

# Analyzing the NYC Subway Dataset

## Questions

### Overview

This project consists of two parts. In Part 1 of the project, you should have completed the questions in Problem Sets 2, 3, and 4 in the Introduction to Data Science course.

This document addresses part 2 of the project. Please use this document as a template and answer the following questions to explain your reasoning and conclusion behind your work in the problem sets. You will attach a document with your answers to these questions as part of your final project submission.

## Section 0. References

Please include a list of references you have used for this project. Please be specific - for example, instead of including a general website such as stackoverflow.com, try to include a specific topic from Stackoverflow that you have found useful.

### REFERENCES

[0] Mann-Whitney U-Test [http://en.wikipedia.org/wiki/Mann%E2%80%93U\\_test](http://en.wikipedia.org/wiki/Mann%E2%80%93U_test)

[1] SciPy Documentation

[2] Statsmodel OLS Documentation

[3] Machine-Learning - Udacity, Coursera Courses, Wikipedia

[4] Rubric

<https://docs.google.com/document/d/1ZWdmlEgtRhreyN7AaiEfoYP70GqxOZqrajWtzlov8HM/pub?embedded=True>

## Section 1. Statistical Test

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?

I used the Mann-Whitney U test to analyze the NYC subway data in order to determine the null hypothesis. The two-tail P value was used because we do not know which data set is higher or lower. The null hypothesis is when the rain has no correlation with ridership. The p-critical value was placed at 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.

The plot indicated that the plot in both rain and no-rain were not normally distributed. Therefore, non-parametric statistics must be used and Mann-Whitney U deals with non-parametric data, while a test like Welch would only be adequate for a normally distributed sample space.

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 with rain:** 1105.4463767458733

**mean without rain:** 1090.278780151855

**U-Statistic:** 1924409167.0

**p-value:** 0.024999912793489721

1.4 What is the significance and interpretation of these results?

Consider the mean and U-statistic. The ratio of the means with and without rain results in about 0.014. The U-Statistic is 1924409167 which is close to the maximum value. The p-value is ~0.025 and if we compare that to the p-critical value (0.05) it meets the criteria, with 95% confidence that the null-hypothesis is indeed false.

## 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**

**Gradient descent using Scikit Learn**

Or something different?

For the Linear Regression, I went with using the OLS Statsmodel to compute the coefficients theta and produce prediction in the regression model.

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

Feature space: rain (0 or 1), precipitation, mean wind speed, hour, mean temperature, and max pressure. The default dummy variables defined on the code were used. It did not seem intuitive to add the dummy variables as part of the feature space.

Originally the output was:

You calculated  $R^2$  value correctly!

Your calculated  $R^2$  value is: 0.318137233709

2.3 Why did you select these features in your model? We are looking for specific reasons that lead you to believe that the selected features will contribute to the predictive power of your model. Your reasons might be based on intuition. For example, response for fog might be: "I decided to use fog because I thought that when it is very foggy outside people might decide to use the subway more often."

Your reasons might also be based on data exploration and experimentation, for example: “I used feature X because as soon as I included it in my model, it drastically improved my  $R^2$  value.”

I added Feature ‘maxpressurei’ because as soon as I places it in my model, my  $R^2$  value increased.

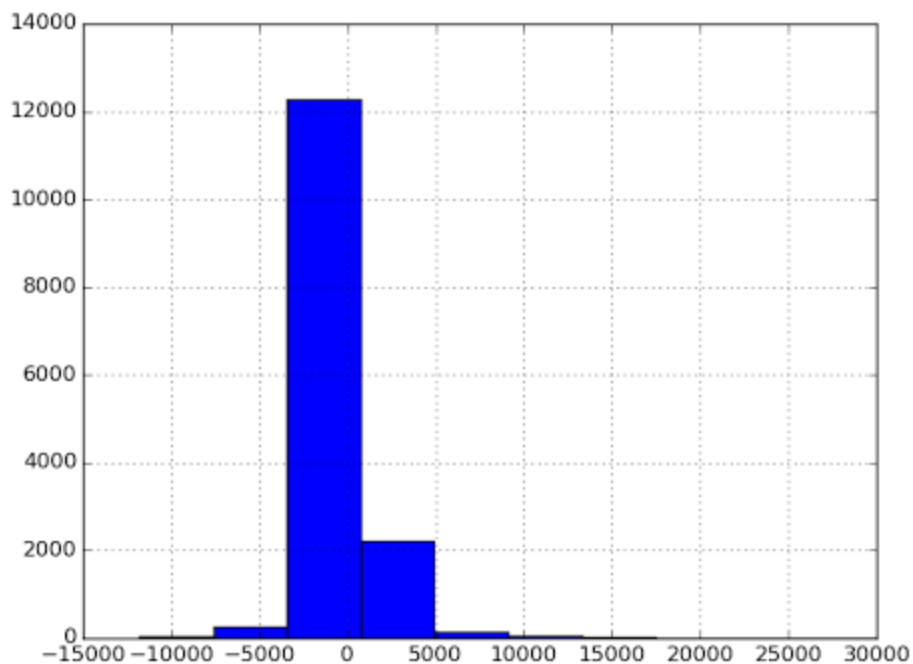
2.4 What are the coefficients (or weights) of the non-dummy features in your linear regression model?

-1.43980597e+03 -1.43291533e+03 -1.57487142e+03 -1.56826017e+03

2.5 What is your model's  $R^2$  (coefficients of determination) value?

Your  $r^2$  value is 0.47924770782

2.6 What does this  $R^2$  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  $R^2$  value?



## Section 3. Visualization

Please include two visualizations that show the relationships between two or more variables in the NYC subway data.

Remember to add appropriate titles and axes labels to your plots. Also, please add a short description below each figure commenting on the key insights depicted in the figure.

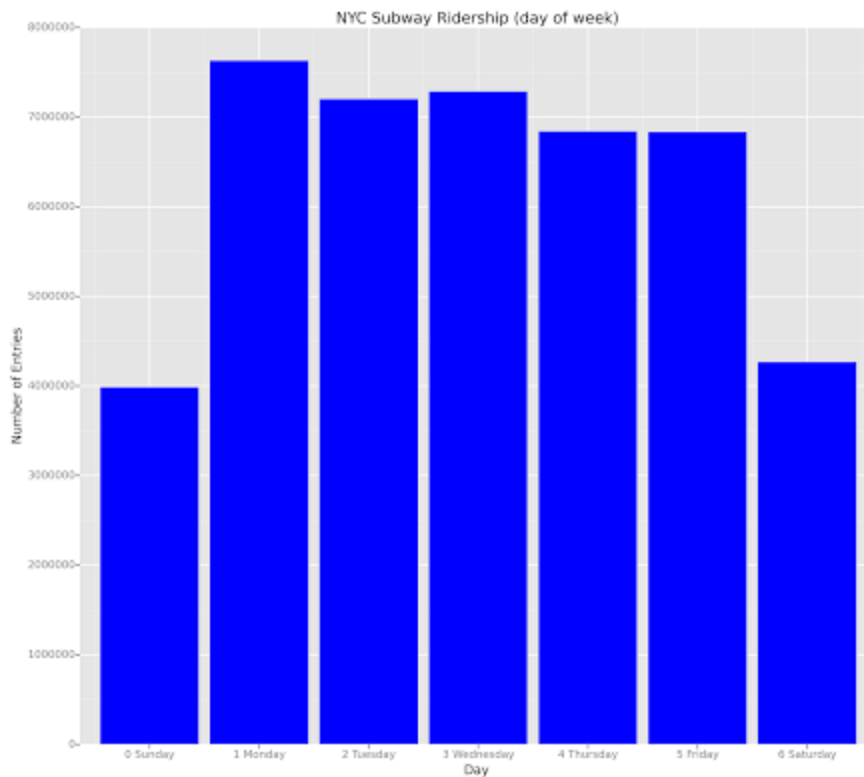
3.1 One visualization should contain two histograms: one of `ENTRIESn_hourly` for rainy days and one of `ENTRIESn_hourly` for non-rainy days.

You can combine the two histograms in a single plot or you can use two separate plots.

If you decide to use two separate plots for the two histograms, please ensure that the x-axis limits for both of the plots are identical. It is much easier to compare the two in that case.

For the histograms, you should have intervals representing the volume of ridership (value of `ENTRIESn_hourly`) on the x-axis and the frequency of occurrence on the y-axis. For example, each interval (along the x-axis), the height of the bar for this interval will represent the number of records (rows in our data) that have `ENTRIESn_hourly` that falls in this interval.

Remember to increase the number of bins in the histogram (by having larger number of bars). The default bin width is not sufficient to capture the variability in the two samples.



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. Some suggestions are:

Ridership by time-of-day

Ridership by day-of-week

## Section 4.

### Conclusion

*Please address the following questions in detail. Your answers should be 1-2 paragraphs long.*

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?

As analyzed from Section 1, by using the Whittney test, (where the p-value was 0.025), it is safe to say that ride the NYC subway when it's 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.

Looking at the linear regression, the positive correlation indicates that rain increases the number of people present on the subway.

## Section 5.

### Reflection

*Please address the following questions in detail. Your answers should be 1-2 paragraphs long.*

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

Dataset,

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

When I inspected the data, there were more entries than exits. It may have been issues with people finding clever ways of leaving the subway station, or other artifacts.

5.2 (Optional) Do you have any other insight about the dataset that you would like to share with us?

[Will elaborate more later]