# 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-Whittney U-Test http://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test
[1] SciPy Documentation
[2] Statsmodel OLS Documentation
[3] Machine-Learning - Udacity, Coursera Courses, Wikipedia
[4] Rubric
https://docs.google.com/document/d/1ZWdmlEgtRhreyN7AaiEfoYP70GqxOZqrajWtzIov8HM/pub?embedded=True
[5]
http://blog.minitab.com/blog/adventures-in-statistics/how-high-should-r-squared-be-in-regression-analysis

# Section 1. Statistical Test

1.1 Which statistical test did you use to analyze the NYC subway data?

I used the Mann-Whitney U test to analyze the NYC subway data in order to determine the null hypothesis.

Did you use a one-tail or a two-tail P value?

The two-tail P value was used because we do not know which data set is higher or lower.

What is the null hypothesis?

From Mann-Whitey "Under the null hypothesis $H_0$, the probability of an observation from the population *X* exceeding an observation from the second population *Y* equals the probability of

an observation from $Y$ exceeding an observation from $X$ : $P(X>Y)=P(Y>X)$ or $P(X>Y)+0.5$ $P(X=Y)=0.5$. A stronger null hypothesis commonly used is "The distributions of both populations are equal" which implies the previous hypothesis."

The null hypothesis is when the two populations are equal, which implies that the rain has no correlation with ridership.

What is your p-critical value?

The p-critical value was placed at 0.05.

**** FIX THIS****

In Section 1.1: Your statement of the null hypothesis for the statistical test is not quite accurate. An exact statement of the null hypothesis can be found in the downloadables from Lesson 3. The downloadable notes about the Mann-Whitney U test can be accessed by clicking on the appropriate link below the video window of any of the Lesson 3 videos.

****

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-Whittney 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 (whitney)`

For the **two-tailed p-value** we get: **0.04999998255**

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 ~0.04999998255 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, hence a type-I error.

# 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.Dummy variables were given for the features 'UNIT' (identification no. / turnstile location). They were defined as a boolean variable (0 or 1).
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.”
**ANSWER:**
I kept rain, precipitation, hour and mean temperature because it was the top performance from what I've explored. I also added Feature ‘maxpressurei’ and as soon as I placed 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?

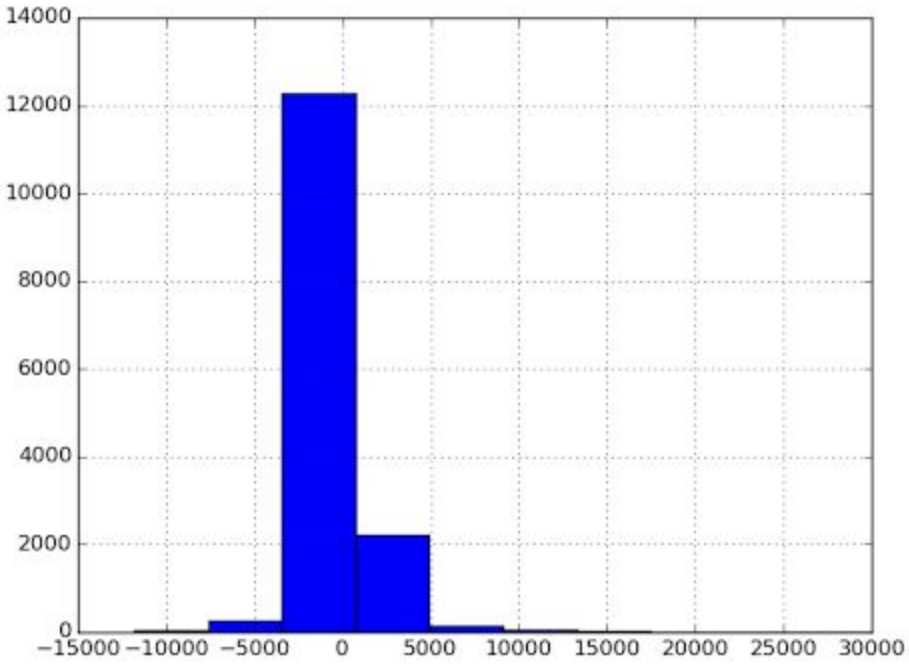
```
-1.43980597e+03  -1.43291533e+03  -1.57487142e+03  -1.56826017e+03
```

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

```
My 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?

R^2 is defined the percentage of variability explained by the model. It is also a quantitative metric of "goodness of fit". The result is 47.9% variation. Despite this variation, it can still be reliable model to use and the histogram below justifies why. However, a more useful insight is observe the change in R^2 in changing slight input variables in the same or similar study.



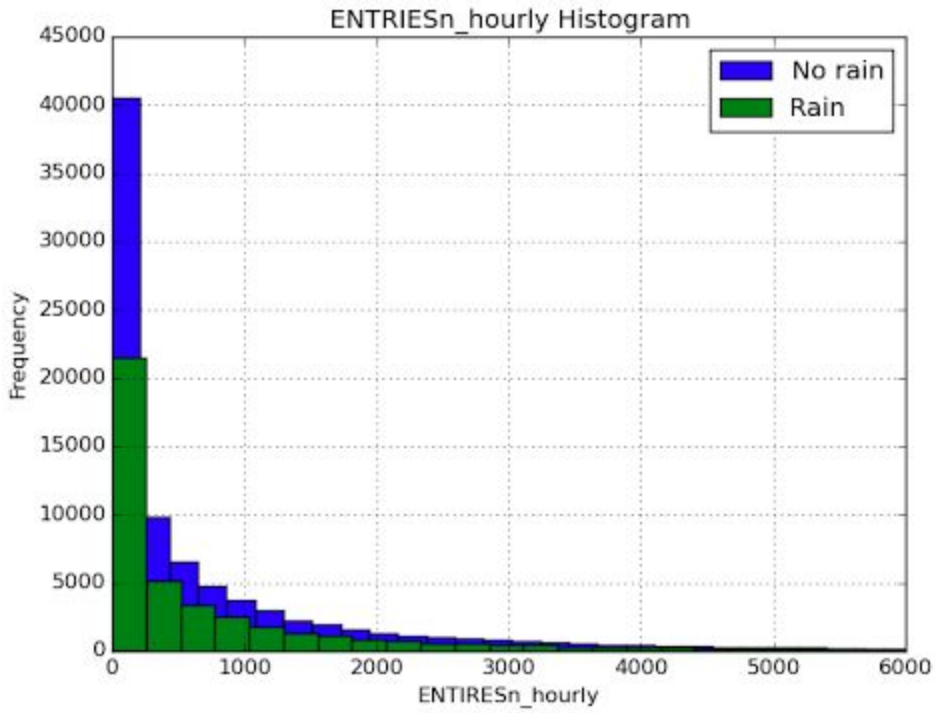

Frequency as a function of residuals.

Because in this use case, we don't have to be meticulous about safety and security, we can rely on this. From the histogram above, the residuals ranged roughly from **+- 0 / 5000.** Which is still a qualtiatively reasonable metric.

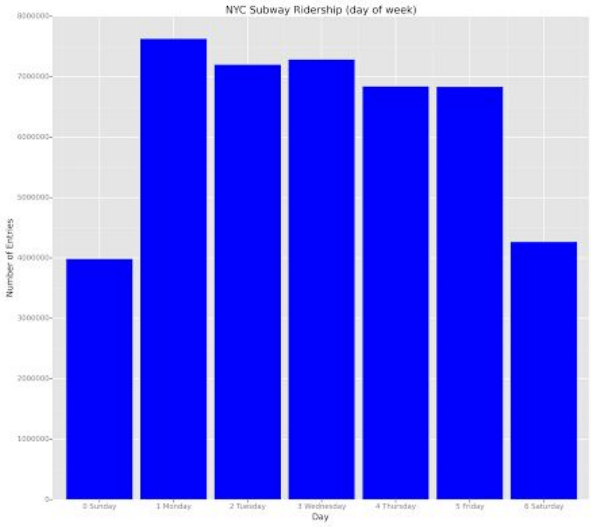
# Section 3. Visualization

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.**

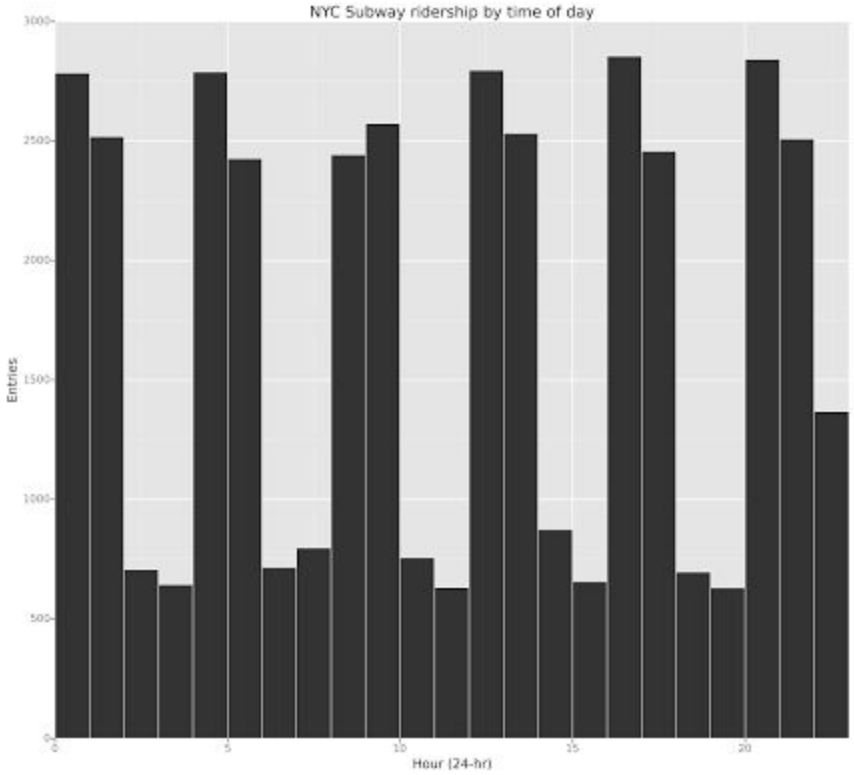
Below is a histogram comparing subway riding frequencies in scenarios with and without rain. Observe that at all times, people will frequent the rain much more when there is NO rain. Note however that these are aggregate values and that there is less frequently rain.




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 day-of-week**
Below is a visualization of the Day of Week. It appears that Monday is the busiest day and that the Weekends (Saturday & Sunday) are the freest.




**Ridership by time-of-day**



Looking at the 24-hr view (12AM-12PM), it's apparent that there's several peaks and troughs throughout the day. It appears as though the peak hours are at Breakfast, Lunch, and Dinner.

# 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), more people ride 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.
Looking at the linear regression, the negative correlation indicates that rain decreases the number of people present on the subway. However, the coefficient for rain in the linear regression model indicates on a negative relation with number of riders. Therefore the results from the linear regression and the statistical test are not in agreement.

# 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. There were also some more regions that were more active than other regions, which is particularly the weakness of examining the problem from an aggregate standpoint, which the Mann-Whitney Model apparently does.

The linear regression model worked for the study, but this model alone won't be as insightful as a learning model. The study could greatly benefit from more data sets, particularly training set, then cross-validation, followed with a test set could have helped improve the data.

Another shortcoming to consider is the assumption of linearity. By definition, linearity in hours implies that that the hour of the day causes ridership to increase as the hour gets later. However, there's a good chance that it is incorrect if you think **ridership oscillates** in the later hours. This would be a case of non-linearity, which is a violation of the model's assumptions.

5.2 (Optional) Do you have any other insight about the dataset that you would like to share with us?
[Will elaborate more later]

In section 5, please expand your shortcoming section, do you think the data is complete? Do you think the duration the data include is sufficient for the analysis you are doing? Do the data include outliers? Can you find shortcomings in the methods you use? Dose linear regression is optimal for such data?