Analyzing the NYC Subway Dataset

By Brian Trippi

For Udacity: Intro to Data Science, Data Analyst Nanodegree March '15 Cohort

Section 0. References

- Wikipedia Entry for the Mann-Whitney U test http://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test
- SciPy.org page on the Mann-Whitney U test http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html
- The Minitab Blog entry, Regression Analysis: How Do I interpret R-Squared and Assess the Goodness of Fit http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit
- Github Yhat ggplot for python resource https://github.com/yhat/ggplot

Section 1. Statistical Test

1.1) In Exercise 3.3 we used a Mann-Whitney U-Test to analyze the NYC Subway Data. We used a one-tail P value, and the null hypothesis is that the two samples are drawn from the same population. The P critical value is 0.0250.

1.2) The Mann-Whitney U-Test test is applicable to this dataset because it does not assume that our data is drawn from a normal probability distribution. In Exercise 3.1, we discovered that ridership numbers for rainy and non-rainy days did not appear to be normally distributed, and that the sample sizes differed significantly, with hourly entries during the rain having many fewer samples.

1.3) The statistical test found that the mean of entries per hour with rain was 1105.45, and the mean of entries per hour without rain was 1090.28. The one-tailed P value was 0.025.

1.4) The P value is the estimated probability that we will mistakenly reject the null hypothesis when the null hypothesis is true. In this case the null

hypothesis is that both of our samples are drawn from the same population. At a 5% significance level the results of our test lead us to reject the null hypothesis, meaning that we believe it is likely that our two samples were drawn from distinct populations. This is good evidence that our mean ridership numbers for rainy and non-rainy days were most likely not drawn from the same population.

Section 2. Linear Regression

2.1) Gradient descent was used to compute the coefficients theta and produce a prediction for ENTRIESn_hourly in the regression model.

2.2) The features used in the model were rain, percipi, fog, and meantempi, with UNIT being used as the dummy variable.

2.3) These features were chosen for the model because they seemed to be the weather related factors most likely to affect ridership numbers. Intuitively, rain would seem to increase ridership numbers since people are probably more likely to ride the subway instead of walking or driving in the rain. The same logic applies to precipi and meantempi. In cases where there is especially heavy rain or snow, or when the temperature is uncomfortably hot or cold, people are probably more likely to ride the subway. Additionally, fog could potentially increase subway ridership numbers if people feel unsafe driving or walking in heavy fog.

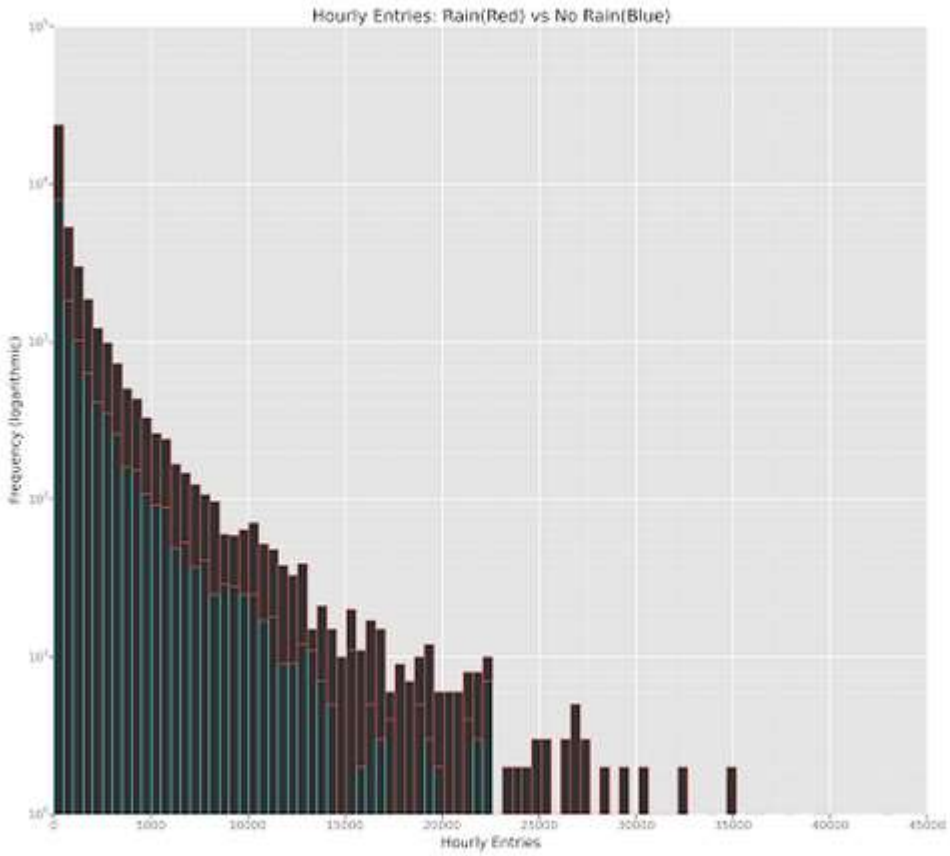
2.4) The coefficients for the non-dummy features in this linear regression model are -7.37 for rain, 62.65 for percipi, -73.33 for fog, and 154.50 for meantempi.

2.5) The model's R^2 coefficient was 0.4263.

2.6) The model's R^2 coefficient of 0.4263 was well above the required value of .20. This seems to indicate a relatively high level of goodness of fit. Of all the variation in the number of entries per hour, our model accounts for over 42% of the variation. This model would appear to be appropriate for the dataset when attempting to predict ridership.

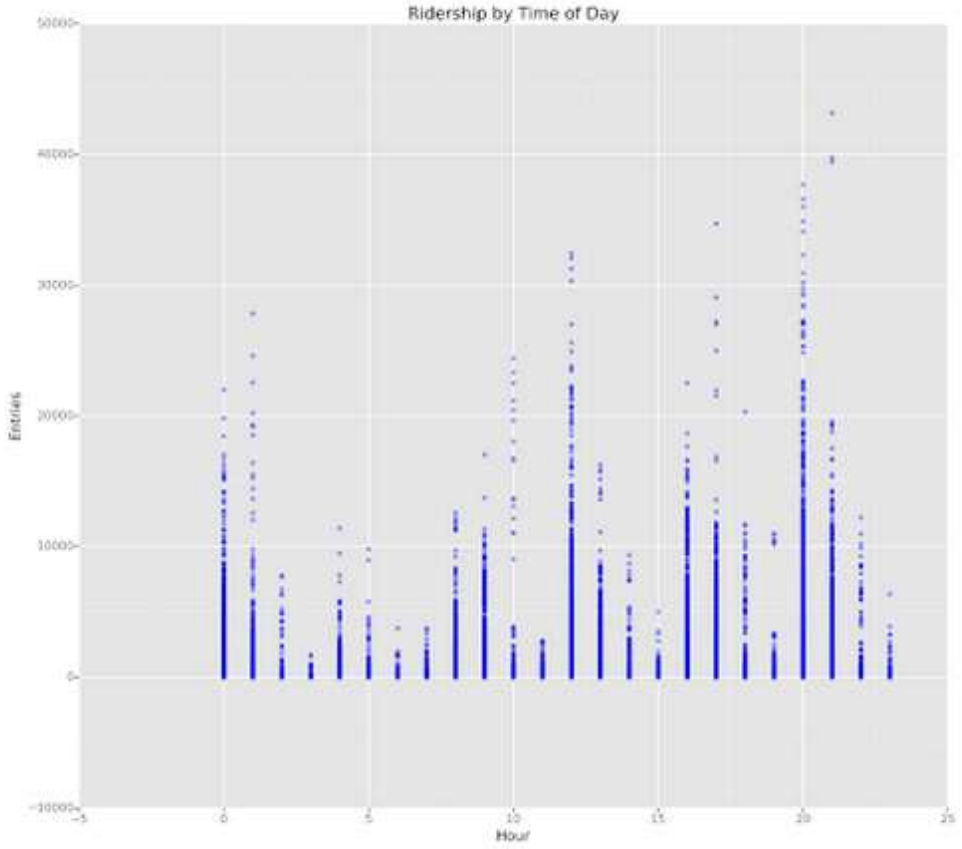
Section 3. Visualization

3.1)



This visualization charts the number of riders per hour on the x-axis, and the frequency that the number of hourly entries occur within our dataset on the y-axis. The major takeaway from this chart is that there are significantly fewer instances of data taken during rainy days than non-rainy days. This chart uses a binwidth of 500. The vast majority of data fell within the first bin, which caused the less frequently occurring values of hourly entries to appear very small relative to the first bin. In order to make the data more readable the visualization was scaled logarithmically along the y-axis which had the effect of making the less frequently occurring hourly entry values easier to view.

3.2)



This visualization is a scatter plot illustrating the number of turnstile entries by time of day. This chart is interesting because it seems to indicate that the NYC subway is most busy during hours that would not be considered conventional "rush hours." The busiest times of day for ridership appear to be focused around 12 pm to 1 pm, and 8 pm to 9 pm. Since there were many data points relative to the size of the points on the visualization, alpha blending was used in an attempt to make the points where entries/hour occurred more frequently appear darker on the chart. That being said, it is possible that the entries/hour during rush hour were consistently within a narrow range which make them appear to be less busy times of day on this chart.

Section 4. Conclusion

4.1) Analysis and interpretation of the data reveals that more people tend to ride the NYC subway when it is raining. This would coincide with the intuitive notion that if it is raining people are more likely to ride the subway instead of walking.

4.2) Statistical analysis showed that the mean number of NYC subway turnstile entries per hour when it was raining was 1105.45, and the mean number of entries per hour when it was not raining was 1090.28. The Mann-Whitney U-Test provided us with good evidence that our ridership numbers for rainy and non-rainy days were most likely not drawn from the same population. Additionally, our linear regression model using gradient descent revealed that over 42% of the variation in subway ridership numbers was due to weather related factors including rain, mean temperature, level of precipitation, and fog. These findings provide further evidence that the higher mean ridership numbers on rainy days was actually due to weather related factors.

Section 5. Reflection

5.1) One shortcoming with this analysis is that the dataset was taken from a relatively narrow range of dates. All of the data in our dataset was taken between the dates of May 1, 2011 and May 30, 2011. Further analysis may reveal that our model is not valid for other months of the year where the weather in New York City is significantly different. Also the weather data for this month may have been atypical when compared to the average weather for this month, which would cause our model to be inaccurate for the month of May in future or past years as well. Additionally, as discussed in Lesson 3, there were several issues that we did not address in our linear regression model such as not including confidence intervals on our parameters, and failure to take measures to detect over/under fitting as well as multiple local minima.