**King County Housing Dataset**

City University of Seattle

CS612 Summer 2020

Group Members:

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

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

Faculty Advisor: Professor Brian Maeng

**Introduction (Anil)**

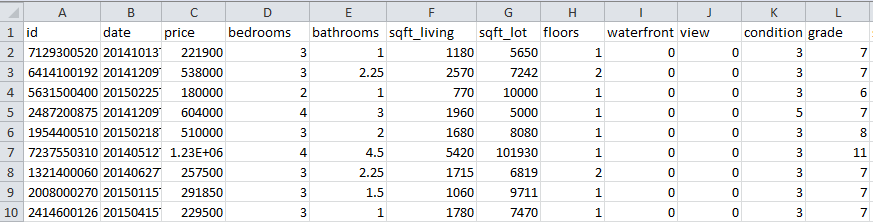
In the past, looking for a house consisted of driving around in neighborhoods to look for “for sale” signs and going to real estate agencies to look at their list of houses. With the advance of technology, today’s real estate listings can be accessed on websites like Zillow. These online real estate 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. This analysis can also be used by construction companies to figure out what types of housing is in high demand.

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

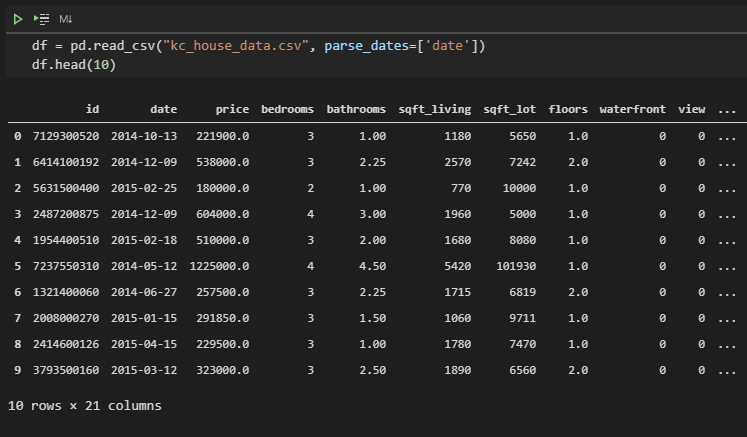
On top of looking for a correlation between different attributes and prices, the distribution of each of these attributes can be used to see the popularity of different values for each attribute. This will show what types of houses are best-selling in an area. This distribution of each attribute will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

**Data (Anil & Minh)**

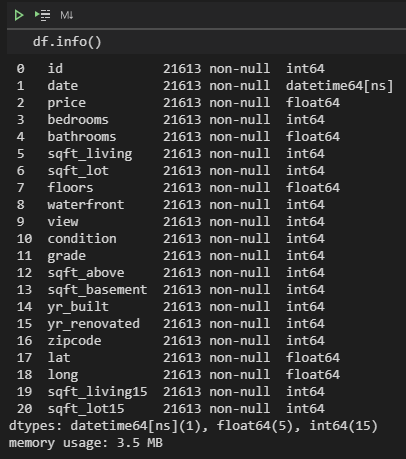
The data is taken from (Kaggle, 2016) in the form of a CSV (Comma Separated Values) file that contains 21614 records. Each record has 21 columns including the id which is the primary key. The primary key is a unique value that allows identification.



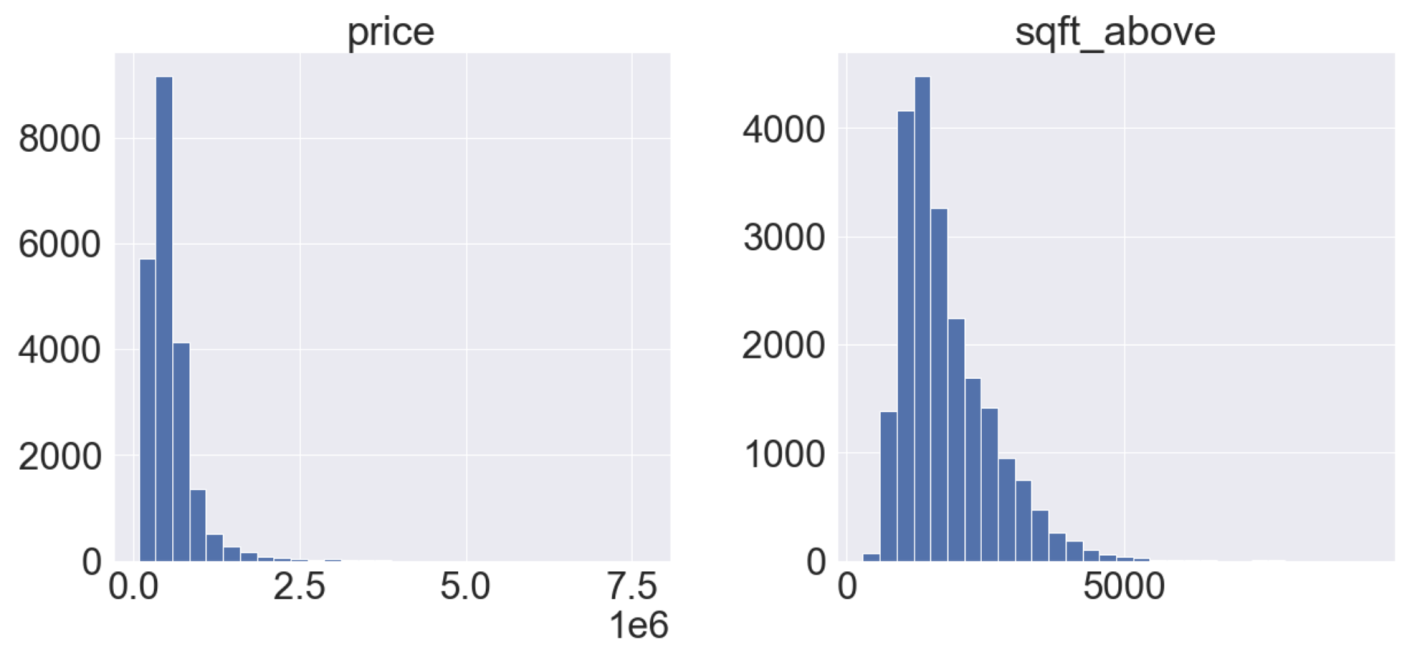
Since the data is two dimensional, a DataFrame is a suitable data type. The CSV file is read as a DataFrame object and then assigned to the variable df. The head function is used to examine the first 10 rows of the data.

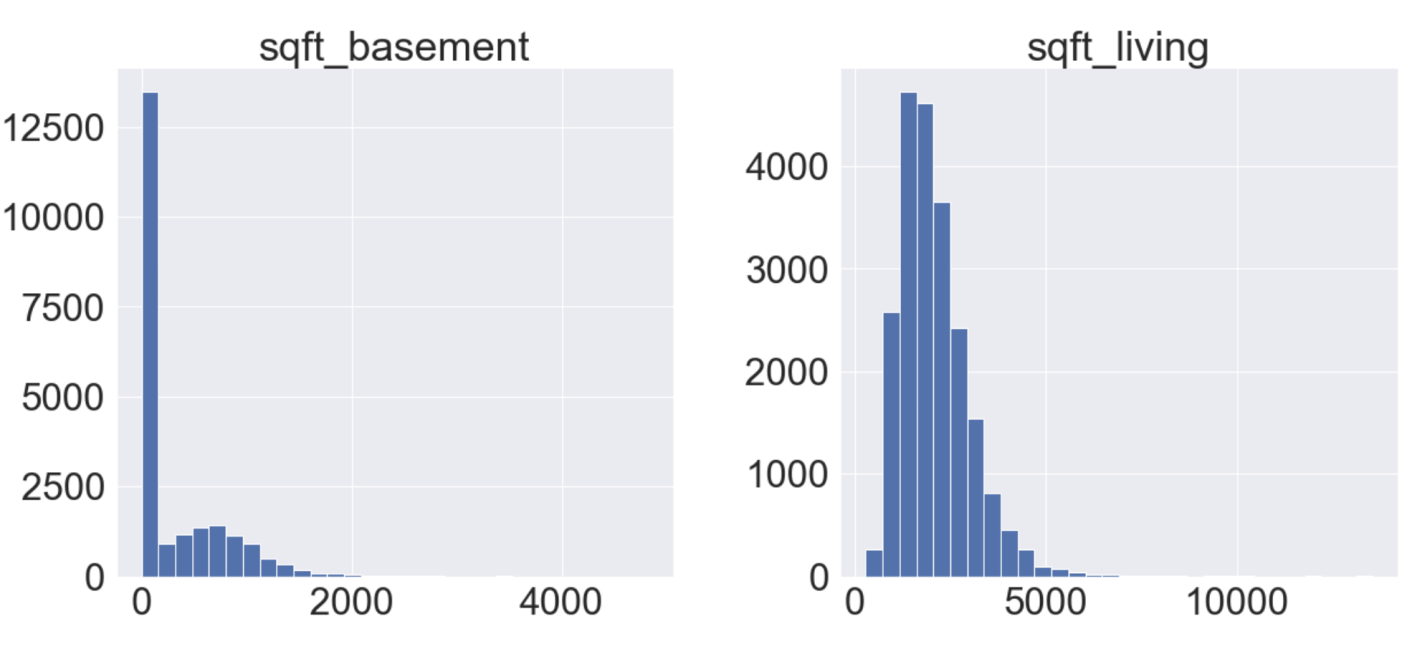


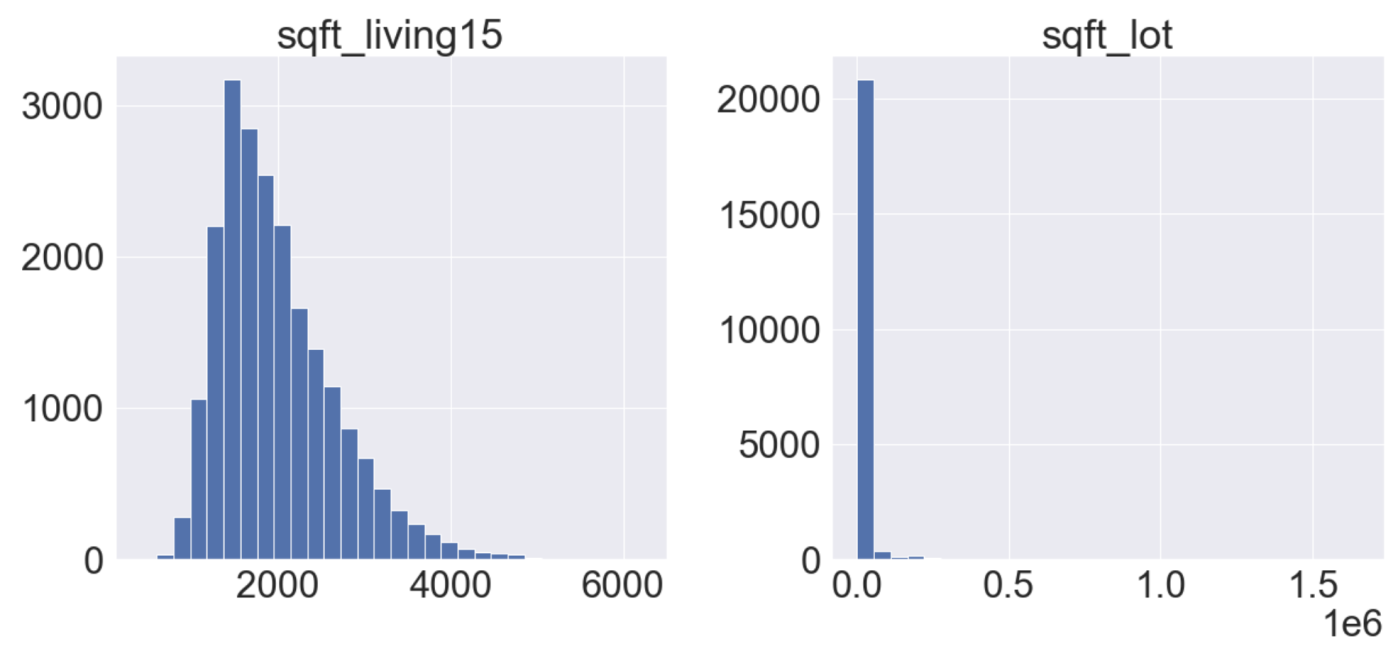
info function gives us an overview of the structure of our data. One interesting observation is how the python DataFrame uses 3.5MB (Megabytes) for the CSV file that is actually 2.39MB and 2.40MB on disk. The size on disk is usually slightly higher than the actual size because of the way files are divided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

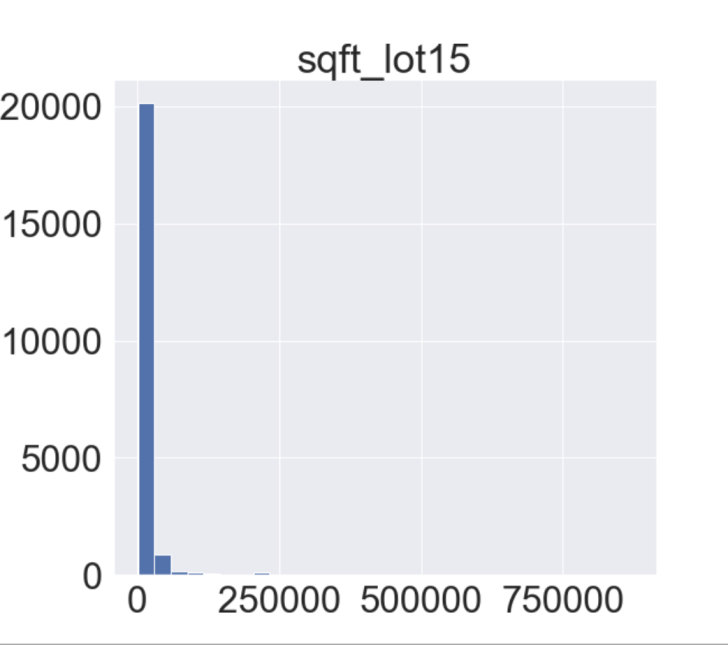


The data contains all numerical values most of these features are continuous while the rest are discrete (both ordinal and nominal). “waterfront” is a nominal binary variable while “bedroom”, “bathrooms”, “grade”, “floor” and “views” are ordinal variable. Variables such as “lat”, “long”, “zipcode” does 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 pretty normal.

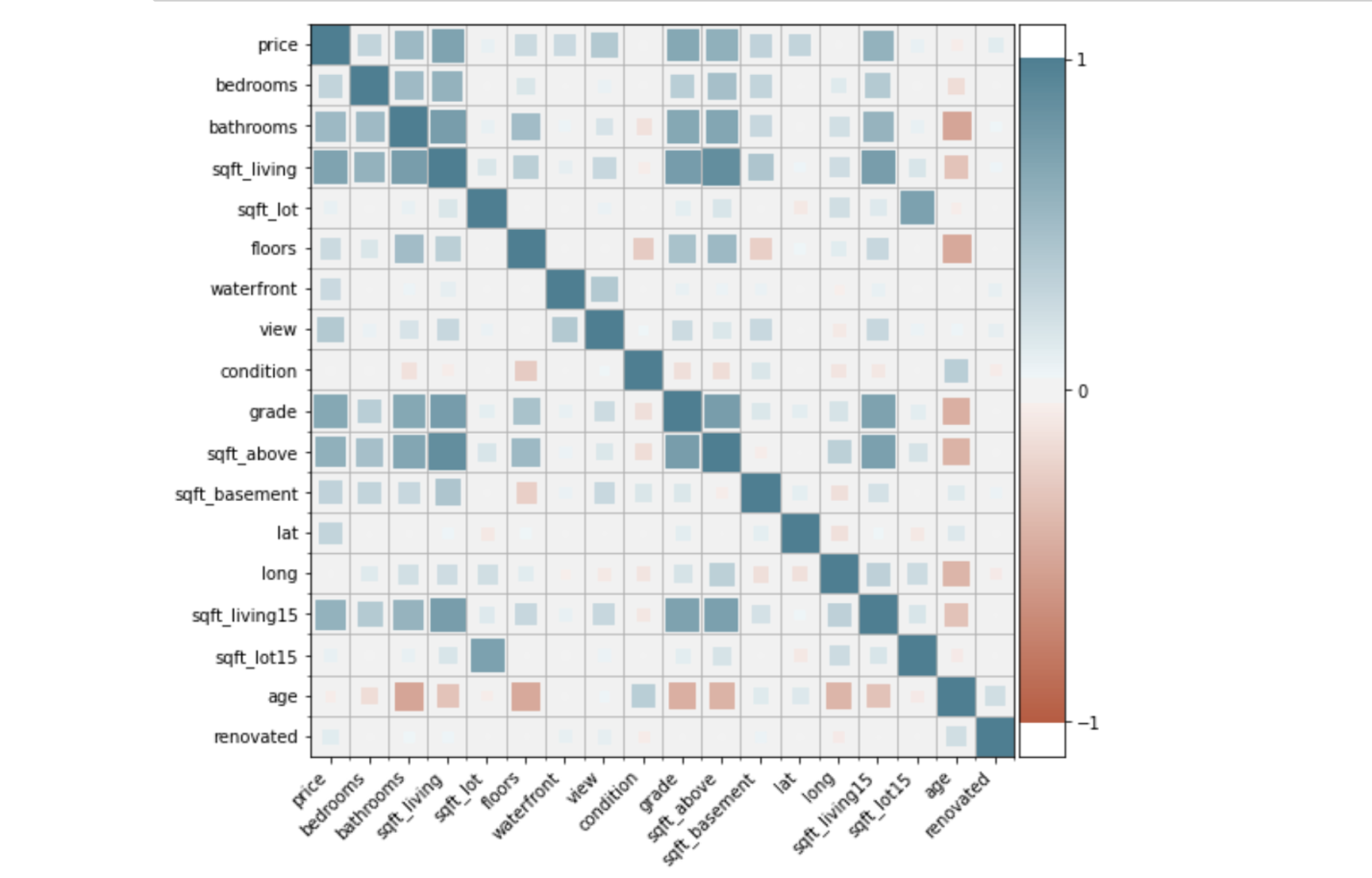
**Literature Review**

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 are the same everywhere. 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”.

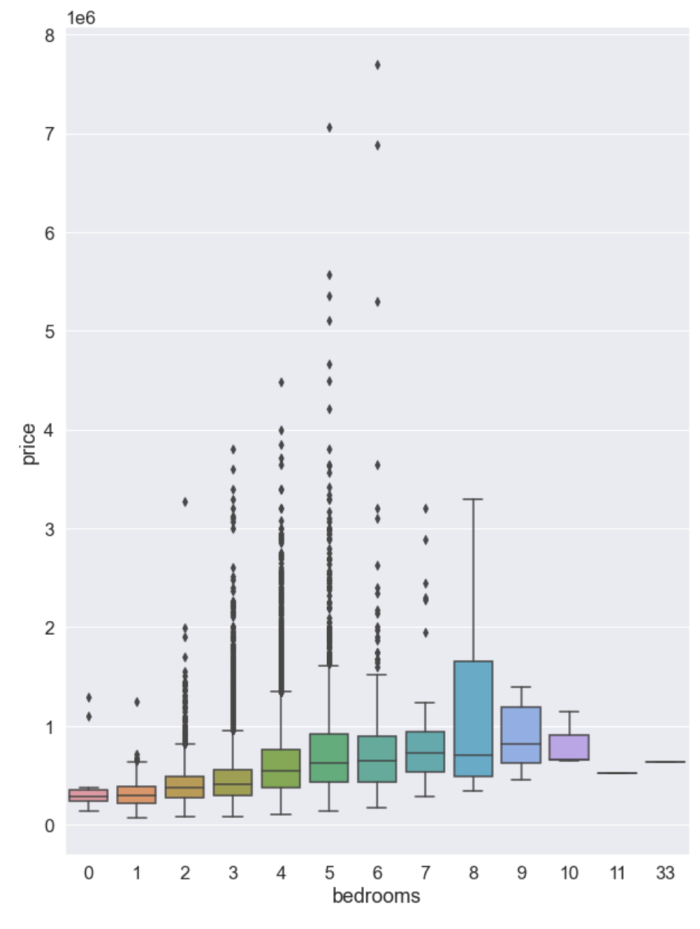
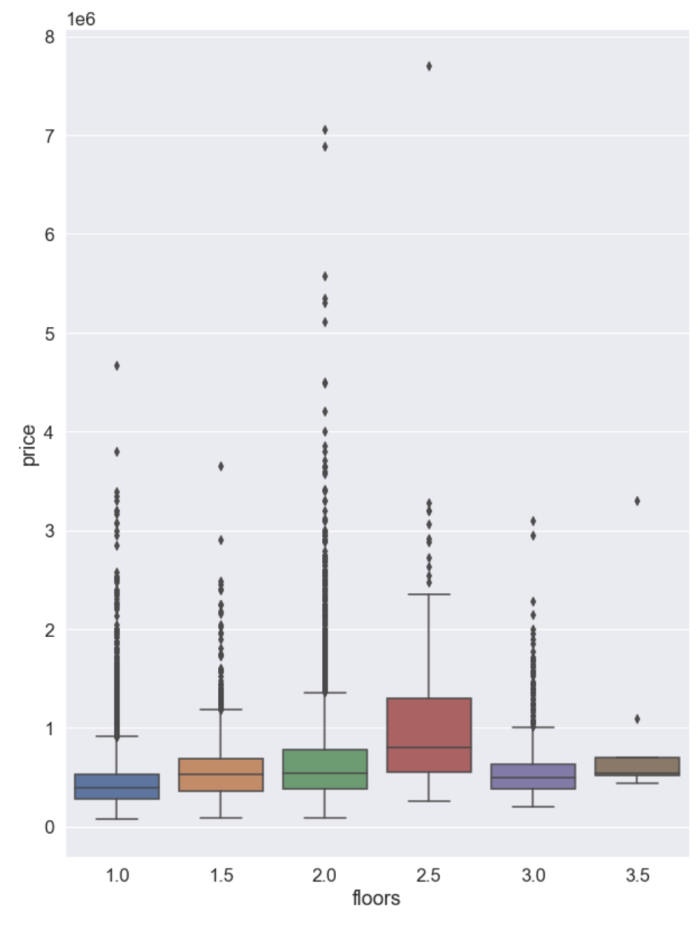
**Methodology**

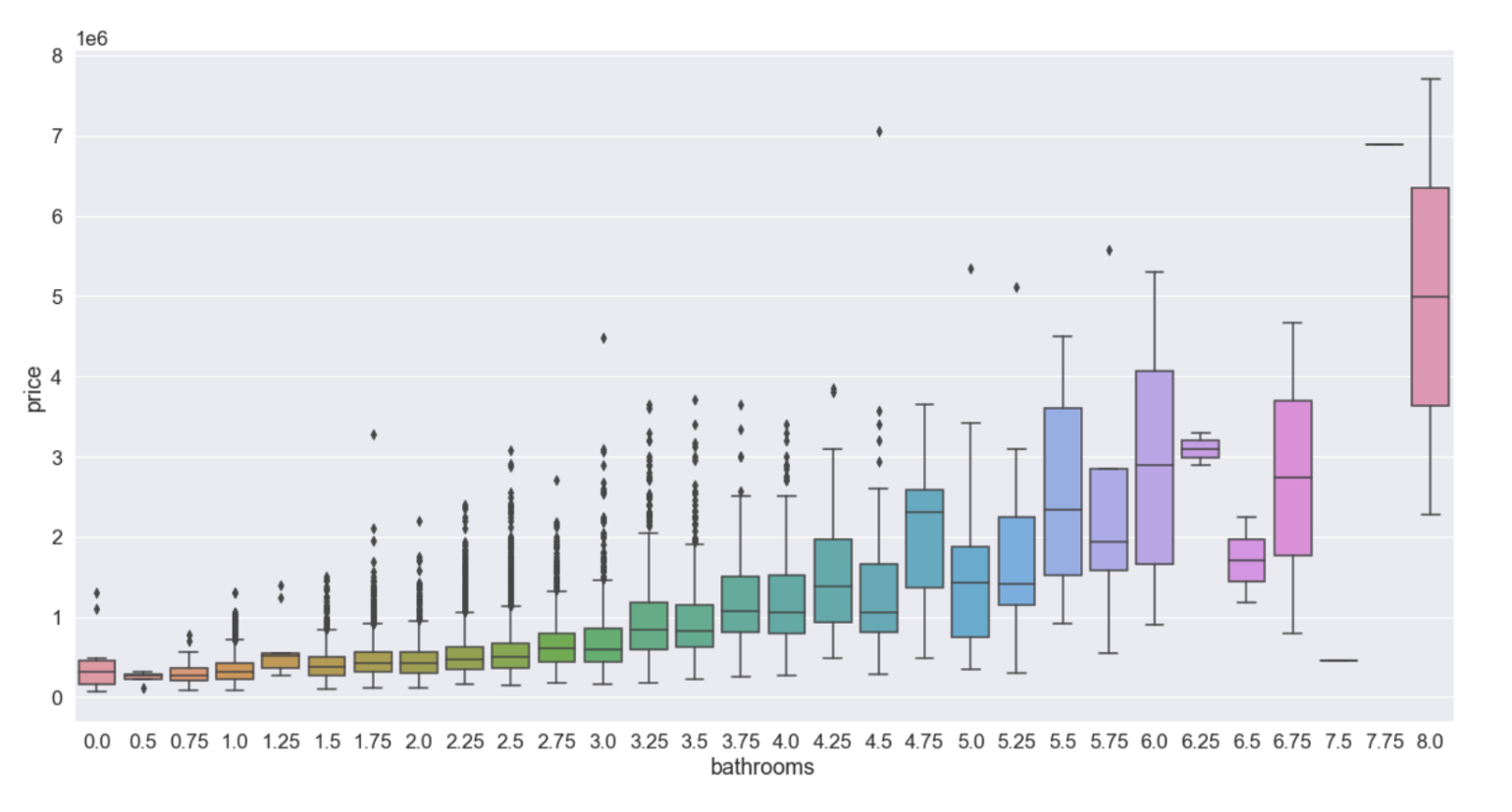
There are total 18 columns in this dataset this is of course after we have excluded unnecessary features. We first use the correlation matrix to look for any notable correlation between each variable with the price house.

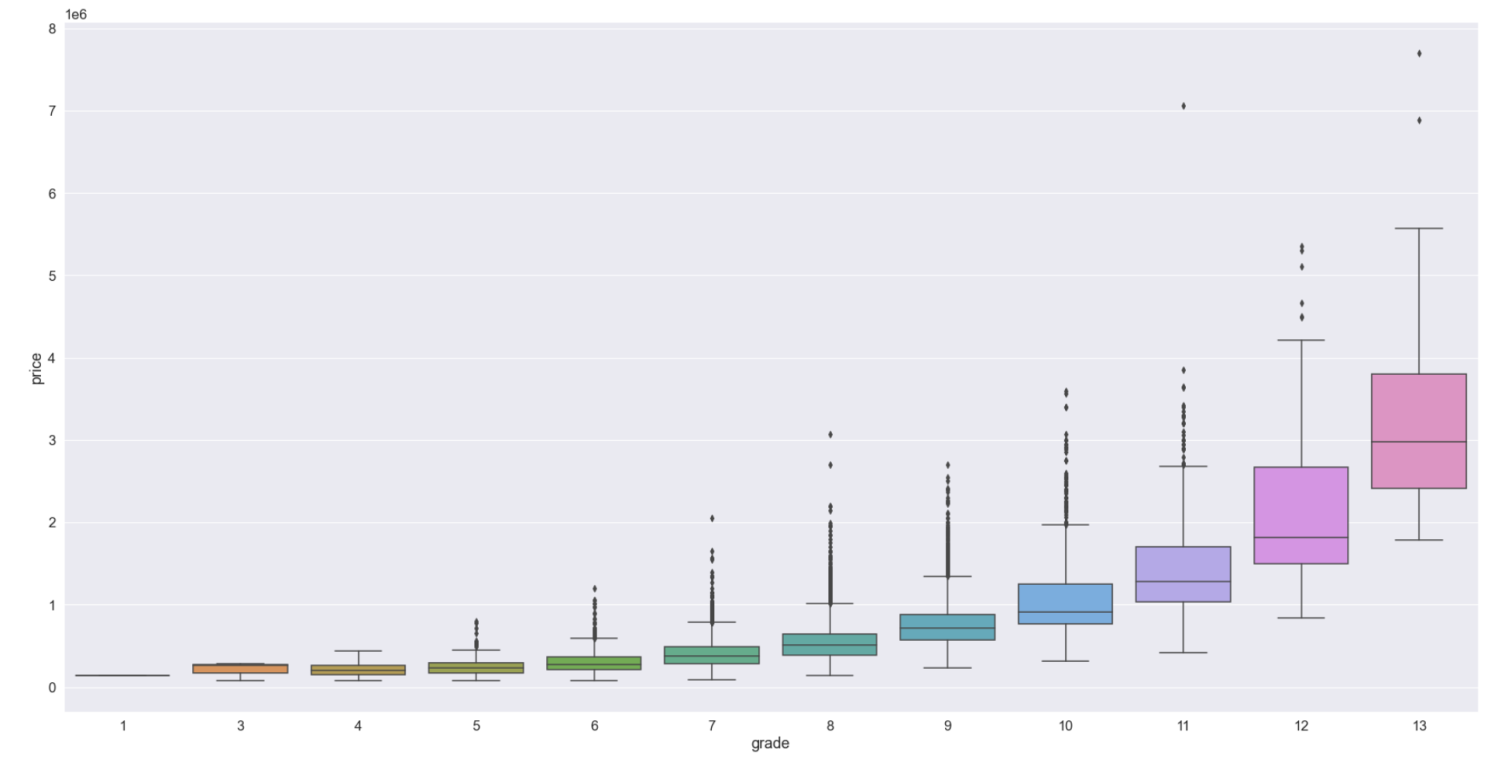


From this correlation map of the numeric features, we can infer 2 things. First, features relating to size seems to influence the prices of the house the most, namely: "bedrooms", "bathrooms", "sqft\_living", "sqft\_lot", "sqft\_above", "sqft\_living15", "grade". Features relating to the quality of the 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 by using the box plot on discrete variables: “floors”, “bedrooms”, “bathrooms” and “grade”







**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>