Analyzing the NYC Subway Dataset

Sam Wigley

October 20, 2015

Overview

The purpose of this project is to analyze the features of the NYC Dataset using machine learning algorithms to see if there is a significant difference in subway ridership on rainy and non-rainy days based on the data in the improved NYC Subway Dataset, which combines NYC Subway turnstile data, and data from weather underground from the same time period. The readings are stored in the file in four-hour bins, with totals on entry and exit counts per bin, per Unit ID, which is effectively the same as a station ID.

The dataset includes several other features related to the station and weather conditions including fog, and wind conditions, and latitude and longitude of the station and weather reading. Another goal is to explore the data to find and show other interesting features that show a correlation with subway ridership, and report some other interesting findings.

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.

- http://blog.yhathq.com/posts/aggregating-and-plotting-time-series-in-python.html
- https://georgemdallas.wordpress.com/2013/10/30/principal-component-analysis-4-dummies-eigenvectors-eigenvalues-and-dimension-reduction/
- https://en.wikipedia.org/wiki/Welch%27s_t_test
- http://scikit-learn.org/stable/modules/sgd.html
- http://docs.ggplot2.org/current/scale_continuous.html
- http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html
- http://stats.stackexchange.com/questions/36064/calculating-r-squared-coefficient-of-determination-with-centered-vs-un-center
- http://stats.stackexchange.com/questions/110380/smoother-lines-for-ggplot2
- http://stattrek.com/statistics/dictionary.aspx?definition=coefficient_of_determination
- http://www.psychstat.missouristate.edu/multibook/mlt08m.html
- http://matplotlib.org/users/pyplot_tutorial.html

- http://blog.yhathq.com/posts/ggplot-for-python.html
- http://matplotlib.org/examples/index.html
- http://chrisalbon.com/python/pandas_apply_operations_to_groups.html
- https://gehrcke.de/2013/07/data-analysis-with-pandas-enjoy-the-awesome/
- http://matplotlib.org/examples/pylab_examples/date_demo_convert.html
- http://matplotlib.org/1.3.0/examples/pylab_examples/legend_demo.html
- http://stackoverflow.com/questions/332289/how-do-you-change-the-size-of-figures-drawn-with-matplotlib
- http://matplotlib.org/users/legend_guide.html#using-proxy-artist
- https://github.com/adam-p/markdown-here/wiki/Markdown-Cheatsheet
- http://strftime.org/

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?

Because of the results shown in the histogram of entry counts over time, we can see that the sets are not normally distributed. In fact it's positively skewed, meaning the most of the samples occur in the lower half of the histogram.

Because the Welch's T-test is not appropriate for testing non-normally distributed data, we performed the Mann Whitney U-Test, which can compare two sets of data regardless of the distribution skew.

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

For this test we are using a two-tailed P value because ridership could be either greater or less on rainy days than clear days. We don't have any determination that ridership will be greater or less on rainy days, so we're performing a two-tailed test.

- What is the null hypothesis?

In this test we are considering the null hypothesis is that ridership on the NYC subway in terms of entries per hour is not significantly different on days where there is rainy weather, when compared with days where there is clear weather.

- What is your p-critical value?

For a two-tailed test of 95% significance we use a p-critical value of 0.05. This means if the value of the p-value returned by the Mann Whitney U-test is less than 0.05, then there is a significant difference in ridership between the rainy and non-rainy sets.

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.

Because the Welch's T-test is not appropriate for testing non-normally distributed data, we performed the Mann Whitney U-Test, which can compare two sets of data regardless of the distribution skew.

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:	2028.20
Without Rain Mean:	1845.54
p-value:	2.74x 10^-06 = 0.00000274

1.4 What is the significance and interpretation of these results?

In the U-test, we got a U value of 153635120.5 and a p value of 2.74 x 10^-6, or approximately 0.00003, well below the 0.05 p-critical. This means that the samples are discernibly different with 95% significance.

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?

I made my first prediction using Statsmodels OLS, but also used Scikit Gradient Descent to make an additional prediction for comparison.

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

I used unit ID, rain, time of day, and day of week as features. Unit ID was added as a dummy variable because it was a categorical, rather than numeric feature. Day of week was encoded to an ordinal integer value.

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 computed day of week because I could see that certain days were obviously busier than others and thought there could be something going on there. Once grouping by day of week, it's apparent this is a good predictor of subway ridership, probably because the majority of the people share a similar work schedule.

I actually chose these purely by trial and error. Unit ID was suggested in the documentation, and it proved to be a high predictor of ridership, probably because certain stations constantly get a higher volume of traffic. I think there is probably a way to plot all of these features against ENTRIESn_hourly and get an idea if there seems to be a visible trend between the two variables, but I actually plugged in values trying to get a better R^2, without getting a lot of luck with most of the other weather indicators. Rain helped, but wasn't a good enough predictor on it's own. Other weather feathers seemed to make little difference.

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

From the linear regression function I got the following values for params:

- rain = 117.225179
- day_of_week = -141.600626
- hour = 123.307487

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

For my linear regression model the value of R^2 is: 0.469

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?

With an R^2 value of 0.47, I would say this model is a fairly good fit, it means almost half of the cause for change in subway ridership is most likely attributed to the features that I selected in the linear regression model, which were only unit id, day of week and hour of day. Just three features can be attributed to almost 50% of the strength of the correlation, so I think this is a relatively good model.

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.

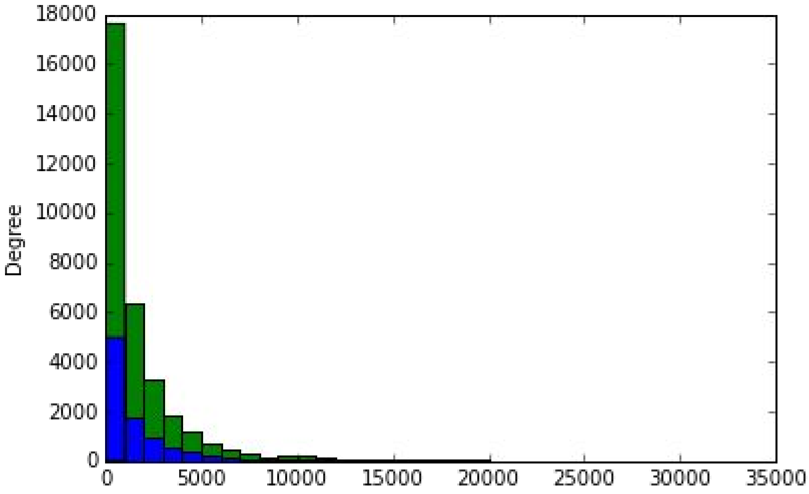


Fig. 1: A histogram showing the frequency of occurrence of levels of ridership, represented by the counting value of ENTRIESn_hourly on rainy and non-rainy days and putting those values into bins of size 1000 riders. ENTRIESn_hourly actually represents a total count of riders at a particular station (UNIT) in a 4 hour period.

> At first, looking at this histogram it appears that somehow ridership is higher on non-rainy days, because of the size of the green bars in the low side of the histogram. What this is actually showing, though is that the greatest occurrence of low ridership is on non-rainy days. The blue bars actually extend higher into the histogram levels on the x-axis, making rainy days have greater occurrences of ridership in the 1500 - 20000 riders per station in a 4 hour period, where non-rainy days have fewer periods in this range.

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

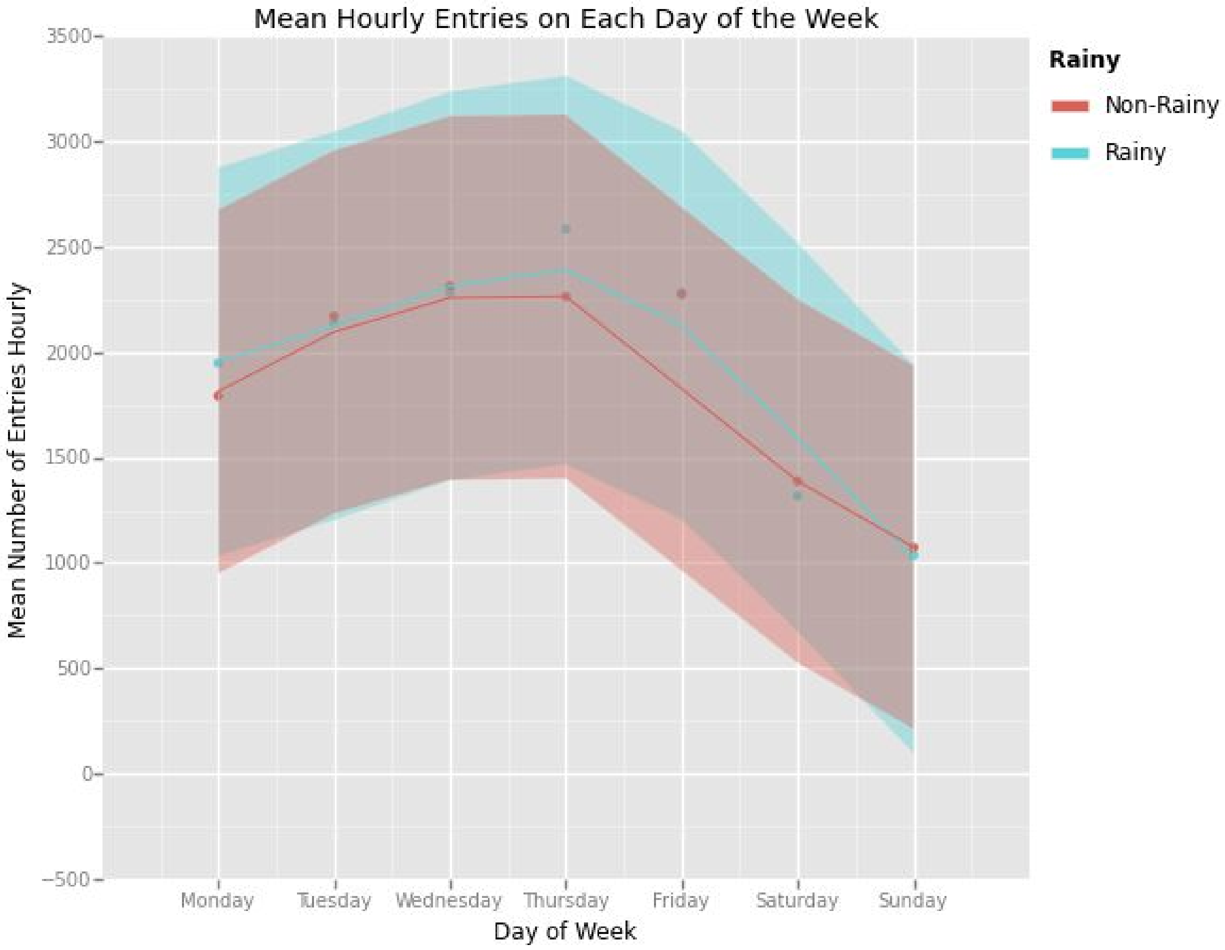


Fig. 2: A chart showing the mean ridership by day of the week.

In this chart, I converted the timestamp into a weekday, and averaged the total ridership at all stations by day of week to see on average how ridership varied by day. The shaded areas and trend lines show the mean and 95% CI of the ENTRIESn_hourly ridership value on each day of the week.

It seems ridership peaks mid-week, and sharply tapers off on weekends. It also seems that toward the end of the week riders are more likely to choose the subway on rainy days, and earlier in the week,the ridership seems more unchanged by rainy weather. Maybe on days when it's harder to get off to work, riders stick by their most ingrained habits? Interestingly on Sundays, when it's completely optional for a lot of people to go out, ridership is slightly lower on rainy days, which probably means riders are more likely to avoid going out all together if they don’t have to.

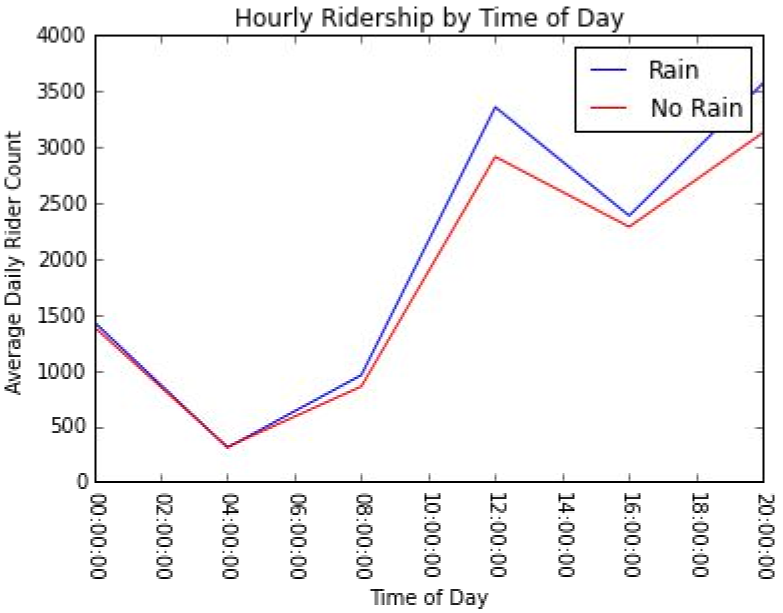


Fig 3: A chart showing mean ridership at all stations by time of day.

In this chart I’m showing mean ridership at all stations by time of day. Notice that the line pivots at 4-hour marks on the x-axis. This is because we only get data in 4 hour blocks. Another interesting point is that earlier in the morning ridership tends to not vary as much between rainy and non-rainy days, but in the afternoon, riders more often prefer the subway on rainy days.

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?

From my analysis here I would say that on days that it is raining, significantly more riders do ride the subway. However most days are not rainy days.

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.

It appears that on rainy days the mean ridership is usually not much greater than non-rainy days. Referencing the chart in Fig 2, the difference seems greatest after mid-week when ridership is at it's peak, and least on weekends when ridership tapers to its lowest levels.

Referencing the chart Fig 3, which shows mean ridership throughout the day, it also seems that mean ridership is usually close to the same between rainy and non-rainy days in the morning when people would be commuting to work, but this starts to diverge to favor more ridership in the afternoons on days when it is rainy, and less in the afternoons on days when it's not rainy. I imagine many people decide to walk home on clear days, but opt for the subway when it’s rainy.

The reason I determined the difference is significant is because our calculated p-value with the Mann-Whitney U-Test was well below the p-critical value of 0.05 with a p-value of 0.00003. Further analysis also shows that on a given weekday, or a given time of day, mean ridership will usually be slightly higher on a rainy day with a mean ridership of 2028, than a non-rainy day, which has a mean ridership of 1846.

Linear regression and gradient descent confirmed that rain, along with a few other predictors made a reasonably good model to predict subway ridership. If rain was removed as a feature in either model, it consistently lowered the R^2 score for each of the predictions. Hour of day, and day of week also proved to be strong predictors, however unit ID appears to have the strongest influence in making predictions with these models.

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.

This dataset lacked resolution in the time of day, so it was hard to see distinct patterns throughout the day. Instead we get four-hour snapshots throughout the day. Because there are only a handfull of samples throughout the day, it makes hourly projections difficult and unclear.

Our technique here could be a little bit more accurate with higher resolution data. The issue is we are marking some samples as a rainy day, when in actuality it only rained for part of the day. On these days we're only getting a partial day's worth of samples, and calling it a rainy day, and likewise we're getting a partial day's non-rainy samples for the same day and also calling it a non-rainy day. I think if we had higher frequency entries, we could better judge rates of ridership by hour and make better short-term and partial-day projections. If we had more data it would be best to either discard days that were partially rainy, or develop a function to return a factor for a partial day's results.

Another issue was that there was not much more than one month's worth of data, so it makes projections based on seasonal trends impossible. In fact it's hard to say with much confidence the difference in rainy versus non-rainy ridership because there just aren't many days in our data, which are labeled rainy days.

Linear regression and gradient descent produced similar results with similar input features. It seemed like one or the other might perform slightly better with a given set of input features. Given this dataset, I think it would be difficult to say that one algorithm consistently performed better than the other. I think it would depend on the training data and the features selected. With the current data and features it appears the Linear Regression method gets slightly better results.

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

I expected a higher predictive value in the latitude and longitude features since UNIT location seems to be a large component in subway ridership, but they didn't seem to improve the result much and, in fact, they lowered the R^2 score when added. By themselves they didn't seem to be great features to use for gradient descent or linear regression.

I created some charts that show ridership relative to latitude and longitude. As it turns out the top sation by mean ridership on rainy and non-rainy days is located at 40.768, -73,982, which when checked on the map is right the corner Central Park and Broadway.

I think more data would definately be helpful, both a higher level of detail time and weather-wise, and also a longer time-period of sample data. This does highlight, though just how much data is needed to get good predictive results and to find interesting combinations of features to make better predictions or assertions about the data. I think just to get started I would want hourly samples from all stations for a four-month time-period. I think this is a factor of 16 times more rows of data.

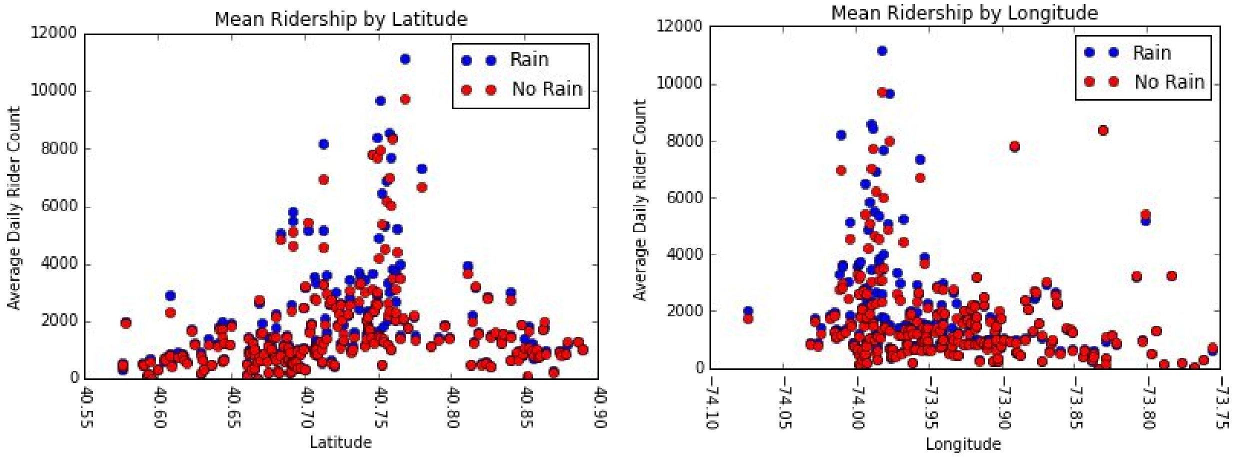
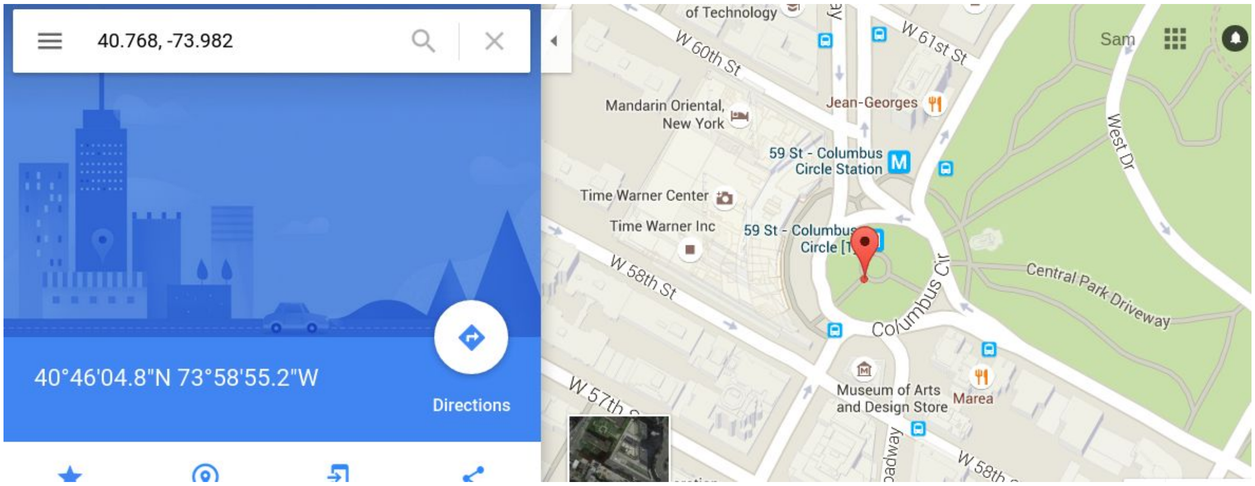


Fig4: Mean ridership by Longitude and Latitude

	latitude	longitude	UNIT	rain	ENTRIESn_hourly
381	40.768110	-73.981891	R084	1	11144.761905
315	40.749533	-73.987899	R022	1	10064.952381
329	40.752247	-73.993456	R012	1	9977.166667
380	40.768110	-73.981891	R084	0	9726.347222
327	40.751849	-73.976945	R046	1	9674.761905

Fig 5: Top stations sorted by mean ENTRIESn_hourly in descending order.



40.768, -73.982

40°46'04.8"N 73°58'55.2"W

Fig 6: Showing the location of the station with the top mean ridership on the map.

This seems reasonable that the busiest station on rainy and non-rainy days would be near a significant destination.