Alex Stoddart
5/2/2015

Section 0 - References

- Section 1:
  - http://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test
  - http://www.graphpad.com/guides/prism/6/statistics/index.htm?how_the_mann-whitney_test_works.htm
  - http://www.statisticslectures.com/topics/mannwhitneyu/
  - http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_Nonparametric/mobile_pages/BS704_Nonparametric4.html
- Section 2:
  - http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit
- Section 5:
  - http://www.weather-and-climate.com/average-monthly-precipitation-Rainfall,New-York,United-States-of-America

Section 1 - Statistical Test

1.1
The test I used was the Mann-Whitney U test. I decided to use a one sided test since we are testing whether ridership is higher when it rains, therefore we want to see if they differ in one direction. The null hypothesis is the frequency of riders is the same on rainy and non-rainy days. My chosen p-critical value is .025.

1.2
The test is applicable because it is non-parametric and can be used on distributions that are not normal. We know that the ENTRIESn_hourly is not normally distributed from the histograms (see section 3.1). A normal t-test requires that the distribution is normal, so we cannot use that in this case.

1.3
P-Value (one-sided): 0.025
With Rain Mean: 1105.446
Without Rain Mean: 1090.279
U Statistic Calculated: 1924409167.0

1.4
We reject the null hypothesis at the 97.5% confidence level that the frequency of riders is the same on rainy and non-rainy days. The results suggest that there is significant statistical difference in ridership between when it rains and when it does not.

Section 2 - Linear Regression

2.1
I decided to use gradient descent for the regression model.

2.2
The features I used were rain, precipitation, and the hour.
A dummy variable was used for the Unit category. This uniquely represents which station the entries came from. This is important because some stations naturally have higher or lower ridership. Taking away this dummy variable causes the R^2 value to drop noticeably.

2.3
I chose the features I did logically. I think that if it is raining, people will be more inclined to use the subway. The amount of precipitation would also be a factor. For instance, a downpour would cause more ridership than a light drizzle. I also think that the later it is (aka the hour increases) ridership will also increase.

2.4
Rain: 1292.800
Precipitation: 152.506
Hour: 645.054

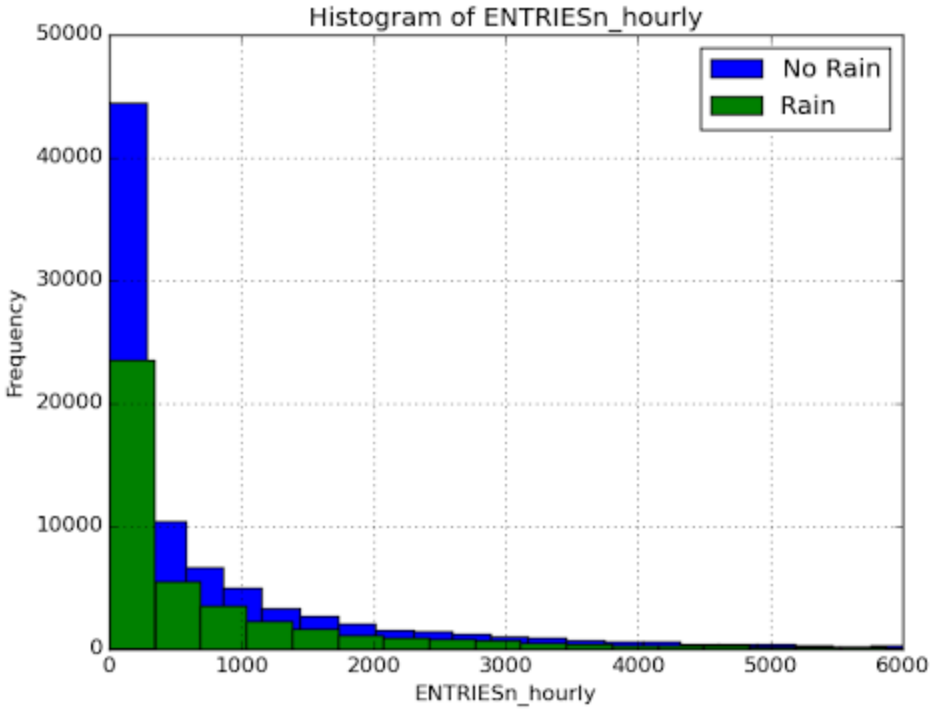
2.5
R^2 : 0.4633

2.6
This R^2 value means that this model explains about 46% of the variability in the data.
Since this is a kind of test of human behavior, and we humans are somewhat unpredictable, being able to explain close to 50% is pretty good.

Although I think this is a decent model given the R^2 value, I suspect that the data is not linear. There are so many things that could affect ridership, and I doubt that it is as simple as a linear relationship.

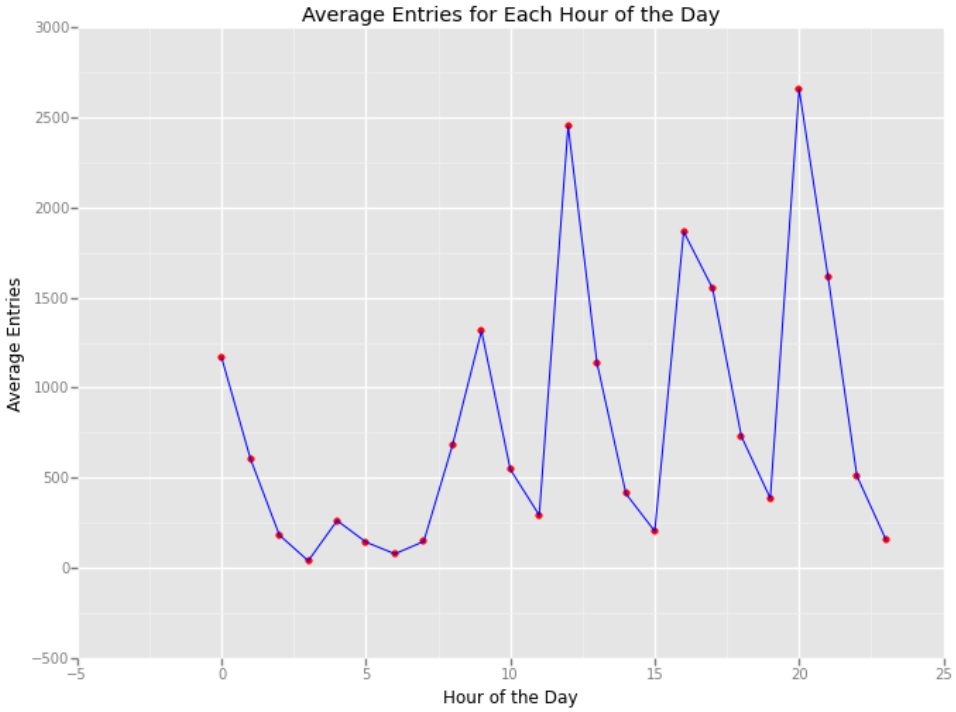
3 - Visualization

3.1



The bin size used for both samples is 150. The sample sizes differ but this graph shows that the distributions are positively skewed.

3.2



This line graph displays the average entries for each hour of the 24 hour day. I added in the points for emphasis so each hour could be viewed more easily. This graph shows trends of high usage during 12 (lunch time), 16 (end of work day), and the highest during 20 (end of happy hour perhaps). What is most surprising to me is that there is no comparatively high trend early in the morning to match the end of the work day spike.

4 - Conclusion

4.1
From my analysis of this data, I would say that yes, more people ride the NYC subway when it is raining.

4.2
The first step that led to my conclusion was my exploratory statistical test comparing the average entries with rain versus the average entries without. The Mann-Whitney U Test showed that the two distributions were significantly different, with the rainy days averaging about 15 more riders. The test reported a 2.5% chance of a type 1 error.

The next step was to dig a little deeper and perform linear regression to test the weight of certain variables on the outcome (the number of entries). I used gradient descent and had rain, the amount of precipitation, and the hour of the day as my variables. I ended with an R^2 value of 46%, the rain variable having the largest weight by far.

5 - Reflection

5.1
Usually, at least from my past experience, the problem with datasets is that they are usually too small. This one does not have an issue with size as far as entries go, but rather with the length of time in which the data takes place. One month is just too small, even though March is one of the rainier months on average in NYC. I think a better dataset would focus on fewer highly traveled stations over more time. This way trends in the data would have more meaning, because more things would be equal.

I wish I could have spent more time on the linear regression. I feel like that could be a course in itself given the importance that it has on data analysis. I tried to get my R^2 value up to just 50% and I could not do that. There is something else about the dataset that I definitely did not discover.