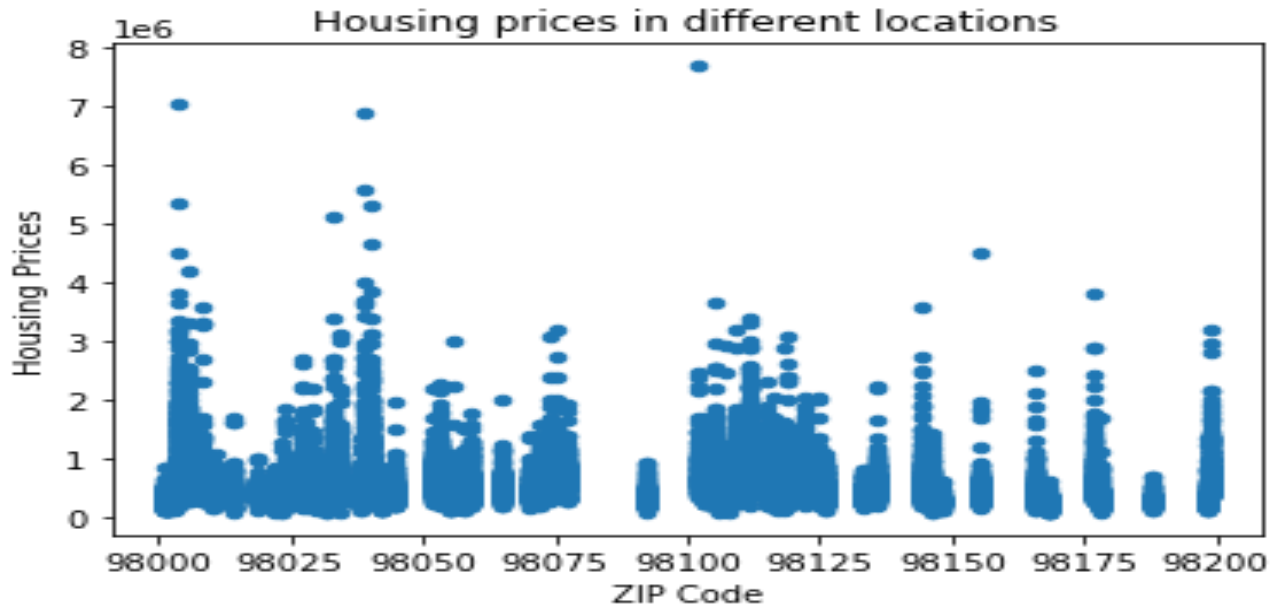
Housing Pricing in King County, USA

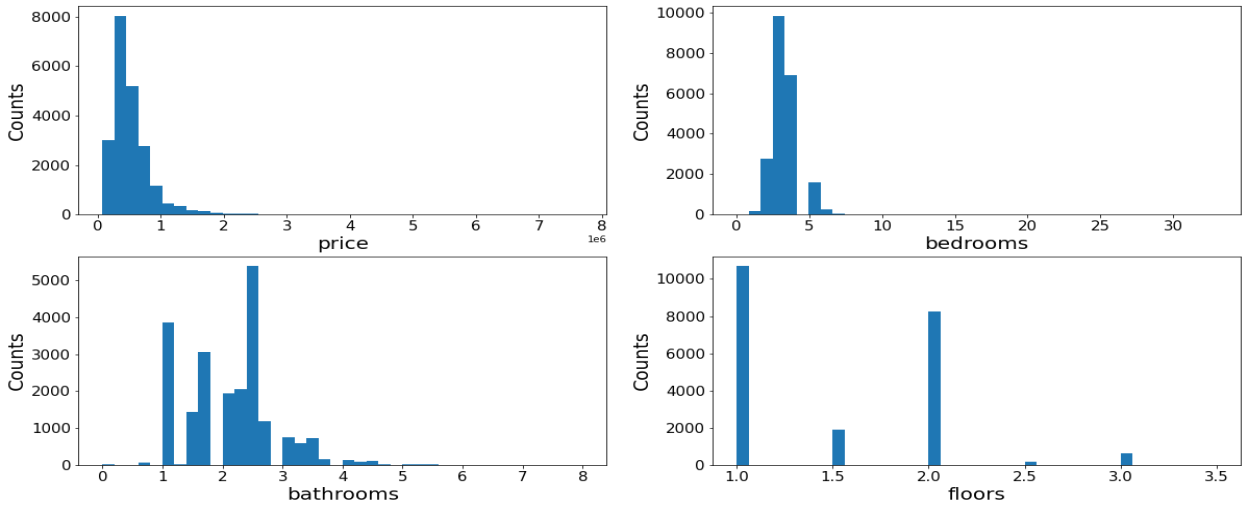
For my project for this term, I have decided to go along with the dataset in Kaggle with the topic "House Sales in King County, USA".

The dataset includes homes sold between May 2014 and May 2015 in King County, Seattle. The total number of observations is 21613. The dataset also contains 20 house features plus the price along with the observations. The description of the 20 features are as follows:

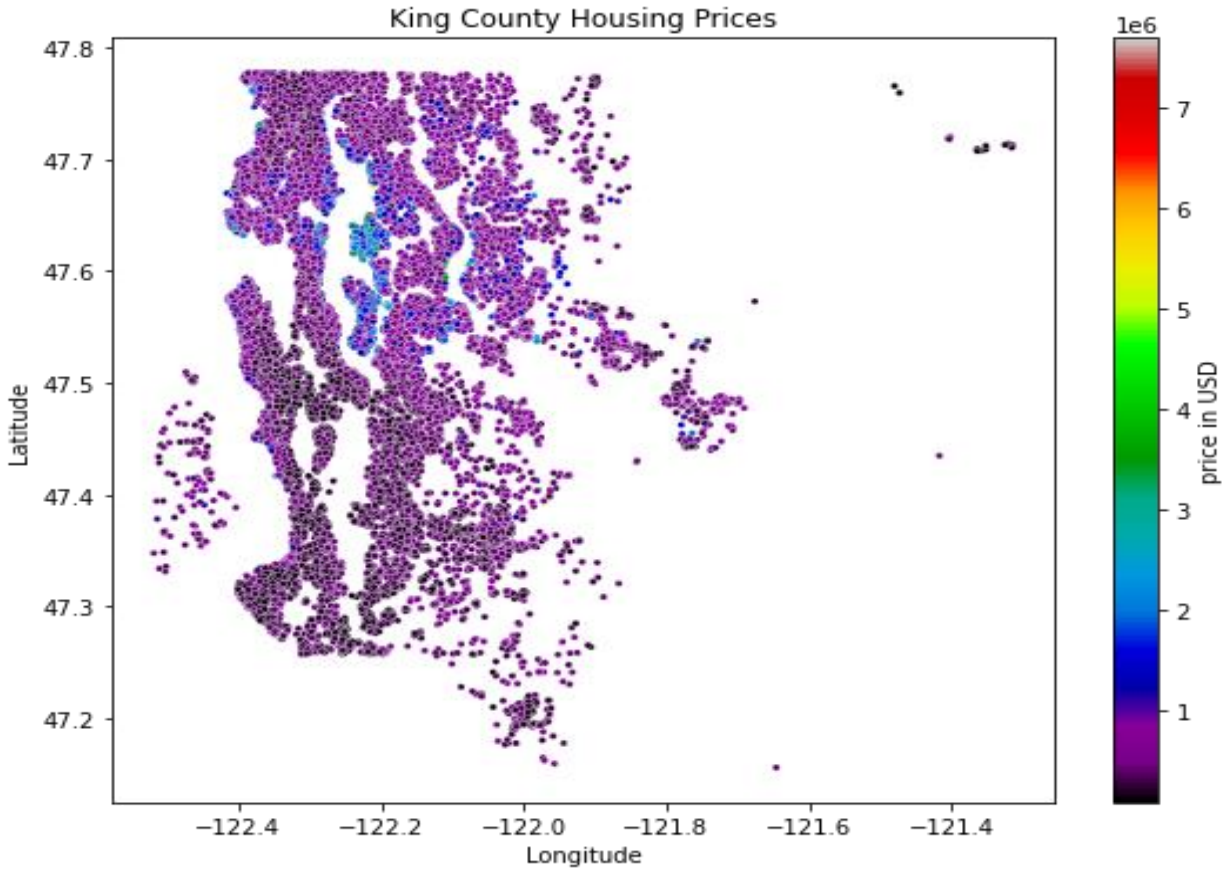
Feature	Description
id	unique numeric number assigned to each house being sold.
date	date on which the house was sold out.
price	price of house.
bedrooms	number of bedrooms in a house.
bathrooms	Number of bathrooms in a house
sqft_living	Square footage of the house
sqft_lot	Area on which the house is located
floors	Floor level of the house
waterfront	If the house has waterfront view (1 means yes and 0 means no)
view	if the house been viewed (1 means yes and 0 means no)
condition	Overall condition of the house on the scale of 1 to 5
grade	Overall garde based on king county grading system on a scale of 1 to 11
sqft_above	Square footage of the house apart from basement
sqft_basement	Square footage of the basement
yr_built	When was the house built
yr_renovated	When the house was renovated
zipcode	Location of the house
lat	Latitude of the location
long	Longitude of the location
sqft_living15	Living room area in 2015
sqft_lot15	Lot size in 2015

Visualization of data

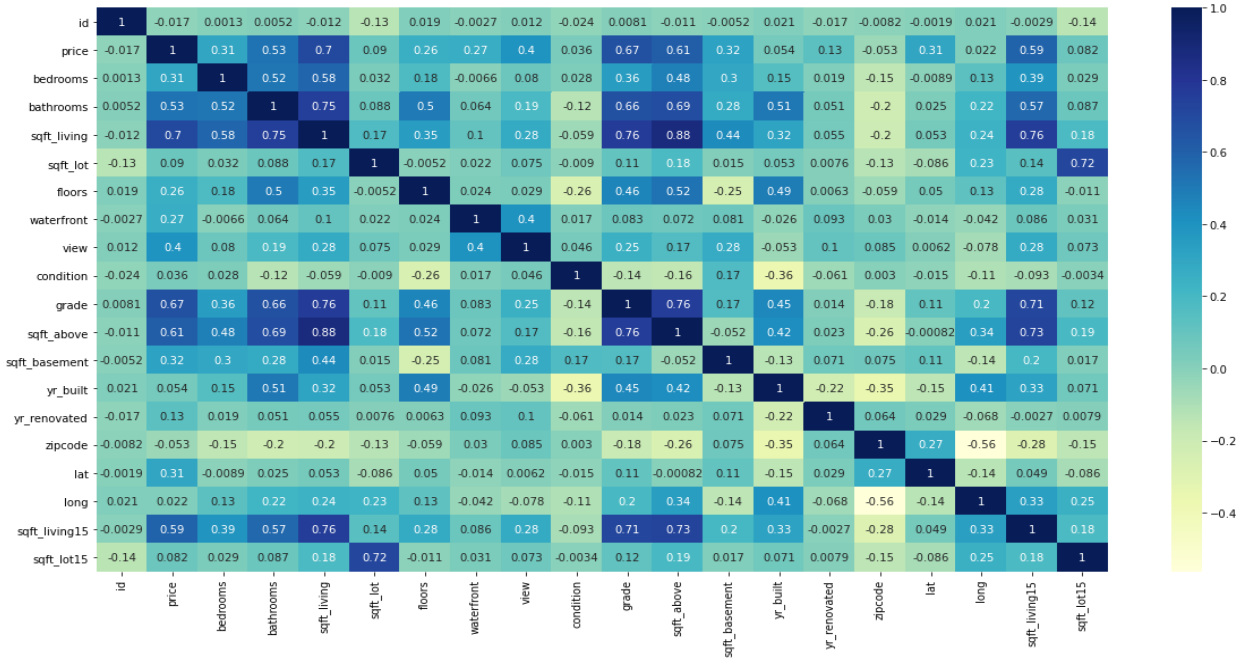




Looking at the location as it is the key in real estate. The task here is to understand the geographical distribution of the houses and the locations where the highest number of house sales were recorded. For the purpose creating scatter plot using latitude and longitude features in the data.



From the map, we can build initial insight but there is no context to help us. It is difficult to draw any meaningful conclusion, so lets look into the heatmap for correlations.



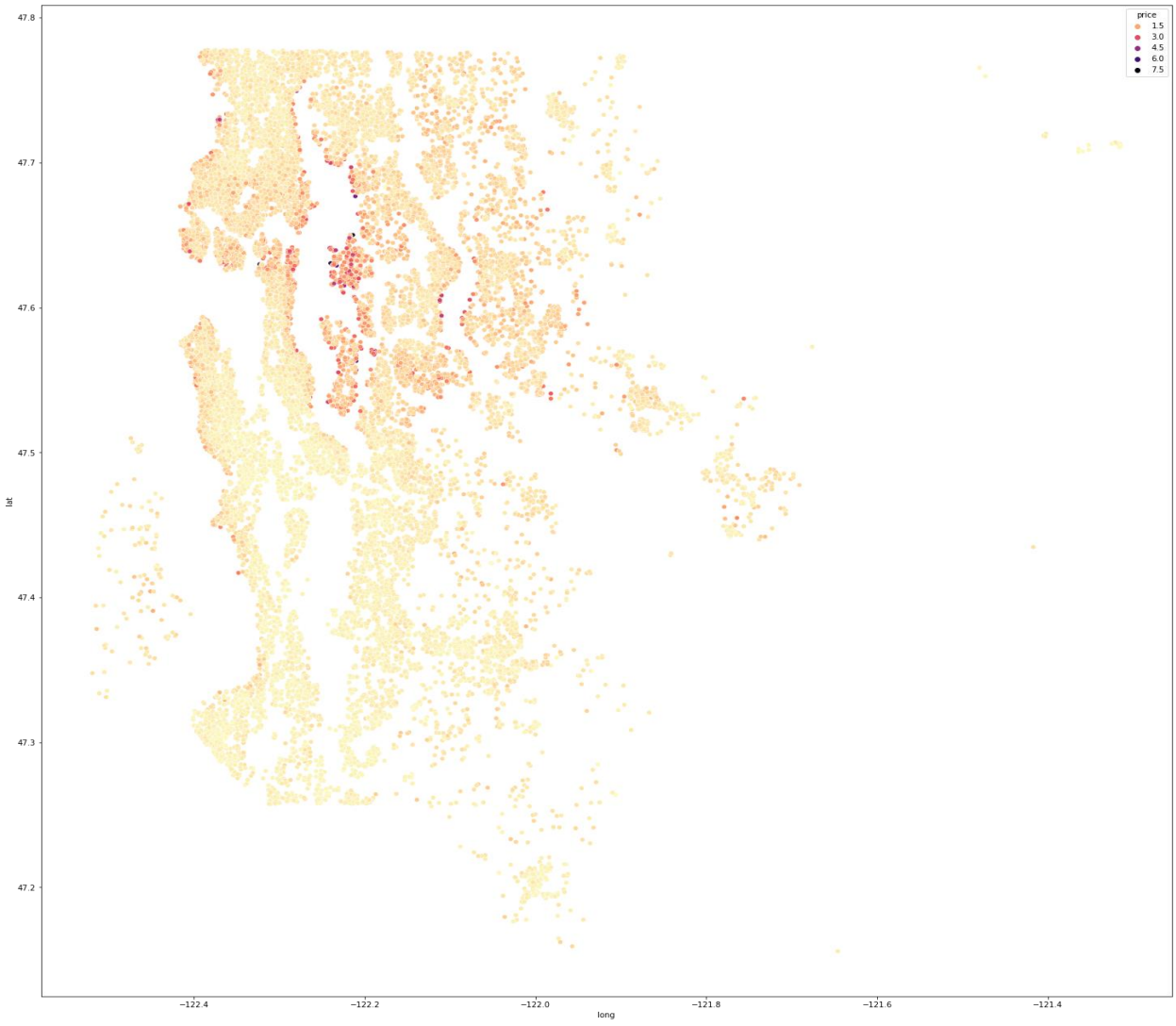
Preliminary observation

After going through the data and the graphs some of the observations of the datasets are as follows:

- By observing the data and graphs, we can interpret that the price is dependent on various features such as number of bedrooms, and bathrooms, square footage, lot area, location.
- Most houses have square foot living in between 500 and 6000 irrespective of other features. The area is directly proportionate with the price.
- Most of the houses sold have 2 to 4 bedrooms.

Detecting Outliers.

Plotting scatter plot using latitude and longitude to detect any location outlier in the data.



Feature Selection/Extraction:

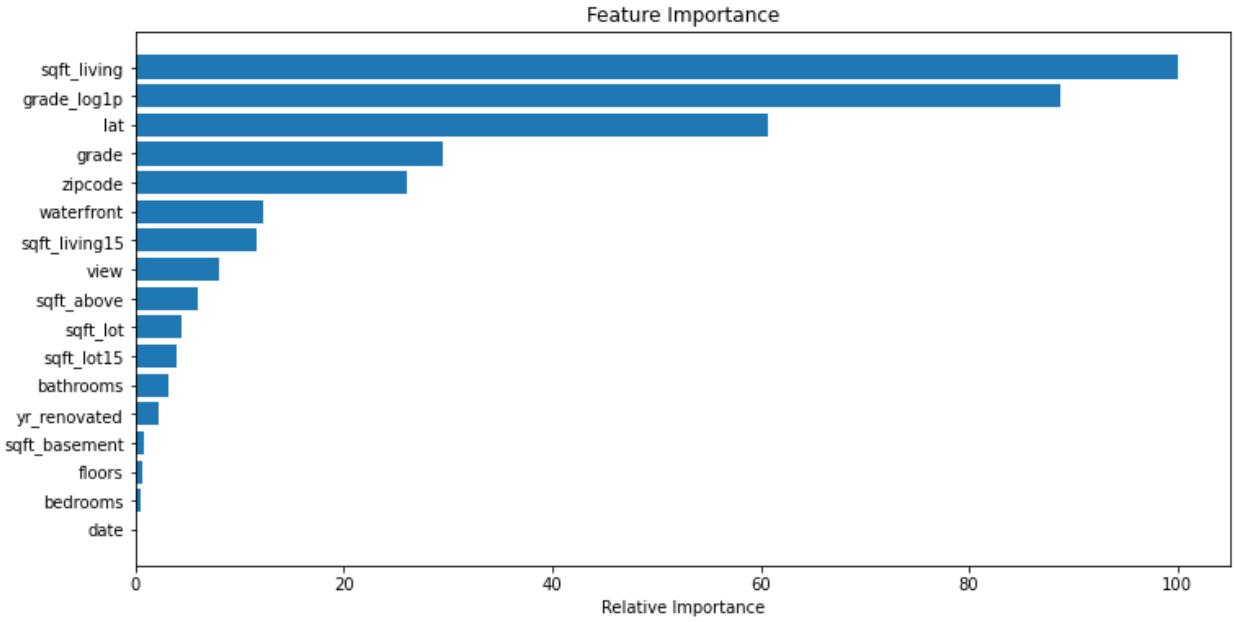
ID may not be of much use. It might be useful to know how many times a unit appears in dataset and use that as a feature but ID on its own is not helpful. Date on the other hand will probably be interesting to check the seasonal trends. I am not sure about view as it only suggests if the building been viewed. Another one that doesn't seem to be of use is sqft_above, as the square footage overall has an impact on the price independent of how it is divided. I am not sure about the zipcode as we have lat and long covering the same information.

Using the visual map of the King County (using lat and long) to check for the outliers, we can see that there are some outliers in some way. The visualization also suggests that the location does have positive impact on the price. Similarly with the Pearson Correlation matrix, sqft_lot and sqft_lot15 are strongly correlated to each other but very weak relation with everything else. It also can be inferred that about 61% of the values for the sqft_basement is zero, so can be converted into binary has or does not have basement.

Modelling.

We are trying to predict the housing prices here and the value to be predicted is continuous, thus we considered linear regression and gradient boosting regressions algorithms. The Coefficient of determination for Linear Regression came to be 68.90% compared to 88.77% for Gradient Boosting. While analyzing both the models we decided that gradient boosting is more applicable in this case.

Feature importance shows significance of predictors of variables on the target after training using GBR.



Conclusion.

The model produced the coefficient of determination R^2 of 88.77%. Square footage has the most relative importance as it increases, the quality of materials increases as well. This helps in increasing the price. But in some locations the square footage didn't play much of a role in price as the size went down the price kept intact.

Although location might be the factor affecting the price in King County houses, it seems that highest valued houses are clustered in and around Seattle, Bellevue and Redmond, the technical employer's hub of the region. Other than these, the prices seemed to go down.