# Analyzing the NYC Subway Dataset

## Section 0. References

- http://docs.ggplot2.org/current/
  The explanations and examples they offer helped me gain a better understanding of ggplot.

- http://spin.atomicobject.com/2014/06/24/gradient-descent-linear-regression/
  Clear explanation of gradient descent

## Section 1. Statistical Test

*1.1* I used the Mann-Whitney U-Test to analyze the NYC subway data. I went with a two-tail test because we are trying to see if there is higher or lower ridership on rainy days. The null hypothesis ($h_0$) is P(x > y) = 0.5. I chose 0.05 as the p-critical value (α).

*1.2* The Mann-Whitney U Test is applicable because the dataset comprises of non-normal and ordinal data.

*1.3*

With rain mean: 2028.196035472
U-value: 153635120.5

Without rain mean: 1845.539438664
p-value: 0.00000274106957

*1.4* Such a small p-value means that we can consider the difference between ridership on rainy and non-rainy days as statistically significant.

## Section 2. Linear Regression

*2.1* I used ordinary least squares method using Statsmodels to compute theta and produce the predictions.

*2.2* The features I used were rain, precipitation, hour, mean temperature, wind speed, day of the week, and fog. Yes I did add the UNIT variable as a dummy variable.

*2.3* I used the following because I thought that:

- Rain: When it is raining people are more likely to take the subway to avoid getting wet and cold.
- Precipitation: More people are likely to take the subway if itis pouring versus drizzling, so the level of rain is important.
- Hour: Work hours correlate with high ridership.
- Mean temperature: Colder days lead to people wanting to stay in the sheltered subway as opposed to the exposed streets.
- Wind speed: A high wind chill is quite uncomfortable and so people would use the subway to avoid it
- Day of the week: The weekdays probably have higher ridership because people are working.
- Fog: When visibility is low people trust the subway more than taking a taxi or driving.

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

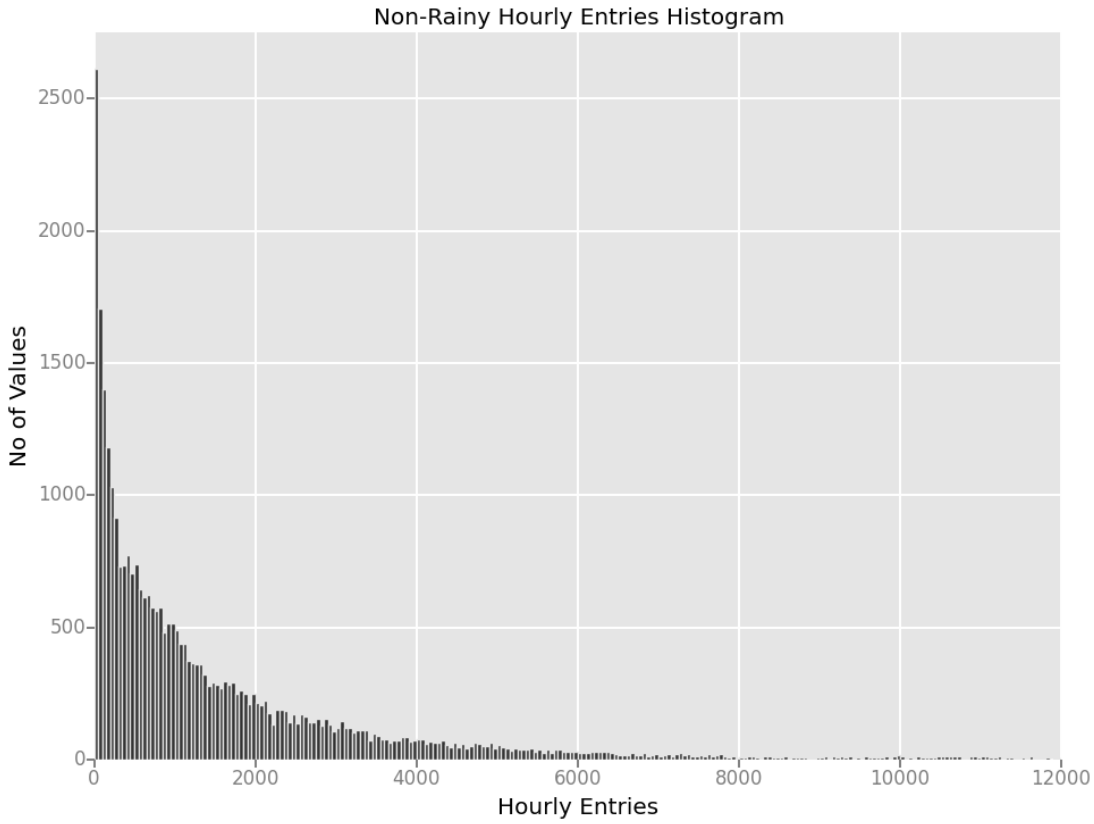
| | | | | | |
|---|---|---|---|---|---|
| rain | 133.038002 | precipi | -533.360637 | hour | 128.140928 |
| meantempi | -14.309909 | wspdi | -33.912844 | day_week | -153.282984 |
| fog | -645.852939 | | | | |

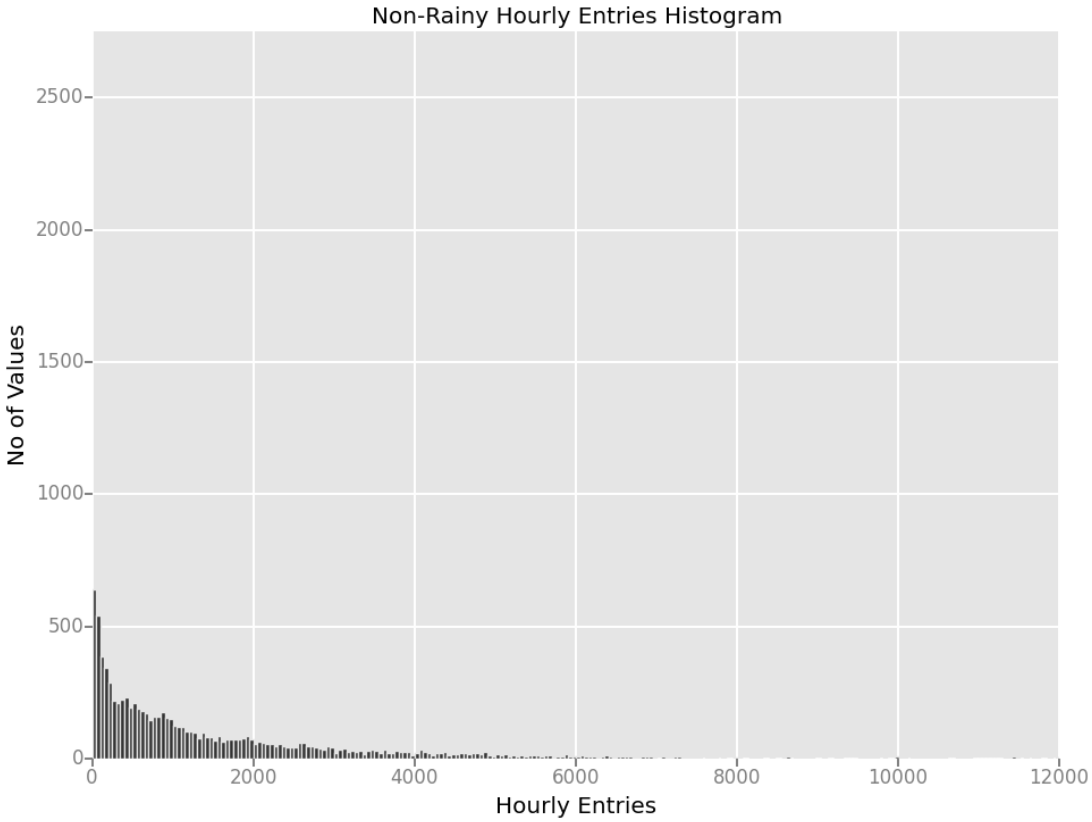
*2.5* $R^2$ = 0.471616197177

*2.6* I think with a $R^2$ of 0.4716 the linear model is for a rough estimate to predict ridership but not one that too much trust should be put in. This linear model only accounts for about 47% of the variance and only 27% of the standard deviation. The linear model does not take into effect seasonality and random outliers.

## Section 3. Visualization

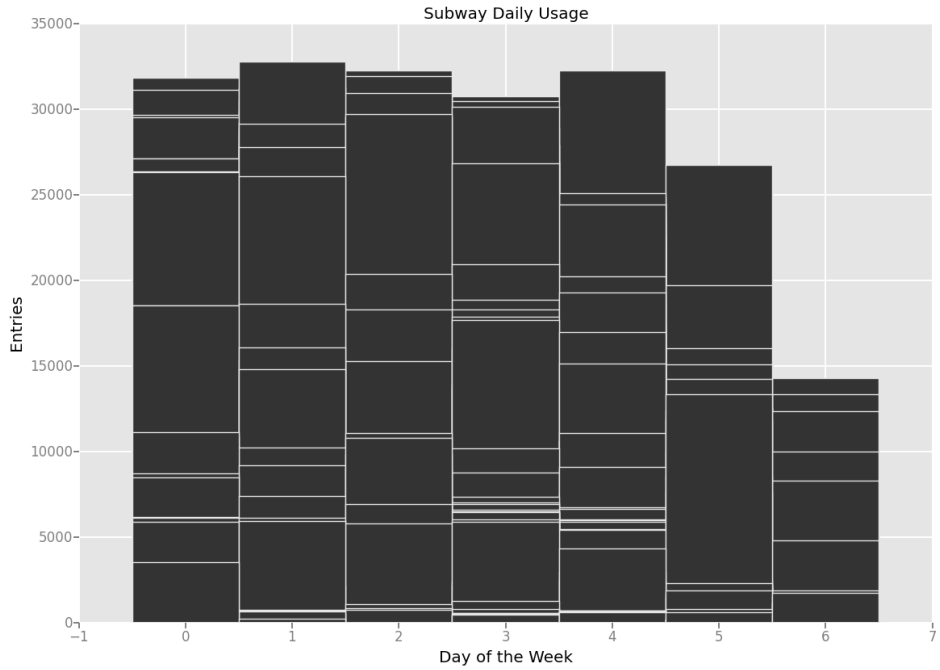
*3.1 two histograms: ENTRIESn_hourly for rainy days and ENTRIESn_hourly for non-rainy days.*






Both histograms are skewed right and appear to be the same, but the difference in mean noted earlier can be seen as a consequence of the smaller difference in range that the rainy day hourly entry values have.

*3.2 Ridership by day-of-week*




I was surprised to see such higher ridership on the weekdays versus the weekends. I wrongly assumed that with more time on their hands people would use the subway more to travel around. People might use the subway on work days and leisurely stroll the streets or use other forms of transportation if they are traveling out of the city.

## Section 4. Conclusion

*4.1* I would say yes and no. More people do ride the NYC subway when it is raining, but it greatly depends on the degree of rain. The rain coefficient is a positive one that leads to a rise in the expected ridership of about 133 extra riders. However the precipitation coefficient is a strikingly high -533, which means that with every quarter inch of rainfall the increase of the rain coefficient is negated by the precipitation coefficient. The bigger issue is that the data has not been 'processed', meaning that it is not fair to compare the data for weekdays and weekends/holidays because people have different habits for when they have to go to work and when they can just stay at home.

## Section 5. Reflection

*5.1* The dataset has an uneven number of rainy versus non-rainy days so it leads to some assumptions being made that decrease the validity of our conclusions. The problem with linear regression is that you assume that the features have a linear relationship with the expected value and this is sometimes false. Also linear regression does not take into effect the possible relationship that the features may have with each other and how that may shift our expected value. I also did not analyze the outliers present in the data to make sure of their validity and whether they benefited or were necessary in our analysis.