# 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.

- https://docs.python.org/2/library/csv.html → Python Standard Library - 13.1 File Formats - CSV
- http://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.fillna.html → pandas.DataFrame.fillna
- http://stackoverflow.com/questions/10982089/how-to-shift-a-column-in-pandas-dataframe → How to shift a column in a Pandas DataFrame
- http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html → scipy.stats.mannwhitneyu
- http://www.randalolson.com/2012/08/06/statistical-analysis-made-easy-in-python/ → Statistical analysis made easy in Python with SciPy and pandas DataFrames
- http://stackoverflow.com/questions/20701484/why-do-i-get-only-one-parameter-from-a-statsmodels-ols-fit → Why do I get only one parameter from a statsmodels OLS fit
- http://blog.minitab.com/blog/adventures-in-statistics/why-is-there-no-r-squared-for-nonlinear-regression → Why don't we use $R^2$ for nonlinear models?
- http://matplotlib.org/api/pyplot_api.html → All matters of pyplot

## 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?

The statistical test used to analyze the NYC subway data was the Mann Whitney U test. I used a two-tailed p-value as the effect was not assumed to be in a particular direction. The null hypothesis, $H_0$, is that the population's subway turnstile entries per hour when raining is equal to the population's subway turnstile entries per hour when not raining. The p-critical value is 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 Mann Whitney U test is applicable to this dataset because it is a nonparametric test and the dataset does not exhibit characteristics of a normal distribution.

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.

*NOTE: I used the improved data set for these calculations.*

The results of this test are as follows:
Mean Entries per Hour (Rain) ≅ 2028
Mean Entries per Hour (No Rain) ≅ 1846
$p < 0.001$

1.4 What is the significance and interpretation of these results?

The results suggest that the population's subway turnstile entries per hour when raining is different than when not raining. We can reject the null hypothesis that the sample means come from the same population.

Further, the direction of the effect shows entries per hour as greater when raining, suggesting that subway ridership is higher in rainy weather.

# 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:

1. OLS using Statsmodels or Scikit Learn
2. Gradient descent using Scikit Learn
3. Or something different?

Gradient descent using 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 in the model included Rain, Hour, Unit and Date. Unit and Date are dummy variables.

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 used Unit as entries per hour varies significantly across different subway stops. Particular entries at Columbus Circle are more heavily trafficked than others and are more heavily trafficked than other stations, say Bushwick Ave. Plotting mean entries per hour across units showed large differences between units. Including this feature in the model drastically improves the $R^2$ value.

I used Date as ridership anecdotally is higher during the week due to worker transportation than during the weekend. Mean entries per hour is roughly 80% higher during the week than during the weekend. Including this feature in the model increases the $R^2$ value moderately.

I decided to use Hour as subway ridership anecdotally would be higher in the afternoon and evening compared to the morning (very different ridership at 4am vs 4pm). A bar chart of total entries per hour by hour of the day seemed to show a trend supporting this assumption. I added this to the model and saw a modest increase in $R^2$.

Rain is included in the model as the question at hand is whether or not ridership is affected by the presence of rain.

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

The weight for Hour is 118.70. The weight for rain is 123.29.

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

The model's $R^2$ is approximately 0.48.

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?

An $R^2$ of 0.49 suggests the model fits the data moderately well and certainly better than the null hypothesis. However, given the non-linearity of the data as discussed in Section 1.2, $R^2$ is not the best measure to ascertain goodness of fit as it assumes linearity in the model. This may in turn skew our reported $R^2$ higher than the data suggest.

# 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 to 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.

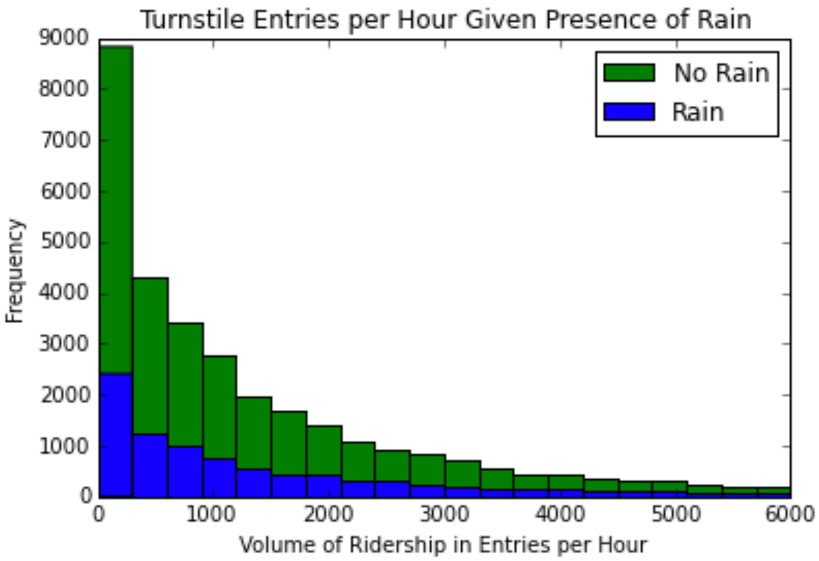

**Figure 1**. The above figure illustrates the frequency of particular volumes of ridership as measured by entries per hour given the presence or absence of rain. While the frequencies of ridership are higher for the no rain condition in each category it is important to note that the frequency of no ridership for the no rain condition is significantly different from that of the rain condition. This suggests that while ridership suffers in the presence of rain, entries per hour are more uniformly distributed. When it is not raining, ridership skews further to the right.

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

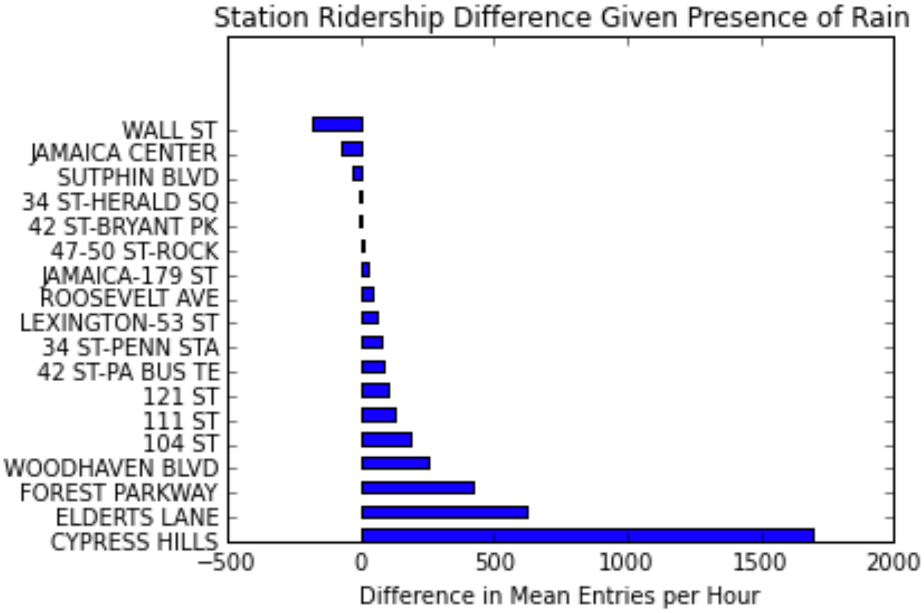

**Figure 2**. Section 1.3 suggests that the presence of rain positively impacts ridership on NYC subways. The figure above illustrates the effect by each subway station by taking the difference between turnstile entries per station when raining and when not raining. A positive number indicates heavier ridership when raining while a negative number indicates higher ridership when not raining. An abbreviated number of stations is shown to illustrate the trend. While most stations show higher ridership when raining, a small number of stations does show higher ridership in the absence of rain. Additionally, the difference is highly variable among stations. It will take further exploration to understand the source of this variance.

# 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?
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.

Based on the results of the Mann Whitney U test, ridership as measured by mean turnstile entries per hour is higher in the presence of rain. Mean entries per hour in the presence of rain (mean ≅ 2028) are significantly higher than mean entries per hour in the absence of rain (mean ≅ 1846; $p < 0.001$). This is evidenced in the above figures.

Figure 1 illustrates the difference in distributions for ridership in the presence and absence of rain. In the absence of rain, the skew is further right showing relatively high frequencies of lower mean entries per hour. In the presence of rain, the skew is less pronounced and frequencies of ridership decay at a slower rate. At higher mean entries per hour, ridership in the rain condition approximates ridership in the no rain condition.

Figure 2 further evidences this trend by showing the effect by subway station. When observing ridership by station, ridership in the presence of rain is generally higher than ridership in the absence of rain.

The gradient descent resulted in a model in which the rain feature had a positive parameter (Beta=123.29) indicating the presence of rain corresponded to an increase in ridership.

Given the trends shown in Figure 1 and Figure 2 as well as the results of the Mann Whitney U test and the gradient descent parameters, the data suggest that ridership increases in the presence of rain.

# 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:

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

There are a handful of potential shortcomings with this particular dataset that beg for further study. First, the dataset only includes data for the month of May 2011. It is unknown the effect of seasonal or annual changes in subway ridership that could mitigate this effect. Second, the rain feature can have varying degrees of intensity.  For instance, a drizzle for 10 minutes may not have much of an impact on ridership but a hurricane for several hours most certainly would have an impact.  In this particular dataset it is only known if there was rain within a particular band of time regardless of the intensity of that rain.

Regarding the analyses used to evaluate the effect of rain on ridership, $R^2$ is not an accurate reflection of goodness of fit.  Given the non-linearity of the data set (see Figure 1), the calculation of $R^2$ is invalid as the $SS_{reg}$ added to the $SS_{err}$ do not equal the $SS_{tot}$.  Instead, the standard error of the regression, S, is more appropriate as it also gives contextual precision with units relevant to the data.

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

For fun, it may be interesting to use some demographics to help predict ridership. Because of the high variability in ridership across stations it is possible that ridership is not affected by rain for those stations in which riders must travel for work or for those areas in which people live close to where they work.