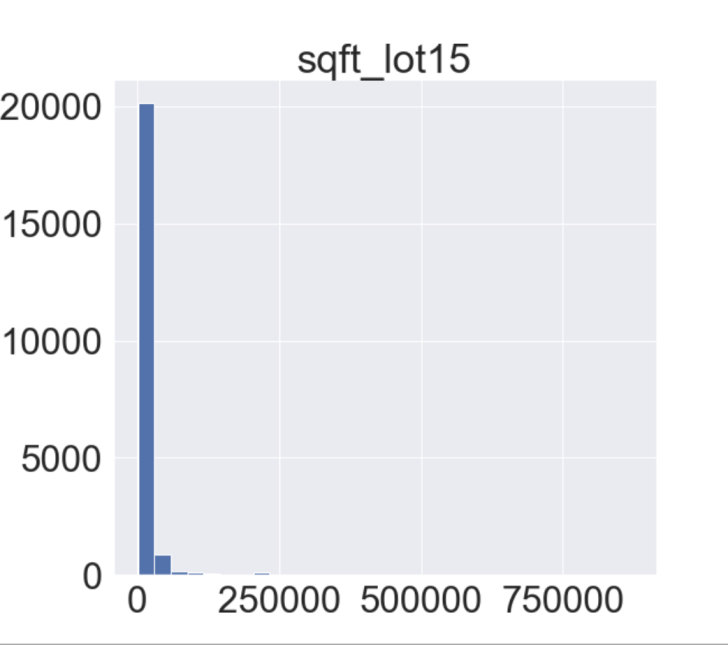


The data contains all numerical values most of these features are continuous while the rest are discrete (both ordinal and nominal). “waterfront” is a nominal binary variable while “bedroom”, “bathrooms”, “grade”, “floor” and “views” are ordinal variable. Variables such as “lat”, “long”, “zipcode” does not represent true numeric value but instead, it represents the location of the property. The remaining variables are all continuous which we can plot using a histogram to check for the normality of the distribution.









We can see that most of these features follow a normal distribution, only “sqft\_basement” having a high concentration around zero. This might suggest that there is a disproportionate number of properties that do not include a basement. Other than that, the data is normal.

**Literature Review**

Analysis of real estate is very common since it is a big and active sector. If the house is priced too high, it likely won’t sell. It the house price is too low, there will be profit loss. Also, the people buying, or renting don’t want to overpay. This need to get a good deal on both sides drives the need for real estate market analysis. This analysis shows the value of other real estate with similar features, therefore determining the value of this real estate (Real Wealth, 2020).

The features that are analyzed are similar since the variables for real estate are similar in different locations and years. These features include “area and neighborhood”, “size or square footage”, “lot size”, “number of bedrooms and bathrooms”, “other rooms”, “number of floors”, “view”, “construction age”, “amenities and features”, “proximity to local amenities”, “recent or notable improvement”.

**Methodology**

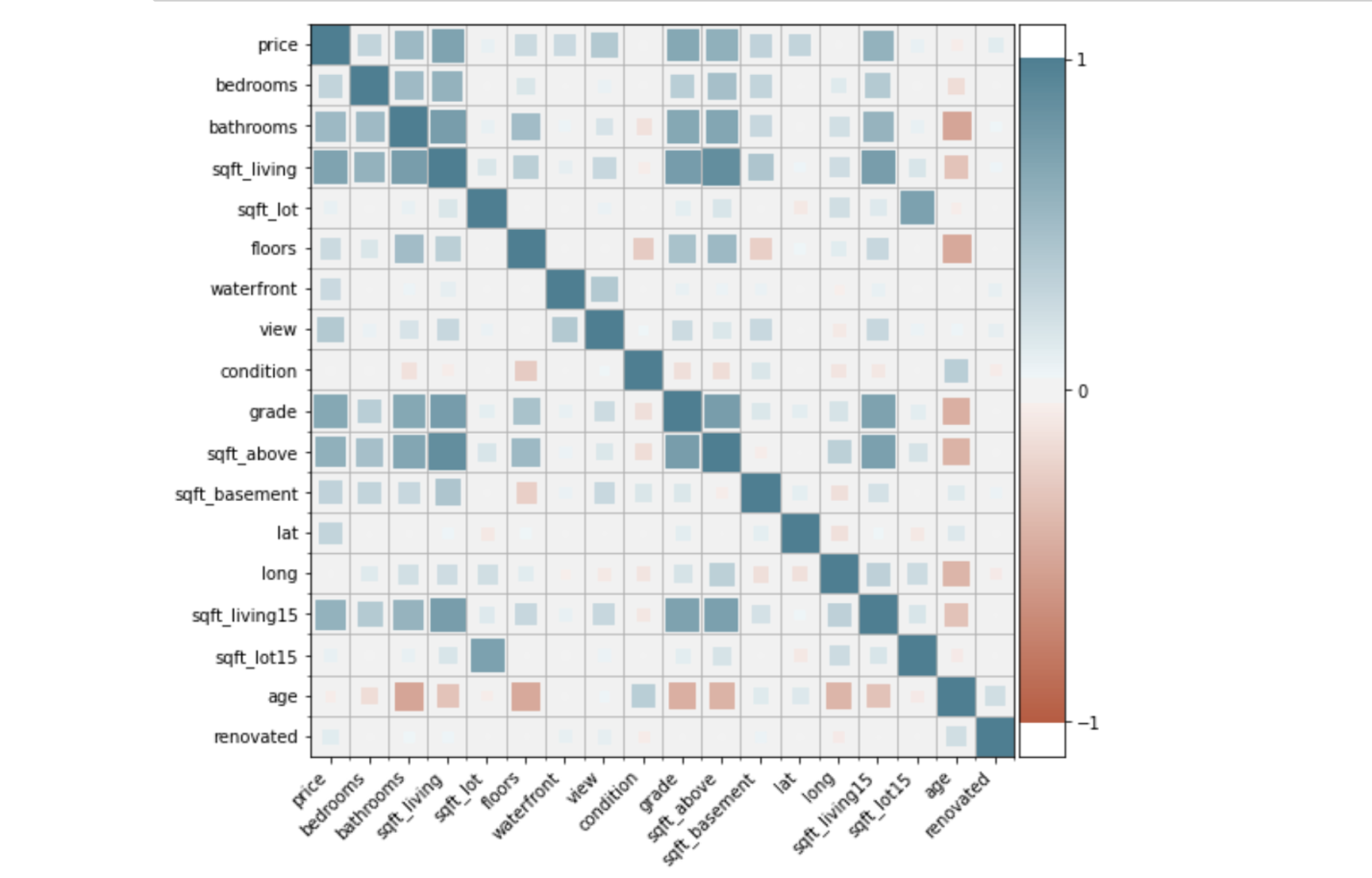
Feature engineering is the first step we took in approaching this dataset is to create new features that might have an influence on the price of the house. Our new columns are “age” which is a numerical feature indicate the age of the house and “renovated” which is a binary variable which indicates whether the house has been renovated ever since it was built. There is a total of 18 columns in this dataset this is of course after we have excluded unnecessary features. All of which are represented numerically, even though some of them are qualitative.

Next, we want to gain insight into the relationship between the variables. The goal is to determine which columns have an impact on the price column as well as how it impacts the price. To answer the question of what, we will use the Pearson correlation matrix to visualize the coefficients of each continuous numerical variable with the price column.

Finally, we will use scatterplot, boxplot and violin plot to dive deeper into the relationship between price and each feature.

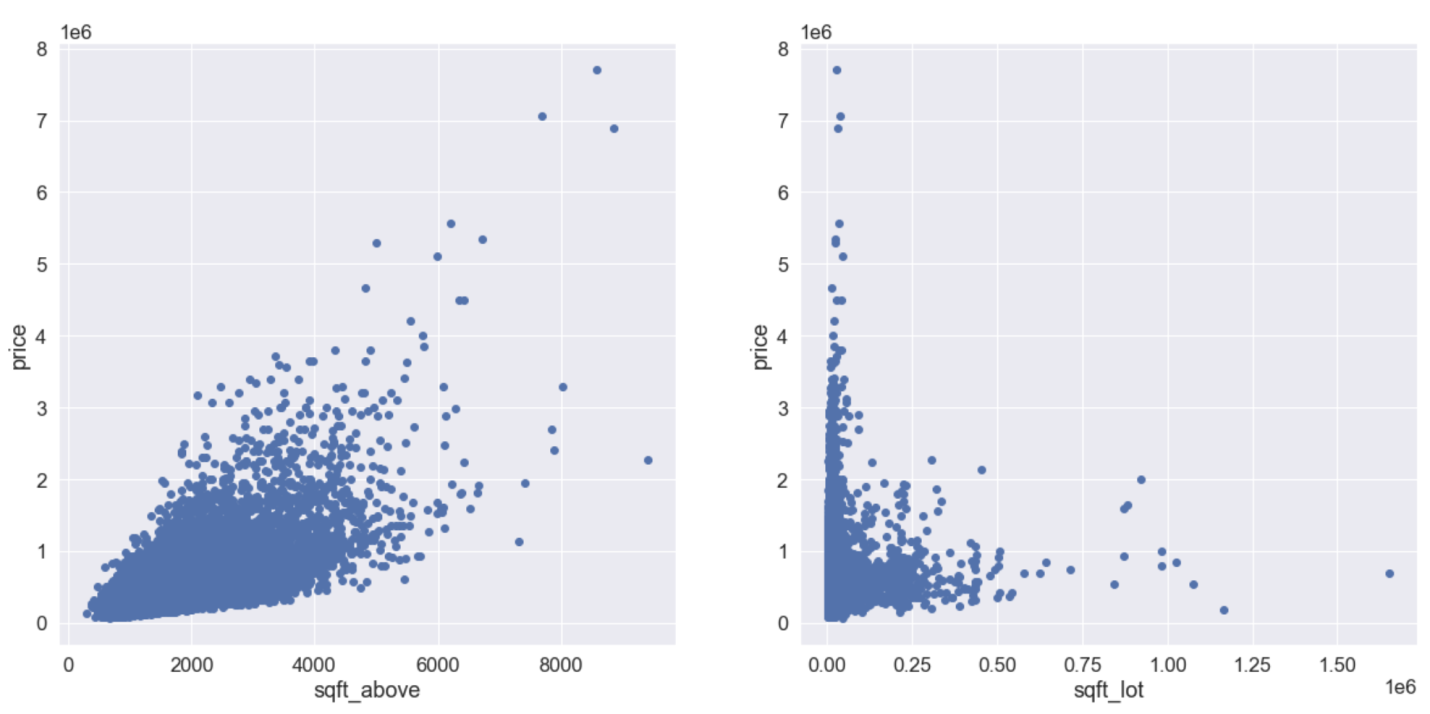
**Result**

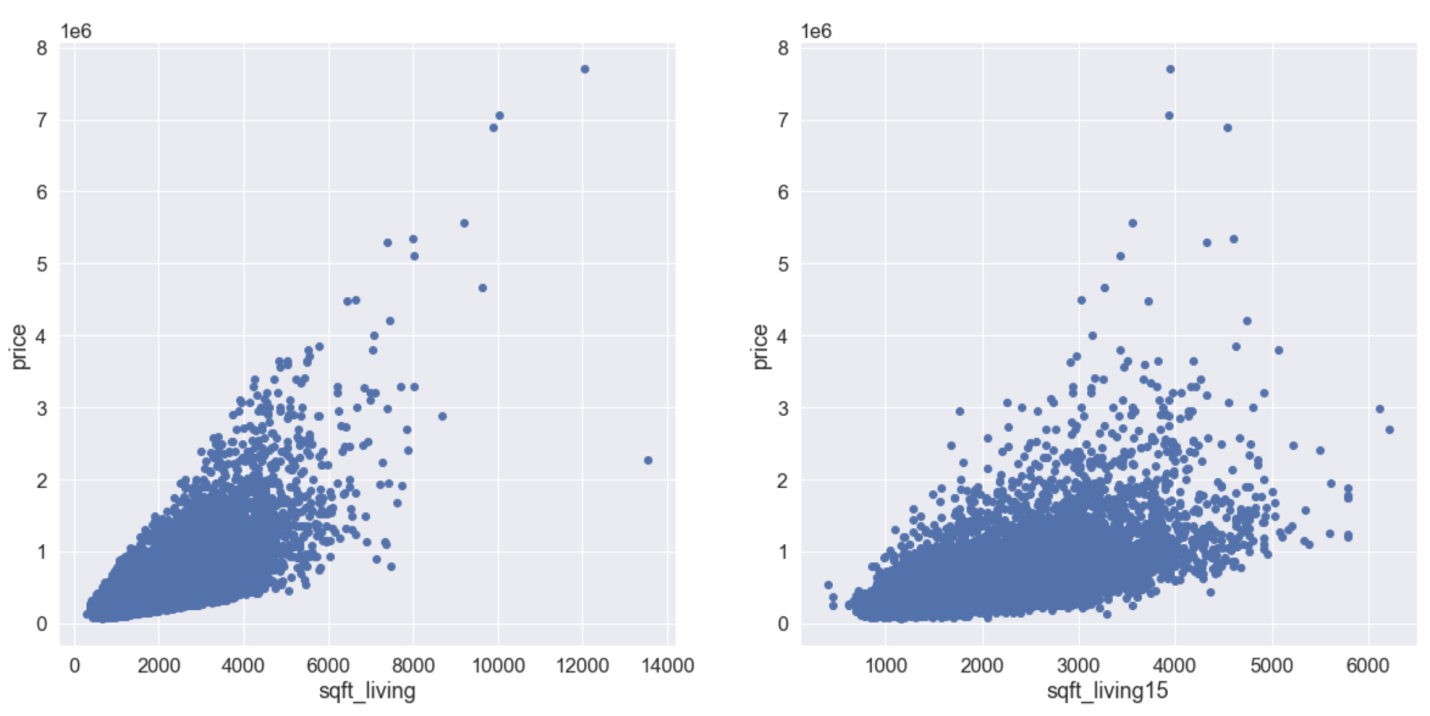
We first use the correlation matrix to look for any notable correlation between each variable with the price of the house.



From this correlation map of the numeric features, we can infer 2 things. First, features relating to size seems to influence the prices of the house the most, namely: "bedrooms", "bathrooms", "sqft\_living", "sqft\_lot", "sqft\_above", "sqft\_living15", "grade". Features relating to the quality of the house such as age or condition has very little influence on the price.

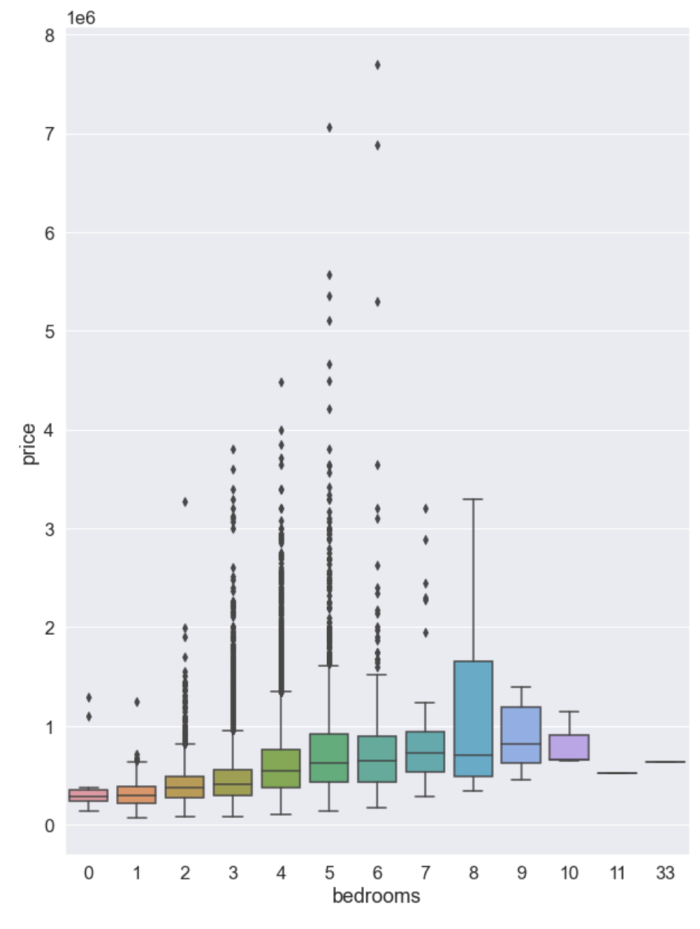
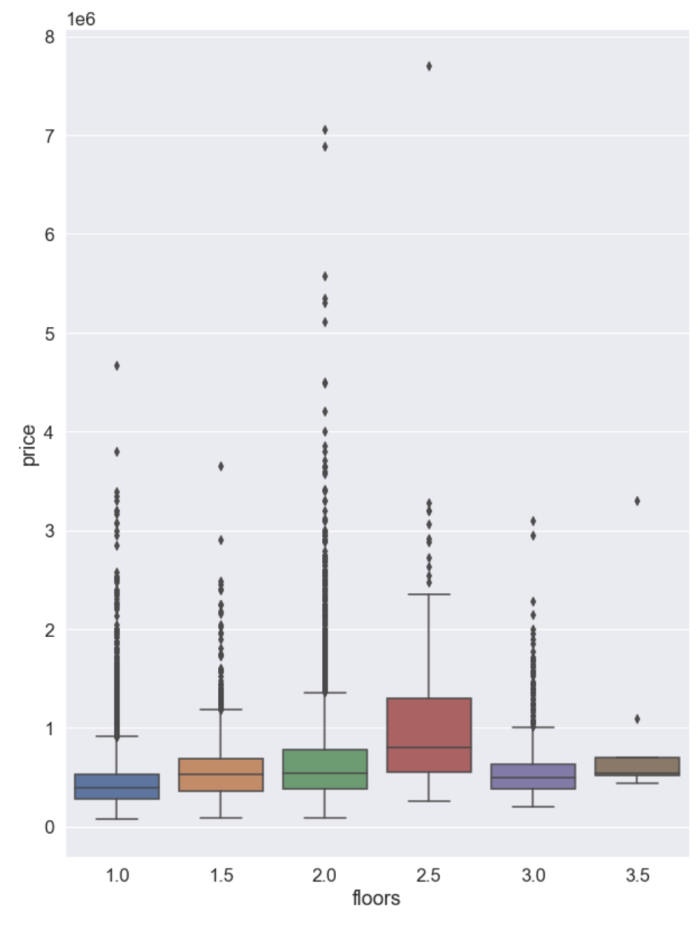
Based on this finding, we will go deeper into how the price of the house change across each of these variables. We will start with continuous variable namely: "sqft\_living", "sqft\_lot", "sqft\_above", "sqft\_living15".

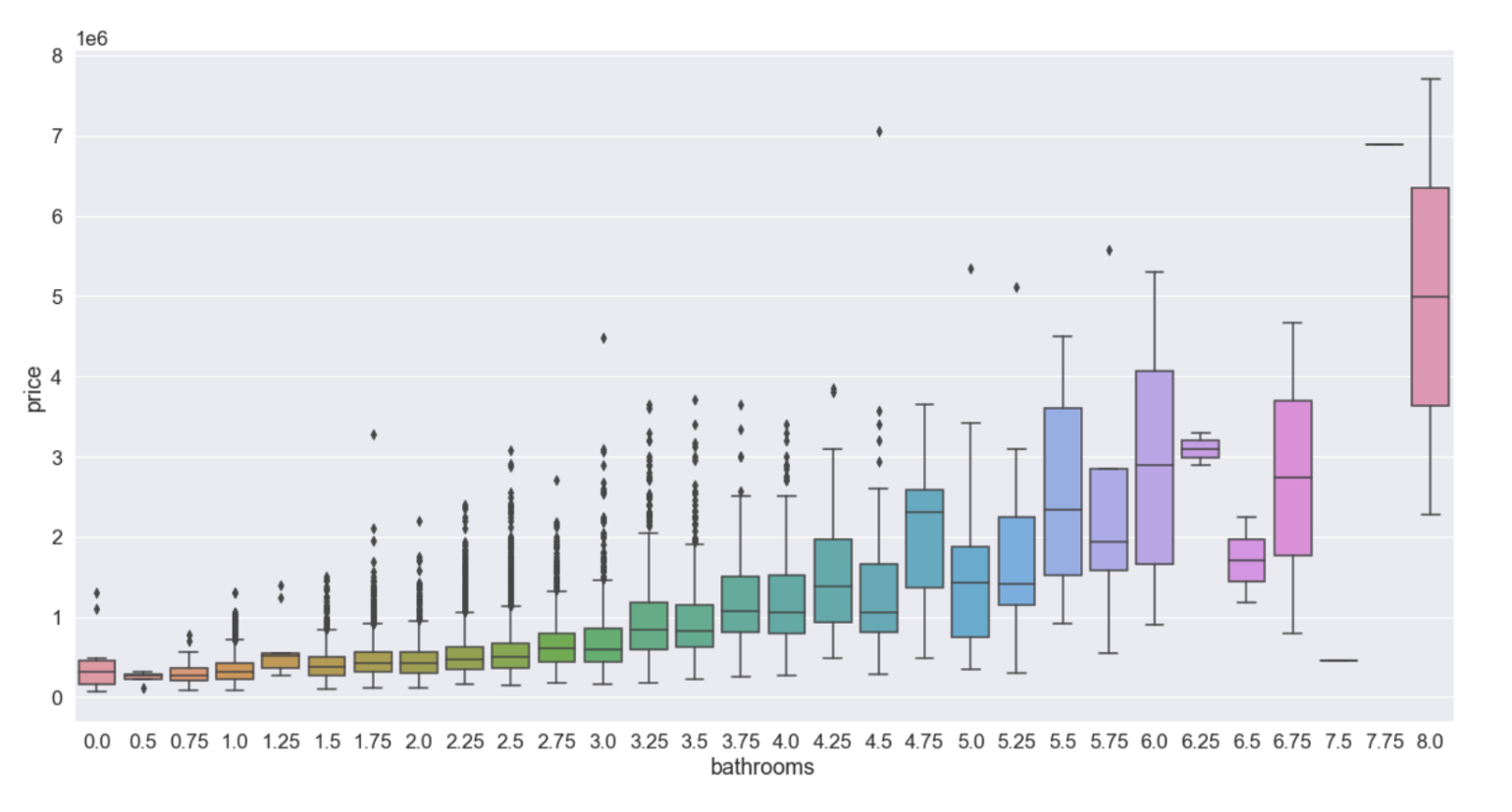


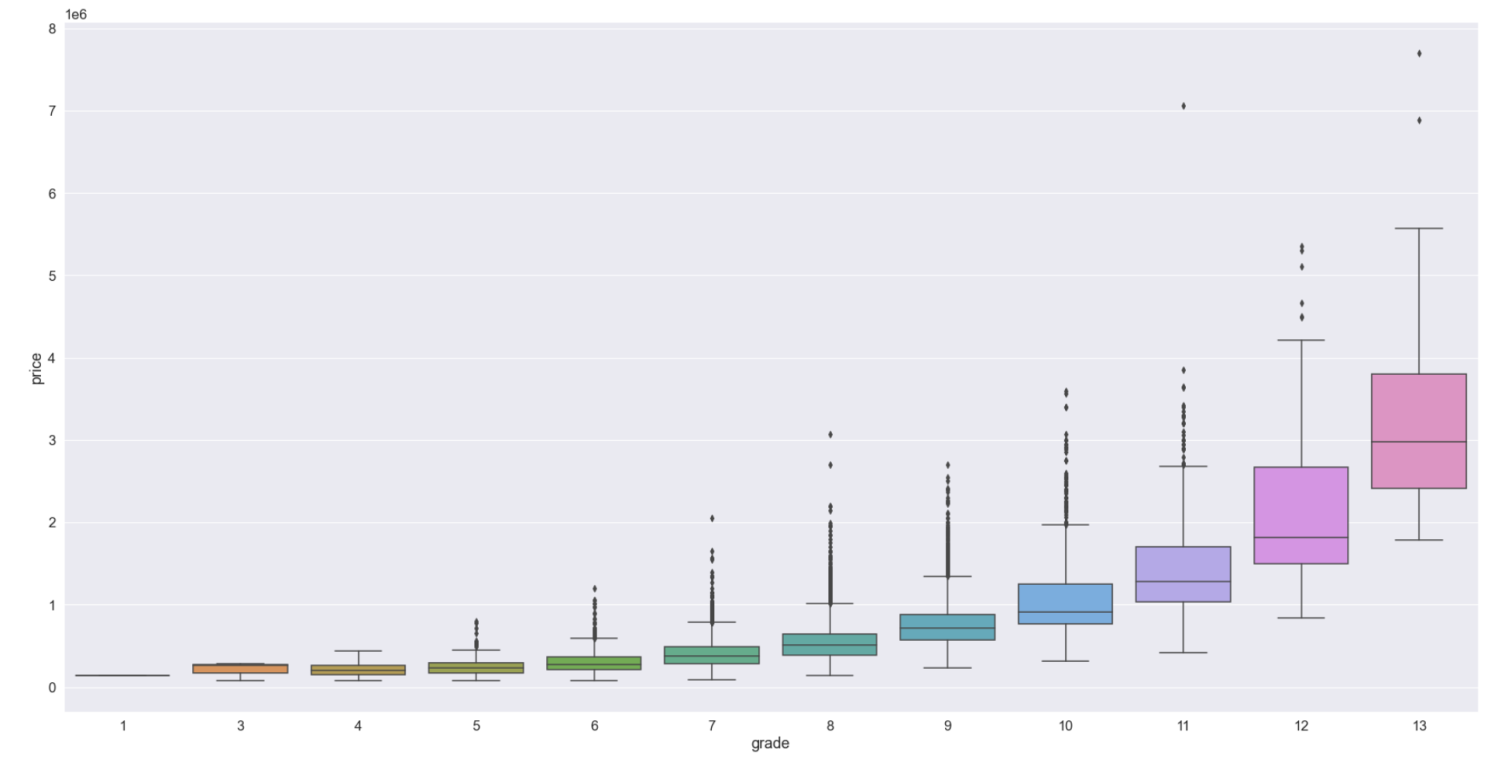


Other than sqft\_lot, there seems to be a trend between price and these features. However, the relationships seem to be very heteroscedastic which means there is a higher variance on the price as the value of the feature increase thus making the prediction less accurate.

Next, we will visualize discrete variables: “floors”, “bedrooms”, “bathrooms” and “grade” using box plot



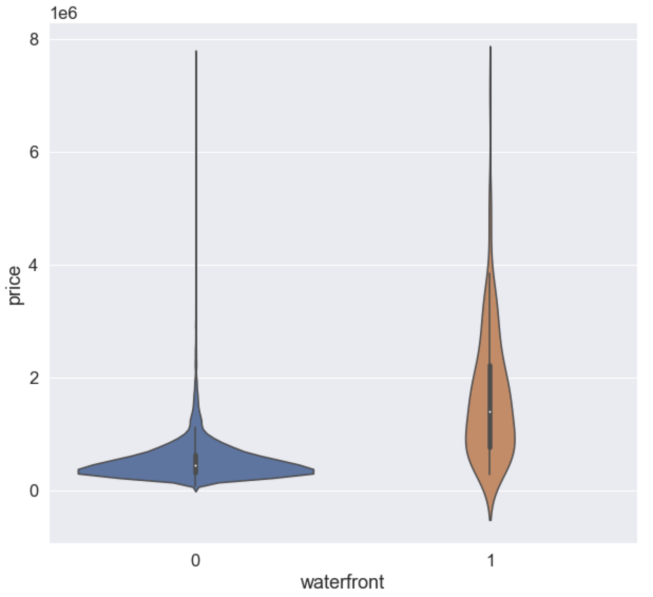
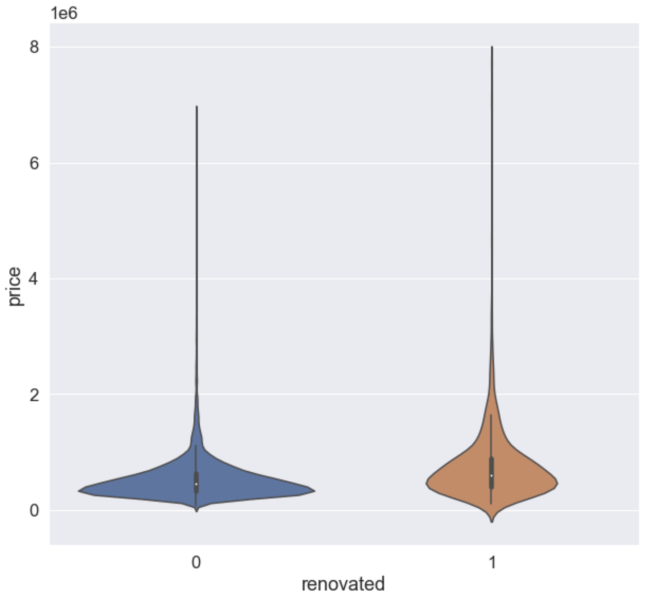


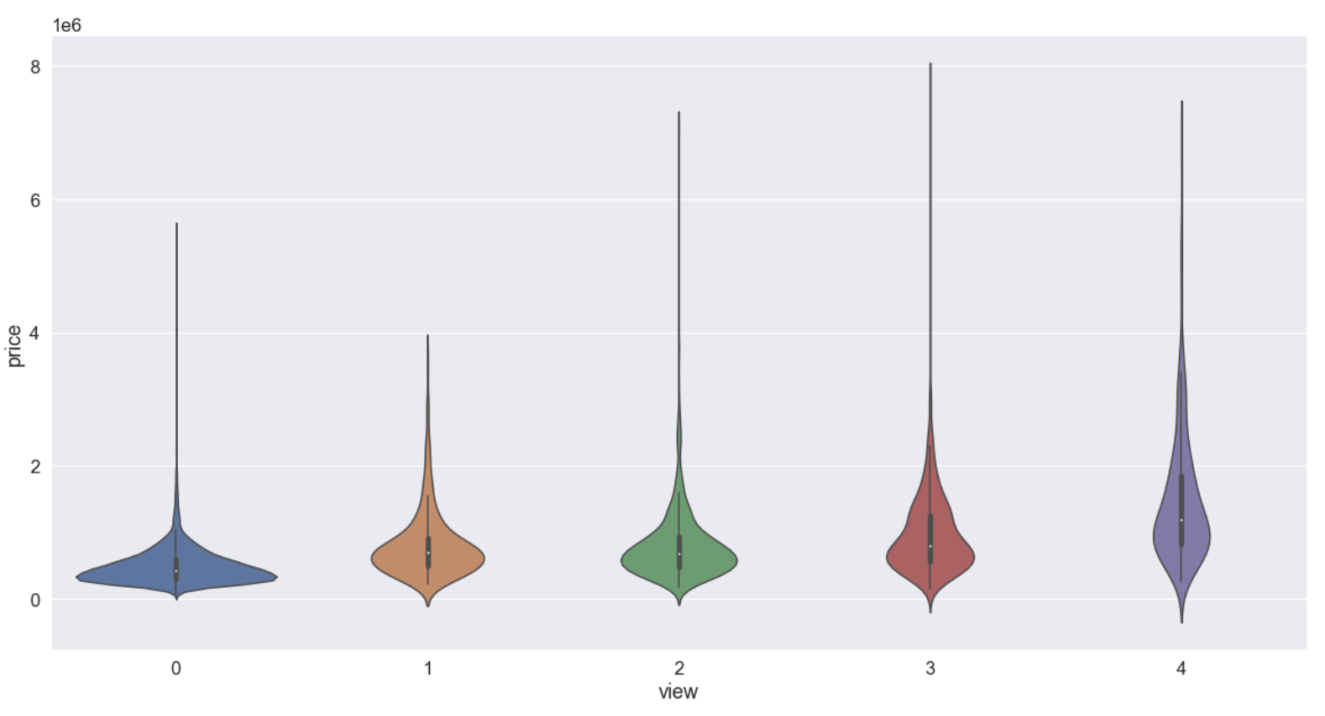


We can see there is a gradual increase in price for 3 variables: “bedrooms”, “bathrooms”, “grade”. This is not presented clearly in floors. We can infer that the size of the house does have an influence on its price.

Next, we are going to use the violin plot to examine other qualitative features:

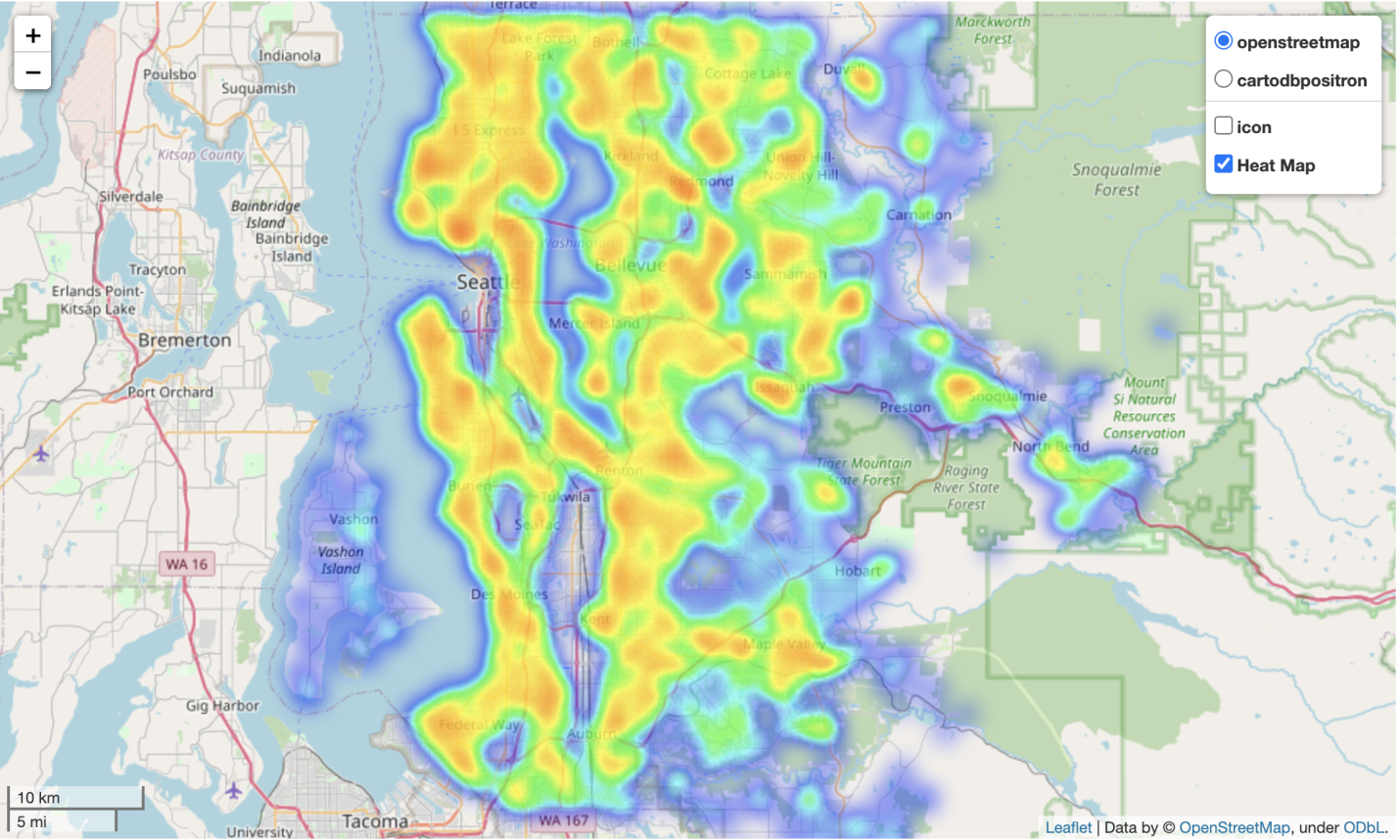
* “renovated” indicates whether the house has been renovated
* “waterfront” indicates whether the house face a body of water
* “view” indicates how many views the house has





The effect of these features on the price is rather less distinctive with the median price remains of the same across each group for “renovated” and “view”. “waterfront”, however, seems to have a very significant impact on the price. This might suggest that it’s not about how the house is built but where it is built that determines the price of the property.

With that knowledge in mind, let’s leverage 2 features “lat” and “long” to visualize how the location has an impact on price using heatmap from the folium package.



There doesn’t seem to be any notable trend from this heat map. However, it does show the house price is much higher from the downtown area of Seattle further to the North compares to the South. We do see a higher concentration of expensive property on the East side of Seattle. There is still a lot of missing area on the heatmap in the downtown Seattle area which mostly includes commercial buildings.

Overall, our finding shows that size features seem to have the most significant impact on the price of the house. The location seems to have an influence but there isn’t any specific trend that can be extrapolated from the data. This might be due to the limited size of the data.

**Conclusion**

In a nutshell, this data helped us discover how size features such as the total area, number of bathrooms, and bedrooms influence the price of the house. It also shows how the price of the property increases significantly if the house is waterfront which hinted at the importance of location in determining the price. However, we couldn’t confirm this using the dataset. It is worth mentioning that the data only contains 21614 records and was collected from the time period of 1 year which makes it very limited.

**References**

Kaggle (2016). House Sales in KingCounty, USA.

<https://www.kaggle.com/harlfoxem/housesalesprediction/activity>

Real Wealth (2020). How to DO a Real Estate Market Analysis.

<https://www.realwealthnetwork.com/learn/how-to-do-a-real-estate-market-analysis>