Analyzing the NYC Subway Dataset

Student Name: Mariano Aguirre

Email: mariano.ap@utexas.edu

Cohort: March 2015

Program: Data Analyst – Nanodegree

Section 0. References

Intro to Data Science notes and materials

Pandas - http://pandas.pydata.org/pandas-docs/dev/index.html

P values - http://www.graphpad.com/guides/prism/6/statistics/index.htm?one-tail_vs__two-tail_p_values.htm

Mann-Whitney U test - http://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test

Dummy variables:
http://en.wikipedia.org/wiki/Dummy_variable_%28statistics%29
http://pandas.pydata.org/pandas-docs/version/0.13.1/generated/pandas.get_dummies.html

Regression – R^2:
http://people.duke.edu/~rnau/rsquared.htm
http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit

Plotting:
https://pypi.python.org/pypi/ggplot/
http://ggplot.yhathq.com/docs/index.html

Section 1. Statistical Test

1.1 Which statistical test did you use to analyze the NYC subway data? Did you use a one-tail or a two-tail P value? What is the null hypothesis? What is your p-critical value?

I used the Mann-Whitney implementation. Since the Mann-Whitney uses as two-tailed model, but returns a one-tailed p value, I double the p-value to get the correct value for the test statistic. The null hypothesis in this case is:

H0: The ridership distribution with rain is equal to the ridership distribution with no rain.

I used a 95% confidence interval, which requires a p-critical value of $p<0.05$.

We can reject the null hypothesis since $0.0498 < 0.05$ and thus the alternate hypothesis is that the distributions are different.

1.2 Why is this statistical test applicable to the dataset? In particular, consider the assumptions that the test is making about the distribution of ridership in the two samples.

The data does not seem normally distributed, as such other tests such as the Welch's T-test is not applicable (Welch's T-test applies to normally distributed data only). However, the Man-Whitney statistical test can be applied to data of unknown distribution.

1.3 What results did you get from this statistical test? These should include the following numerical values: p-values, as well as the means for each of the two samples under test.

Mean of X (ridership with rain) = 1105.446
Mean of Y (ridership with no rain) = 1090.279
The p-value returned was approximately $p=0.0249$ but this is for a one-tailed p-value. The correct value to use is double, so $p=0.0498$.

1.4 What is the significance and interpretation of these results?

Using the Mann-Whitney statistical test, we can compare if the distributions of X and Y are the same (null hypothesis). Since the p-value (adjusted for a two-tailed distribution) is 0.048 and less than our critical p-value of 0.05 ($p < 0.05$), we can reject the null hypothesis and assume that the distributions are different. In other words, we can say that there is less than a 5% probability that a random sample of X will be the same as a random sample of Y. We can also infer from this and the means of X and Y that ridership is slightly higher during rainy days.

Section 2. Linear Regression

2.1 What approach did you use to compute the coefficients theta and produce prediction for ENTRIESn_hourly in your regression model:

a. *Gradient descent (as implemented in exercise 3.5)*

b. *OLS using Statsmodels*

c. *Or something different?*

The coefficients for theta were estimated using the gradient descent algorithm. The stepping value (alpha) and the number of repetitions were adjusted to achieve convergence to a solution. Care must be put in setting up the right stepping value (alpha) to make sure that the algorithm does not get stuck in a local minimum causing you to miss the optimum solution.

2.2 What features (input variables) did you use in your model? Did you use any dummy variables as part of your features?

For the data given, I limited my analysis to rain, hour of the day and the mean temperature. Dummy variables are used whenever there is a qualitative or categorical variable in the data. For example, in estimating the price per square feet in real state, additional characteristics such as having a pool or no pool, two stories vs. one story, are all examples of categorical variables that require the use of dummy variables in order to include them in a regression model. In this particular case, rain is the categorical variable, since it either rain or did not rain. Since rain was already categorized as a 1 for rain and 0 for no rain, there was no need to add another dummy variable for rain.
However, the exercise in problem 3.5 uses the pandas.get_dummies function, which creates multiple dummy variables for all sorts of categorical combinations between rain, hour, and mean temperature.

2.3 Why did you select these features in your model? We are looking for specific reasons that lead you to believe that the selected features will contribute to the predictive power of your model.

- *Your reasons might be based on intuition. For example, response for fog might be: "I decided to use fog because I thought that when it is very foggy outside people might decide to use the subway more often."*

- *Your reasons might also be based on data exploration and experimentation, for example: "I used feature X because as soon as I included it in my model, it drastically improved my R^2 value."*

I included rain because based on experience, people are more likely to seek shelter and get out of the rain. A person may decide that if its not raining, they could walk to their destination. The subway still allows them to reach their destination away from the rain.

I also included the time of day since ridership could be highly correlated to working schedules. People may be heading to work or coming home from work. Taking a trip to a restaurant or the supermarket.
Finally, I pick the mean temperature as another indicator of potential ridership. People during the day may use the subway more likely if it is particular a hot day, whereas people may decide to walk if the temperature is comfortable.

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

The non-dummy coefficients are the first 3 values in the theta_gradient_descent array.

Rain coefficient = 12.096
Hour of the day coefficient = 452.136
Mean Temperature coefficient = -61.191

Typically the y-intercept comes first but the code added the column of 1's at the end of the array, making the last coefficient the y-intercept.

2.5 What is your model's R^2 (coefficients of determination) value?

The R2 value of my model based on rain, hour of the day and mean temperature was 0.463598.

2.6 What does this R^2 value mean for the goodness of fit for your regression model? Do you think this linear model to predict ridership is appropriate for this dataset, given this R^2 value?

We are trying to predict ridership usage that is highly dependent on human behavior and responses to a multitude of variables. The calculated R^2 is able to account for about 46% of the variance by just using three variables (rain, hour of the day and temperature). Still, we cannot account for the other 54% of the variance. Additional analysis of the data could uncover other variable combinations that account for some of the 54% but not all of it. When trying to account for human behavior R^2 values less than 50% are not uncommon (see reference).

Section 3. Visualization

Please include two visualizations that show the relationships between two or more variables in the NYC subway data.

Remember to add appropriate titles and axes labels to your plots. Also, please add a short description below each figure commenting on the key insights depicted in the figure.

3.1 One visualization should contain two histograms: one of ENTRIESn_hourly for rainy days and one of ENTRIESn_hourly for non-rainy days.

- *You can combine the two histograms in a single plot or you can use two separate plots.*
- *If you decide to use to two separate plots for the two histograms, please ensure that the x-axis limits for both of the plots are identical. It is much easier to compare the two in that case.*
- *For the histograms, you should have intervals representing the volume of ridership (value of ENTRIESn_hourly) on the x-axis and the frequency of occurrence on the y-axis. For example, each interval (along the x-axis), the height of the bar for this interval will represent the number of records (rows in our data) that have ENTRIESn_hourly that falls in this interval.*
- *Remember to increase the number of bins in the histogram (by having larger number of bars). The default bin width is not sufficient to capture the variability in the two samples.*

Figure 1 - Histogram of hourly entries for rain and no rain

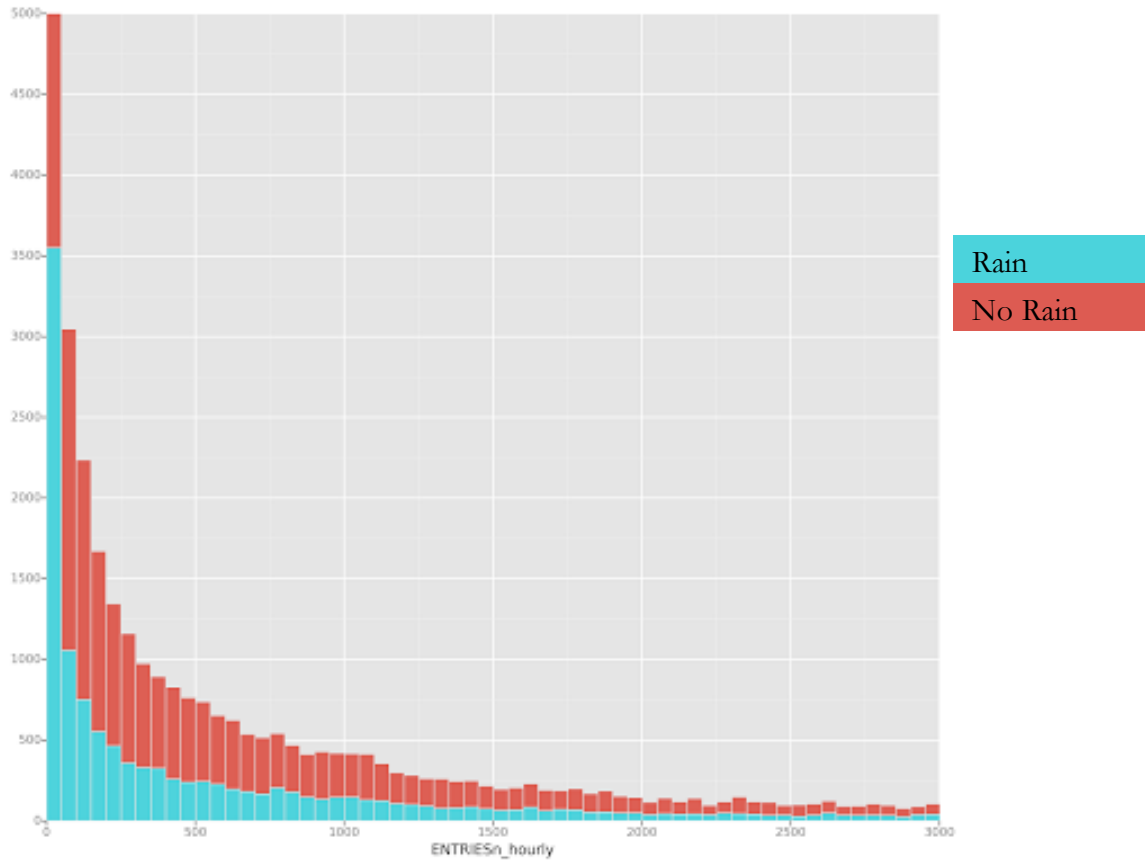

Figure 1 shows a histogram of ridership for both rain and no rain scenarios. As can be seen from the graph, ridership volume is significantly higher with no rain compared to ridership with rain. However, we are looking at the total sum and this could lead to misleading conclusions. The more appropriate visual is to normalize the data using percentages (e.g. rain days over total days) to correctly compare the results between two conditions.

3.2 One visualization can be more freeform. You should feel free to implement something that we discussed in class (e.g., scatter plots, line plots) or attempt to implement something more advanced if you'd like. Some suggestions are:

- *Ridership by time-of-day*
- *Ridership by day-of-week*

Figure 2 - Ridership by time of day

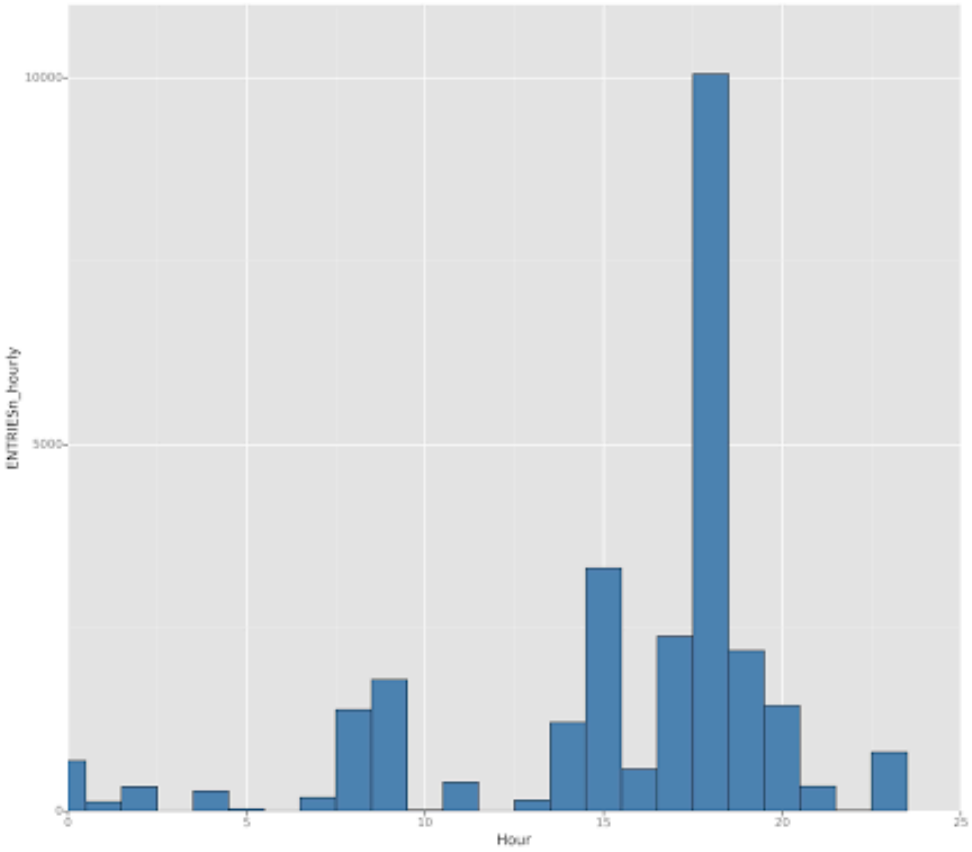

Figure 2 shows ridership by time of day. A quick look at this bar graph and one can see spikes around morning (8-9am) likely when people go to work and around 6pm, which could represent people leaving work or getting dinner.

Section 4. Conclusion

Please address the following questions in detail. Your answers should be 1-2 paragraphs long.

4.1 From your analysis and interpretation of the data, do more people ride the NYC subway when it is raining or when it is not raining?

When looking only at the mean of ridership for both the rain and no rain cases, the mean suggests that ridership is higher when it is raining. This alone is not sufficient to determine if ridership is higher during rain. One must look at the results of the statistical test and visuals of the data normalized in percentages to correctly make the right comparisons. Given that our statistical test disproved the null hypothesis, we can say with 95% confidence that indeed, ridership is slightly higher during rainy days.

4.2 What analyses lead you to this conclusion? You should use results from both your statistical tests and your linear regression to support your analysis.

The distributions with rain and no rain are clearly distinct based on the Mann-Whitney statistical test results. This means that there is a clear difference with a 95% confidence interval that the ridership is higher under the rain scenario. The linear regression exercise provided a way to explore which features contributed more to predicting ridership in the subway. The R2 for all three variables (rain, hour and mean temperature) was 0.46. However, using rain alone as the only feature can account for explaining 42% of the variance in ridership (R^2=0.4248). Through linear regression we know that rain is strongly correlated with ridership.

Section 5. Reflection

Please address the following questions in detail. Your answers should be 1-2 paragraphs long.

5.1 Please discuss potential shortcomings of the methods of your analysis, including:

1. *Dataset,*
2. *Analysis, such as the linear regression model or statistical test.*

The accuracy of the conclusions is partially dependent on the quality of the data. If there is missing data, one does not know if it is missing random or data specific to one particular feature. For example, sensors collecting data could fail in certain stations when it is raining skewing the data in favor of no rain days.
If the dataset is highly non-linear, a linear regression model may not accurately fit the data correctly and lead one to the wrong conclusions if used.
If the dataset does not have enough samples, a statistical test may not have enough accuracy.

5.2 (Optional) Do you have any other insight about the dataset that you would like to share with us?

With additional time it would have been interesting to study if the amount of rain (precipitation) affects ridership in any particular fashion. For example, would heavy rain actually result in higher ridership versus light rain? Also, the partial data from the online website did not have any data for thunder (all values were zero). Would thunder have impacted ridership behavior? Are certain subway stations more likely to be over-utilized when raining? Answers to these questions could help improve the subway system in NYC by making changes to the stations, frequency of stops, etc.