Analyzing the NYC Subway Dataset

By Frank Lettiere

Section 0 – References

The only references I used were links that were provided either in the class videos or problem set sections in the Intro to Data Science coursework. The following are the specific links I used to complete my work.

http://docs.scipy.org/doc/numpy/reference/generated/numpy.dot.html

https://dev.mysql.com/doc/refman/5.1/en/counting-rows.html

http://goo.gl/HBbvyy

http://docs.python.org/2/library/datetime.html#datetime.datetime.strptime

http://pandas.pydata.org/pandas-docs/stable/visualization.html#histograms

http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html

http://docs.scipy.org/doc/numpy/reference/generated/numpy.mean.html

http://www.itl.nist.gov/div898/handbook/pri/section2/pri24.htm

http://docs.scipy.org/doc/numpy/reference/generated/numpy.mean.html

http://docs.scipy.org/doc/numpy/reference/generated/numpy.sum.html

https://pypi.python.org/pypi/ggplot/

https://docs.python.org/2/library/logging.html

Section 1 – Statistical Tests

After plotting the data and realizing that the distributions did not follow normal distribution for either the ridership during rainy or nonrainy days, I used the Mann-Whitney U test to analyze the data sets. This test does not assume that the data is drawn from any particular distribution.

The null hypothesis was that the two different data sets came from the same population. I used a two-tailed p-test and a p-critical value of .05.

The Mann-Whitney U test was applicable here because although the distributions were not normal, the observations were independent, that is the rain data and the nonrain data are independent of each other, and the hourly entries is ordinal in nature. With those two assumptions and knowing that we wanted to test whether or not the two different data sets came from the same distribution, the Mann-Whitney U test was the logical choice.

The mean ridership for rainy and nonrainy days were 1105.45 and 1090.28 hourly entries respectively, rounded to the nearest hundredth place. The U value for the test was 1924409167.00 and the p-value was exactly 0.0500, rounded to the nearest hundredth and ten-thousandth places respectively. Without rounding the p-value was just under the p-critical value of .05.

Given these values we can, strictly speaking, reject the null hypothesis and conclude that the two data sets do not come from the same distribution, with 95% confidence. However it is interesting to note that the p-value is right at the critical level meaning if we did choose a p-critical value any small than .05 the test would not have been able to reject the null hypothesis.

Section 2 – Linear Regression

I used the Gradient Descent approach to compute the theta coefficients and produce a prediction for ENTRIESn_hourly in my regression model. The features that I choose to use were minimum temperature ('mintempi'), whether or not it rained ('rain'), the hour of the day ('Hour'), and the mean wind speed ('meanwindspdi'). I also used the dummy variable 'UNIT' and included it in my features dataframe as well.

I ended up using this specific set of features both based on my intuitional thoughts on what factors would best predict ridership on a subway and then based on trial and error efforts with different combinations of features in an attempt to maximize my R^2 value. I wanted to stick to the features I thought were the most important and keep the list as concise as possible. Although we are given a lot of data here that doesn't necessarily mean that all of it is useful in predicting the ridership on the subway. My list of features is my best attempt to optimize the important data in the prediction and leave out data that would only make the calculation more complicated without garnering better results.

My coefficients for my nondummy features are as follows…

'mintempi' - -5.8672e+01

'rain' – 1.5840e+01

'Hour' – 4.6774e+02

'meanwindspdi' – 5.1745e+01

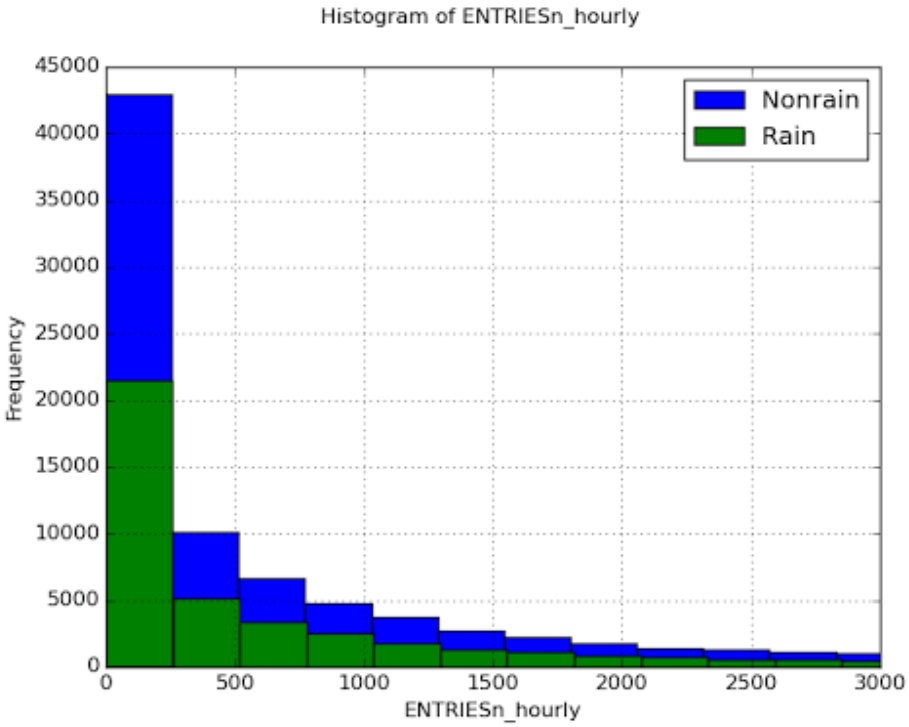
My R^2 (coefficient of determination) value is .4647

This R^2 value means that 46.47% of the variation in ENTRIESn_hourly is explained by my linear model, made up of the list of features and dummy variables discussed above.

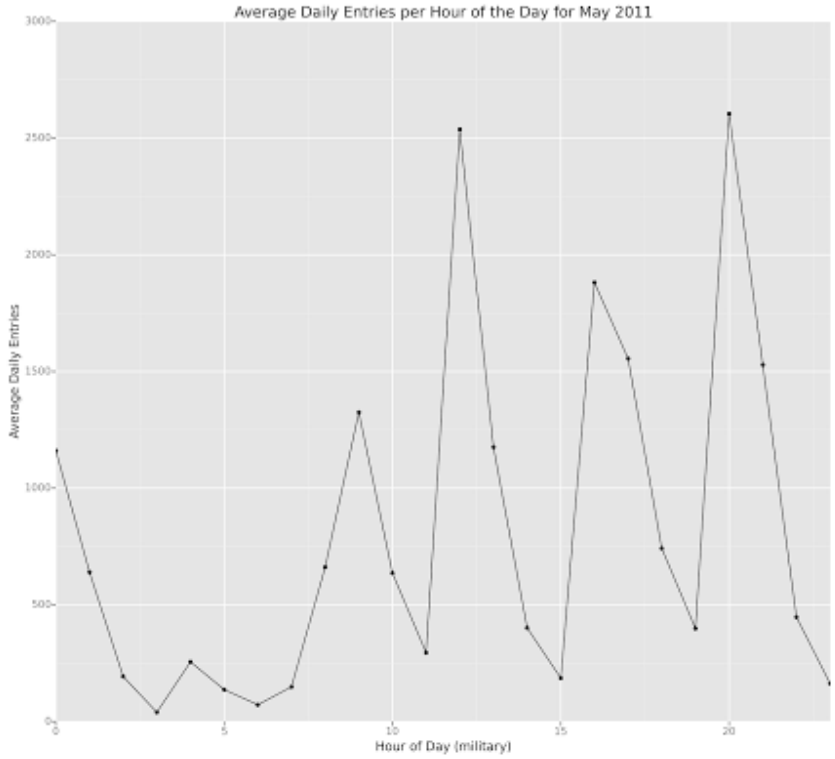
I believe this linear regression model is appropriate to predict the data set since we are using the individual turnstile units to predict the total number of riders. The sum of these turnstiles would be the most useful data in predicting the overall ridership.

Section 3 - Visualization

The first visualization here compares the frequency of hourly entries on rainy days vs nonrainy days. The main takeaway from this graph was that is it clear that neither of the distributions are normal, in fact they are both positively skewed. This is what lead to the decision to use the Mann-Whitney U test in our Section 1 analysis, rather than Welch's T test.



The second visualization shows the average daily entries per hour of the day for the entire data set, that is all of May 2011. The x-axis is the hour of the day (military time) and the y-axis is the average daily entries for the month. What is interesting to note here is the waves of high entries and then low entries that persist throughout the day. From the first peak at 9 we see a consistent pattern of 2-3 hour decline (10-11, 13-15, 17-19, 21-23) in entries before a steep increase (12, 16, 20, and 0). Also, as probably expected, the time of day with the consistently lowest number of entries is 2-7.



Section 4 - Conclusion

In conclusion, the results here lead me to believe that people ride the subway in NYC more when it is raining than when it is not raining. This is based off the lower average hourly entries on nonrainy vs rainy days, about 1090 vs 1105. Although these means are not quite as far apart as I originally thought they would be, looking at the first visualization to decide to conduct the Mann-Whitney U test and creating a linear regression model further backed up my conclusion.

The Mann-Whitney U test was able to reject the null hypothesis, just barely, that the two data sets came from the same population. Therefore, we can reasonably conclude that the data set from rainy days is showing more frequency in ridership overall than the data set from non rainy days. Furthermore, when we used linear regression and gradient descent to try to predict ridership going forward, we used rain as one of our features to help in prediction. Since the coefficient for rain in our regression equation was positive, 1.5840e+01 to be exact, that is telling us that the rain variable has a positive effect on the number of hourly riders on the subway.

For all these reasons I can confidently conclude that people ride the subway more when it is raining outside in New York City.

Section 5 – Reflection

There are some potential shortcomings in the methods used in this analysis of the NYC subway data. First of all, with the data set itself, we are only using data from a one month long period of one year. I would argue that this data set is a little too specific if we want to make a general statement about NYC subway ridership. I think we would need to look at samples of data from different months and even different years to get a little better look at the overall data.

One other problem that I see is in using a linear regression model to make our predictions. While I think that a linear model would work fairly well for most normal predictions, I do think it would not give an accurate prediction for extreme situations on both sides of the curve. On the low side, using linear regression employs a y-intercept that would likely be too high to estimate ridership on a day when any sort of emergency is in place. For example, if there is a natural disaster or terrorist threat a lot of people will likely not travel and subway stations may be closed off. In the other extreme case, the growth of ridership cannot possibly stay linear forever. Even if the subway was free and everyone wanted to ride it, there is only a finite number of people in the city and only a finite amount of space on the trains. Thus as ridership increases it will eventually only do so at a decreasing rate. Thus a linear model will eventually fail.

With that being said I still believe that the analysis done here is valid and in most cases would be useful in predicting future ridership or making claims about the tendencies of people to ride the subway depending on the weather conditions.