# Section 0. References

1. [Mann-Whitney U Test](https://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test)
2. [Shapiro Wilk Test](http://docs.scipy.org/doc/scipy-0.15.1/reference/generated/scipy.stats.shapiro.html)
3. [Mann-Whitney null hypothesis under unequal variance](http://stats.stackexchange.com/questions/56649/mann-whitney-null-hypothesis-under-unequal-variance)
4. [Linear Regression](https://en.wikipedia.org/wiki/Linear_regression)

# Section 1. Statistical Test

## 1.1 Which statistical test did you use to analyse 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?

Mann-Whitney U test is used to analyse the NYC subway data. A two-tail p-value is used as it is not hypothesised that any one of the two data set has higher or lower mean than the other.

H0: There is no significant difference in the groups of data (Rain has no effect on the ridership of NYC subway)

P-critical value or the significance level used is 0.05 (α ≤ 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.

Both the groups of data are not normally distributed (can be seen by plotting a histogram in Section 3.1). Additionally, Shapiro-Wilk test was performed on both the groups of data, a p-value very close to zero was obtained which means that the data is not normally distributed. Welch’s two sample t-test assumes that data given is normally distributed therefore it is not applicable in this case. Hence, a non-parametric test such as Mann-Whitney U test is more suited.

## 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.

|  |  |
| --- | --- |
| Mean (Entries with rain) | 1105.446 |
| Mean (Entries without rain) | 1090.279 |
| U-statistic | 1924409167.0 |
| p-value (two-tailed) | 0.05 |

## 1.4 What is the significance and interpretation of these results?

When we look at the means of both the groups of data, we cannot infer that there is a significant difference in them. Additional statistical tests are required to find out whether there is a significant difference or not. When Mann-Whitney U test was performed on both the groups of data, a p-value of 0.05 is obtained, which is within the p-critical value or significance level of 0.05. Hence, we have a moderate evidence against the null hypothesis that there is no difference in both the groups of data and we can infer that rain affects the ridership on the NYC subway.

# Section 2. Linear Regression

## 2.1 What approach did you use to compute the coefficients theta and produce prediction for ENTRIESn\_hourly in your regression model:

OLS using Statsmodels or Scikit Learn

## 2.2 What features (input variables) did you use in your model? Did you use any dummy variables as part of your features?

Features: ‘meantempi’, ‘rain’, ‘meanwindspdi’

Dummy variables were used for ‘UNIT’ which was used as a category. All the distinct stations in ‘UNIT’ were converted to dummy variable and added as a feature. Another set of dummy variables are used with ‘Hour’.

## 2.3 Why did you select these features in your model?

‘rain’, ‘meantempi’ and ‘meanwindspdi’ variables are related to weather. Adding these variables also increased the R2 value that ranged from slight to significant.

‘Hour’ variable was added to model as ridership in a subway depends on the hour of the day (Section 3.2) and there was also a significant increase in the R2 value. ‘Hour’ was converted to dummy variable as hours in a day are categorical in nature.

Adding other variables as features had very marginal effect or no effect on the R2 value, therefore they were not added to the model.

## 2.4 What are the parameters (also known as "coefficients" or "weights") of the non-dummy features in your linear regression model?

|  |  |
| --- | --- |
| **Features** | **Coefficients** |
| Constant | 1557.775429 |
| meantempi | -5.290304 |
| Rain | 2.204533 |
| Meanwindspdi | 25.136729 |

## 2.5 What is your model’s R2 (coefficients of determination) value?

R2 = 0.50178844198400041

## 2.6 What does this R2 value mean for the goodness of fit for your regression model? Do you think this linear model to predict ridership is appropriate for this dataset, given this R2 value?

R2 is a value that measures goodness of fit in a regression model. R2 value of 1 signifies that the model perfectly fits the data. R2 value also explains the percentage of variance explained by the linear model. The R2 value obtained is 0.501788 i.e. 50.1788% of variance is explained by the linear model. To further analyse the goodness of fit, the histogram of residuals can be plotted. If the residuals are normally distributed, we can assume that the choice of our model is appropriate.

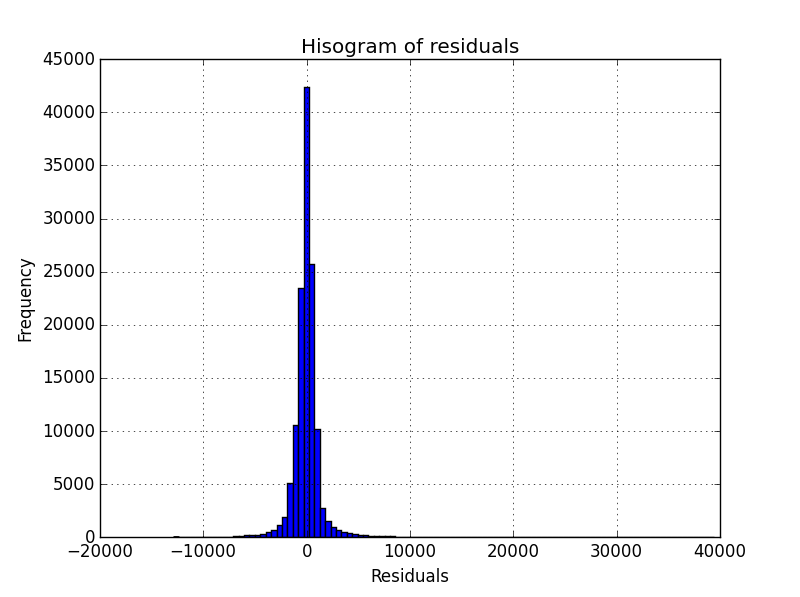


Fig 1: Residual Histogram

From the histogram, it is clear that the residuals are normally distributed. From the value of R2 and histogram of residuals, we can infer that the linear model to predict subway ridership is appropriate.

# Section 3: Visualization

## 3.1 One visualization should contain two histograms: one of ENTRIESn\_hourly for rainy days and one of ENTRIESn\_hourly for non-rainy days.

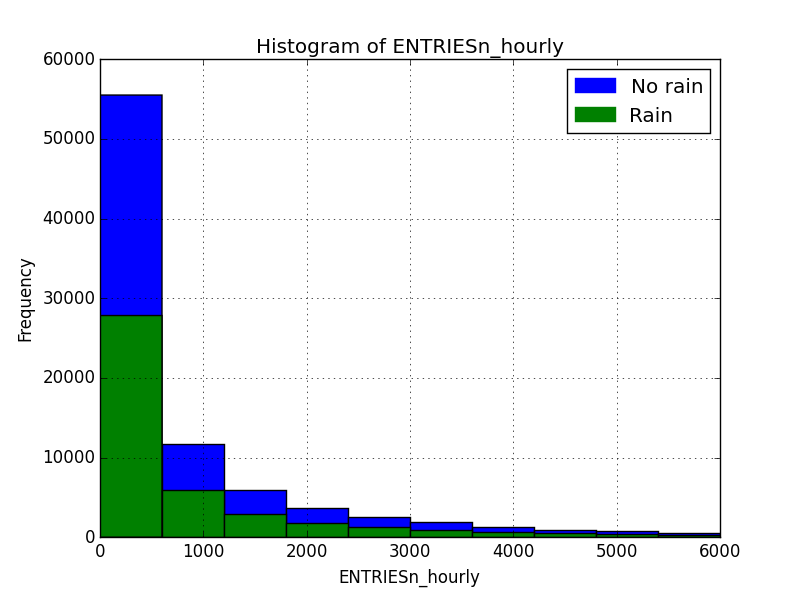


Fig 2: Histogram of subway entries

This plot shows the histogram of subway entries on the days when it rained and on the days when it did not rain in the same plot. From the plot, it can be seen that number of days when it rained is less than the number of days when it didn’t.

## 3.2 One visualization can be more freeform. You should feel free to implement something that we discussed in class (e.g., scatter plots, line plots) or attempt to implement something more advanced if you'd like.

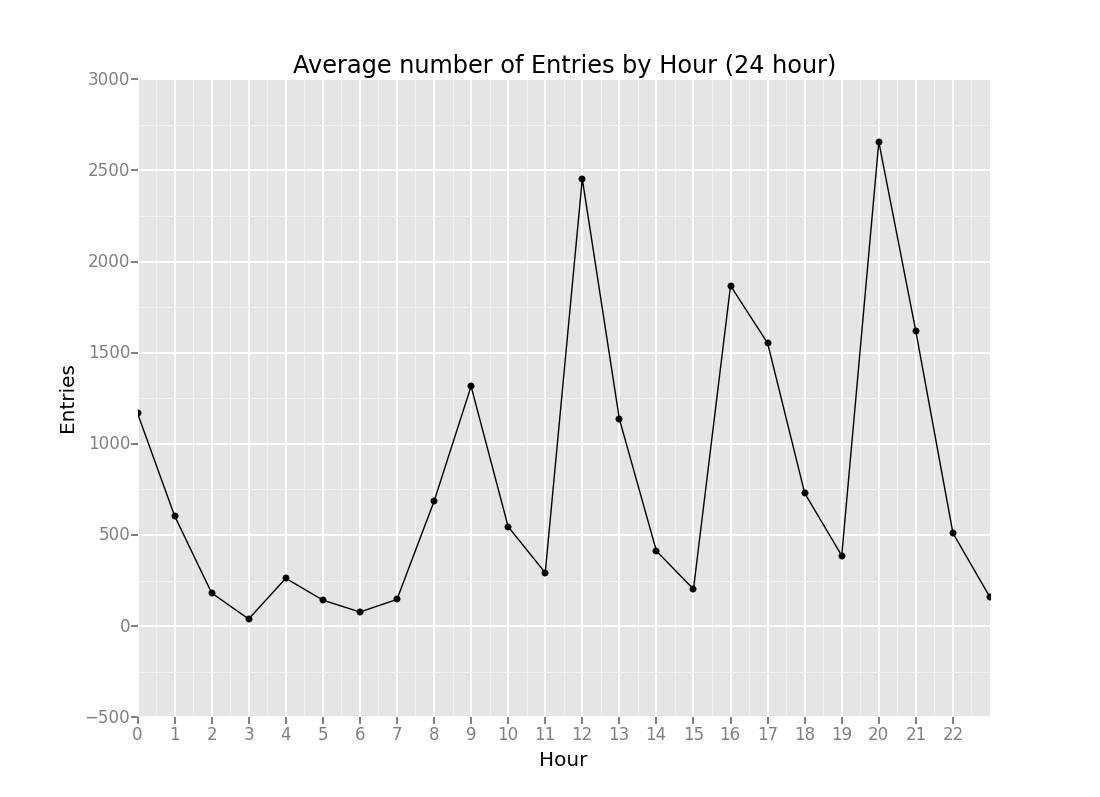


Fig 2: Average subway ridership by hour

This plot shows the average number of subway ridership from all stations by hour of the day. The subway ridership depends on the hour of the day. There were low number of riders during the early hours of the day and there were several spikes of high number of riders during the day (8-10 AM, 12 AM, 4-5 PM, 8-9 PM)

# Section 4: Conclusion

## 4.1 From your analysis and interpretation of the data, do more people ride the NYC subway when it is raining or when it is not raining?

From the analysis and interpretation of data, it can be concluded that more people ride on the NYC subway when it is raining.

## 4.2 What analyses lead you to this conclusion? You should use results from both your statistical tests and your linear regression to support your analysis.

Starting with plotting of histogram of entries at NYC subways, it is clear that the number of rainy days is less than the number of non-rainy days. When we compare the means of the two sets of data, there is not much difference. Mean of entries on rainy days (1105.446) is more than the mean of entries on non-rainy days (1090.279). Comparing means does not provide us with sufficient evidence of difference in sets of data. Mann-Whitney U test is done to establish that the two sets of data are from different population (p-value = 0.05).

A linear model was constructed to predict the subway ridership which included weather parameters and hour of the day as variables. The R2 value was approximately 0.50 and the residuals were normally distributed. The coefficient of rain in the linear model is positive, which means that the predicted subway ridership increases when there is rain. Hence, it can be concluded that more people ride in the NYC subway when it rains.

# Section 5: Reflection

## 5.1 Please discuss potential shortcomings of the methods of your analysis, including:

1. Dataset

The dataset includes all the weather parameters that can be included to predict the number of riders in the subway. To get a complete picture of the subway ridership or to get a more accurate prediction model, some more information can be added, such as whether the day was weekday, weekend or a public holiday etc.

1. Analysis, such as the linear regression model or statistical test.

The linear model can be improved by incorporating non-linear transformations of the variables. This will help in increasing the accuracy of the predictions, but the problem of over-fitting should be carefully looked at.