1.1 Which statistical test did you use to analyze the NYC subway data? Did you use a one-tail or a two-tail P value? What is the null hypothesis? What is your p-critical value?

Statistical Test: Mann Whitney U Test

P value: Two tailed P value, as we want to know in which situation NYC subway is most used, that is during rainy days or non rainy days (Checking the relationship in the both direction). Due to that, one tailed P value is not appropriate as it check the statistical significant in one direction of interest.

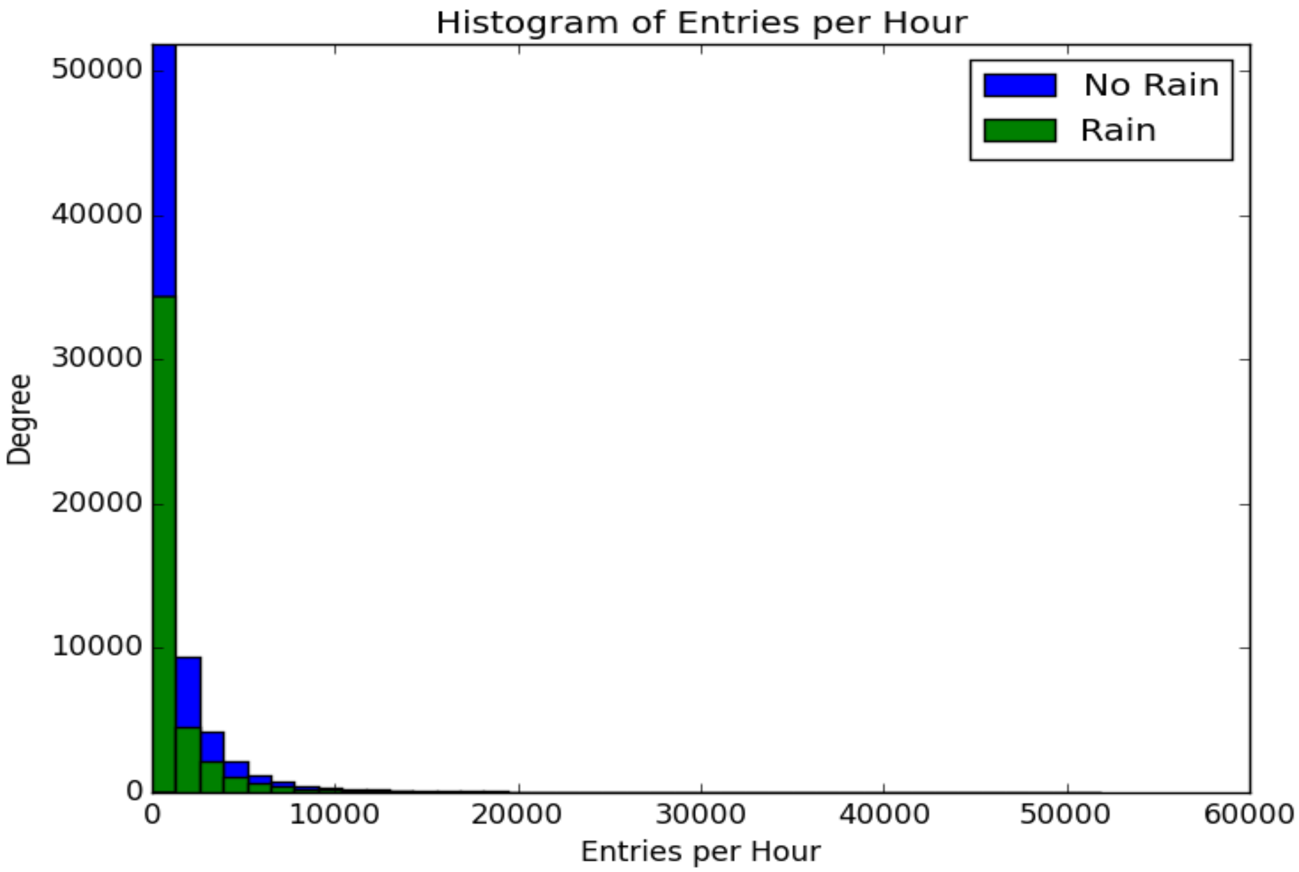
Null Hypothesis: The distributions of both populations (number of entries on rainy day and number of entries on non rainy day) are equal.

P-critical value: 0.05

1.2 Why is this statistical test applicable to the dataset? In particular, consider the assumptions that the test is making about the distribution of ridership in the two samples.

Reason: After examine the distribution of the given two samples (1. numbers of people used New York subway on the rainy days 2. number of people used New York subway on the non rainy day) using histogram; it is clear that the distribution is not normally distributed (Which is positively skewed). Due to that the non-parametric test - Mann Whitney Test will be useful for the statistical analysis because it does not assume that the given dataset is normally distributed.

Histogram:



1.3 What results did you get from this statistical test? These should include the following numerical values: p-values, as well as the means for each of the two samples under test.

Results:

The Mann-Whitney U test returns the following values.

1. The Mann-Whitney statistical value: **1924409167.0**
2. One-sided p-value assuming an asymptotic normal distribution: **0.024999912793489721**

Sample 1 (Numbers of entries on rainy days) mean: **1105.44637675**
Sample 2 (Numbers of entries on non rainy days) mean: **1090.27878015**

Sample 1 (Numbers of entries on rainy days) median: **282.00**
Sample 2 (Numbers of entries on non rainy days) median: **278.00**

1.4 What is the significance and interpretation of these results?

Here due to two-sided test, value of p should be multiply by 2. Now **p = 0.049998255869794442**

The calculated p-value is less than the critical p value (0.05), which rejects the null hypothesis and stats that the distribution of both the samples (number of entries on rainy day and number of entries on non rainy day) are not equal.

Along with Mann-Whitney U test statistics, the higher mean and median of sample1 (numbers of entries on rainy days) states that the numbers of entries on rainy days is higher than the number of entries on non rainy days.