# Analyzing the NYC Subway Dataset

## QUESTIONS

Wesley Strange
AT&T | WS2234@ATT.COM

# Table of Contents

Overview ..... 2
Section 0. References ..... 3
Section 1. Statistical Test ..... 4
1.1 ..... 4
1.2 ..... 4
1.3 ..... 4
1.4 ..... 4
Section 2. Linear Regression ..... 5
2.1 ..... 5
2.2 ..... 5
2.3 ..... 5
2.4 ..... 5
2.5 ..... 5
2.6 ..... 5
Section 3. Visualization ..... 6
3.1 ..... 6
3.2 ..... 7
Section 4. Conclusion ..... 8
4.1 ..... 8
4.2 ..... 8
Section 5. Reflection ..... 8
5.1 ..... 8

## 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.


"Data - What to Do with It?" CSV Reader/Writer Tutorial. Web. 28 May 2015. <https://docs.google.com/document/d/1S4Gk42ZPBKAUZh7IPbqyzq_vk18oCvcniVcA4byBXCE/pub>.

"8.1. Datetime — Basic Date and Time Types." 8.1. Datetime. Web. 28 May 2015. <http://docs.python.org/2/library/datetime.html#datetime.datetime.strptime>.

"Mann–Whitney U Test." Wikipedia. Wikimedia Foundation. Web. 28 May 2015. <http://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test>.

"Scipy.stats.mannwhitneyu." Scipy.stats.mannwhitneyu — SciPy V0.15.1 Reference Guide. Web. 28 May 2015. <http://docs.scipy.org/doc/scipy-0.15.1/reference/generated/scipy.stats.mannwhitneyu.html>.

"Pandas.DataFrame.fillna." Pandas.DataFrame.fillna — Pandas 0.16.1 Documentation. Web. 28 May 2015. <http://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.fillna.html>.

"Pandas.DataFrame.shift." Pandas.DataFrame.shift — Pandas 0.16.1 Documentation. Web. 28 May 2015. <http://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.shift.html>.

"Yhat/ggplot." GitHub. Web. 28 May 2015. <https://github.com/yhat/ggplot/>.

"Plotting." Plotting — Pandas 0.16.1 Documentation. Web. 28 May 2015. <http://pandas.pydata.org/pandas-docs/stable/visualization.html>.

"GraphPad Statistics Guide." GraphPad Statistics Guide. Web. 1 July 2015. <http://graphpad.com/guides/prism/6/statistics/index.htm?one-tail_vs__two-tail_p_values.htm>.

## 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?

- Mann-Whitney U-Test.
- Two-tail P value.
- Null Hypothesis: The “entries with rain” and “entries without rain” datasets come from the same population.
- P-critical value: 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.

- Mann-Whitney U-Test is a non-parametric test that does not assume that the data is drawn from any particular underlying probability distribution.
- The NYC Subway Data was not normally distributed. The Mann-Whitney U-Test is a more efficient test than the t-test on non-normal distributions.

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.

- With_rain_mean: 1105.4463767458733
- Without_rain_mean: 1090.278780151855
- Mann-Whitney U-Test Statistic: 1924409167.0
- One sided p-value: 0.024999912793489721
- Two sided p-value: 0.049999825586979442

1.4 What is the significance and interpretation of these results?

- The two sided p-value (0.049999825586979442) is less than the p-critical value (0.05). Therefore, we can conclude that the Null Hypothesis is false and that the ridership is different when it rains vs. when it does not rain.

## 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?

- Linear Regression OLS Least squares model.

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

- Hour
- Mintempi (minimum temperature)
- Rain
- Meanwindspdi (mean wind speed)
- Dummy variables were used for the feature 'UNIT'.

2.3 Why did you select these features in your model?

- Hour: I chose hour because the plot "Average Ridership by Hour" showed that there was a strong correlation between ridership and time-of-day. Plus it drastically improved my R2 value once I plugged it in.
- Mintempi: I chose minimum temperature because I thought that people would be more likely to use the subway when the temperature is colder.
- Rain: I chose rain because I thought that people might decide to use the subway when it is raining instead of walking. Plus it improved my R2 value once I plugged it in.
- Meanwindspdi: I chose mean wind speed because I thought people would be more likely to use the subway if it is windy outside.

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

- Hour: 65.3855425
- Mintempi (minimum temperature): -10.4647333
- Rain: 55.5085936
- Meanwindspdi (mean wind speed): 25.3476672

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

- Your r^2 value is 0.479988742549

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?

- The R2 value of 0.479 means the linear model accounts for 48% of the variance. This value is higher than the 0.4 value we were trying to achieve by this model. This model is appropriate for the purposes of predicting NYC Subway ridership.

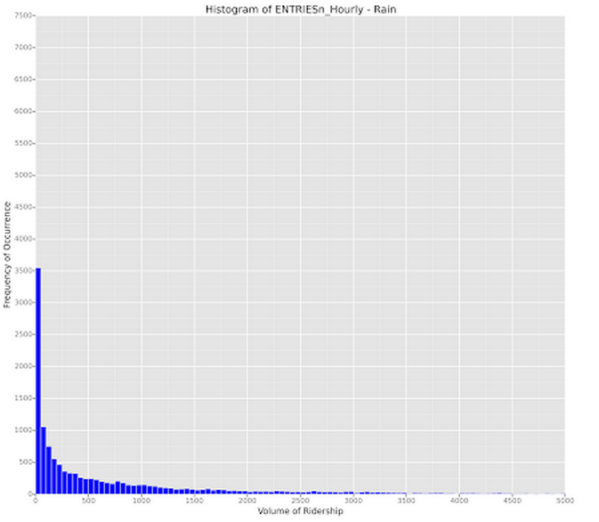
## Section 3. Visualization

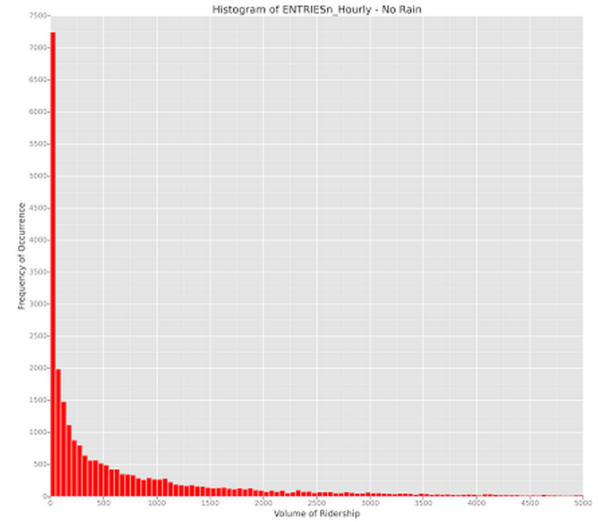
Please include two visualizations that show the relationships between two or more variables in the NYC subway data. Remember to add appropriate titles and axes labels to your plots. Also, please add a short description below each figure commenting on the key insights depicted in the figure.

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.



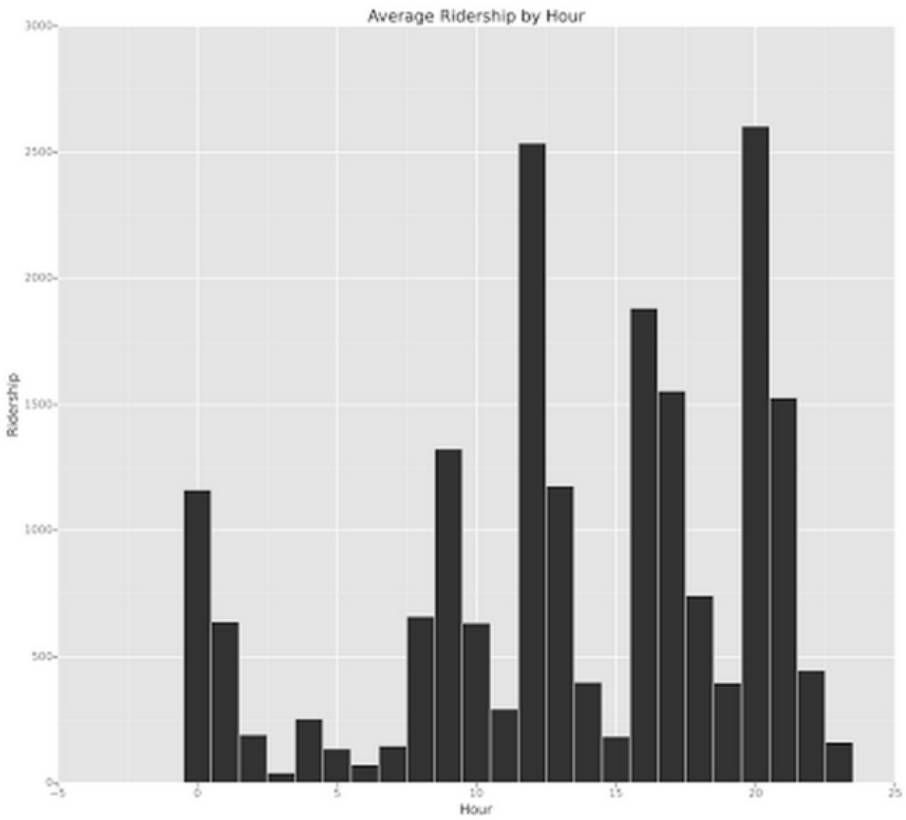




This visualization compares the histograms for non-rainy days and rainy days. The **RED** histogram representing the Volume of Ridership on non-rainy days. The **BLUE** histogram representing the Volume of Ridership on rainy days.

Key Insights

The distributions for both the non-rainy days and rainy days are not normally distributed. There were less rainy days than there were non-rainy days. We cannot draw a conclusion from the graphs on whether or not the ridership is greater when it rains.

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.




This plot illustrates the Average Volume of Ridership by time-of-day (Hour).

Key Insights

There are a few spikes in ridership throughout the day occurring around 9AM, 12PM, 4PM, 8PM, and Midnight, with the spikes at 12PM and 8PM being the highest. The spikes at 9 AM and 4 PM can most likely be attributed to worker's commuting to and from work. The spikes at 12 PM, 8 PM, and Midnight are harder to explain without knowing more about the riders.

Why are more people riding the subway at 12PM and 8PM than during the normal work rush hour? Are people going out to eat during these times? Is the spike at Midnight due to people heading home from the bars?

Overall, it's clear that there's a strong correlation between ridership and time-of-day.

## 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?

- I concluded that more people ride the NYC subway when it is raining. The Mann-Whitney U test confirmed that there was a statistical difference between the rain dataset and no-rain dataset. Then, taking into account that the mean for the rain dataset was 1.3% higher than the mean for the no-rain dataset, it was clear that more people ride the NYC subway when it rains.

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.

- I performed the Mann-Whitney U test on the rain / no-rain datasets. The Mann-Whitney U test returned a p-value of 0.025 (less than the p-critical value of 0.05), which tells me that there is a statistical difference between the rain/no-rain datasets. Taking into account that the mean for the rain dataset (1105) was 1.3% higher than the mean for the no-rain dataset (1090), I can confirm with high confidence that more people ride the NYC subway when it rains.

  The results from the Linear Regression model also support my conclusion that more people ride the subway when it rains. The model produced a positive coefficient for rain (55.5). The positive coefficient for rain tells me that rain will have a positive linear effect on ridership. Therefore, we can predict that the ridership will increase when it is raining.

## 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.

- There are a few shortcomings with the data sets used to complete the analysis. Firstly, the original dataset does not have a formal time period for reporting the riders hourly. The time is reported randomly versus the enhanced data set that has a formal time structure for reporting every 4 hours (00:00:00, 04:00:00, 08:00:00, 12:00:00, 16:00:00, 20:00:00). Secondly, the weekday feature would be a good enhancement to the data set since the week day commute plays a large role in expected ridership. Lastly, the data set only spans one month out of the year (May). I would expect the NYC Subway ridership to fluctuate throughout the year depending on the season. For example, I would expect the ridership to increase during the winter months when there are freezing temperatures and snow is on the ground, but then decrease during the spring/summer months when the weather is more pleasant.

  There could also be a problem with collinearity depending on the features selected when performing the Linear Regression analysis. For example, if you selected both the ‘rain’ and ‘precipi’ features in your model, then the R^2 results could be thrown off since they are highly correlated with each other.