Capstone Project 1

Data Wrangling

Data Type

First, check the data type of "date" column in the dataframe, then converted "date" to datetime format using Python Pandas to_datetime function.

```
INPUT:  df['date'].dtype                    OUTPUT:  dtype('O')
INPUT:  df.date = pd.to_datetime(df.date)   OUTPUT:  datetime64[ns]
```

Outliers Detection

I looked into the distribution of house prices from 2014 to 2015 in King County using Empirical Cumulative Distribution Function (ECDF) and distribution plot. To plot ECDF, we need to sort house prices in ascending order.

```
INPUT:  x = np.sort(historical_price.iloc[:,0])   OUTPUT:  (array([75000., 78000., 8000., ..., 7700000.])
```

The ECDF plot tells us that most of the house prices are < 1 million dollars and there are 3 houses with prices over 6 million dollars, but we can only say houses with prices over 6 million dollars could be outliers.

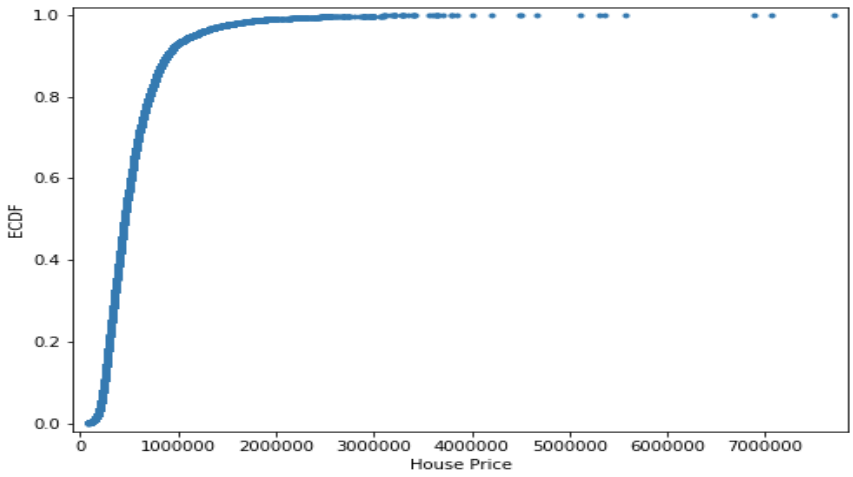


Figure1 ECDF Plot with x-axis "House Price" and y-axis "Probability".

In addition to ECDF plot, I used histogram and Python Seaborn library to check the distribution of house prices to detect outliers. From Figure 2&3, house prices are skewed to the right and deviated from the normal distribution.

Figure 2 Histogram with positively skewed distribution

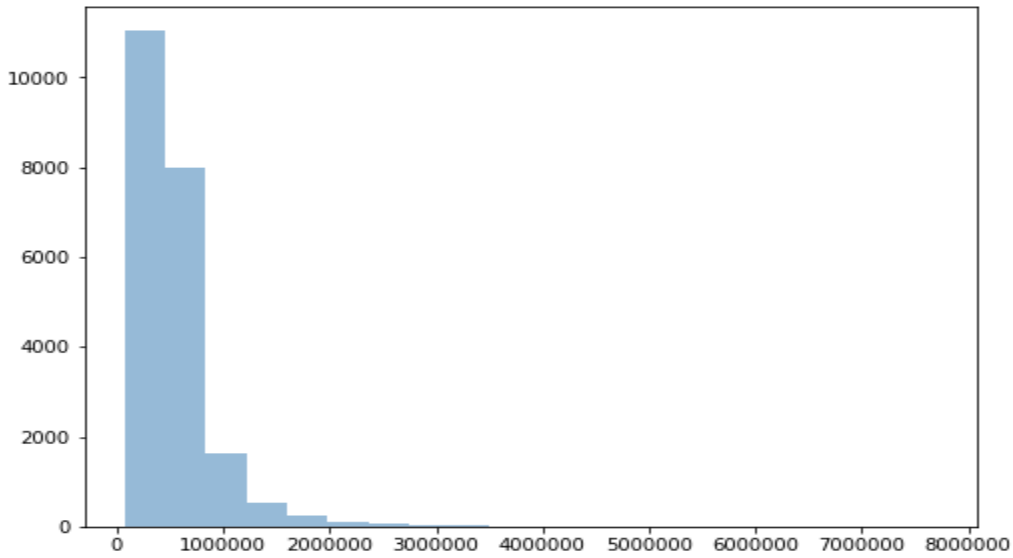
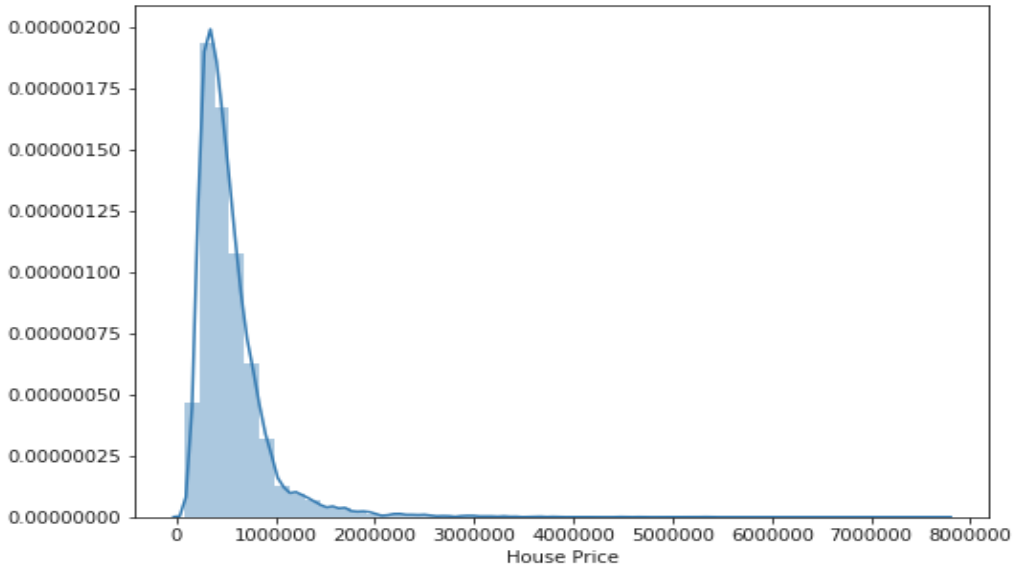


Figure 3 Seaborn Distribution plot



Deal with Outliers

Before further identifying the causes of outliers, I will choose to keep the data first. Since house prices are often influenced by many factors. Furthermore, houses with extreme prices may be an indication of helpful information.