Capstone Project 1: Milestone Report

Predicting House Price

Part 1: Problem Statement

Over the past few decades, the housing price has shown increments as times goes by. It is no doubt that the longer the period of time the higher the price of housing. This is a big issue to be taken into consideration. It is considered an important issue because house is one of the important basic human needs. Therefore, being concerned about the housing price is a must and the factors that contribute to the increase of housing price need to be determined.

The following are the objectives of this proposal:

1. To determine the causes of increase in housing price.

2. To predict the price of housing price in the future.

3. To identify the effect of increase in housing price.

Background

By doing prediction on the housing price, the factors that lead to the increase of housing price can be determined. The housing price in the future can be forecasted as well. The importance of doing this proposal on the housing price is that, it will give positive impact to the government. It can help the government to predict or estimates the price of housing in the future. These can wider the awareness of the price of housing which is now in high rate. This research will also can be an advantage for the housing developers in determining the housing price by identification of the factors that lead to the housing price estimation. Moreover, this prediction is also important to individuals who will purchase house in the future. This will give awareness to the individual regarding the housing price and help them to figure out the best decision in purchasing a house.

Dataset

The dataset for this project is downloaded as a csv file from Kaggle .This dataset contains house sale price for king county, in Seattle WA. It includes homes sold between May 2014 and May 2015. The dataset set consists of 21613 observations and 19 features plus the house price and ID columns. There were some columns that contained a high percentage of missing values. Those with high percentage of cardinality and missing values were later dropped in the wrangling phase. The default type for majority of the features were of object dtype but was later converted to an appropriate format to be computationally efficient. Several columns needed to be dropped due to the high number of missing values. After the dataset is properly cleaned and formatted I would move onto the exploratory analysis and feature generation phase. The dataset took in 778 KB in memory when loading into a Pandas data frame. Variable types included: int and Date.

Part 2: **Data wrangling**

This dataset contains house sale prices for King County, which includes Seattle. It includes homes sold between May 2014 and May 2015.Let's gets started by reading the dataset I'll be working with and deciphering its variables. For this Capstone project, I'll be using a Kaggle dataset house price patterns. Kaggle is a great community of data scientists analyzing data together – it's a great place to find data to practice the skills covered in this project.

The dataset contains a detailed set of information about house pricing features and the main problem statement here is to determine the house price that should continue to sell, and how to predict it. The file contains the observations of house sale prices for King County, which includes Seattle. It includes homes sold between May 2014 and May 2015. The end solution here is to create a model that will predict house price – I'll perform EDA on this data to understand the data better.

Let's analyze the dataset and take a closer look at its content. The aim here is to find details like the number of columns and other metadata which will help us to gauge size and other properties such as the range of values in the columns of the dataset.

All the exploratory data analysis like shape, info, head and describe has been performed to get an overview of the features and to look on the target variables.

Dataset contains: **
Id: a notation for a house
Date: Date house was sold
Price: Price is prediction target
Bedrooms: Number of Bedrooms/House
Bathrooms: Number of bathrooms/House
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 to 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 coordinates
Sqft_Living15: Living room area in 2015(implies-- some renovations) this might or might not have affected the lotsize area
Sqft_Lot15: lotSize area in 2015(implies-- some renovations)

Missing values:

I also used the drop cleaning function to reduce the dataset by dropping columns that won't be used during the analysis.

To check if there are any null values in the dataset:

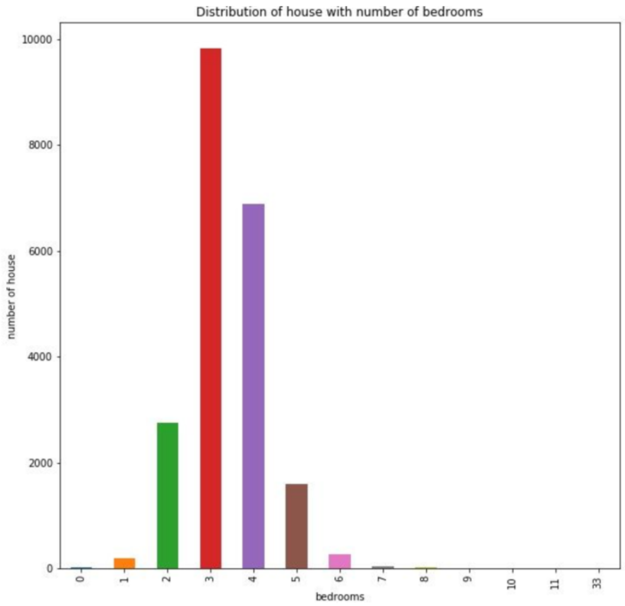
The answer to these questions is important for practical reasons because missing data can imply a reduction of the sample size. This can prevent us from proceeding with the analysis. Moreover, from a substantive perspective, we need to ensure that the missing data process is not biased and hiding an inconvenient truth. The good news is our data set contains no missing values.

Outliers Detection:

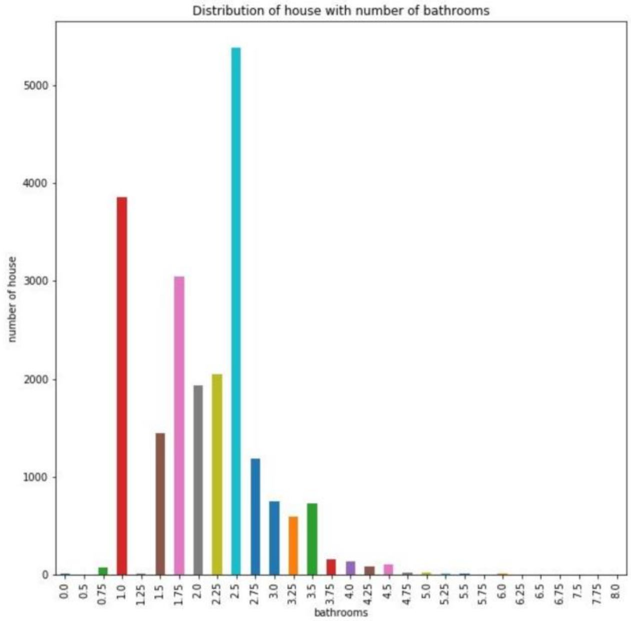
Outliers were detected and analyzed using the outlier box plots. From the outlier box plot I inferred that the data consist of many outliers for the target and price variables. However the outliers for the price variable corresponded to outliers for Numbers of bedrooms, number of bathrooms and square feet living.

Part 3: Exploratory Data Analysis and Inferential Statistics

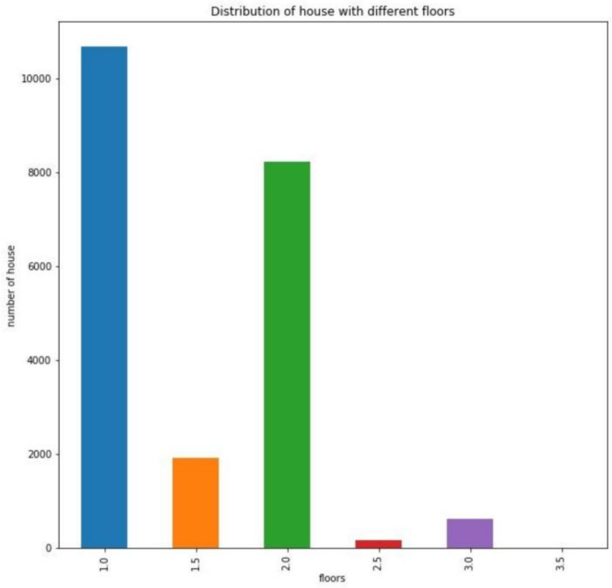
From the dataset the most expensive house is priced at 7,700,000 and the least expensive house is priced 75,000. The average or mean price of a house is 540,088.14 and the median price of a house is 450,000. I plotted the histogram of price and it is right skewed. This indicates that there might be some outliers in the dataset. I also plotted a boxplot of price to check for outliers and found there are some outliers. Then I checked for outliers for other features like the number of bedrooms, the 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 in 2014 and 2015. There were 14633 houses sold in the year 2014 and 6980 houses sold in the year 2015.



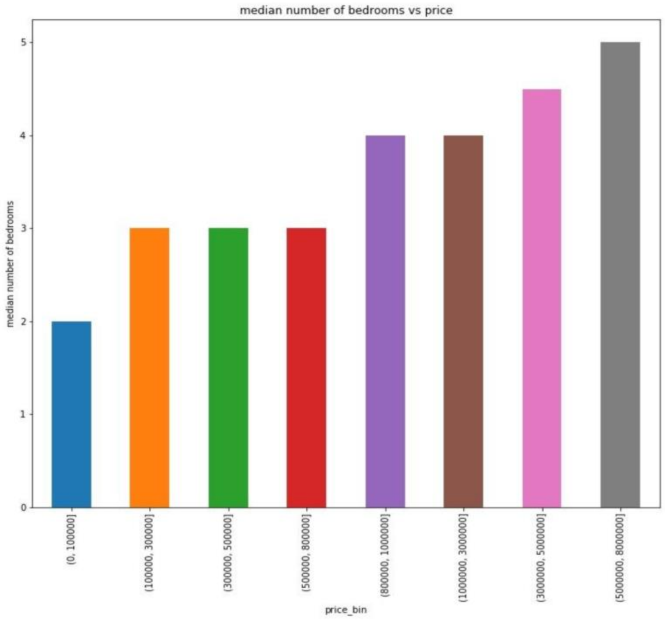
The distribution of the house with the number of bedrooms plot tells us that most of the hou
have 3 bedrooms.



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

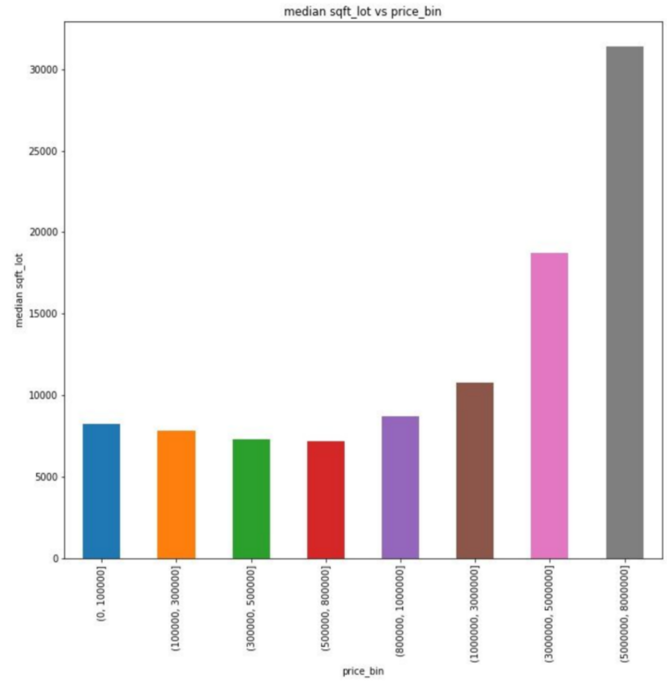


Grouping the house by floors tell us that most of the houses have 1 floor.

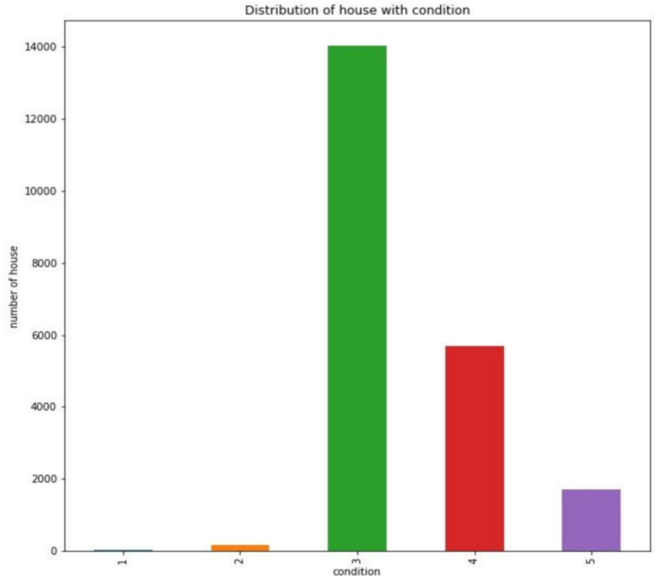


The plot shows the median number of bedrooms and price_bin is directly proportional to each other and tells us that more expensive houses have a maximum median number of bedrooms. Houses priced 5,000,000 to 8,000,000 has median 5 bedrooms.

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



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



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

To gain a sense of the relationship of the features with each other and with house prices, the target variable, I tried to use a diverse set of data visualization tools, including the following: boxplots, histogramplots, scatter plots and correlation plots. The first EDA I performed was to examine the distribution of the home sale prices. The price of a house is dependent on various factors like size of area, how many bedrooms, location and many other factors. I conducted a hypothesis test to check if there is no significant correlation between a number of bedrooms and price. The p-value for the hypothesis test is less than the level of significance 0.05, so I rejected the null hypothesis. So I support that there is a correlation between a number of bedrooms and price. I also conducted a hypothesis test to check the correlation between a number of bathrooms and price. The p-value for the hypothesis test is less than the level of significance 0.05, so I reject the null hypothesis and suggest that there is a correlation between a number of bathrooms and price.

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