Predict House Price

Capstone Project 2 Milestone Report

Introduction

A price of a house is dependent on various factors like size or area, how many bedroom, location, price of other houses, and many other factors. Real estate investors would like to find out the actual cost of the house in order to buy and sell real estate properties. They will lose money when they pay more than current market cost of the house and when they sell for less than current market cost. The banks also want to find the current market price for house, when they use someone's house as a collateral for loans. Sometime loan applicant over value their house to borrow maximum loan from the bank. Banks and financial institutions also provide mortgage loan to home buyers. Local home buyers can also predict the price of the house to find out if a seller is asking for too much. Local seller can also predict their house price and find out how much is a fair market price.

Descriptive Data Analysis

The dataset for this project is downloaded as a csv file from Kaggle website. This dataset contains house sale prices for King County, which includes Seattle. It includes homes sold between May 2014 and May 2015. The data set consists of 21613 observations and 19 features plus the house price and the id columns.

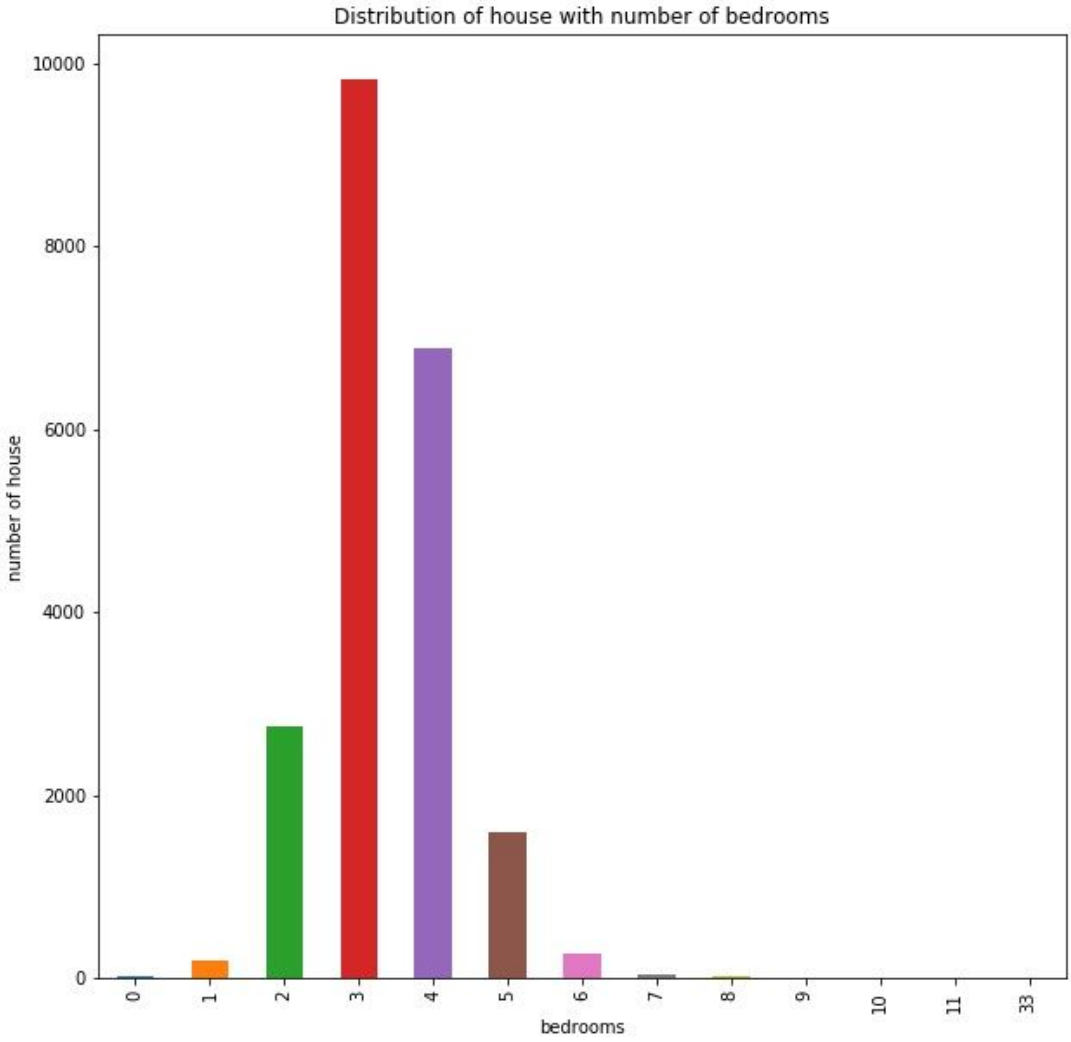
https://www.kaggle.com/harlfoxem/housesalesprediction

The columns are as follows:-
id: a notation for a house
date: Date house was sold
price: Price is prediction target
bedrooms: Number of Bedrooms
bathrooms: Number of Bathrooms
sqft_living: square footage of the home
sqft_lot: square footage of the lot
floors: total floors (levels) in house
waterfront: House which has a view of a waterfront
view: Has been viewed
condition: How good the condition is (Overall)
grade: overall grade given to the housing unit, based on King County grading system
sqft_above: square footage of house apart from basement
sqft_basement: square footage of the basement
yr_built: Built Year
yr_renovated: Year when house was renovated

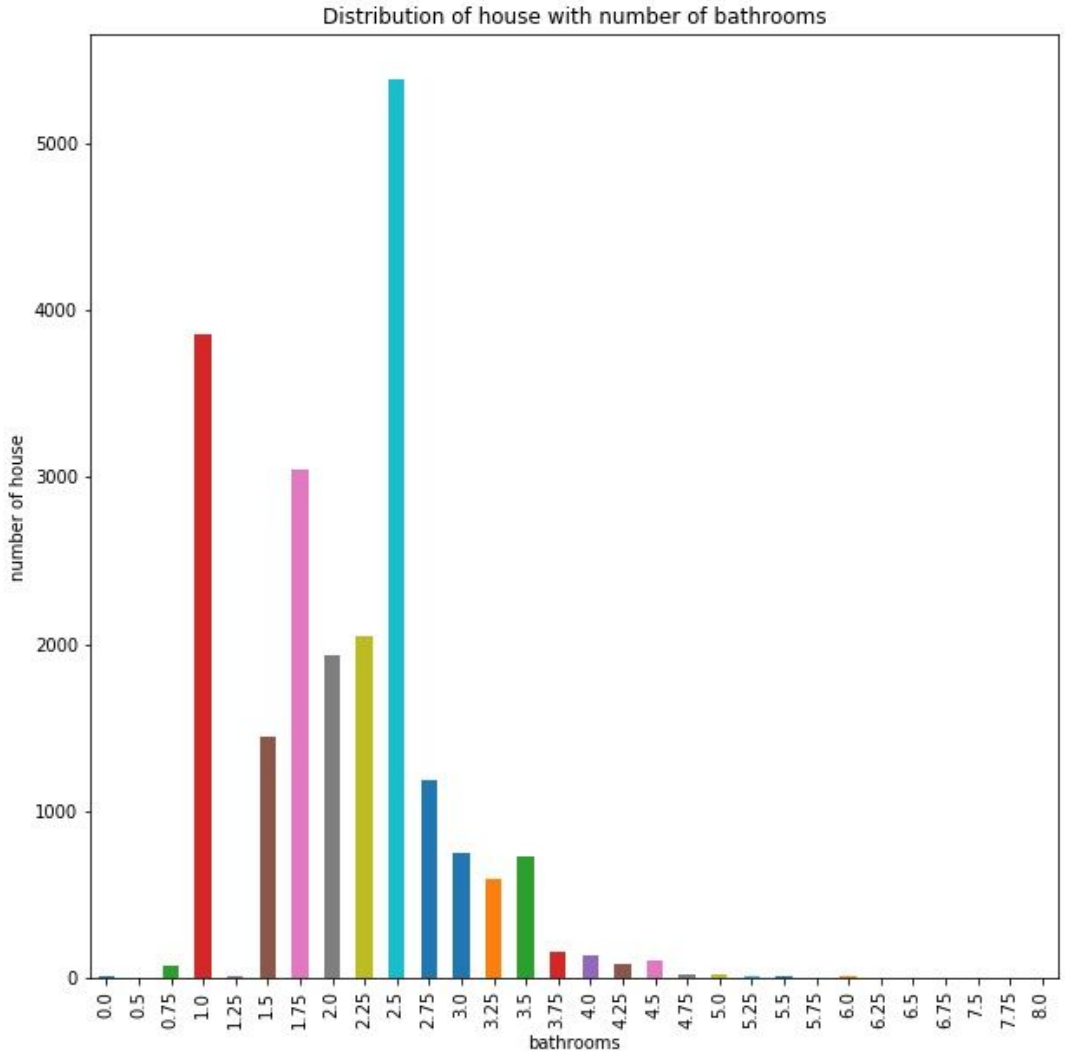
zipcode: zip
lat: Latitude coordinate
long: Longitude coordinate
sqft_living15: Living room area in 2015(implies -- some renovations) This might or might not have affected the lot size area
sqft_lot15: lot size area in 2015(implies--some renovations)

I imported the data from the csv file and converted them into pandas dataframe. From our dataset, I found that the most expensive house is priced 7,700,000.00 and the least expensive house is priced 75,000.00. The average or mean price of a house is 540,088.14 and the median price of a house is 450,000.00. I plot the histogram of price and found the dataset is right skewed. This indicates that there might be some outliers. I also plotted a boxplot of price to check for outliers in the dataset and found that there are some outliers in our dataset. Then, I checked for outliers for other features like number of bedrooms, number of bathrooms, sqft_living, sqft_lot, floors, condition, grade, sqft_above, sqft_basement. Except for the column number of floors, all other columns show that there might be some outliers. Then, I converted the date column from string to datetime data type. Then, I checked how many houses were sold on 2014 and how many on 2015. There were 14633 houses sold on year 2014, and 6980 houses sold on year 2015.

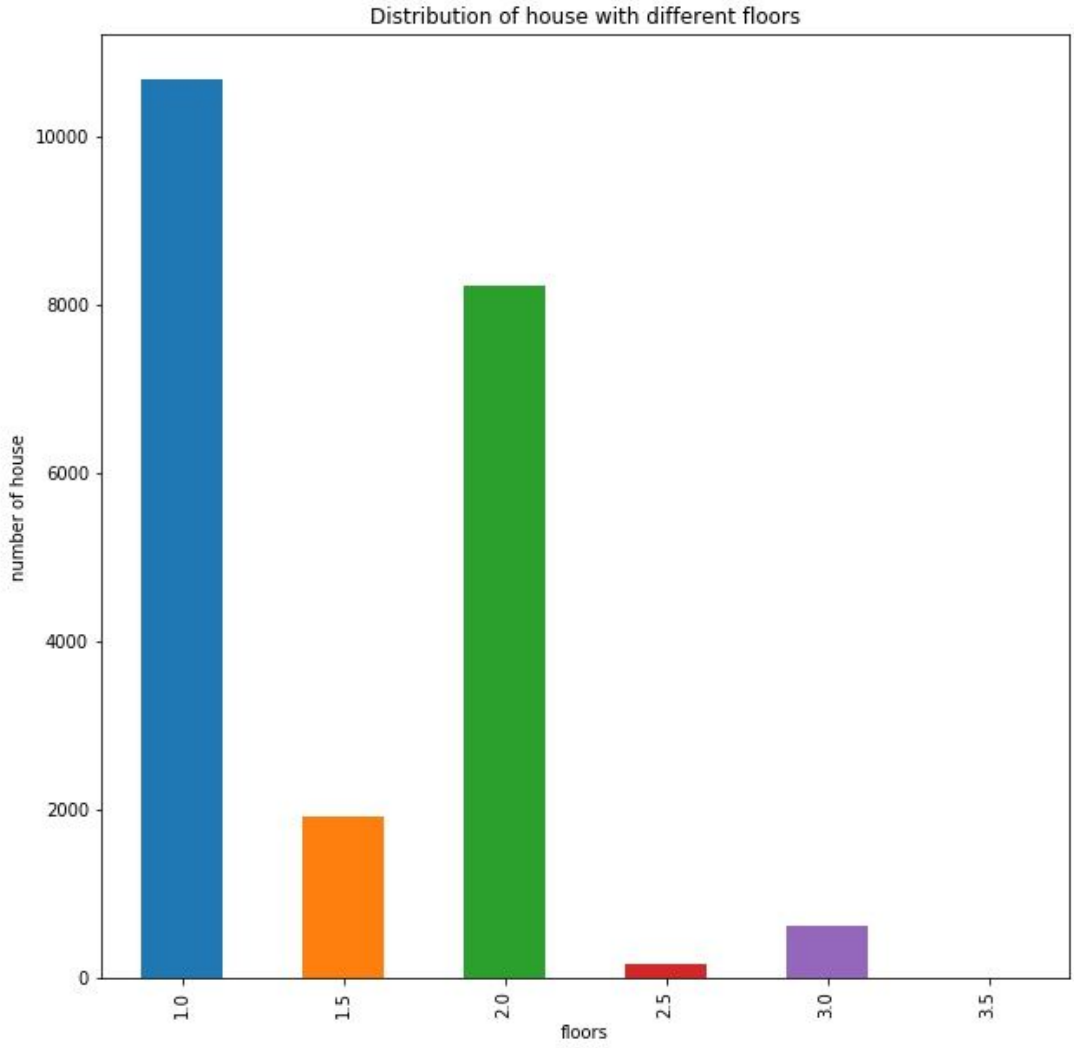
I plotted bar plots with grouped data to see the distributions and demographics of the data.



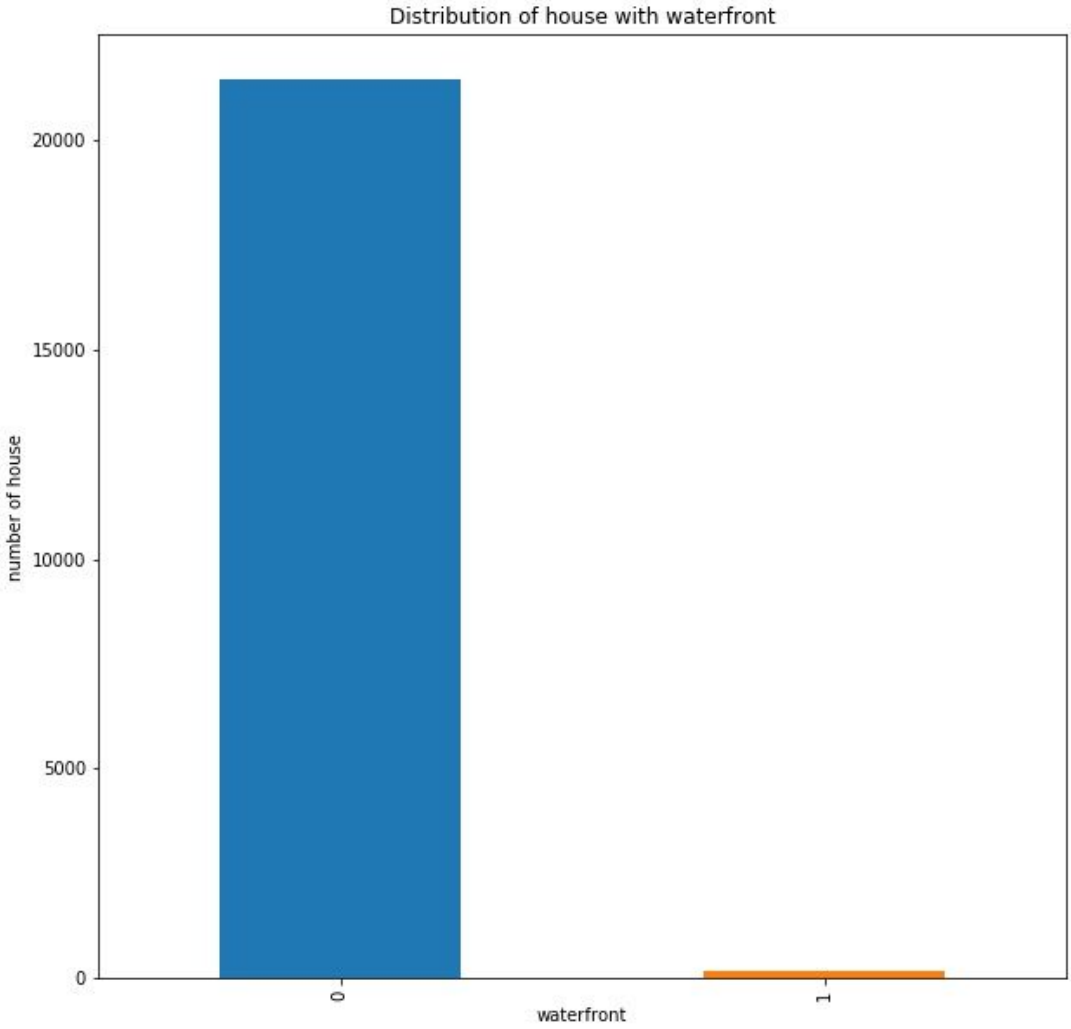
The distribution of house with number of bedrooms plot tells us that most of the houses have 3 bedrooms.



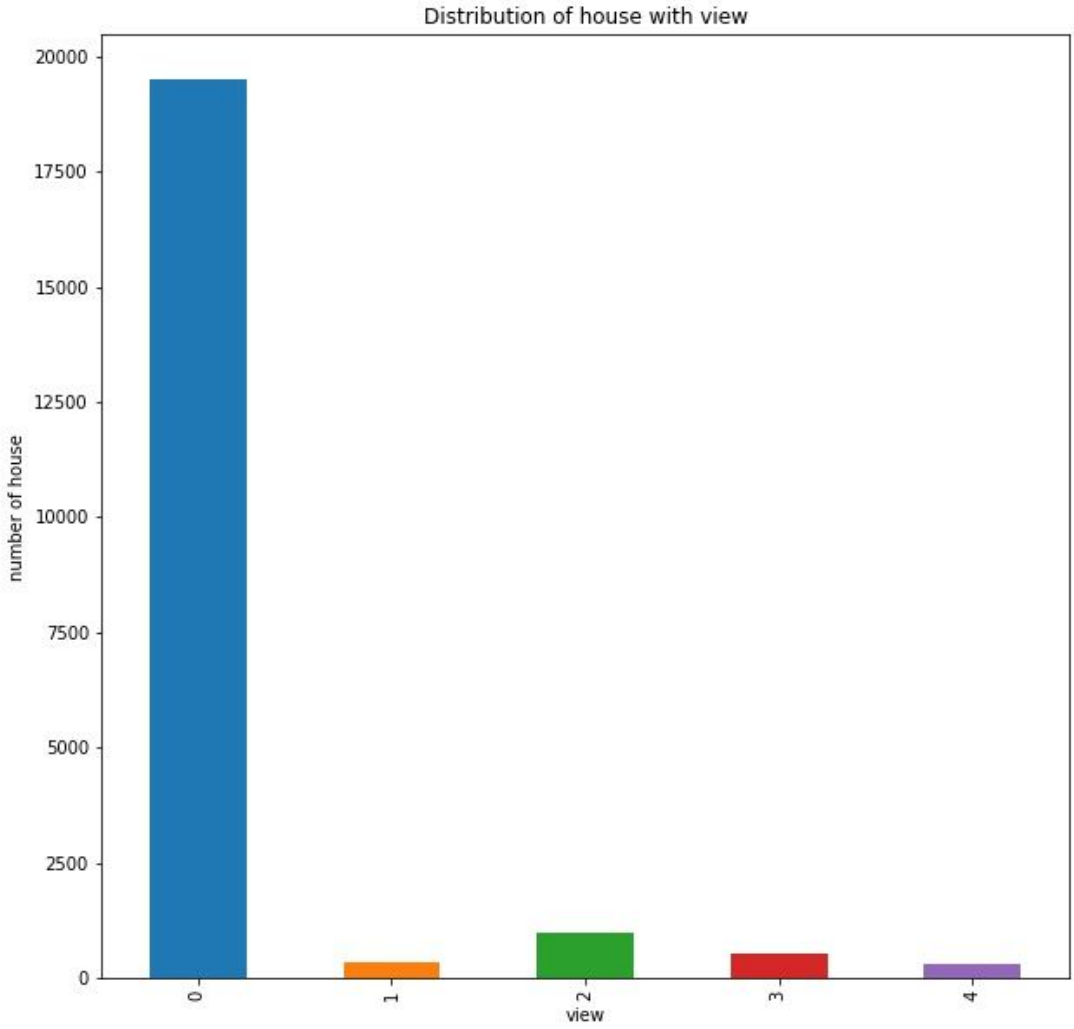
The distribution of house with number of bathrooms plot tells us that most of the houses have 2.5 bathrooms.



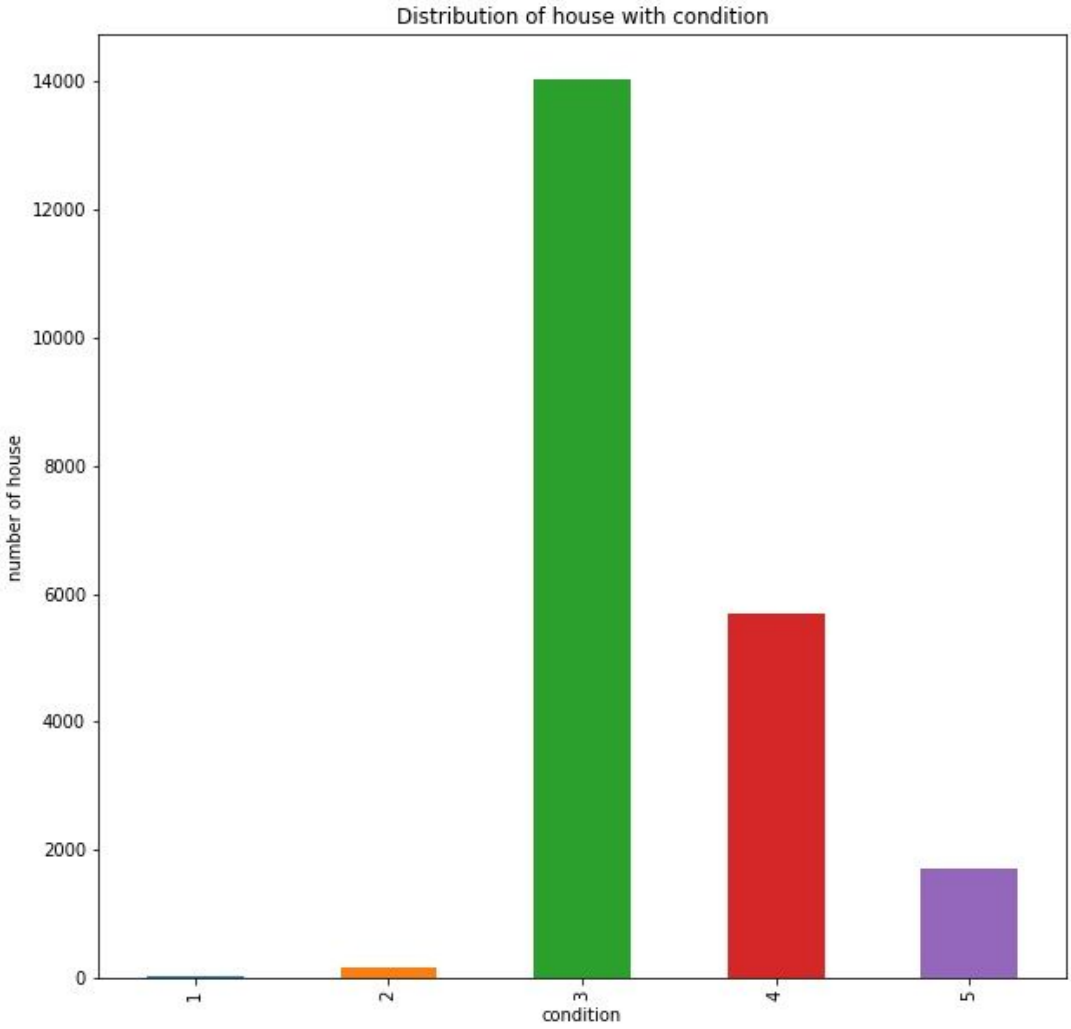
Grouping the house by floors tell us that most of the houses have 1 floor.



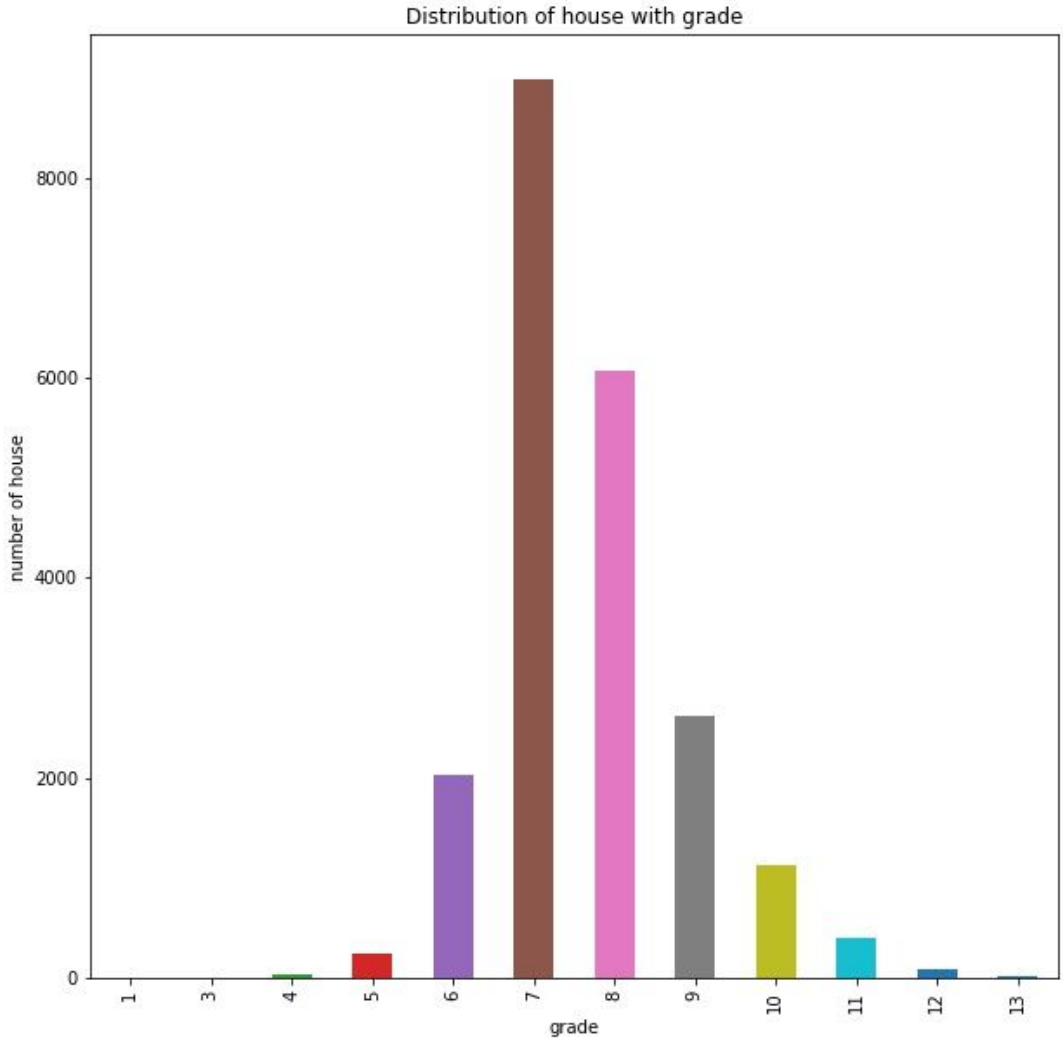
Grouping the house by waterfront tells us that most houses do not have waterfronts.



The distribution of house with view bar plot tells us that most of the house have 0 views. It means that the houses are sold to the first viewer of the house.

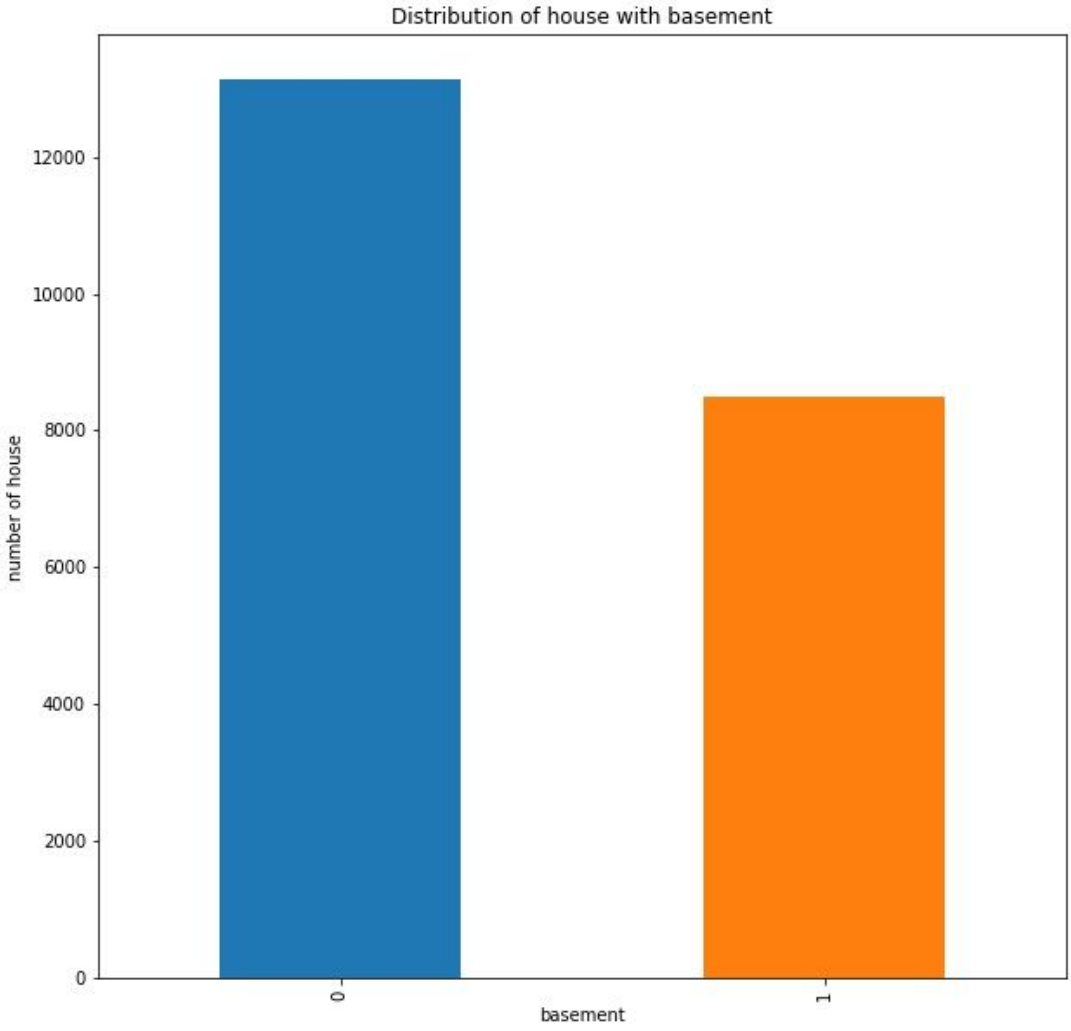


Grouping the house by condition tells us that most of the house have 3 points for condition.

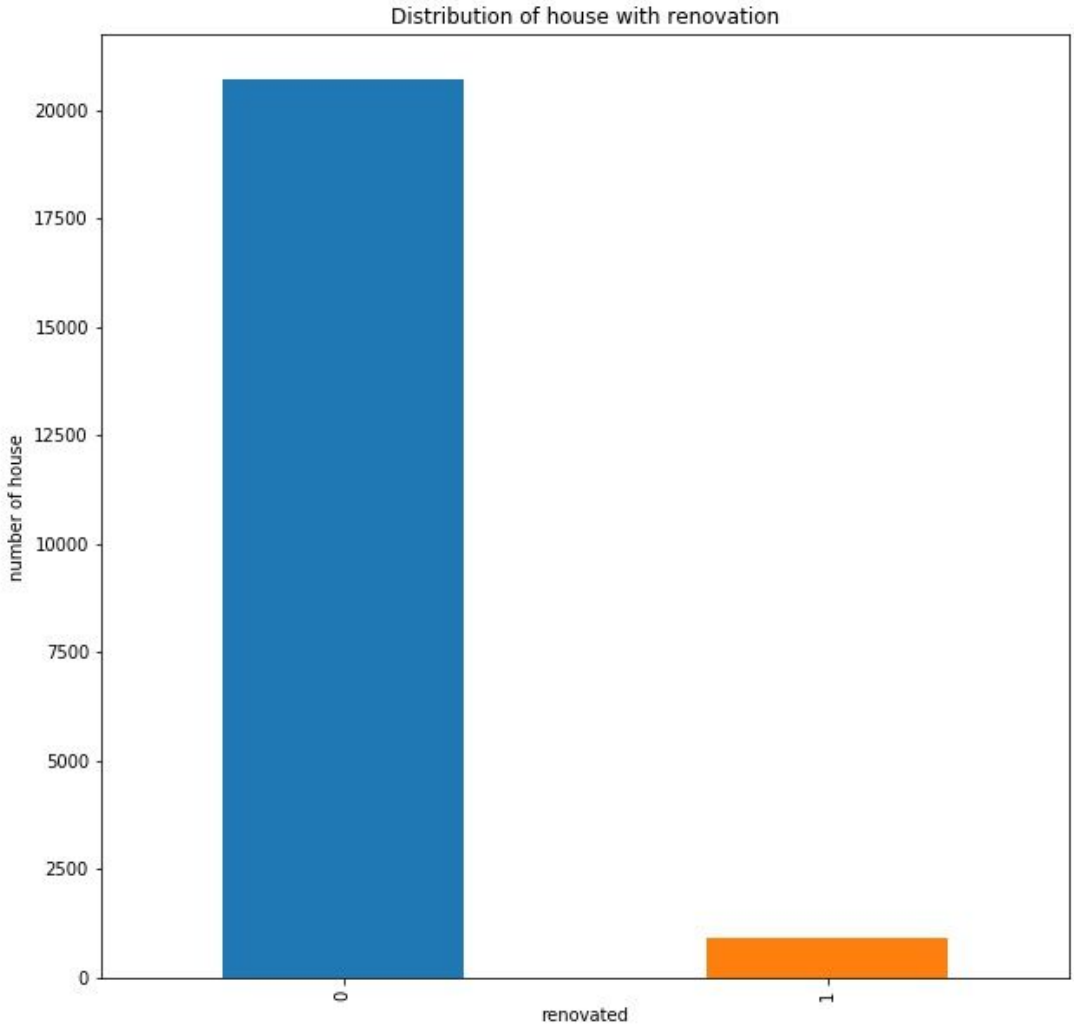


Grouping the house by grade tells us that most of the house have 7 points for grade.

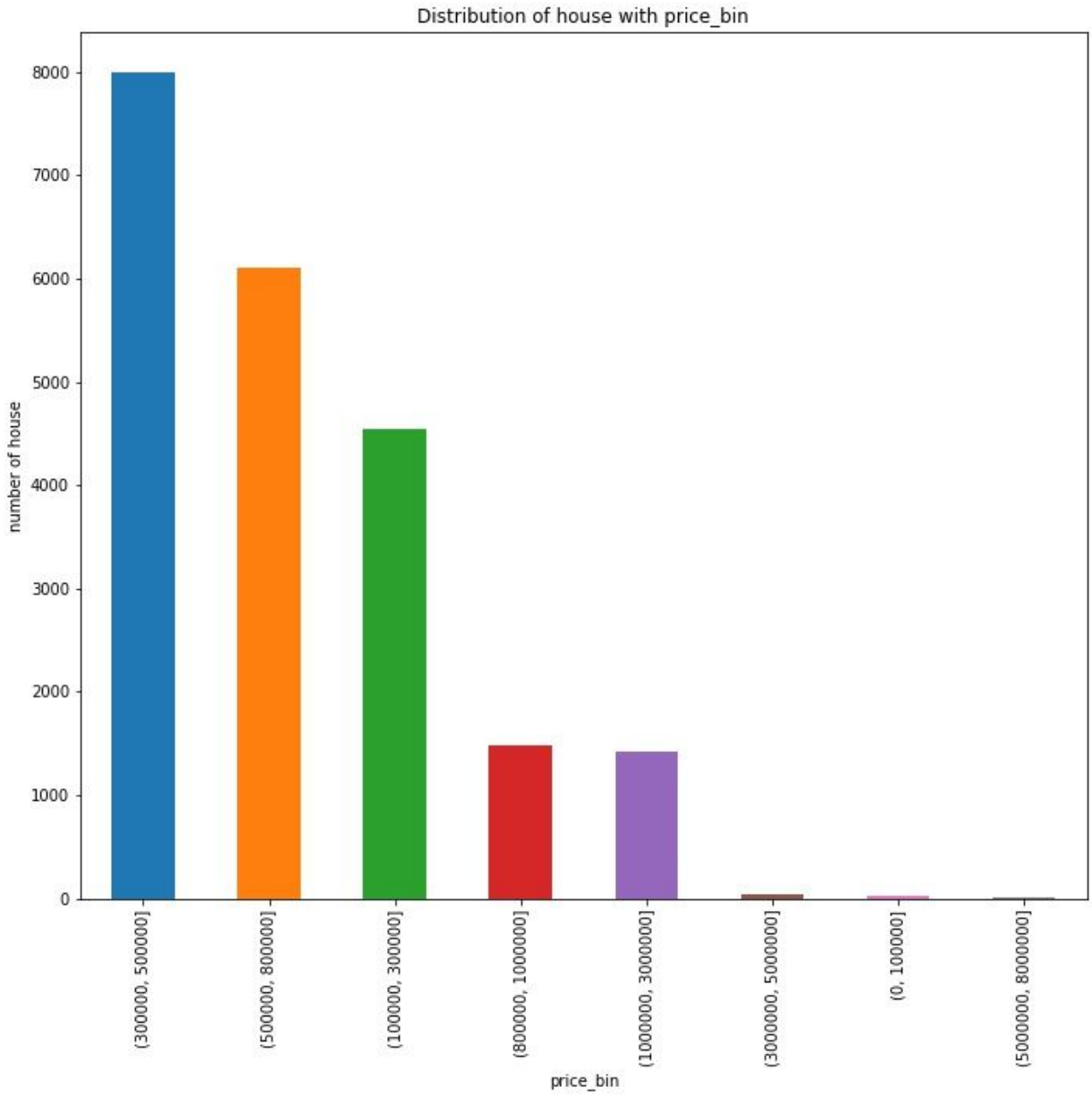
Then, I created a new column named basement to hold boolean data of house with or without the basement.



I created a bar plot and found most of the houses does not have basement.

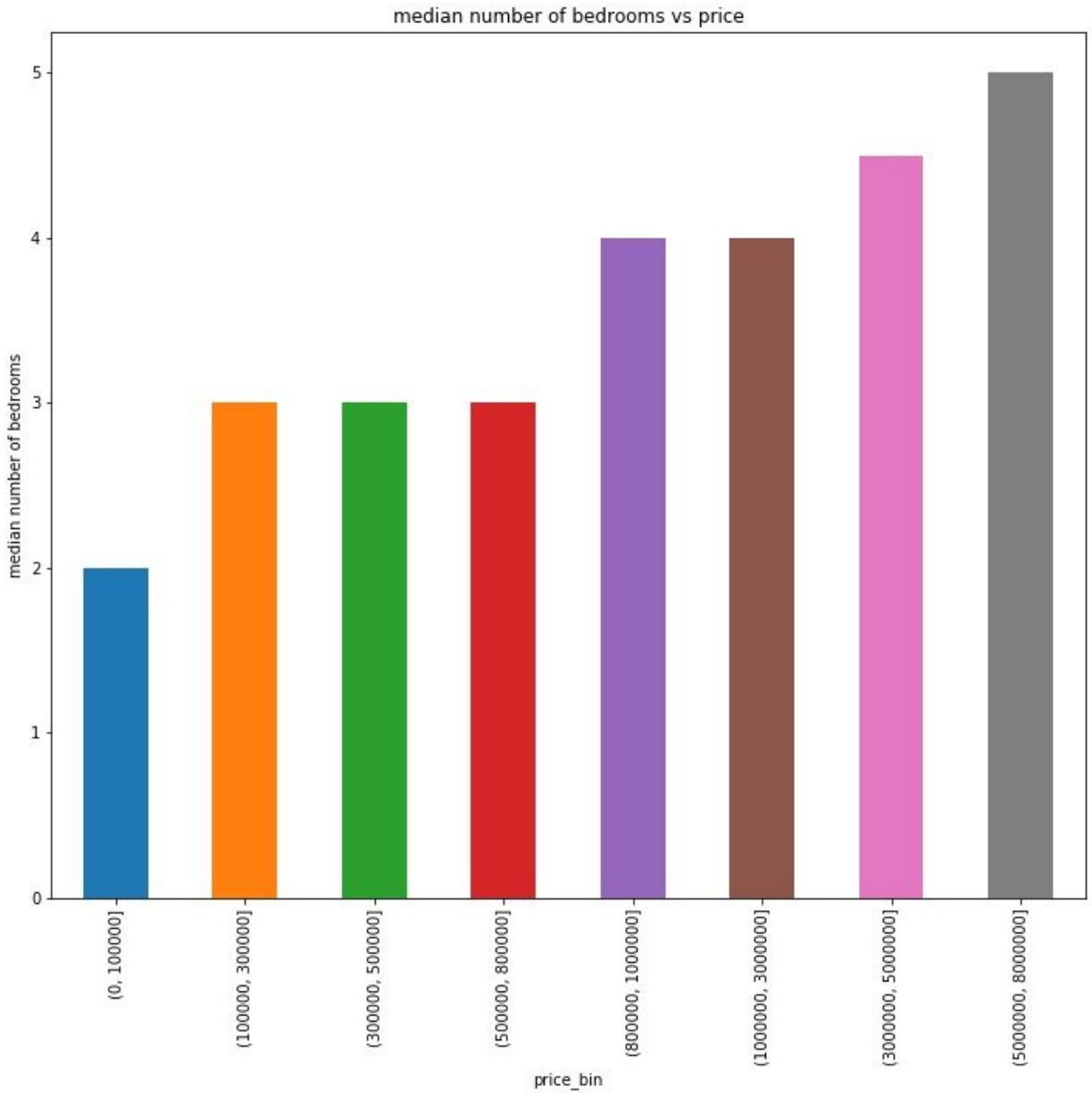


Grouping the house by renovated or not renovated, I found that most of the houses are not renovated.



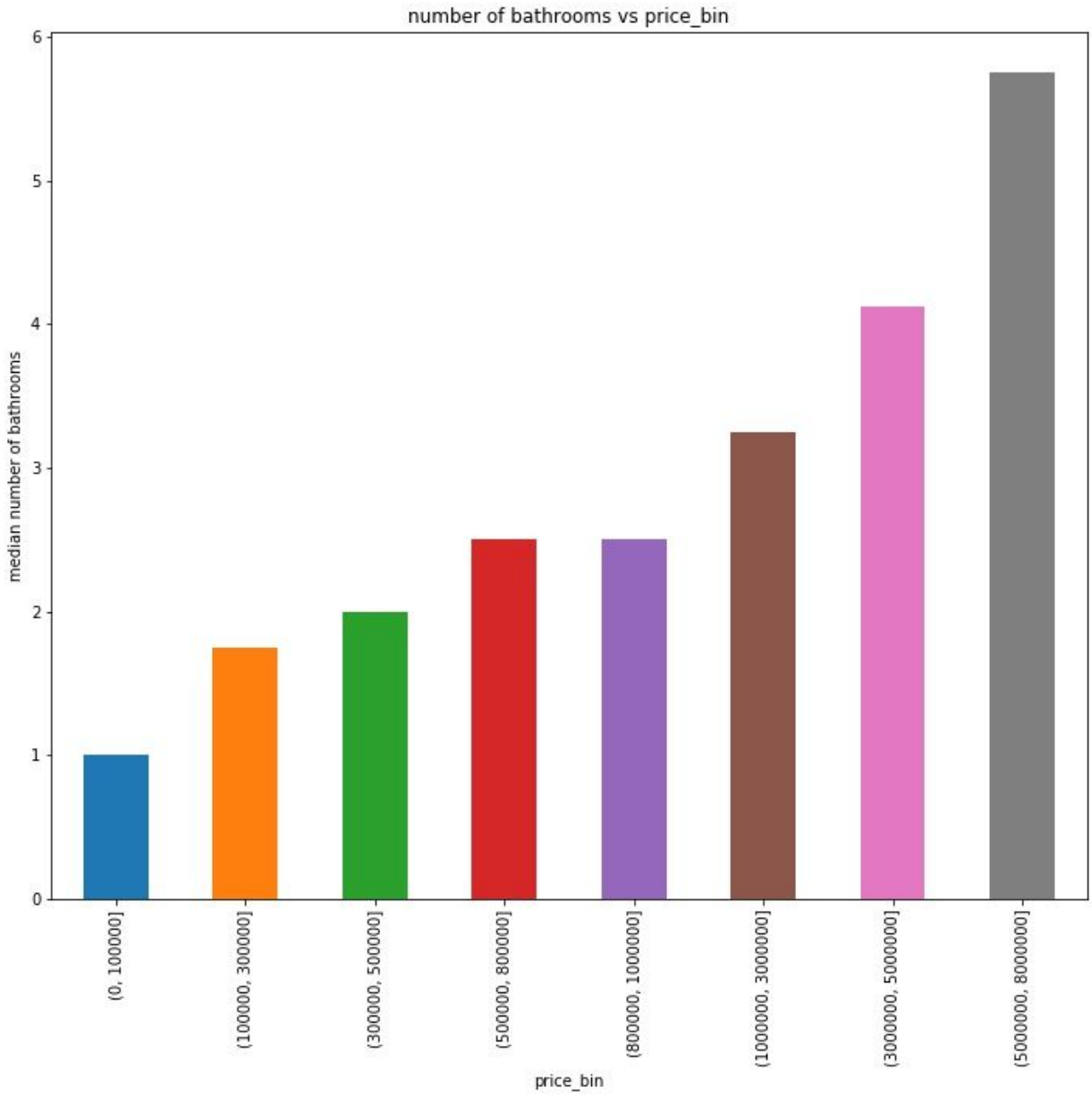
Grouping the house by price_bin, I found that most of the house are priced between 300,000 to 500,000.

Then, I plot bar plot of median number of bedrooms vs price_bin.

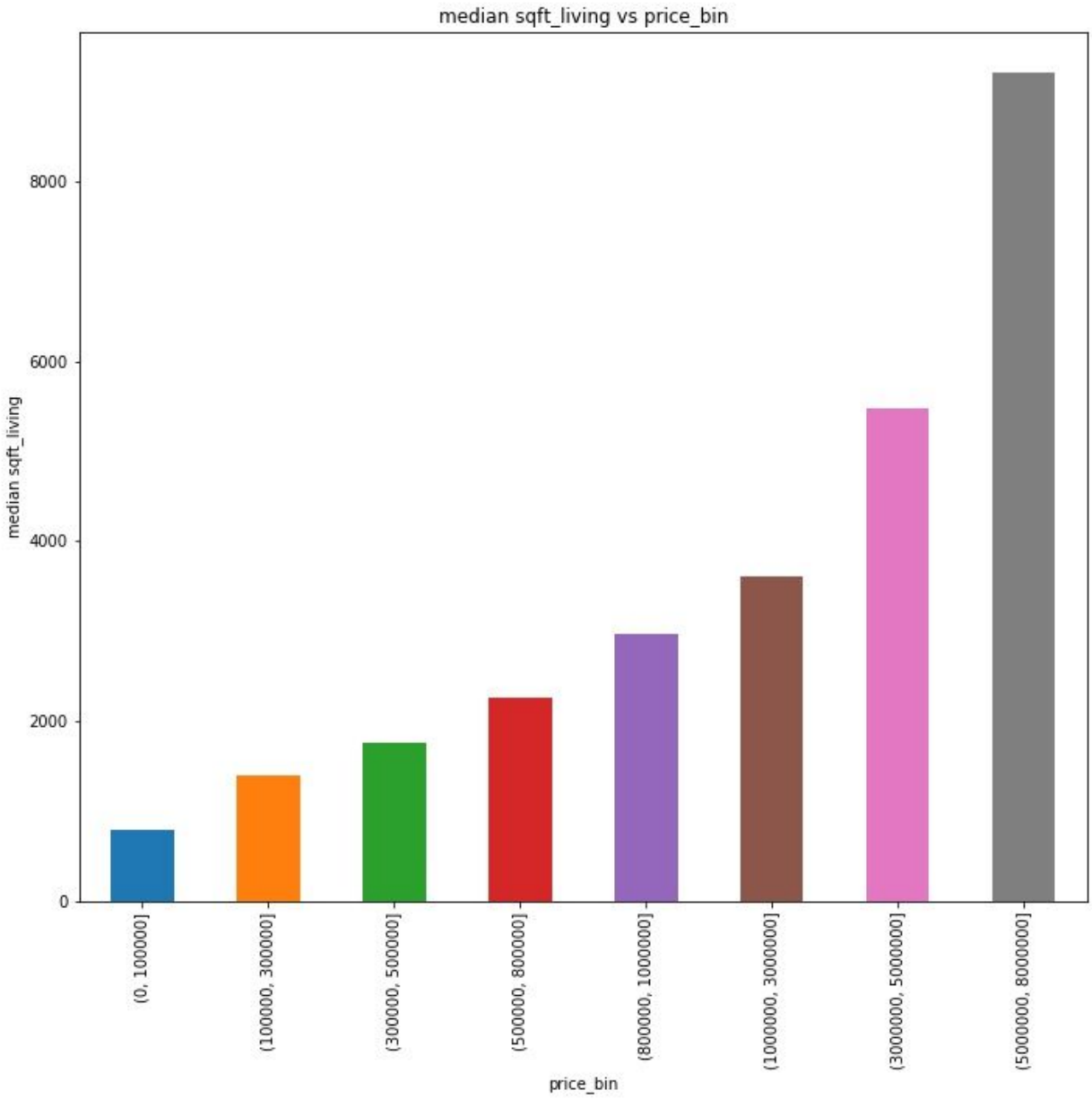


The plot shows median number of bedrooms and price_bin is directly proportional to each other and tells us that more expensive houses have more median number of bedrooms. Houses priced 5,000,000 to 8,000,000 has median 5 bedrooms.

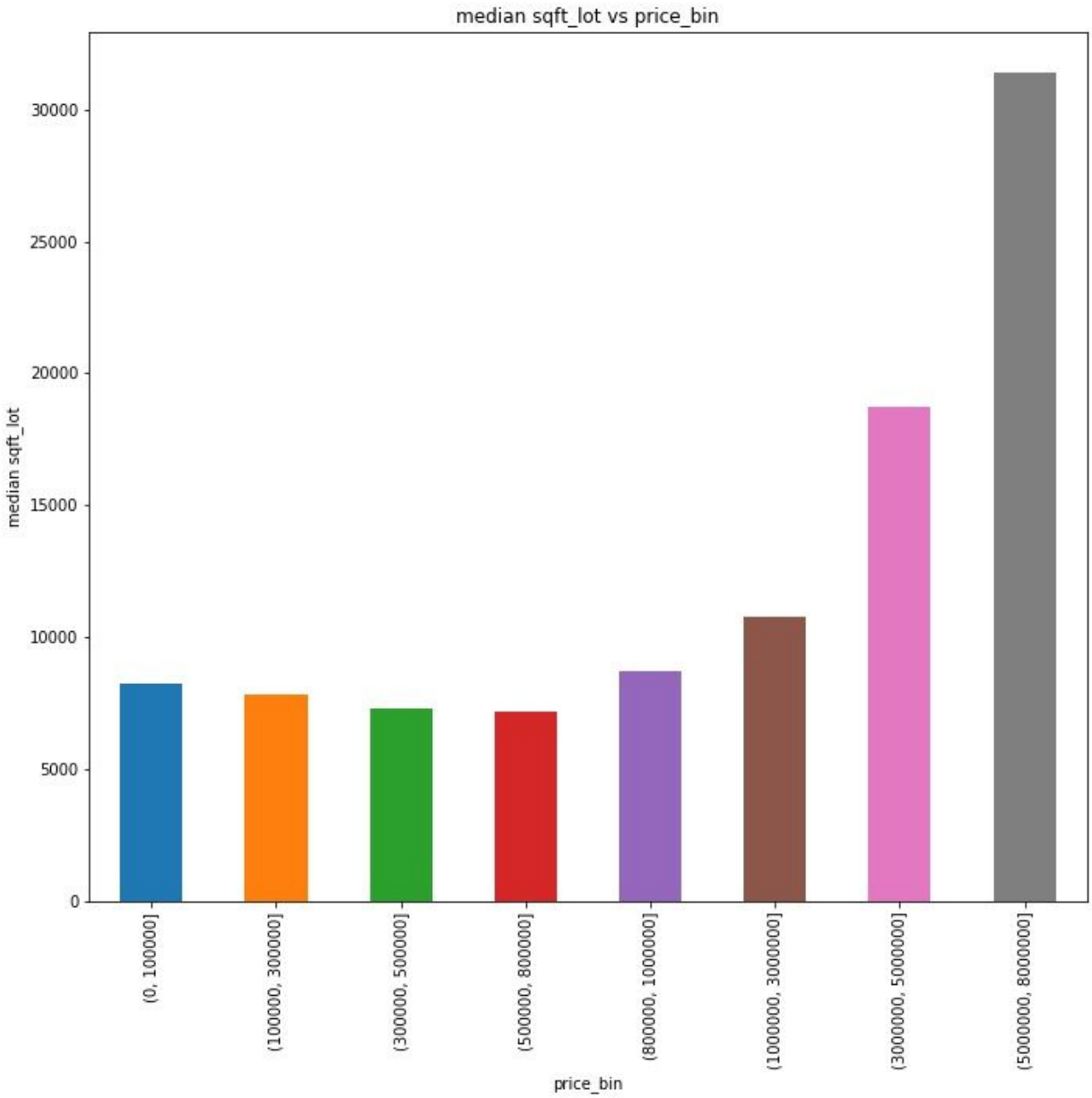
Then, I plot another bar plot of median number of bathrooms vs price_bin.



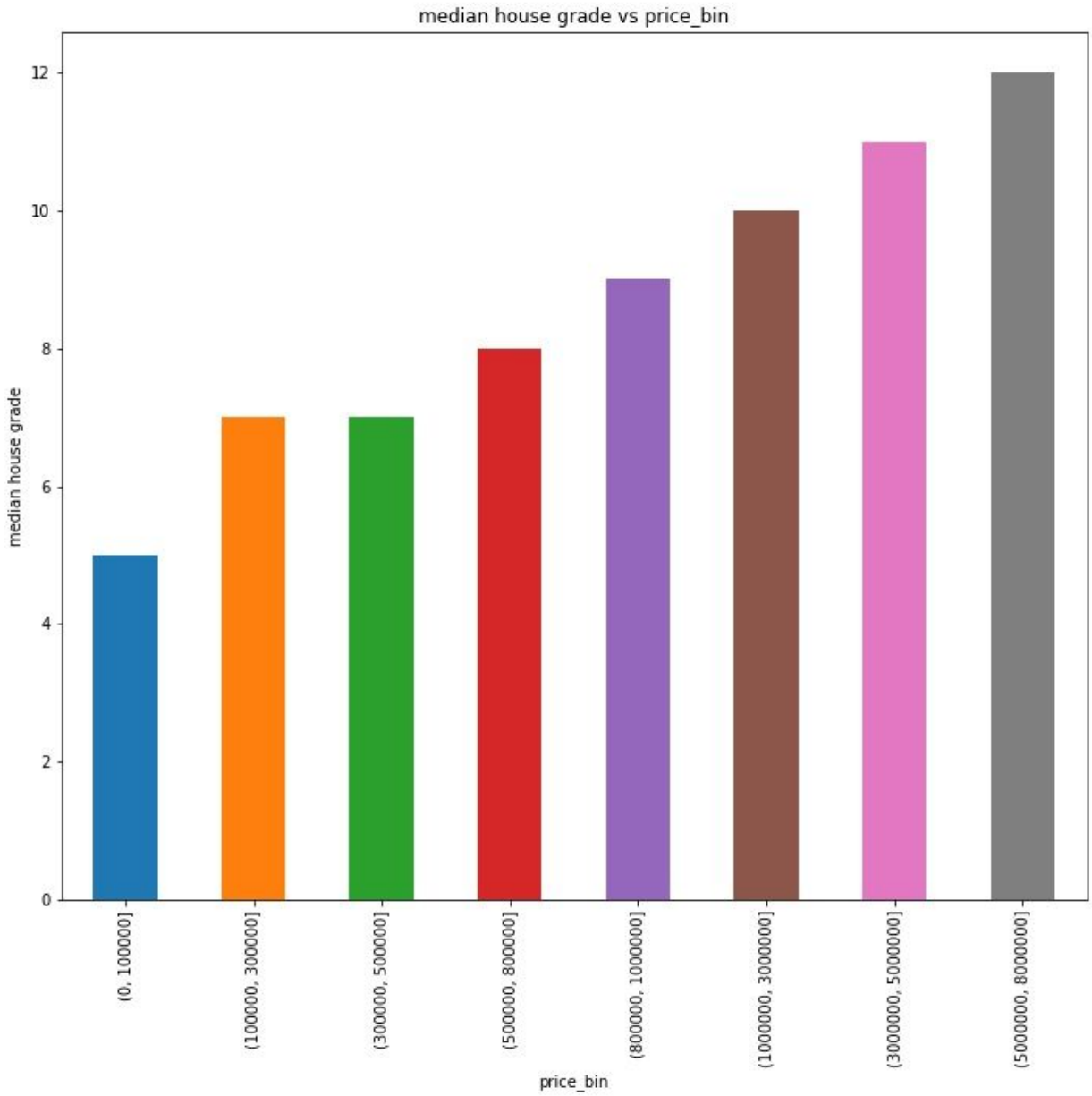
The plot shows median number of bathrooms and price_bins are directly related to each other and tells us that more expensive houses have more median number of bathrooms.



Sqft_living vs price_bin plot tells us that more expensive houses have larger sqft_living and less expensive house has smaller sqft_living.



Sqft_lot vs price_bin plot shows that houses priced 0 to 500,000 has inverse relation with median sqft_lot and houses priced 500,000 to 8,000,000 has direct relationship with median sqft_lot.



Median house grade vs price_bin plot shows direct relationship between median house grade and price_bin.

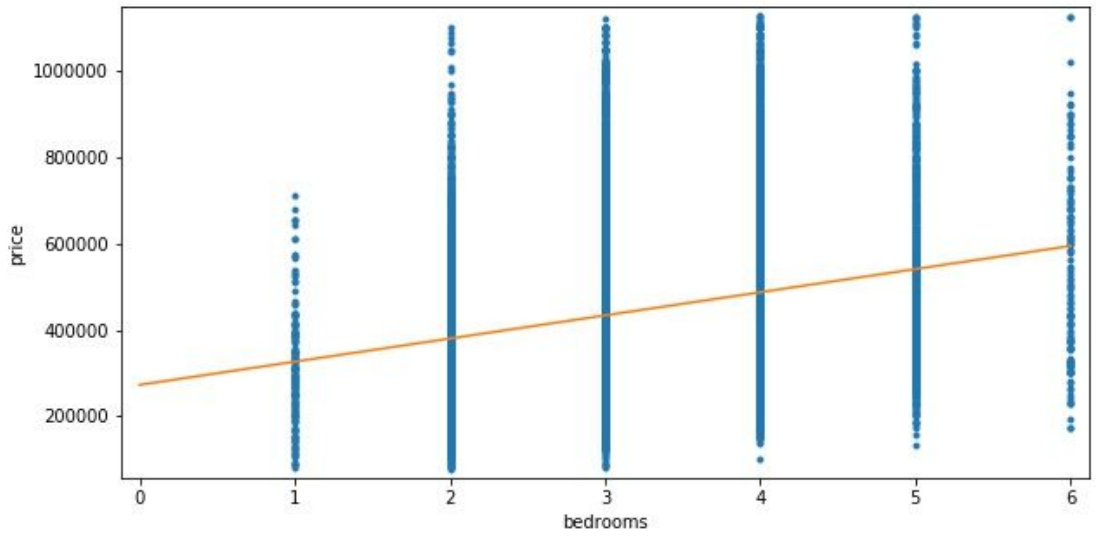
I also drew scatter plot to find if there is any correlation between price and other features.Price vs number of bedrooms shows a positive correlation between price and number of bedrooms. It also shows us some outliers data. Price vs number of bathrooms shows some positive correlation with some outliers. Price vs sqft_living also shows positive correlation. Price vs sqft_lot shows some positive correlation but many outliers. Price vs number of floors, price vs waterfront, price vs view, and price vs condition shows very weak correlation. Price vs grade, price vs sqft_above, price vs sqft_basement, price vs sqft_living15, and price vs sqft_lot15 shows a positive correlation.

Data Wrangling and Cleaning

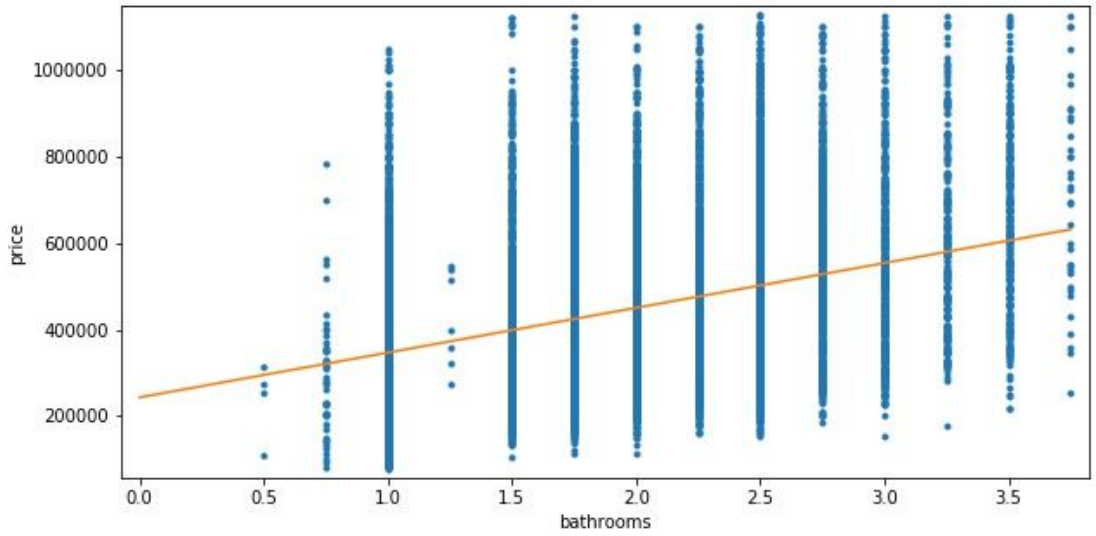
I created a copy of the dataframe df called new_df in order to perform data wrangling. First, I created a function named remove_outliers which takes two input as parameters. One is the string name of the column, and another is the name of the dataframe. Inside the function, interquartile range is calculated. Then, we will locate the outliers using the upper and lower limit values. Then, we will drop the observation which have outliers. I removed outliers from price, bathrooms, sqft_living, sqft_lot, sqft_above, sqft_basement, sqft_living15, and sqft_lot15. For the column bedrooms, there was a bad data 33 which looks like a mistakenly entered data. I checked to see what the features of the house with 33 bedrooms were. Comparing median sqft_living and median sqft_lot size, I found that the mistake data must be 3 instead of 33. I plotted a box plot for the column bedroom to see if there are any other outliers. The plot show that there are some outliers in the data set. I decided to remove outliers with more than 6 bedrooms and less than 1 bedrooms because few houses have more than 6 bedrooms and less than 1 bedrooms. Majority of the houses have more than 1 bedroom and less than 6 bedrooms. After data wrangling, we have 16,496 observations in our dataset.

Data Story

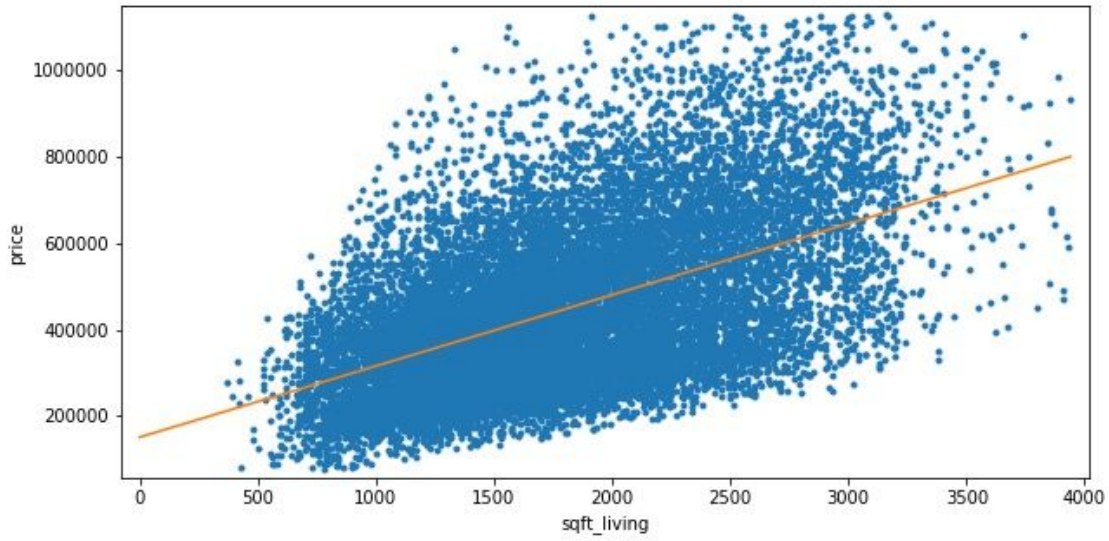
I defined a function which take three parameters as input i.e. x_input, y_input, and data. This function will plot scatter plot with regression line. I drew scatterplot of price vs most of the features from the dataset.



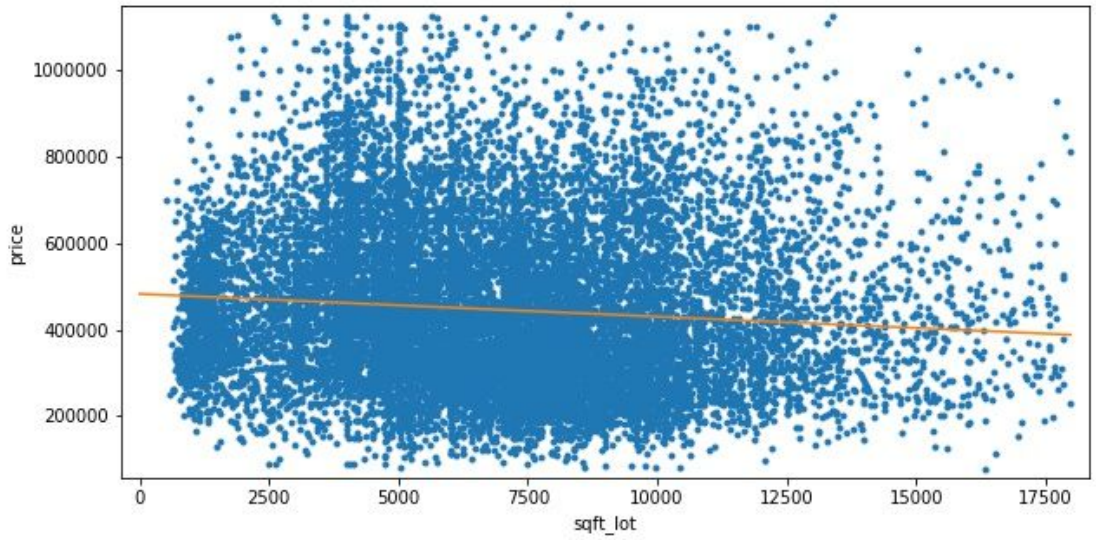
Price vs bedrooms scatter plot show there is a positive correlation.



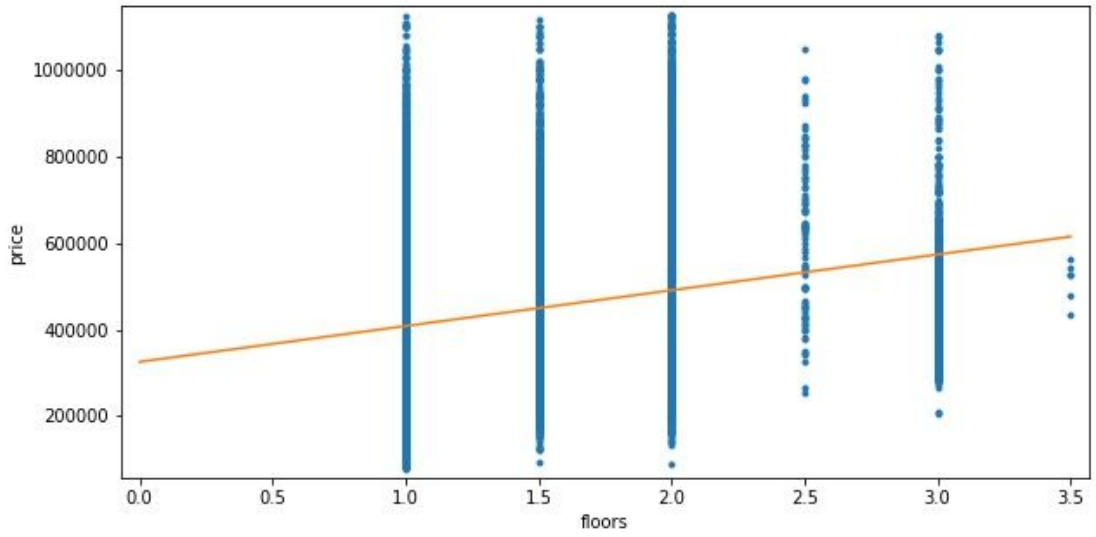
Price vs bathrooms scatter plot also shows there is a positive correlation.



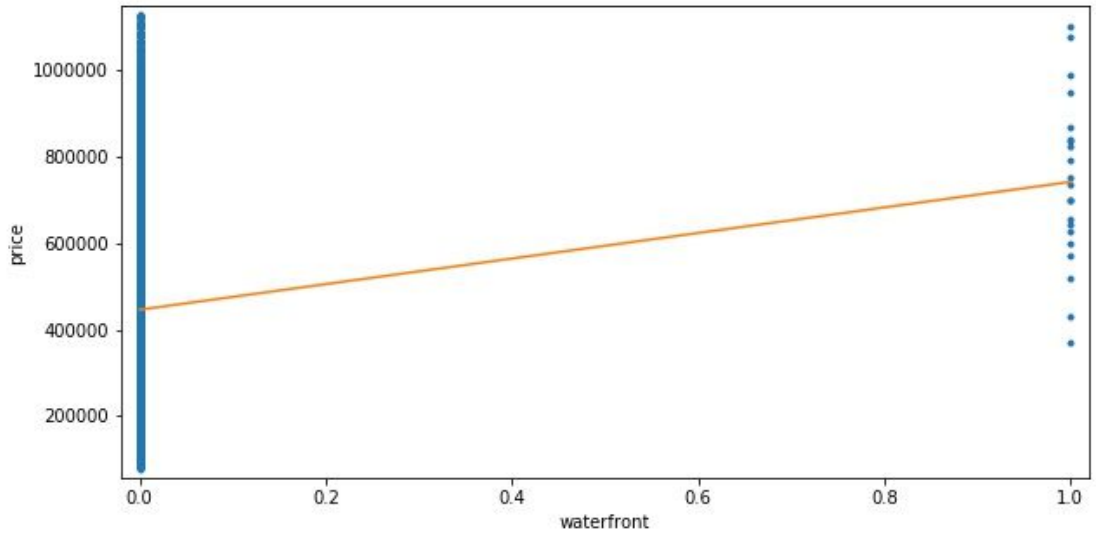
Price vs sqft_living plot also shows there is a positive correlation.



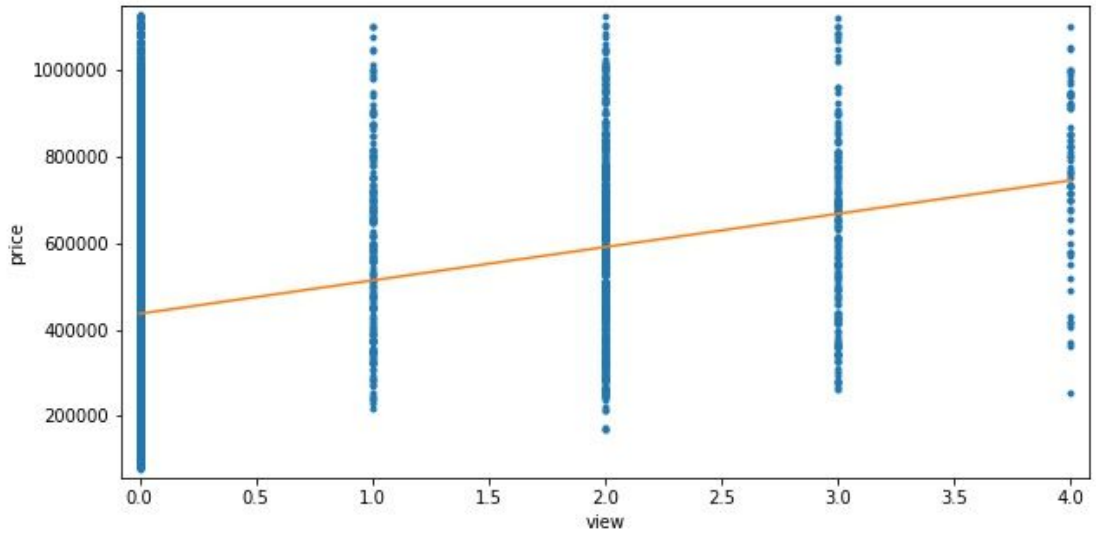
Price vs sqft_lot scatter plot shows weak negative correlation.



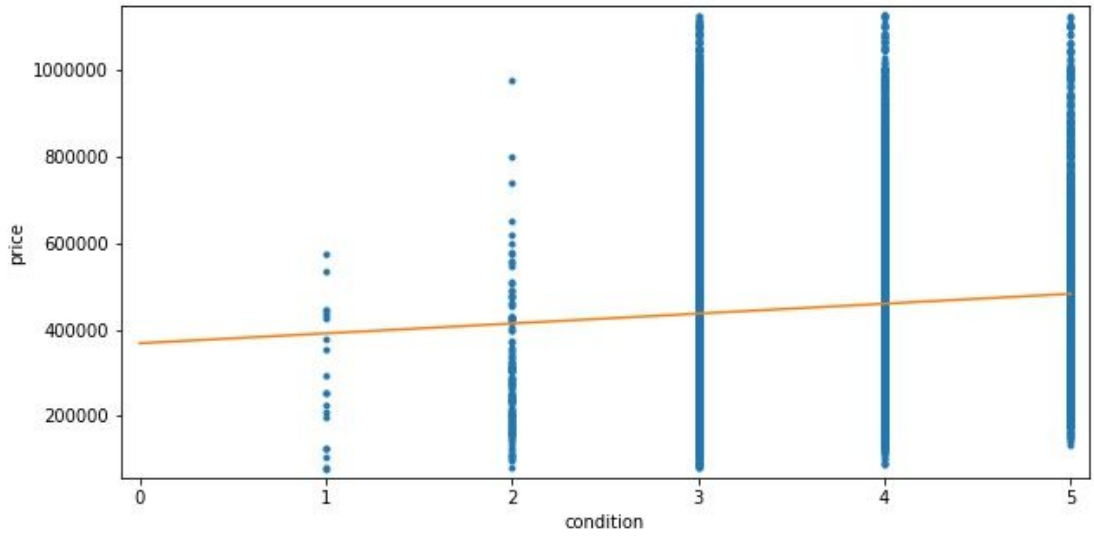
Price vs floors scatter plot shows positive correlation.



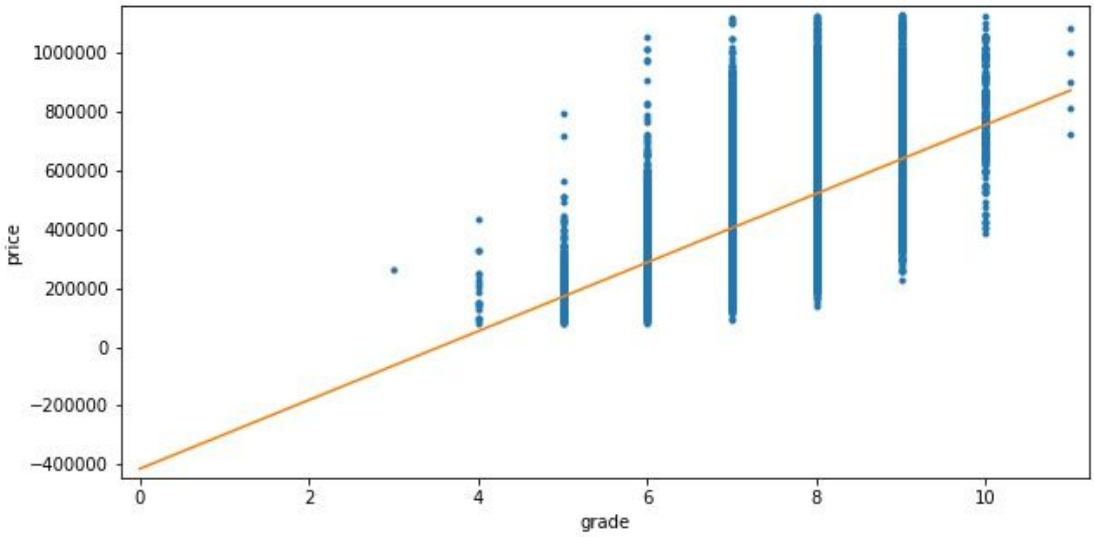
Price vs waterfront scatter plot shows positive correlation.



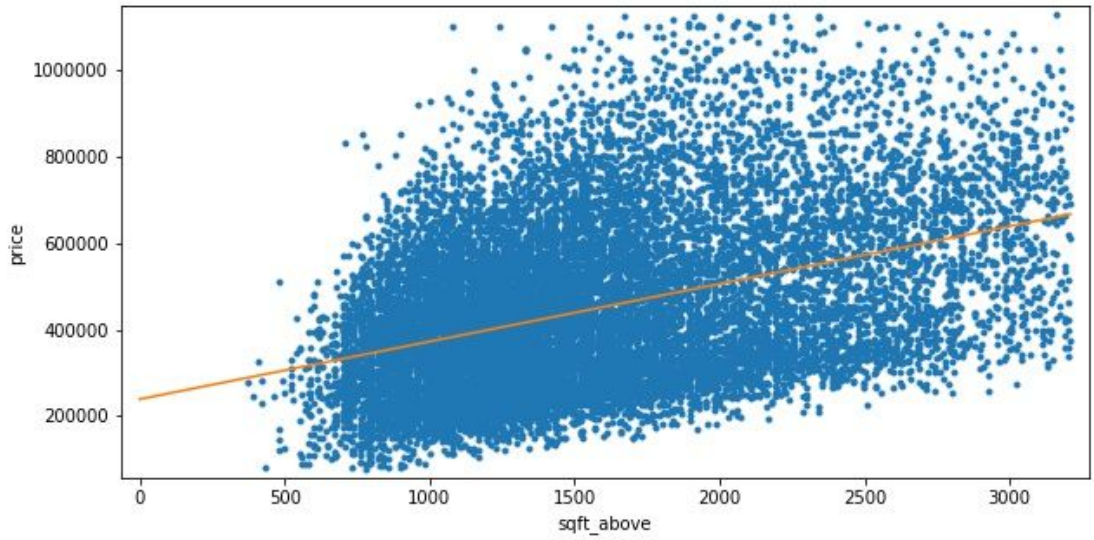
Price vs view scatter plot shows positive correlation.



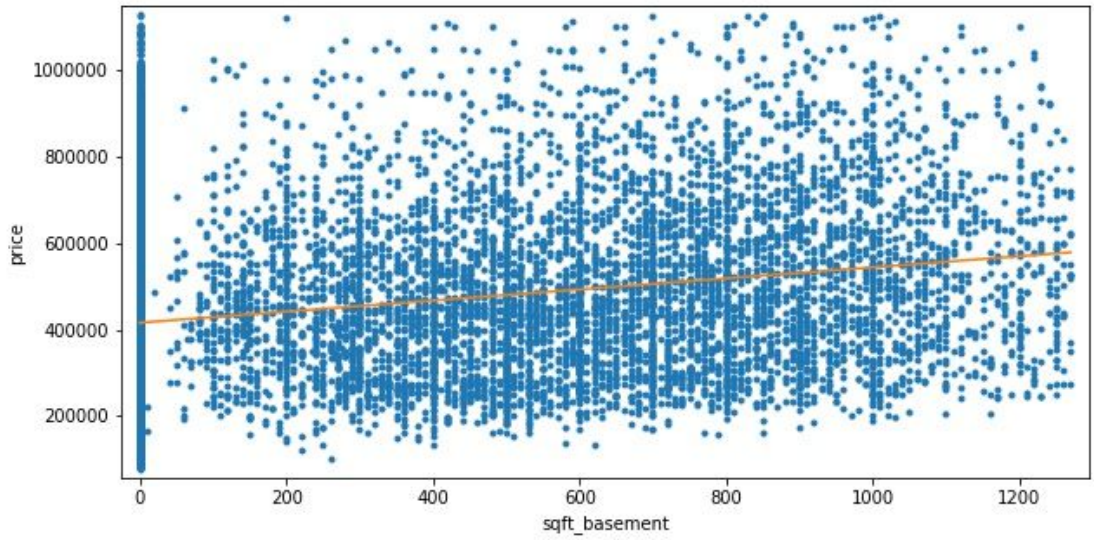
Price vs condition scatter plot shows a weak correlation.



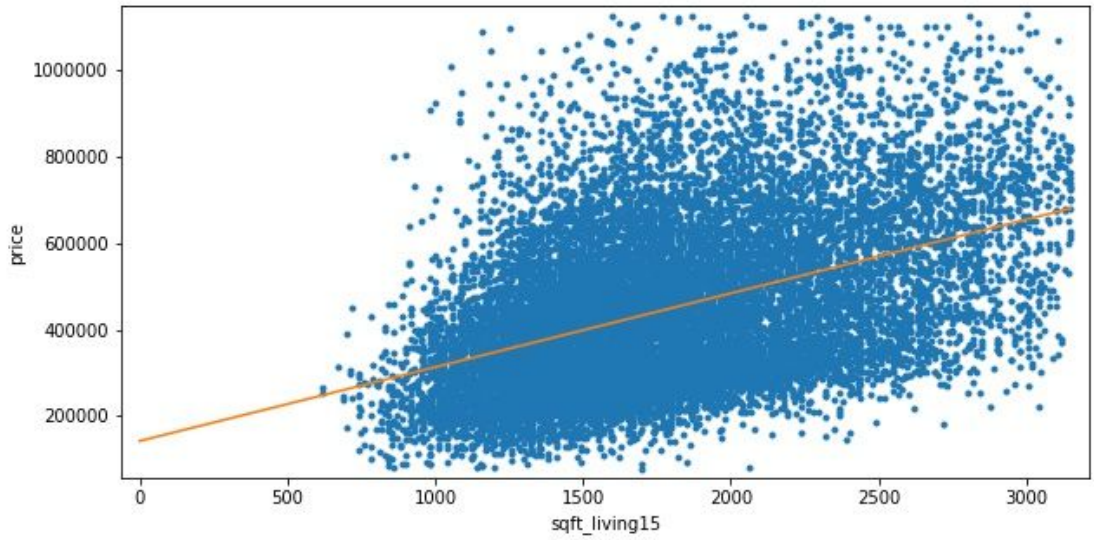
Price vs grade scatter plot shows positive correlation.



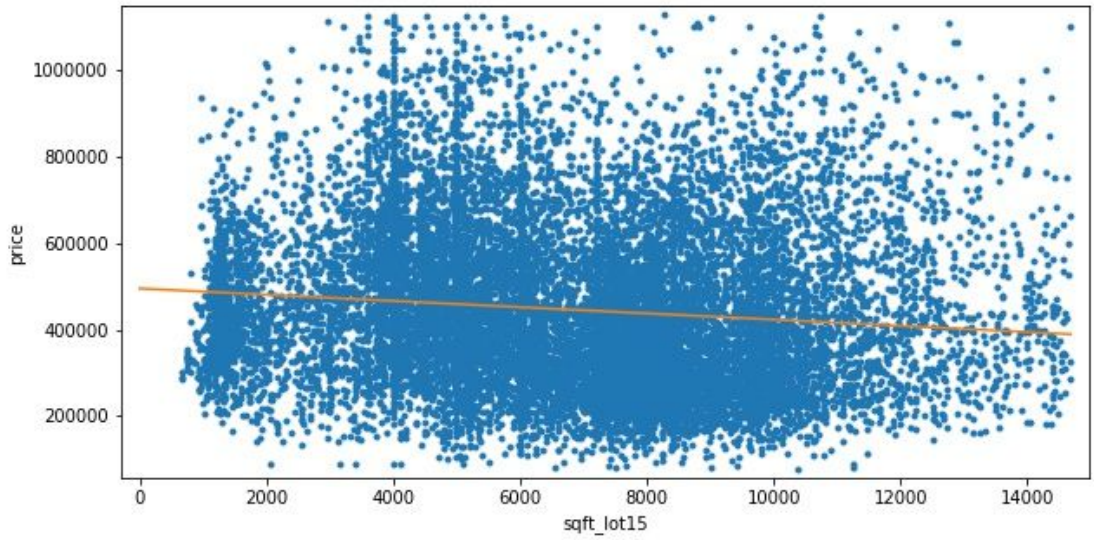
Price vs sqft_above scatter plot shows positive correlation.



Price vs sqft_basement shows weak correlation.

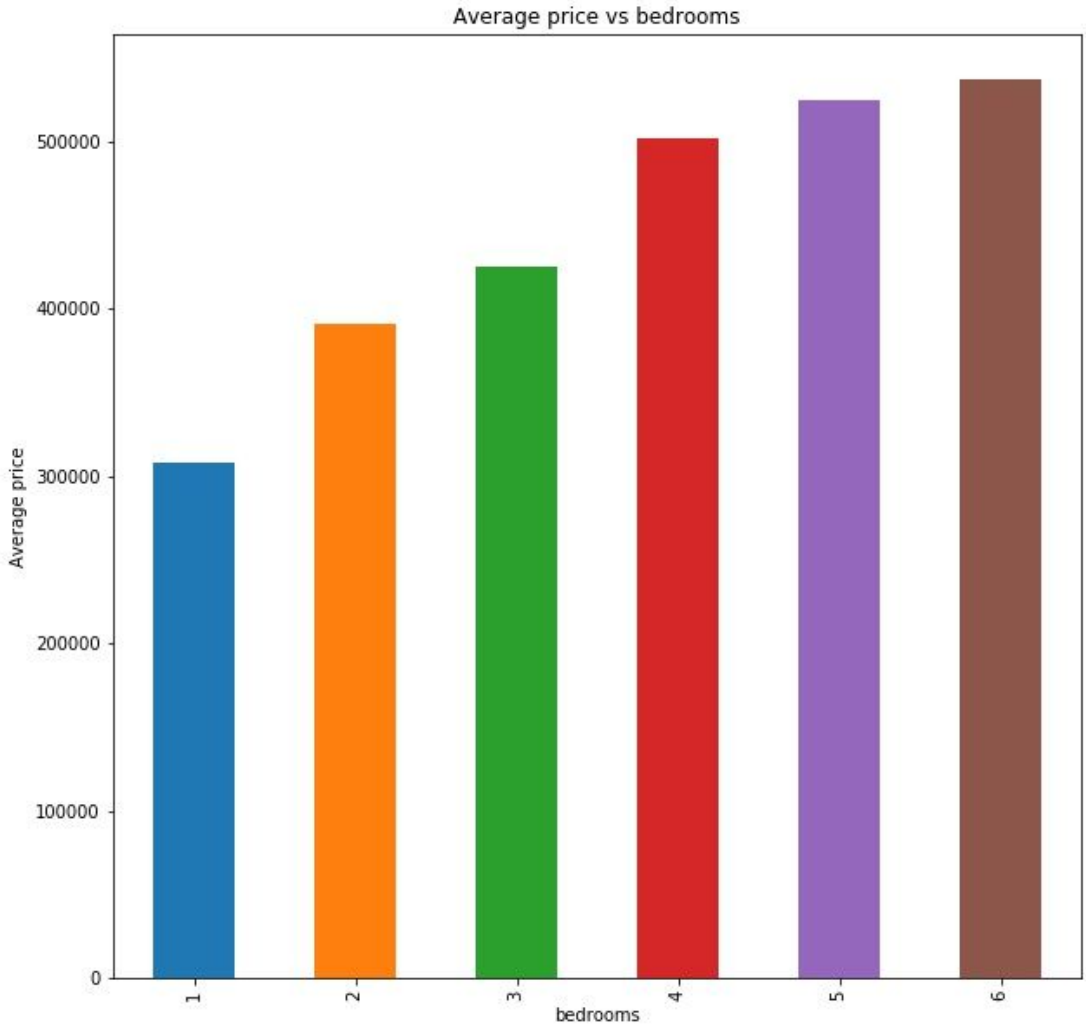


Price vs sqft_living15 scatter plot shows positive correlation.

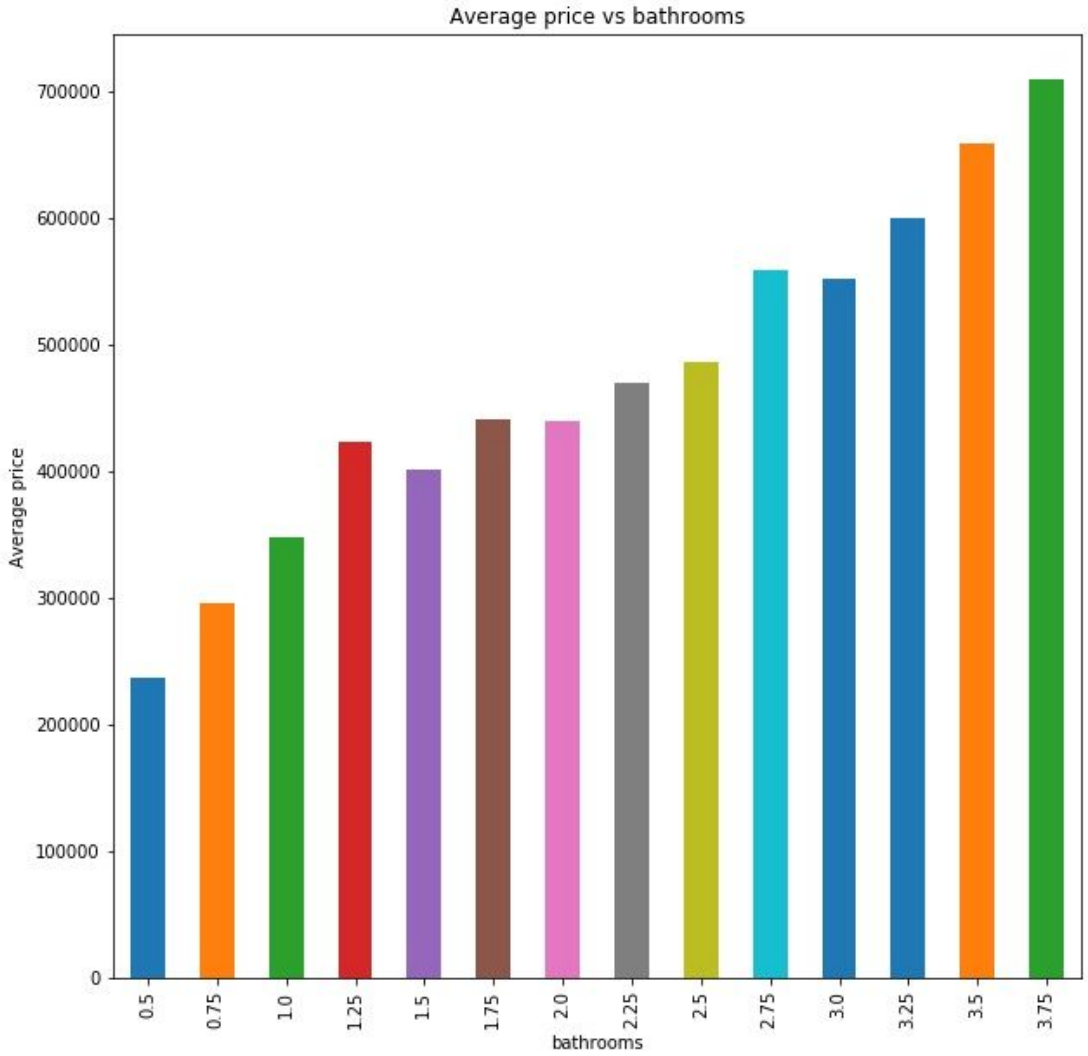


Price vs sqft_lot15 scatter plot shows weak negative correlation.

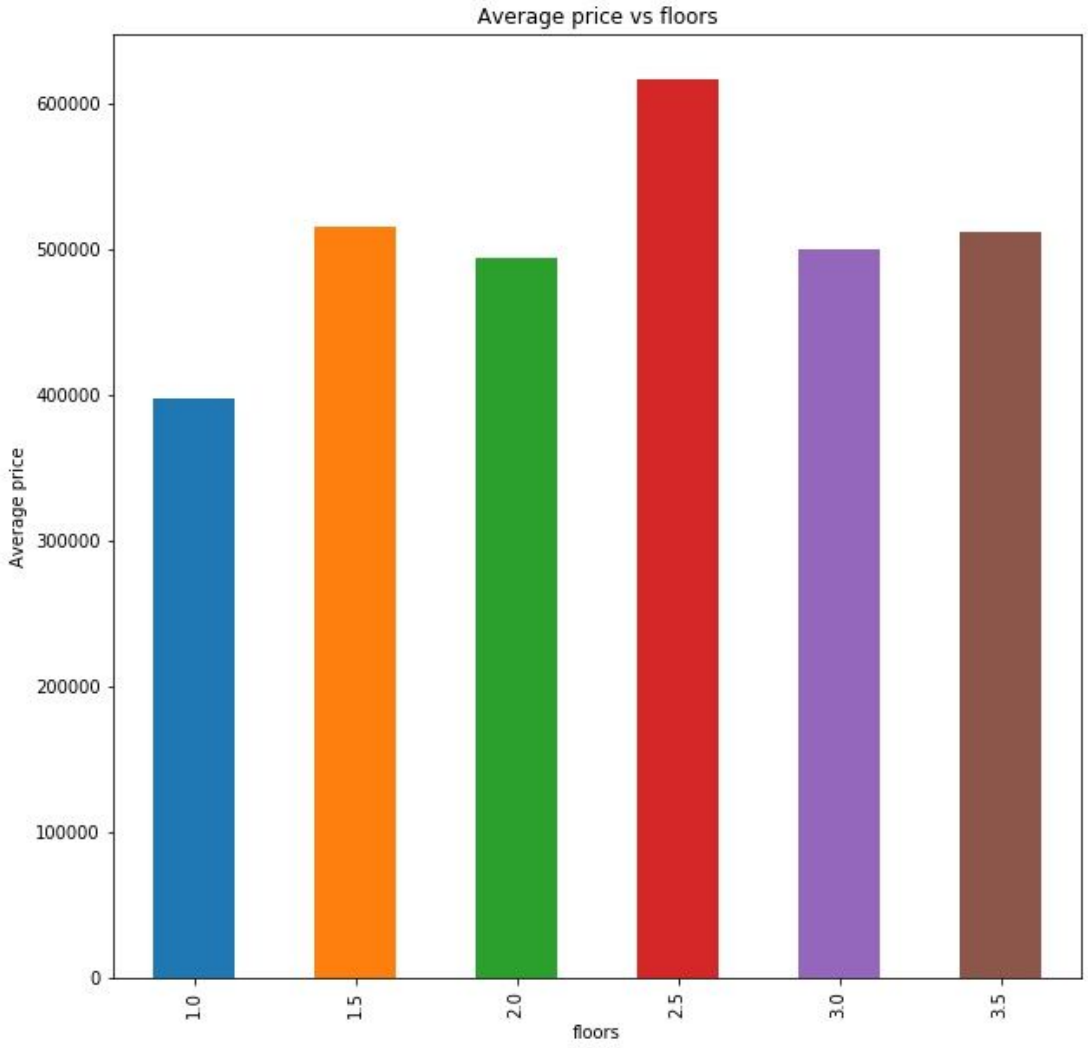
I defined another function named bar_chart which will draw a bar plot using a column name and dataframe as parameter input.



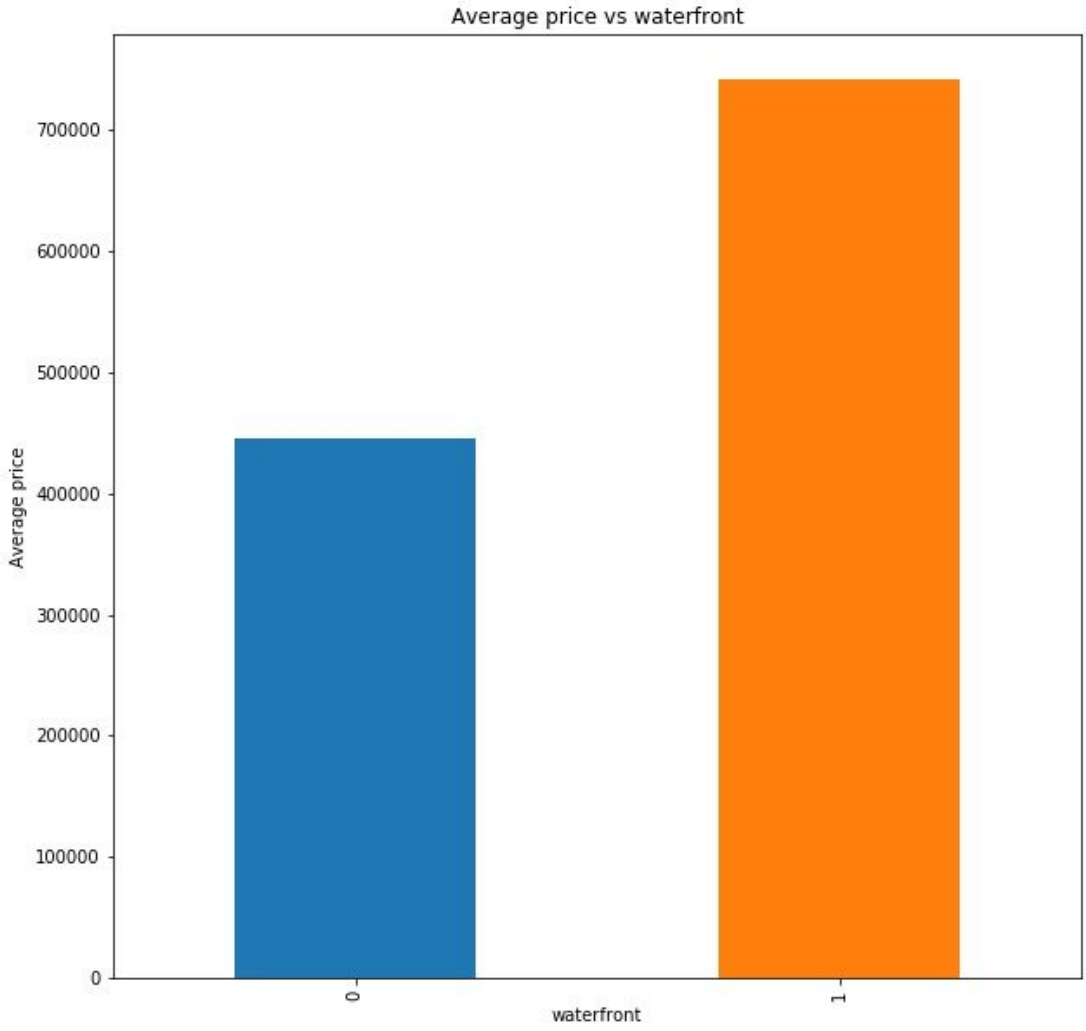
The bar plot average price vs bedrooms shows average price and bedrooms is directly proportional to each other. House with 6 bedrooms has the highest average price.



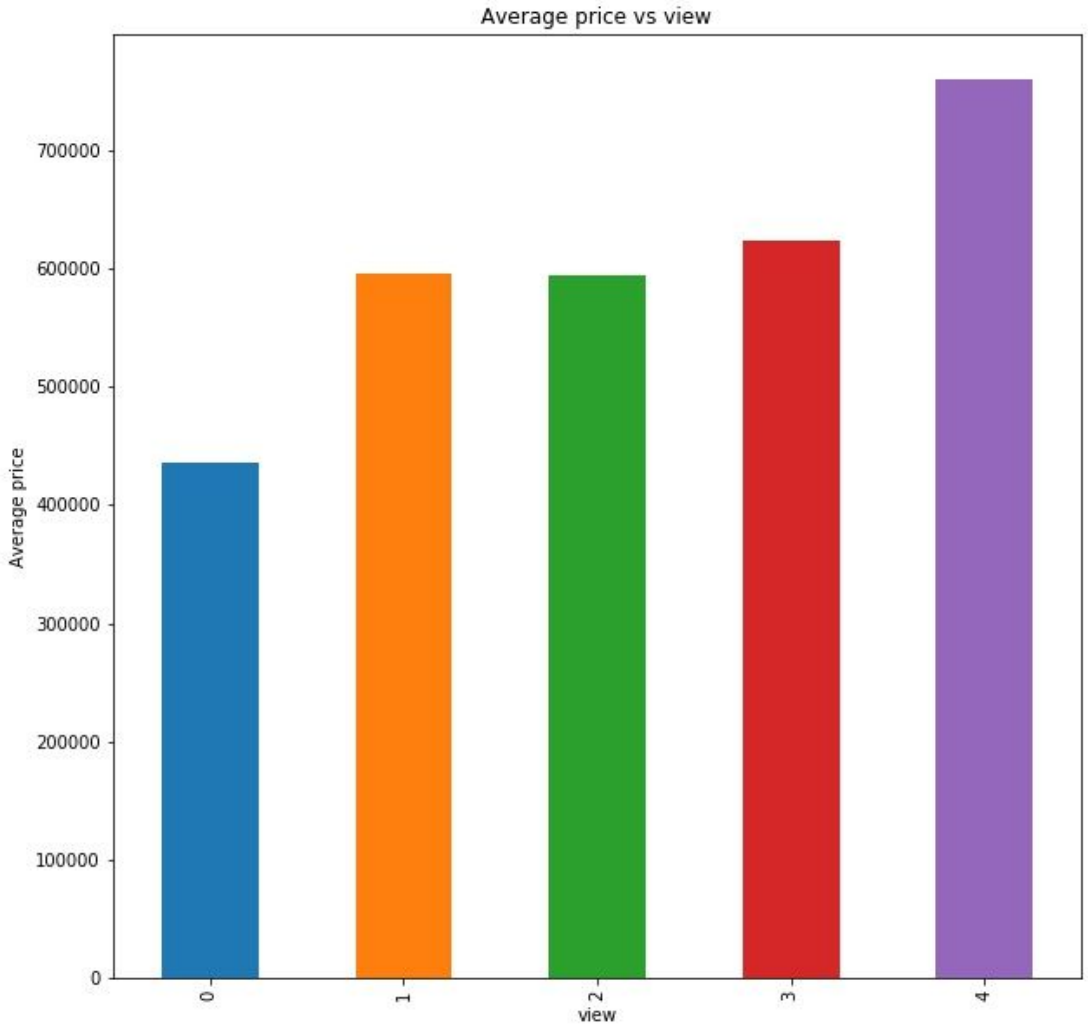
The bar plot average price vs bathrooms shows average price and bathrooms is also directly proportional to each other. House with 3.75 bathrooms has the highest average price.



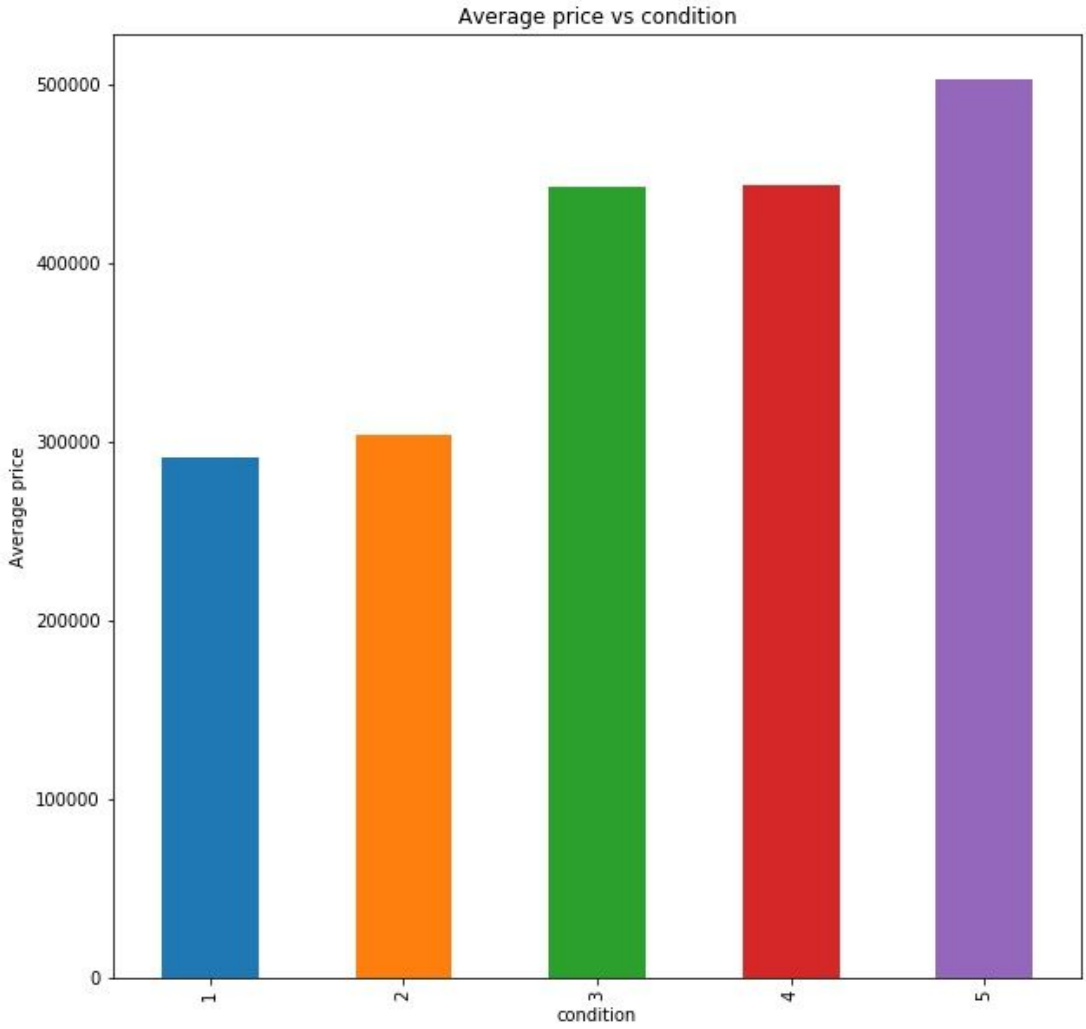
The bar plot average price vs floors shows house with 2.5 floors have the highest average price.



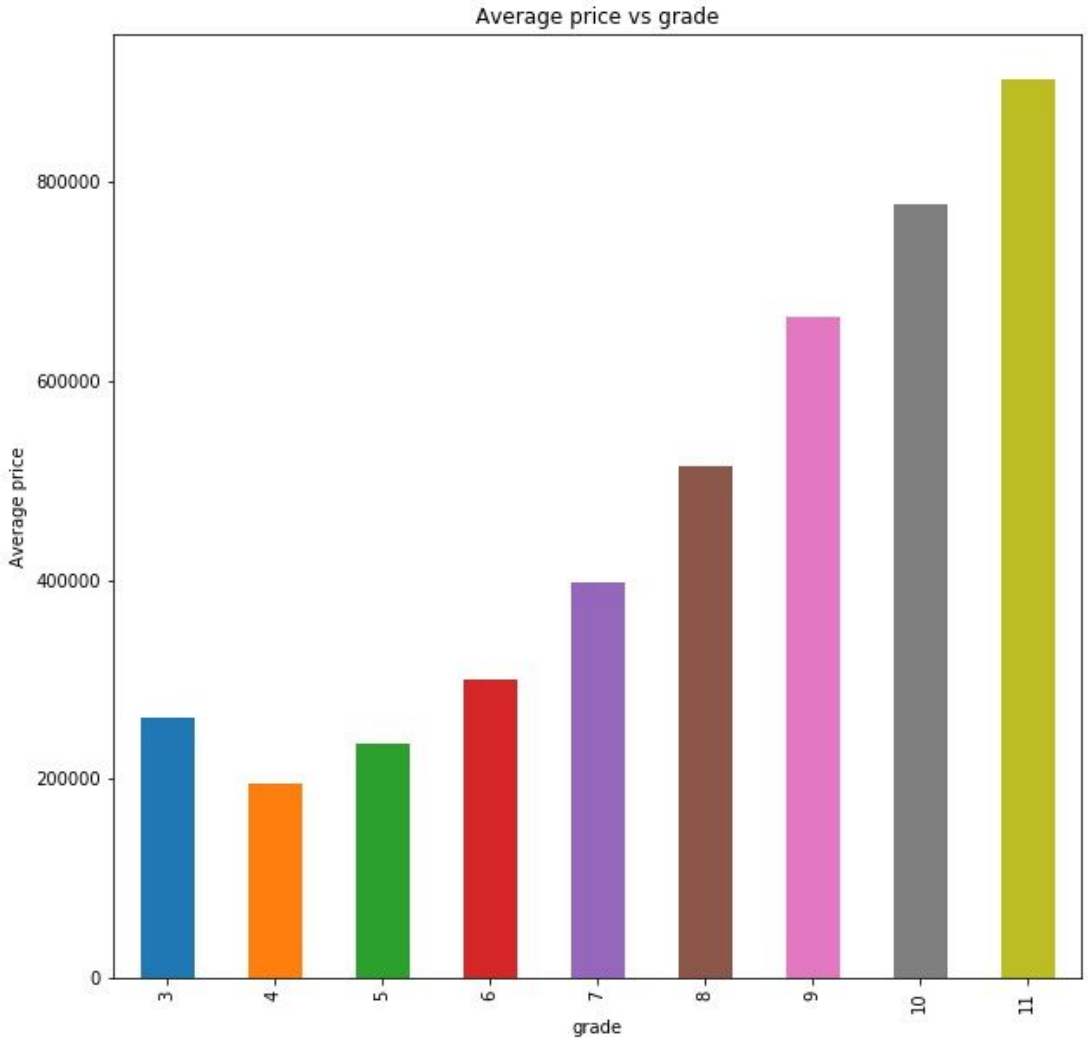
The bar plot average price vs waterfront shows house with waterfront has the highest average price.



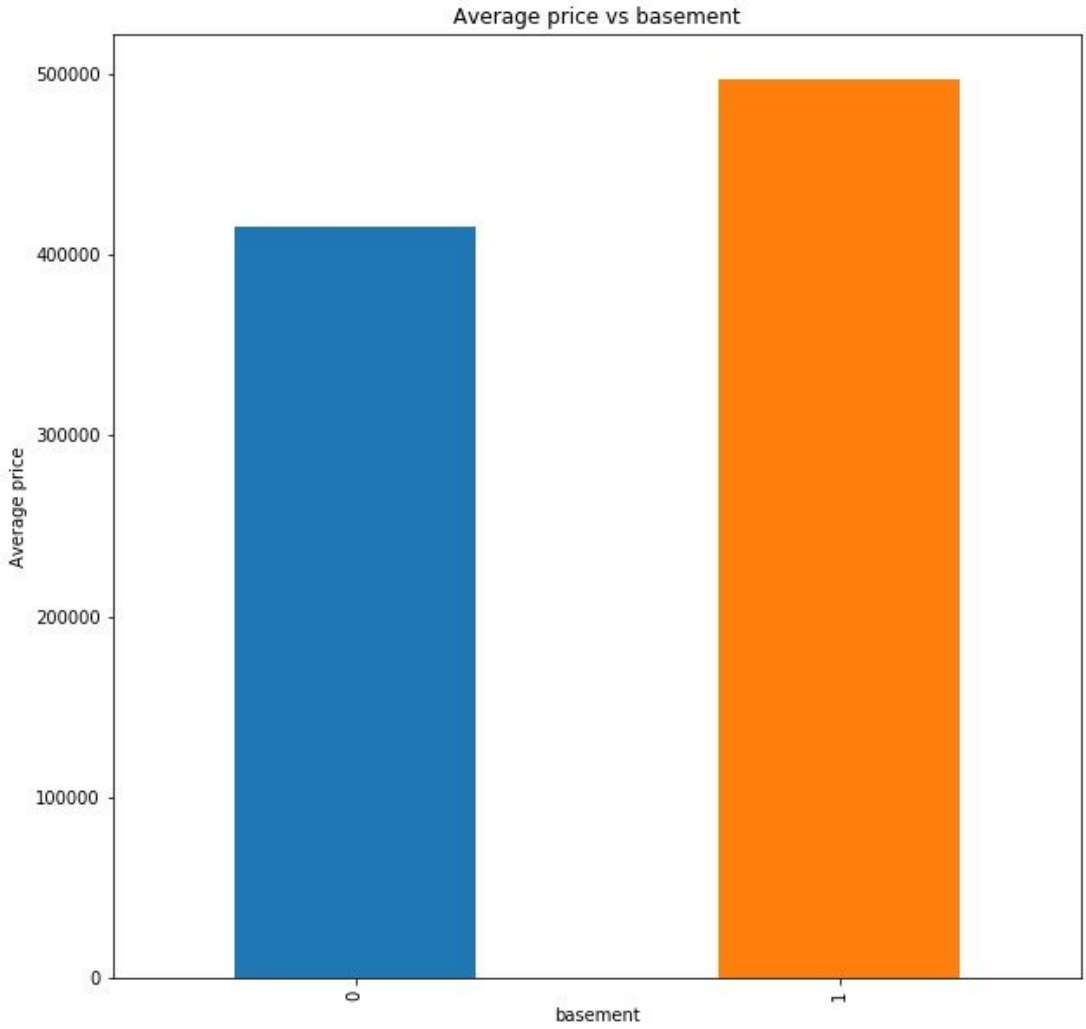
The bar plot average price vs view shows house with 4 view has the highest average price.



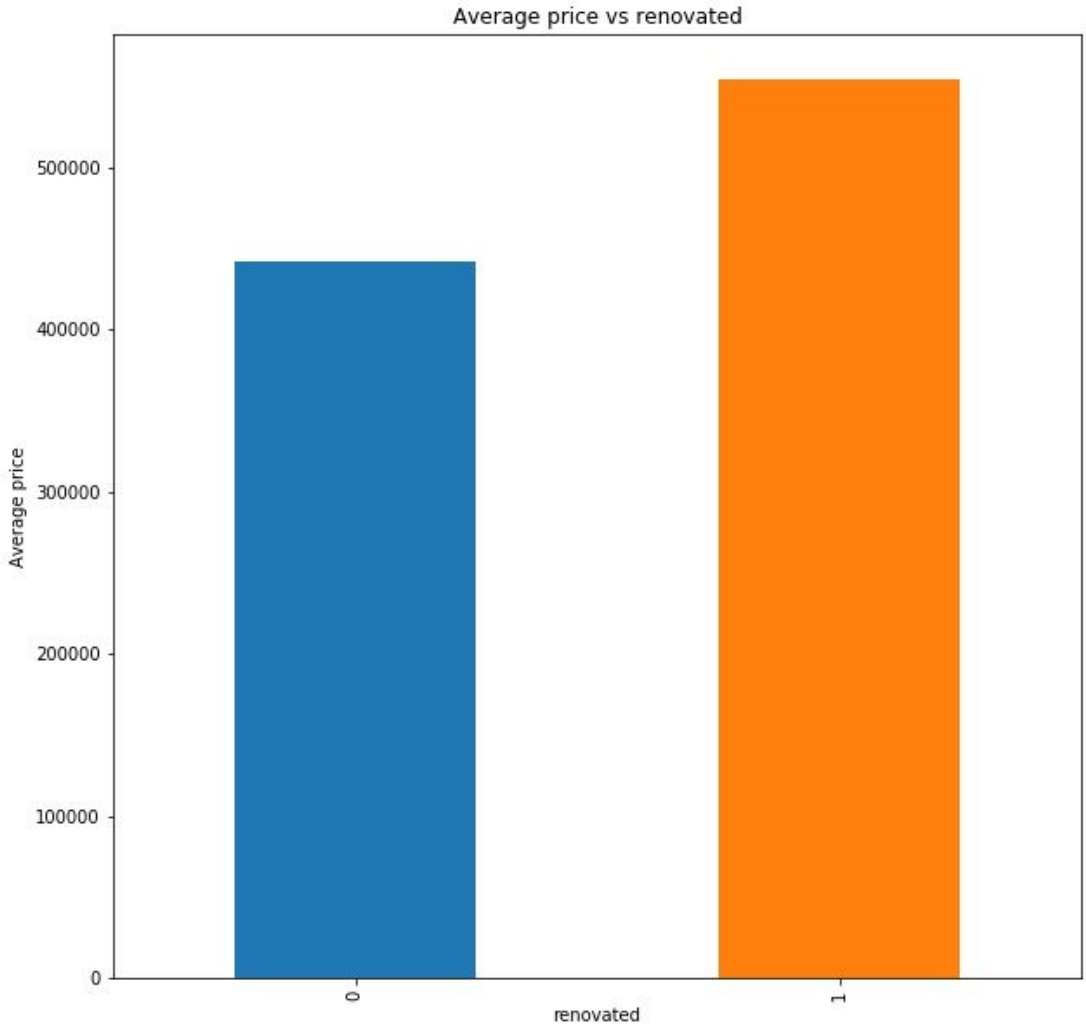
The bar plot average price vs condition shows house with condition 5 has the highest average price.



The bar plot average price vs grade shows house with grade 11 has the highest average price.



The bar plot average price vs basement shows house with basement has the highest average price.



The bar plot average price vs renovated shows house which are renovated has the highest average price.

Inferential Statistics

I conducted a hypothesis test to check if there is no significant correlation between number of bedroom and price. The p-value for the hypothesis test is less than the level of significance 0.05, so we reject the null hypothesis. So we support that there is a correlation between number of bedrooms and price. I also conducted hypothesis test to check correlation between number of bathrooms and price. The p-value for the hypothesis test is less than the level of significance 0.05, so we reject the null hypothesis and suggest that there is a correlation between number of

bathrooms and price. Similarly I conducted a hypothesis test to check correlation between sqft_living and price. The p-value for the hypothesis test is less than the level of significance 0.05, so we reject the null hypothesis and suggest that there is a correlation between sqft_living and price. I also conducted a hypothesis test to check if there is correlation between grade and price. The test suggest that there is a correlation between grade and price. I also conducted a hypothesis test to check if there is no statistical importance between mean house price and number of bedroom less than 3 and greater than 3. The p-value for the test was greater that the level of significance 0.05, so we fail to reject the null hypothesis. This suggests us that there is no statistical importance between mean house price and number of bedrooms less than 3 and greater than 3.