Springboard Capstone Project 1

Nancy Mao | dongxiunan.mao@gmail.com

Beijing House Pricing Prediction

3. Statistical Inference Analysis

3.1 Goals

From the storytelling (EDA) section, we had the following discoveries:

1. Community Average Price seems to have a positive correlation with the dependent variable.
2. Number of bedrooms, living rooms, kitchens, and bathrooms have a positive correlation with the total price as houses with more rooms would be larger with more sq. meters.
3. Among the binary variables elevator, fiveYearsProperty and subway, only subway seems to have a big impact on the house pricing.

In this section, we will perform statistical tests to answer the following questions and validate if the correlations found in section 2 are statistically significant.

1. Independent variables that are significantly correlated with the target variable
2. Correlations between pairs of independent variables
3. If the binary variables (elevator, fiveYearsProperty and subway) have significant impact on the house price and also validate the findings above from the storytelling section.

3.2 Correlations - Independent variables vs. target variable

Below shows the correlation efficient between the independent variables and the target variable pricePerSqMeter:

	Pearson's *r* vs. pricePerSqMeter
pricePerSqMeter	1.000
communityAverage	0.685
DOM	0.299
followers	0.258
kitchen	0.025
buildingFloors	0.023
Cid	0.000
Lat	-0.051
bedRoom	-0.073
bathRoom	-0.079
ladderRatio	-0.085
livingRoom	-0.124
Lng	-0.154
squareMeters	-0.166
constructionTime	-0.215

By defining "strong" correlation between two variables as the absolute value of their correlation coefficient is higher than or equal to 0.5, we found only communityAverage has a strong correlation with the target variable.

3.3 Correlations between pairs of independent variables

	Pearson's *r*	**p-value**
squareMeters_vs_bathRoom	0.736	0.000
squareMeters_vs_bedRoom	0.724	0.000
squareMeters_vs_livingRoom	0.618	0.000
bedRoom_vs_bathRoom	0.548	0.000
livingRoom_vs_bathRoom	0.522	0.000

The above table shows Pearson correlation coefficients for strong correlated independent variables. Here again, I assume "strong" correlation between two variables as the absolute value of their correlation coefficient is higher than or equal to 0.5.

The number of bedrooms, living rooms, and bathrooms shows a high positive correlation with the square meters of the house. But the number of kitchens does not have the strong positive correlation with the square meters. Also, DOM and number of followers of the house on the website also shows a high positive correlation.

The highly correlated independent variables should be treated carefully when building regression models due to multicollinearity.

3.4 Binary Variable (elevator, fiveYearsProperty and subway) and pricePerSqMeter

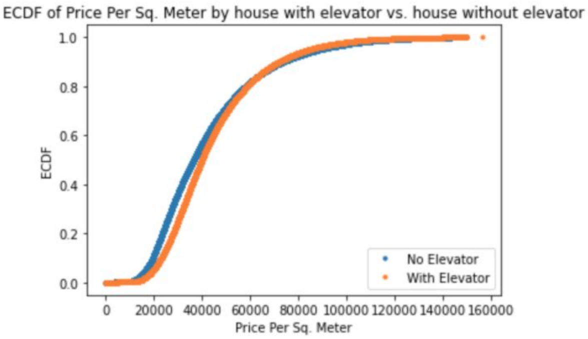
From the previous exploration with the data set, we found that among the three binary variables, subway seems to have a more notable impact on the target variable pricePerSqMeter than than elevator and fiveYearsProperty.

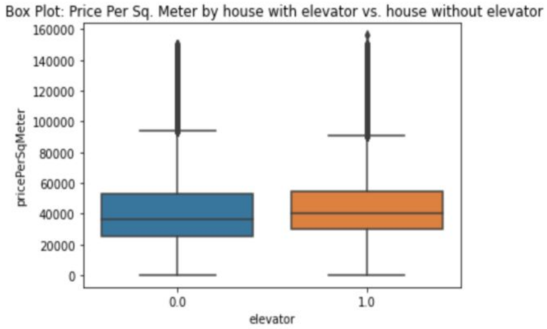
We will use frequentist and simulation approaches to understand statistically how different the means of the target variable are and whether these differences could be explained through chance.

For each of the binary variables, I used a permutation test and a bootstrap test to measure 1) if the distributions of the house price by the two groups are significantly different and 2) if the difference of means of house price are significantly different. Then I used the functions from scipy.stats to do a Mann-Whitney test and a Welch's t-test to validate the results from the permutation test and the bootstrap test.

3.4.1 Elevator vs. pricePerSqMeter

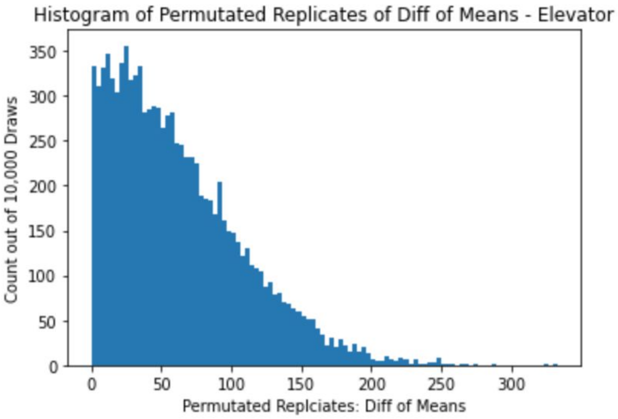
- ECDF and Box Plot
  - Firstly I created the ECDF charts and the box plot to visualize the distribution of the pricePerSqMeter by houses with elevator vs houses without elevator.





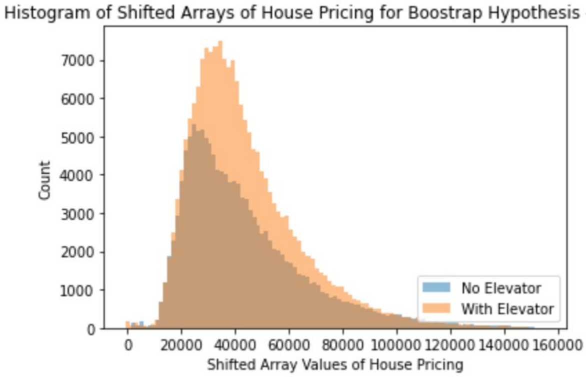
  - ○ From the ECDF charts, we can see there are a slightly larger proportion of the houses without elevators where the price is below around 60,000.

- Permutation Test - Simulating the null hypothesis that prices of houses with elevators and without elevators have identical distributions. The goal is to understand how likely we would have calculated a difference of means as extreme as the current value.
  - ○ Results (10,000 permutation replicates)
    - ■ `Empirical Diff of Mean: 2220.323518352743`
    - ■ `Proportion of permutation replicates with value at least as extreme as the empirical diff of means`
      `p-value = 0.0000`

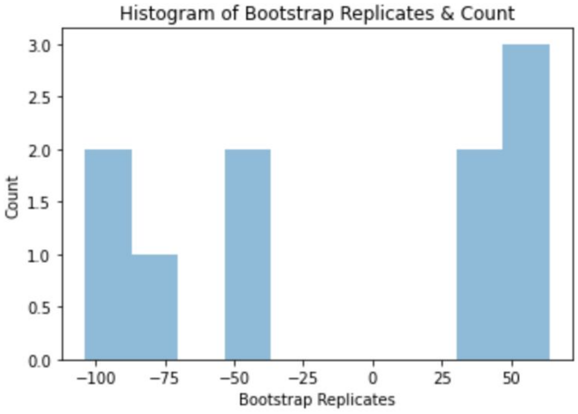


- Bootstrap Test - Simulating the null hypothesis that prices of houses with elevators and without elevators have identical means where the distributions are not necessarily identical. The goal is to understand how likely we would have calculated a difference of means extreme as current value by shifting both groups to have the same mean.
  - ○ Results (10,000 bootstrap replicates)

        - Mean of Concatenated Data: 43567.725202547575



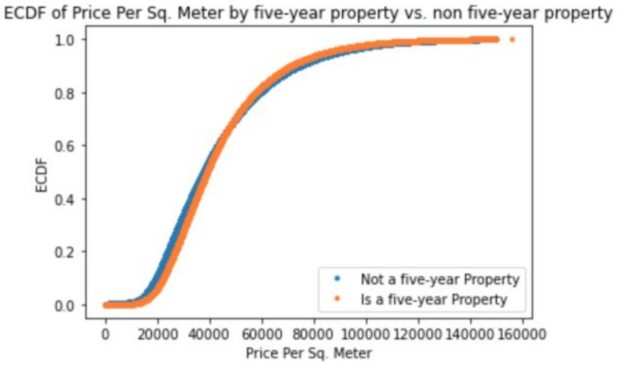
        - Proportion of bootstrap replicates with value at least as extreme as the empirical diff of means
          p-value = 0.0000

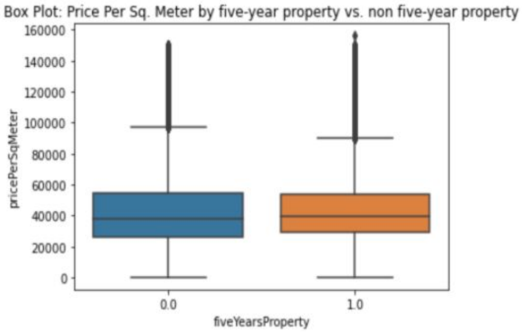


- Mann-Whitney and Welch's T-Test
    - Results:
        - MannwhitneyuResult(statistic=10880479016.5, pvalue=0.0)
        - Ttest_indResult(statistic=-28.013667945055506, pvalue=1.971996620377344e-172)
- Conclusion: All of the four tests rejected the null hypothesis with p-values almost being zero which indicates that elevator has a statistically significant impact on the target variable

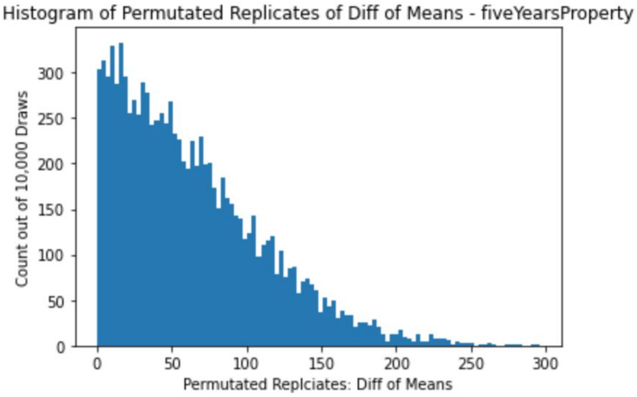
3.4.2 fiveYearsProperty vs. pricePerSqMeter

- ECDF and Box Plot

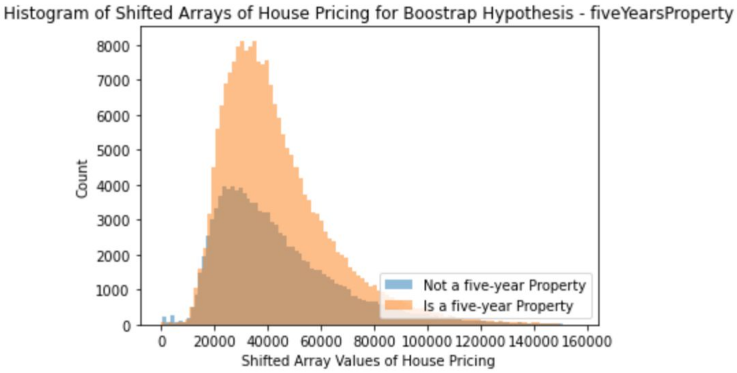




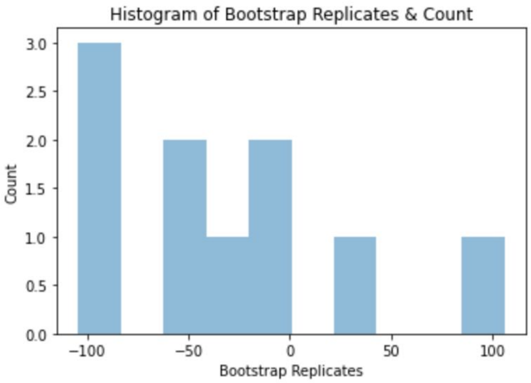
- Permutation Test
  - Results (10,000 permutation replicates)
    - Empirical Diff of Mean: 638.8027132151547
    - Proportion of permutation replicates with value at least as extreme as the empirical diff of means
      p-value = 0.0000



- Bootstrap Test
  - Results (10,000 bootstrap replicates)
    - Mean of Concatenated Data: 43567.725202547575



    - Proportion of bootstrap replicates with value at least as extreme as the empirical diff of means
      p-value = 0.0000



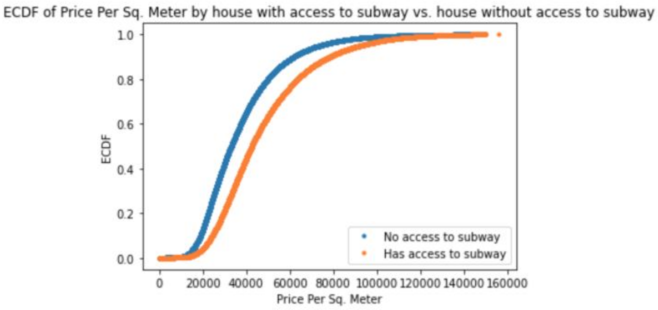
- Mann-Whitney and Welch’s T-Test
  - Results:

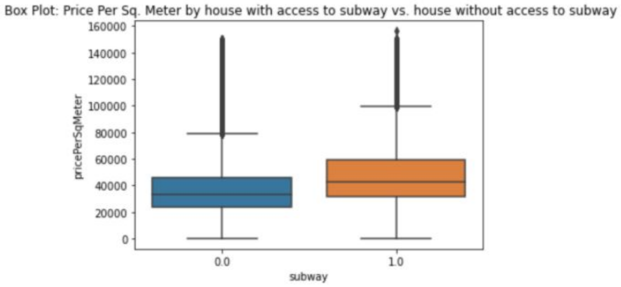
    -
      ```
      MannwhitneyuResult(statistic=10866017339.0,
      pvalue=2.1382780878680677e-116)
      ```

    -
      ```
      Ttest_indResult(statistic=-7.6850894427140055,
      pvalue=1.5354933157064842e-14)
      ```
- Conclusion: All of the four tests rejected the null hypothesis with p-values almost being zero which indicates that fiveYearsProperty has a statistically significant impact on the target variable

3.4.2 subway vs. pricePerSqMeter

- ECDF and Box Plot

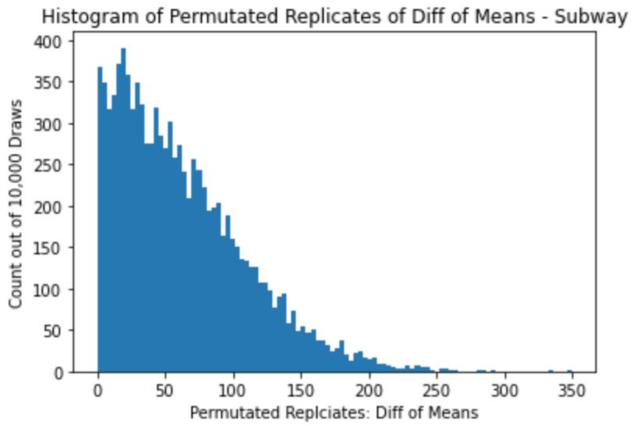




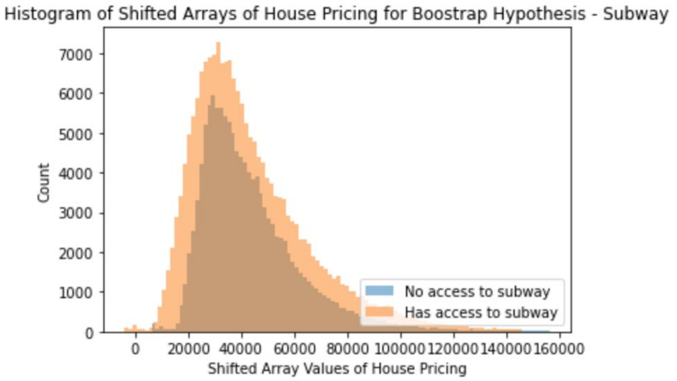
- Permutation Test
  - Results (10,000 permutation replicates)

    -
      ```
      Empirical Diff of Mean: 10253.661140687305
      ```

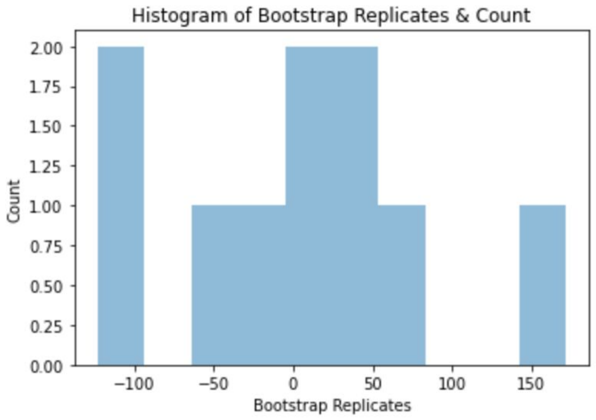
    -
      ```
      Proportion of permutation replicates with value at least
      as extreme as the empirical diff of means
      p-value = 0.0000
      ```



- Bootstrap Test
  - Results (10,000 bootstrap replicates)
    - `Mean of Concatenated Data: 43567.725202547575`



    - `Proportion of bootstrap replicates with value at least as extreme as the empirical diff of means`
      `p-value = 0.0000`



- Mann-Whitney and Welch’s T-Test
  - Results:
    - `MannwhitneyuResult(statistic=8366121538.5, pvalue=0.0)`
    - `Ttest_indResult(statistic=-138.93892088114305, pvalue=0.0)`

- Conclusion: All of the four tests rejected the null hypothesis with p-values almost being zero which indicates that subway has a statistically significant impact on the target variable