Analyzing the NYC Subway Dataset

Harrison Hocker

References:

Mann-Whitney U Test: http://www.statisticssolutions.com/mann-whitney-u-test/ and http://docs.scipy.org/doc/scipy-0.15.1/reference/generated/scipy.stats.mannwhitneyu.html

Linear Regression: http://en.wikipedia.org/wiki/Gradient_descent, http://mathworld.wolfram.com/MethodofSteepestDescent.html, and http://scikit-learn.org/stable/modules/sgd.html

Visualization: http://stackoverflow.com/questions/22599521/how-do-i-create-a-bar-chart-in-python-ggplot, http://stackoverflow.com/questions/25347621/geom-hist-in-python-ggplot, http://ggplot.yhathq.com/docs/geom_histogram.html

Statistical Test:

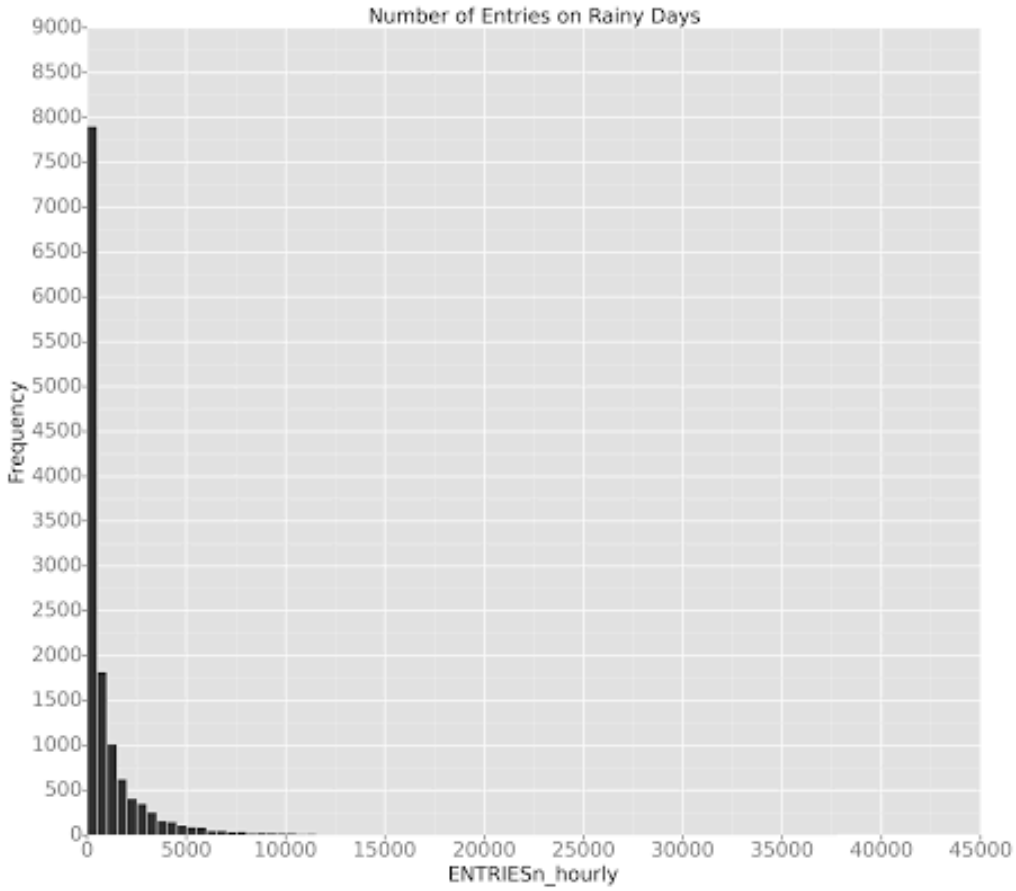
1. I used a Mann-Whitney U test to analyze the NYC subway dataset. The problem was to compare the number of days that occurred with and without rain. The null hypothesis is that there is no difference in how people ride the subway whether it is raining or not. I used a 2-tailed test because I no reason to believe that rain should increase or decrease ridership. The p-critical value is the standard 0.05.
2. Because we can't make any real assumptions about the distribution of rainy days in New York we should use a non-parametric test. Non-parametric tests are in general more applicable because they make no real assumptions about how the data is distributed. To contrast, a t-test assumes the data follows a Gaussian distribution but the same can't be said about weather. The SciPy Mann-Whitney U implementation by default returns a 1-tailed P-value which gives more power to detect 1 of 2 possible differences: more people ride the subway when it rains or less people ride the subway when it rains. So using a 1-tailed test will give us more power to determine if the null hypothesis (the number of days from each distribution is the same) holds. The resulting P-value was 0.025, which is less than the standard critical value of 0.05 so we assume the number of riders on rainy days is more than than the number of riders on sunny days.
3. The p-value is 0.025. The average number of hourly entries into the subway when it rains is 1105.4 and 1090.3 when it's not raining.
4. The conclusion is that more people ride the subway when it rains which is statistically significant with a p-value less than or equal to the critical value.

Linear Regression:

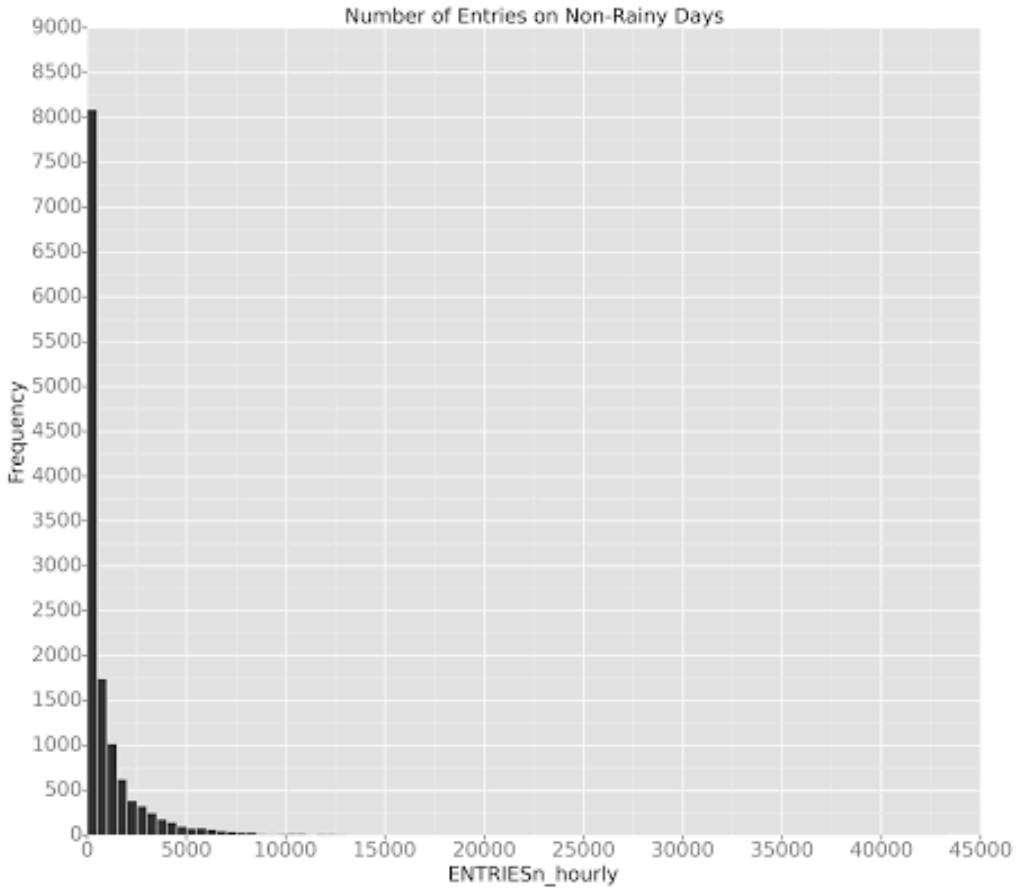
1. I used gradient descent to perform the linear regression.
2. The features used in the model were rain, precipitation, hours, and initial mean temperature. I also used a dummy variable derived from Unit column.
3. I selected these features because rain and precipitation should encourage more riders to take the subway. The time of day should also affect how many riders are in the subway. For example, more people will ride during commutes to and from work than will be riding at 3am. Lastly, the temperature should also affect how many people will ride the subway. For example, if the weather is too hot or too cold this should discourage people from taking the subway.
4. The weights for the rain, precipitation, hours and mean temperature were 2.9, 14.7, 467.7 and -62.2 respectively.
5. The R squared value is 0.46.
6. The R squared value indicates that the variables included in the model do indeed predict how people ride the subway when it's a rainy or non-rainy day. It does not account for all of the behavior but the turnstile data can provide some insight into the decision making process for whether someone will ride the subway or not. There are multiple factors that come into play when determining if someone will take the subway. For example, most people will not miss work if the weather is bad but they might consider missing a movie or some other recreational activities, hence some people might not even go outside. Given the specific R squared value, I would say that weather does have a significant impact on subway usage.

Visualization:

1. Histograms of the frequency of riders on Rainy and Non-Rainy Days.



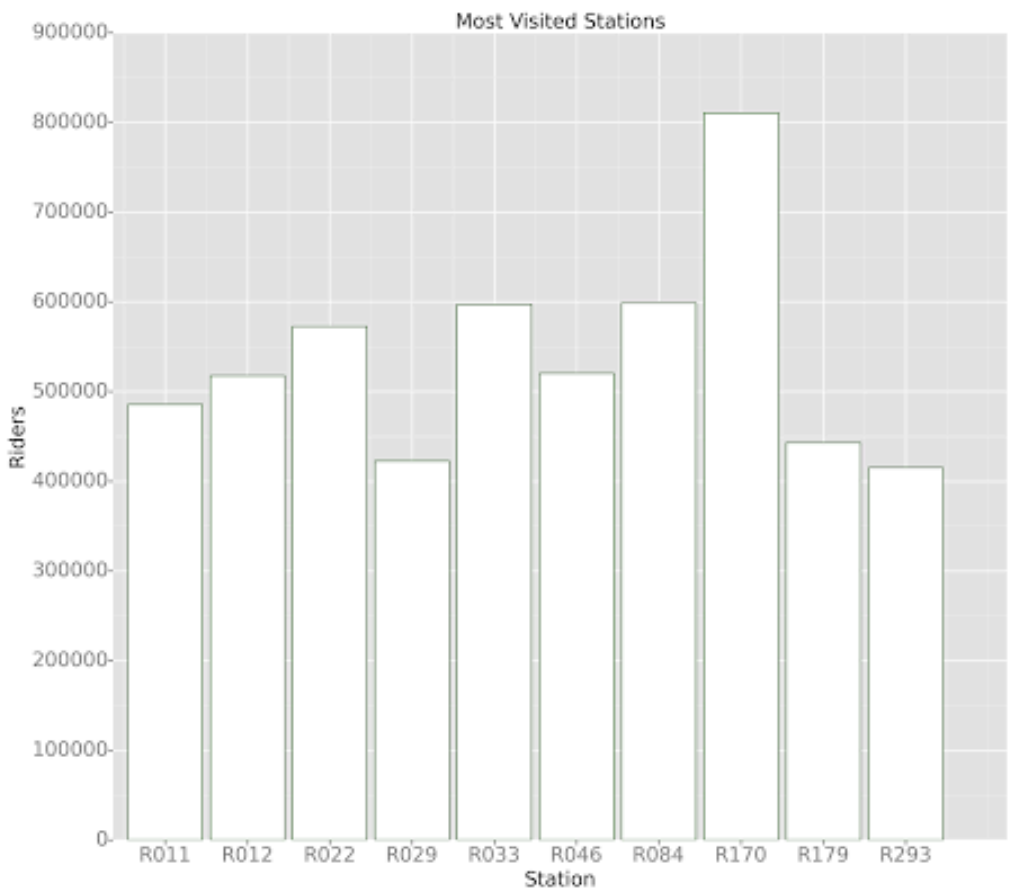
This figure shows that there are just over 8,000 occurrences where there are no entries into the subway on a rainy day.



This figure shows that there are just under 8,000 occurrences where there are no entries into the subway on a non-rainy day.

When compared to the Non-Rainy day plot, the Rainy day plot shows that there are more entries into the subway because the number of times where there were no subway entries was less frequent on rainy days than non-rainy days. Combined with the fact that the average number of hourly entries (1105) on rainy days is higher than the average number of hourly entries (1090) on non-rainy days I concluded that more people ride the subway when the weather is bad.

2. Visualization of the top-10 most used stations.



This plot identified the top-10 most frequently visited subway stations in New York. Thus if we wanted to collect additional information about how rain affects subway ridership, these stations would be the best locations to survey.

Conclusion:

1. From the analysis, I have concluded that more people ride the NYC subway when it is raining that when it is not raining. The visualizations comparing the number of entries on rainy and non-rainy days clearly show a trend where there are fewer occurrences where no people enter the subway i.e. the leftmost bar in the rainy day plot is shorter than the leftmost bar in the non-rainy day plot.
2. The plots suggest that a trend does exist in the data; however, they are not conclusive. To substantiate the above claim, I performed a 1-tailed Mann-Whitney U test to see if there was any difference in how people ride the subway on rainy versus non-rainy days. Using a 1-tailed test yielded a p-value of 0.025 indicating that there is a difference in how people ride the

train. Even if we use the 2-tailed p-value we still arrive at a value of 0.05, which is equal to the standard critical value. The regression analysis showed a correlation coefficient of 0.46, which suggests that there is a moderate relationship in the data. Taken together, the analysis shows that more people ride the subway when it rains.

Reflection:

1. One of major shortcomings of the dataset is that is does not address the possibility that there are just less people willing to leave their homes when it is raining. If less people leave their homes when it the weather is bad then it should come as a surprise that there are more people on the subway. Also, the dataset clumps all rain conditions together and does not separate out light versus heavy rain. One of the shortcomings of the regression analysis used is that it assumes the relationship is linear. During the analysis no transforms were performed on the input data to explore any non-linear relationships in the data.