Analyzing the NYC Subway Dataset

Short Questions

Section 1

Statistical Test

1.1 The data has been analyzed using first a two-tail test to check whether the difference in means is due to chance or not, and another test to compare whether the fact that more people use the subway in rainy days is due to chance or sampling error or not.
Both tests have been done at a 95% significance level (p-critical is 0.05).

For the Welch's t-test, the null hypothesis is that the mean of passengers in rainy days and non rainy days are equal. The non-null hypothesis is that the means are different. The number of passengers is the number of entries per hour (*ENTRIESn_hourly*) and a day is considered to be rainy if *rain* is equal to 1. That is,

$$H_0 : \mu_A = \mu_B ,$$
$$H_1 : \mu_A \neq \mu_B \quad ,$$

where μ_A and μ_B are the means of the *ENTRIESn_hourly* for rainy days and non-rainy days respectively.

For the second test, the Mann-Whitney U Test has been used to test whether it is more likely that more people take the subway in a rainy day or not:

$$H_0 : P(x_A > x_B) = 0.5 ,$$
$$H_1 : P(x_A > x_B) \neq 0.5 \quad ,$$

where x_A and x_B are the *ENTRIESn_hourly* for rainy days and non-rainy days respectively.

1.2
For the first test, it makes sense to use a two tailed test. Since we are comparing means, we need to check whether the first mean is significantly larger than the second, or if the first mean is significantly smaller than the second.
For the second test, we just want to know whether x_A is significantly larger than x_B or not, but not both, so it makes sense to do a one-tailed test.

1.3
The mean for non-rainy days is 1850.959 and for rainy days is 2260.001. The Welch's t-test returned a p-value of 4.242e-12 and a t-value of 6.95.
The Mann-Whitney U test returned a p-value of 2.74e-07 and a U value of 153635120.5.

1.4
Since the p-value returned is much smaller than p-critical the null hypothesis can be rejected, concluding that the observed data is not due to chance, therefore it is very probable that in a rainy day more people take the subway than in a non-rainy day.

Section 2

Linear Regression

2.1 I’ve used Gradient descent to predict entries hourly for rainy days.

2.2 The input variables are the entriesn_hourly for rainy days (rain=1).

2.3
I’ve did some tests using a meantempi value of 0.016 or more, foggy days and cold days but I didn’t find any significant difference, so I sticked to using rainy vs non-rainy days.

2.4

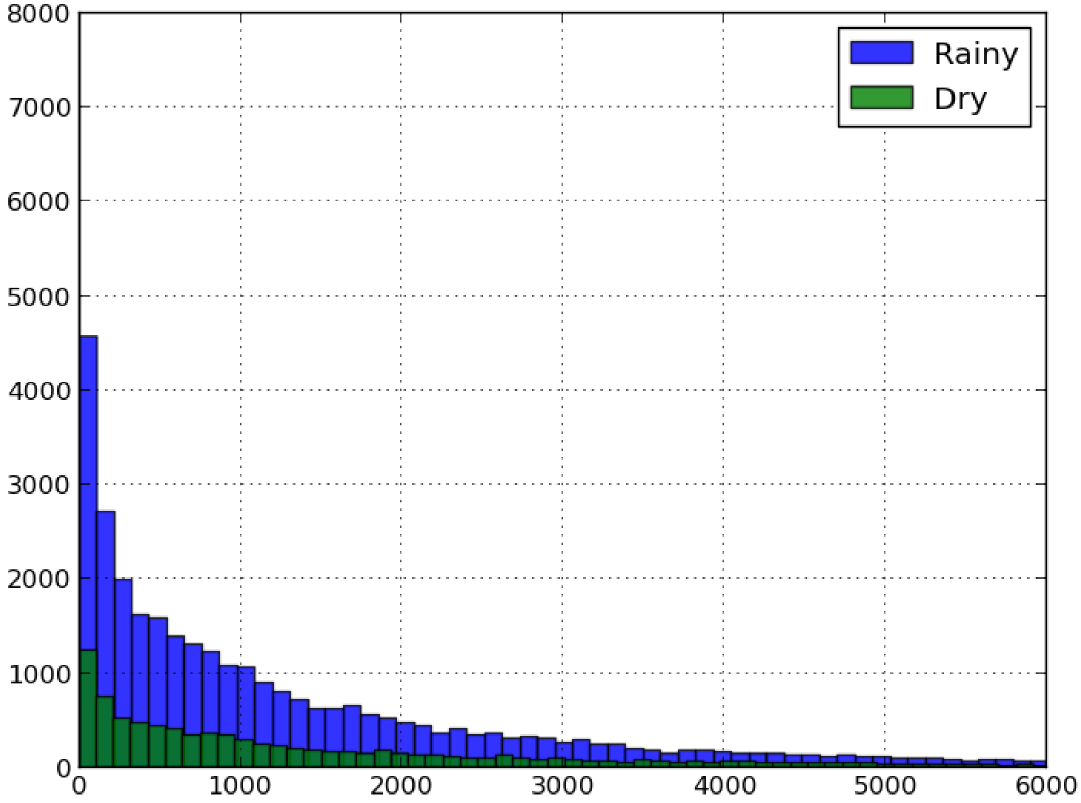
2.5
The value for R2 is 0.456

2.6
The value for R2 is 0.456, this means that the model account for about 46% of variability in the values respect the fit line. Given this value, the model can give a general idea of the difference between rainy and non rainy days, but it might be possible to make better predictions with extra data (holidays, sports or other social events that might influence ridership).

Section 3

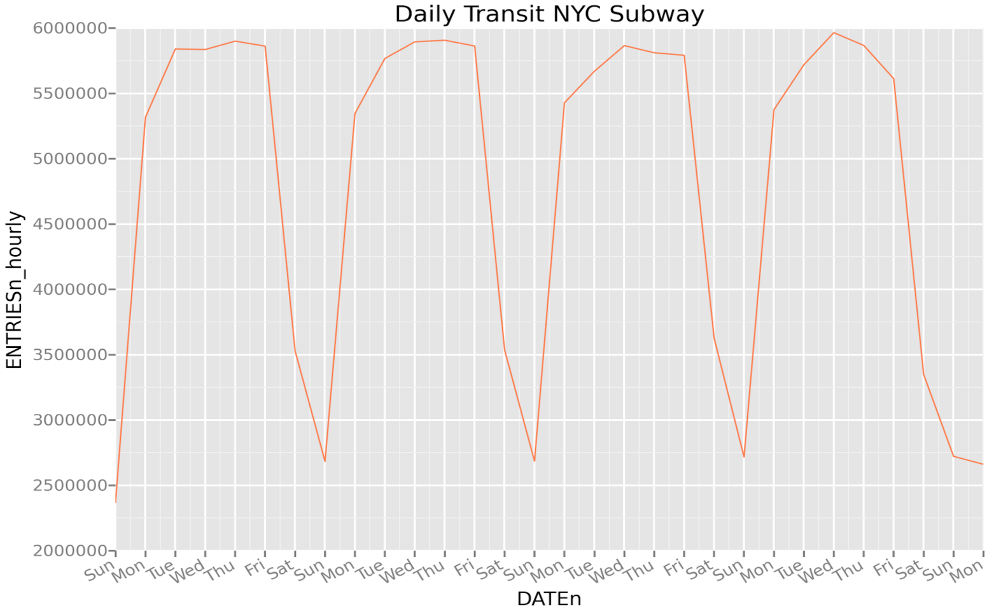
Visualization

3.1

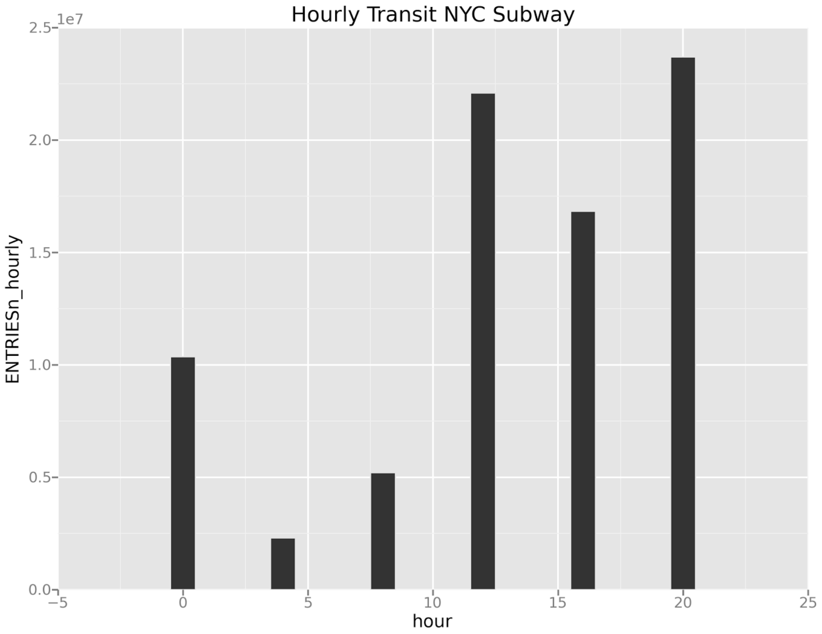


3.2

Plotting the number of entries per hour for a 3 week period helps in having an idea about the weekly pattern of commuters.



Hourly ridership bar plot



Section 4

Conclusion

4.1
More people ride the subway in a rainy day

4.2

The results of the tests indicate that it is not casual that the mean of the number of entries per hour is greater in a rainy day than in a non-rainy day.

Section 5

Reflection

5.1

The dataset can give an insight in to what is more likely to happen according to a set of meteorological data, but it doesn't take into account social events like sports, demonstrations, holiday season, tourist destinations etc. which probably influences ridership. For instance in the weekly plot it can be seen that new yorkers use the subway primarily during the week.

Analyzing the data using Welch's t-test and, after the null hypothesis has been rejected, apply the MW U-test to see how likely it is that the number of passengers will be larger in rainy days, is a simple but reasonable way of getting an idea about the behaviour of NYC subway users. The r_squared value might not be great, but it already gives an approximation good enough to conclude that the increase in ridership during rainy days is not due to chance.