Section 1. Statistical Test

1.1 I run the Mann-Whitney U test to analyse and investigate the data of the NYC Subway. We have two samples according to two different weather conditions: rainy day or not rainy day. This non-parametric test checks whether two samples taken in consideration had the same distribution (null hypothesis) or either one of the two is likely to generate higher values. This translates in setting the null hypothesis that the probability of the first sample distribution being higher than the second one is equal to 0.5, whilst the alternative hypothesis is that such a probability is different than 0.5, and as a result the two samples do not have the same distribution. I considered a two-tailed p-value as otherwise we would assume that rain will not be associated with lower ridership, and the critical significance level was fixed at 5%.

1.2 We can use Mann-Whitney U test because we have two samples and we are interested in studying their distributions and how they compare to each other. Under the null hypothesis, the two samples come from the same distribution, i.e. they have equal probability to observe an element from each distribution exceeding an element from the other distribution; whilst under the alternative hypothesis one sample has larger values than the other one. The U test statistic is computed as the sum of ranks in one of the samples.

1.3 We obtained a one-tailed p-value of 0.0249999, thus being a two-tailed p-value of approximately 0.05. We conclude that we are 95% confident to reject the null hypothesis that the two samples have the same distributions. Moreover, in our test, we got a U-value of 1,924,409,167, and the average entries during rainy and non-rainy days are 1105.4464 and 1090.2788, respectively.

1.4 The test provides enough evidence to conclude that the two distributions appear to be (statistically) significantly different, since our two-tailed p-value is below our two-tailed significance level of 0.05.

Section 2. Linear Regression

2.1 I used both the Gradient Descent method and the OLS method.

2.2 In the OLS the variables used were ‘rain’, ‘precipi’, ‘Hour’, ‘meantempi’ and ‘UNIT’. ‘UNIT’ is a dummy feature in my regression, as it is a variable made of a string and a numerical part. I introduced it to assign a category or group to each element in the dataset.

2.3 Those variables provide a good contribution as we can expect weather variables, such as , ‘rain’, ‘precipi’ do play a role in describing the Subway use as people are more likely to use it to shelter themselves from the precipitations. The same intuition can be a good driver to take into account other variables such as ‘Hour’ (beginning and end of the working time are known to be the busiest periods of public transports). ‘rain’ was the only binary variable used: it assigns the value 1 in case of rainy day and 0 otherwise. It may have an importance in our analysis as we may expect weather affects the Subway usage. ‘meantempi’ indicates the average time of a ride in the Subway for each user. ‘UNIT’ is a dummy variable as described in 2.2.
Lately such a use got confirmation in terms of statistical goodness-of-fit measured by our R-squared measure. In fact, as it will be reported next, those variables could explain a high percentage of the prediction.

2.4 The estimated coefficients for the independent variables above are 23.3261, 36.0905, 62.2629 and -10.9566, respectively, meaning that all variables have a proportional impact on the subway entries. Nevertheless, their impacts are not all the same. The coefficients relative to ‘Hour’ and

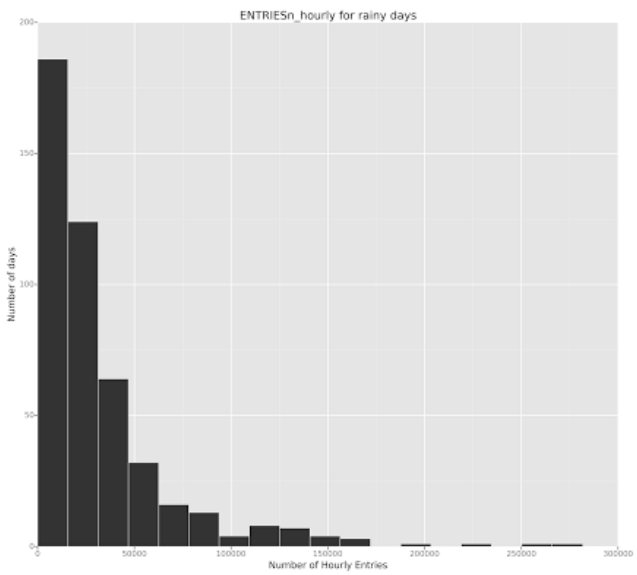
'meantempi' are statistically significant, as their p-values lie below the significance level of 0.05; whereas we observe that the t-statistics for 'rain' and 'precipi' are not very high and their p-values are higher than the significance level. We conclude 'rain' and 'precipi' are not statistically significant. Finally, 'meantempi' is the only variable, whose coefficient is negative, meaning it affects negatively the Subway number of entries.

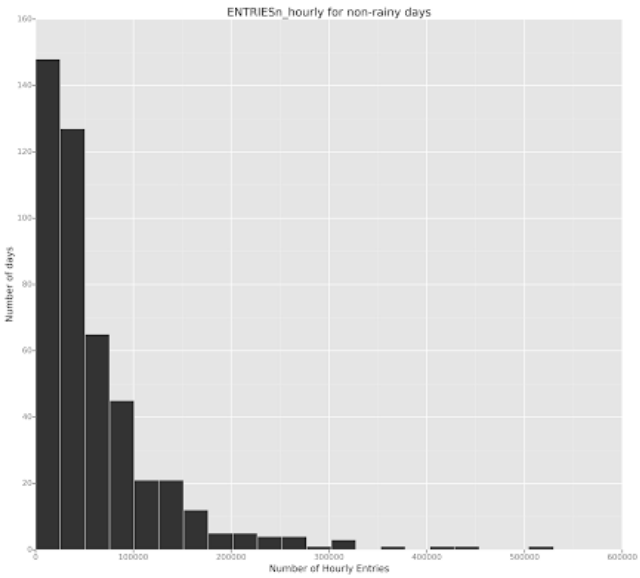
2.5 My R-squared is approximately 31.81% using the gradient descent and 48.3% using the OLS. The R-squared is a model's measure for the goodness-of-fit. It indicates the contribution of the input variables to explain the input variable, or, stating it differently, it measures how well the regression line approximates the real data point. The biggest drawback of this measure is that adding up new input variables always provides a "positive" contribution to the R-squared, meaning that we will always have an (artificially) higher R-squared, without getting penalised for each variable added.

2.6 The R-squared I got is not very high. This means that the linear model used has a quite large error and it is not the most appropriate approximation for the given data set. However, my goal is mainly to identify a relationship between the NYC Subway use and some variables, therefore, for that I would mainly rely on the p-values obtained for each variable rather than on the magnitude of the R-squared. Indeed, I found out already two important statistically significant relationships that are not offset by the poor performance of other measures like the R-squared. Moreover, to have a better view on the output predictions one should also delve into residuals analysis and prediction intervals and find out which models may convey narrower intervals. The presence of fat tails, indeed, may induce the conclusion that there are some very large residuals affecting the adequacy of our linear regression model, which has the underlying assumption that residuals are normally i.i.d. (*independent and identically distributed*).
Look at my answer above about further criticism to the measure of goodness-of-fit.

Section 3. Visualization

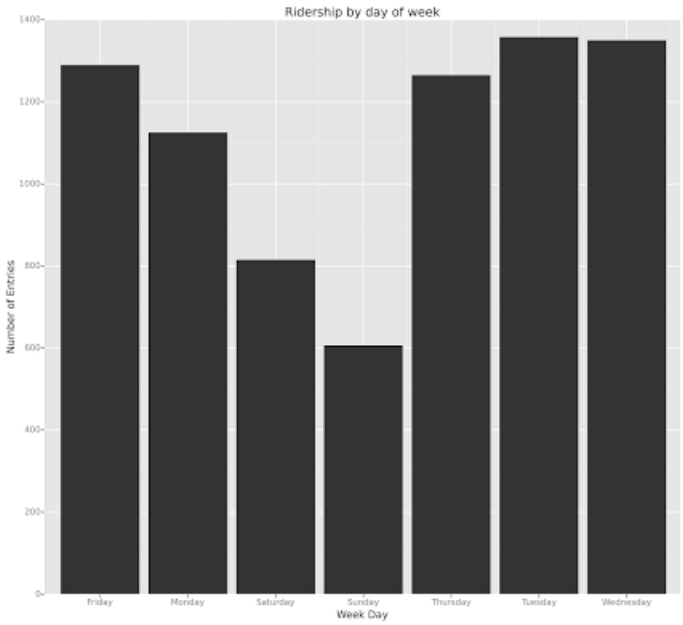
3.1 Histogram of rainy days (left); Histogram of non-rainy days (right).





The two plots of *Section 3.1* show the two histograms for the two samples under consideration: Subway usage in rainy (*Plot 1*) and non-rainy (*Plot 2*) days. In order to compare the two results, we should compare the average hourly entries given the amount of rainy and non-rainy days, respectively. In fact, *Plot 2* shows more days but a less frequent Subway usage.

3.2 Plot 2: Ridership by day of the week. On weekends the Subway is more heavily used.



Section 4. Conclusion

4.1 More people use NYC Subway during rainy days than non-rainy days. This can be easily observed looking *Plot 1* in *3.1* and comparing it to *Plot 2* in *3.1* and taking into account the fact that the sample size of rainy days is smaller than the one in non-rainy days. Indeed, on average, the Subway usage is higher in *Plot 1*. Furthermore, statistical analysis supporting the fact that the two samples have different distributions (as reported in *Section 1*) and the variables 'rain' and 'precipi' in our linear regression have positive coefficients on the number of entries, although there are not statistically significant. Another support of our conclusion comes from the fact that the average Subway usage during rainy days is higher. Further statistical details are to be discussed in the following later on.

4.2 A significant analysis leading to this conclusion was the histogram comparison between the two samples. I generated two histograms based off every instance of entry data for rainy and non-rainy days. In fact, comparing the two *Plots* in *3.1* we can immediately see how the first sample (*Plot 1*) outnumbers the second sample (*Plot 2*). The R-squared shows a significant degree of multiple linear correlation between the covariates and the output in our prediction (number of entries), and the chosen coefficients are statistically significant given the results from our test statistics.

Besides that, one can also look at the mean and the median to see that the amount of passengers during rainy days is higher.

Section 5. Reflection

5.1 One important shortcoming has already been in *Section 2*.

5.1.1 Concerning the dataset, we do not know which was the sampling period considered and how it might change across time and/or season. It would be certainly help to run the same analysis on data from other sources. One of the drawbacks of the dataset is also that we do not have more precise information about clients' behaviours, for instance how does the destination change across time between different city areas as temperature and other atmospheric conditions change. It would, indeed, be very interesting to analyse whether travellers are more likely to go from the suburbs to downtown during the morning and vice versa during the evening or at night. In addition, it would be interesting to analyse such a distinction between weekdays and weekend days.

5.1.2 The linear model and the t-test make several assumptions on the data. The stronger one is on the parametric family adopted. In fact, the residuals are assumed to be normally distributed. Linearity is sometimes a simplistic approximation as the minimising values are linear with respect to the observations on the output variable and for very sparse dataset it is not capable to capture correctly data patterns.

5.2 Look at 5.1.1.