William Autry
Project 2

Intro to Data Science: Analyzing the NYC Subway Dataset

Section 0. – References

http://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test

https://statistics.laerd.com/spss-tutorials/mann-whitney-u-test-using-spss-statistics.php

http://www.statisticssolutions.com/mann-whitney-u-test/

http://socserv.mcmaster.ca/jfox/Courses/SPIDA/dummy-regression-notes.pdf

http://www.ats.ucla.edu/stat/mult_pkg/faq/general/tail_tests.htm

Section 1. – Statistical Test

What is our independent variable? What is our dependent variable?
Independent Variable: Rainy/Non-Rainy Weather
Dependent Variable: NYC Subway Ridership

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 chose to use the Mann-Whitney U-test to compare distributions without prior knowledge as to which data set will be higher/lower. The Mann-Whitney U-test ignores the sizes of the samples, because it considers a single draw from each distribution.
- I used a two-tail P value.
- Null Hypothesis: The distribution of ridership shown on rainy days will be comparable to non-rainy days. The null hypothesis asserts that if a single draw from each distribution is chosen, then the probability that the draw from distribution A is bigger than one from distribution B is 50/50.

  Given random draws x from population X (Rainy) and y from population Y (Non-Rainy), the standard two-tailed hypotheses are as follows:

  H_0: $P(x > y) = 0.5$
  H_1: $P (x > y) \neq 0.5$

- P-Critical Value: 0.05

  If the p value from the test is less than this value, we can reject the null.

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 distribution of ridership is not a normal distribution, therefore we must use the Mann-Whitney U-test which is non-parametric and does not rely on assumptions.

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 (w/ Rain): 1105.446
- Mean (w/o Rain): 1090.279
- P-Value: 0.049

1.4 What is the significance and interpretation of these results?

- These results are significant because our alpha for the test is 0.05 and the p-value is 0.049 which is less than 0.05. With these results we can then reject the null hypothesis on the basis that the ridership is statistically different.

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?

- I used Gradient Decent & (OLS) Ordinary Least Squares

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

- Rain: ‘rain’, Precipitation: ‘precipi’, Hour: ‘Hour’, Mean Temperature ‘meantempi’, & Dummy Variable ‘UNIT’.
- Yes, the dummy variable ‘UNIT’ for the turnstile identifier.

2.3 Why did you select these features in your model? We are looking for specific reasons that lead you to believe that

- I used the selected features, because I felt that they all correlated to rainy weather and contribute to the predictive power of determining a trend in Subway Ridership.

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

- Rain: 0.266
- Percipi: 0.000447
- Hour: 0
- Mean Temp: 1.214563E-16

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

- 0.46

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?

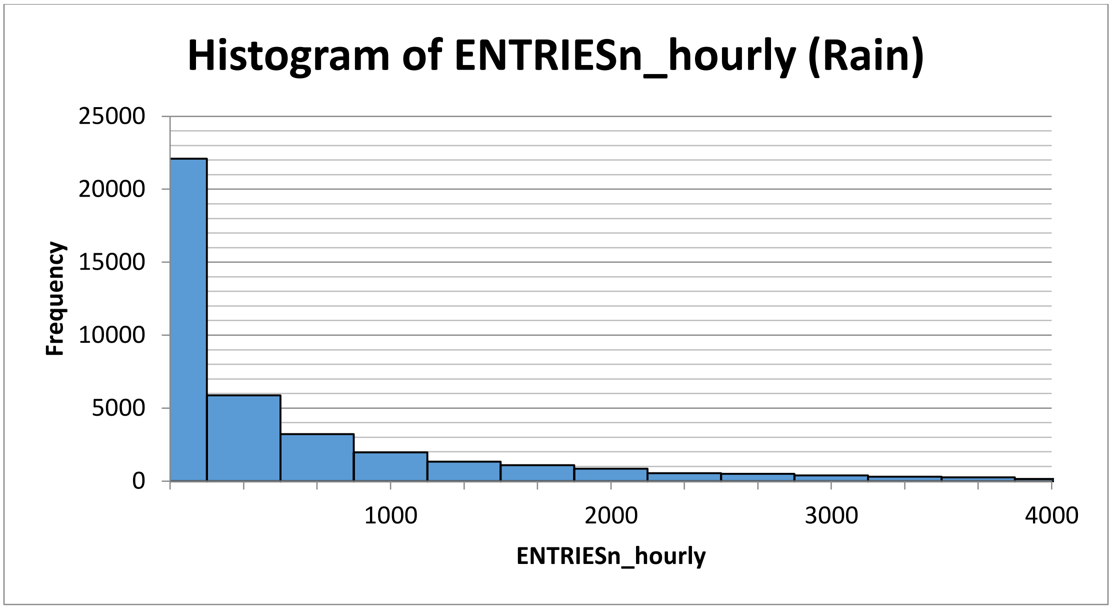
- The 0.46 value means that we can predict NYC Subway Ridership with 46% accuracy.
- Yes, it is appropriate for this dataset since we can explain nearly half of the data variability. Within the context of what we are ultimately trying to determine, this is adequate.

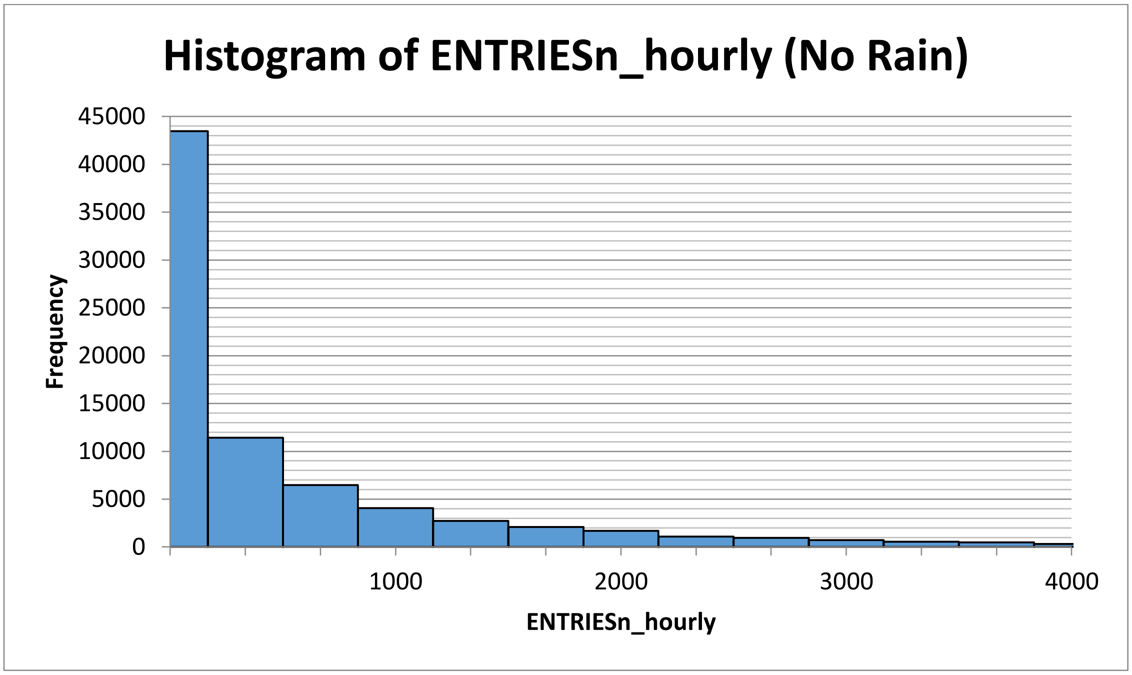
Section 3. – Visualization

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.

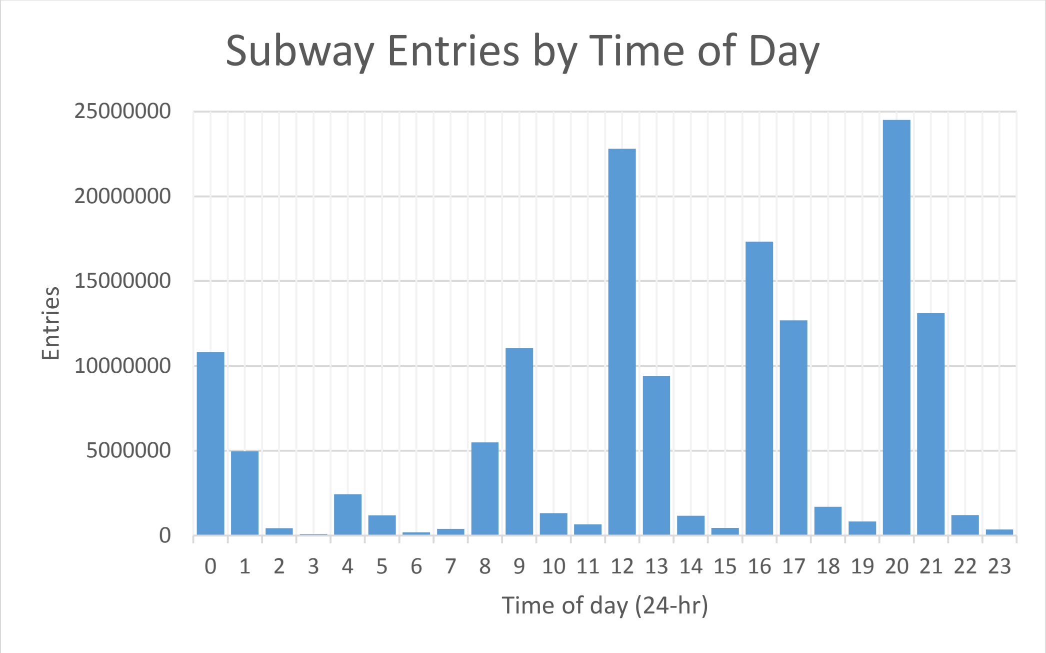
3.1 One visualization should contain two histograms: one of ENTRIESn_hourly for rainy days and one of ENTRIESn_hourly for non-rainy days. You can combine the two histograms in a single plot or you can use two separate plots.





From the two histograms above, one might conclude that rain negatively affects ridership; however, visualizations alone cannot determine this.

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: This graph shows the peak hours of NYC Subway Ridership to be around 11-noon and 8-9 PM.

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?

Yes, from my analysis, the results show that more people ride the NYC Subway when it *is* raining. The Mann-Whitney U test compares the distributions of ridership for rainy and non-rainy days. The results are significant because our alpha for the test is 0.05 and the p-value is 0.049 which is less than 0.05. With these results we can then reject the null hypothesis on the basis that the ridership is statistically different. Rain contributes to a change in the frequency of NYC Subway Ridership.

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 Mann-Whitney U-Test statistical analysis results are significant because our alpha for the test is 0.05 and the p-value is 0.049 which is less than 0.05. The OLS regression model was a good way to identify trends. The visualization of the turnstile data in the form of histograms is inconclusive in determining rain's effects on ridership. The histograms show similar distribution shapes, however the frequencies are different due to the difference in samples (44104 vs. 87847).

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:

Dataset, Analysis, or statistical test.

My R^2 value is 0.46, one possible shortcoming is that I may have underestimated the need for more accuracy in my analysis. Another possible shortcoming of the dataset is that there are 144532327 entries in comparison to 117026133 exits, which is a difference of 27506194. This difference cannot be explained. The final shortcoming is in my statistical test where I could have used a Shapiro-Wilk test instead of the Mann-Whitney U test.