Analyzing the NYC Subway Dataset

Bindu Paul

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 U test to analyze my data. I used a two-tailed p-value to analyze my data because if ridership is greater or less on rainy days both matter. H_0: P(Entries during Rain > Entries during No Rain) = 0.5. H_1: P(Entries during Rain > Entries during No Rain) ≠ 0.5. P-value = 0.05 (this is the value returned from the MWU test multiplied by 2 for a two-tailed test).

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 Mann-Whitney U test is applicable to this dataset because the assumption made is that the samples are non-parametric and are not normally distributed.

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.

U = 1924409167, Z-score = -142, P-value = 0.05, P-critical = 0.05 or 95% confidence interval, Mean ENTRIESn_hourly rain = 1105, Mean ENTRIESn_hourly no rain = 1090.

1.4 What is the significance and interpretation of these results?

Analyzing the data shows my P-value and P-critical are the same so I would accept the null hypothesis. Randomly pulling from both samples, 50% of the time, entries during the rain would be greater than entries during no rain.

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 the OLS using Statsmodels

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

I used “Hour” and “rain” as input variables. I also used dummy variables labeled “UNIT” as apart of my features. I also ran another one with only “Hour” as the input variable.

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.

I used “Hour” with and without “rain” features because I wanted to analyze how much of an impact rain had on the R^2 value. I used the dummy variable because it increased my R^2 value drastically.

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

Hour = 59.483324; rain = 16.28

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

R^2 = 0.45756 with rain, R^2 = 0.45754 without rain

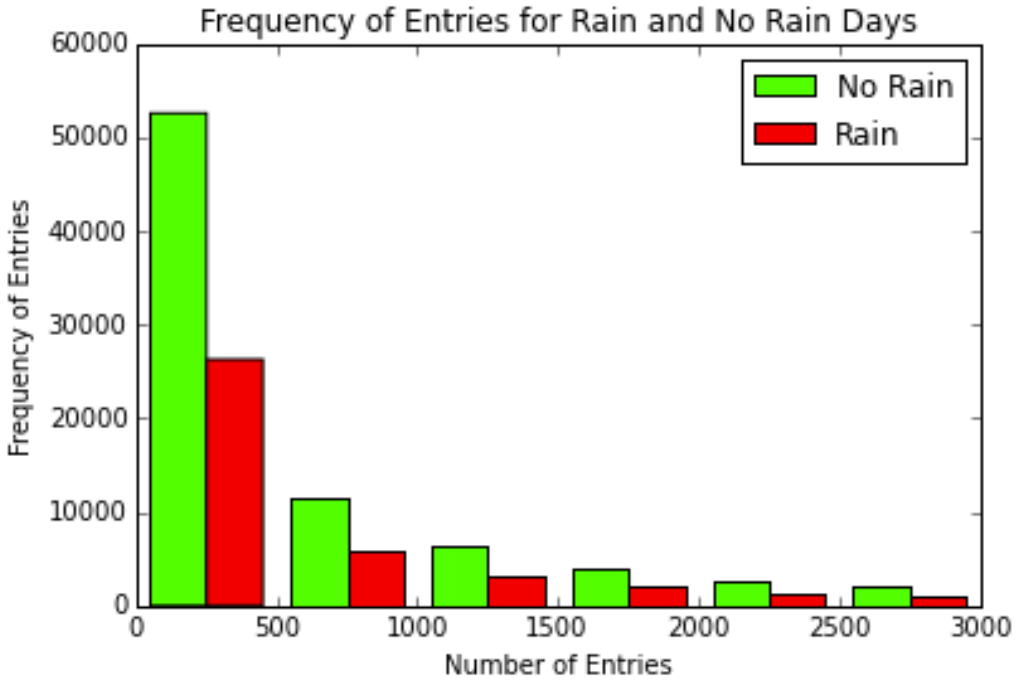
2.6 What does this R^2 value mean for the goodness of fit for your regression model? Do you think this linear model to predict ridership is appropriate for this dataset, given this R^2value?

My R^2 value of 0.45756 does not show the strongest relationship to predicting entries hourly. Using Hour and rain as features, only account for 45.76% of the variability; the other 54.24% is affected by other features and I would not use this model to predict ridership during rain.

Section 3. Visualization

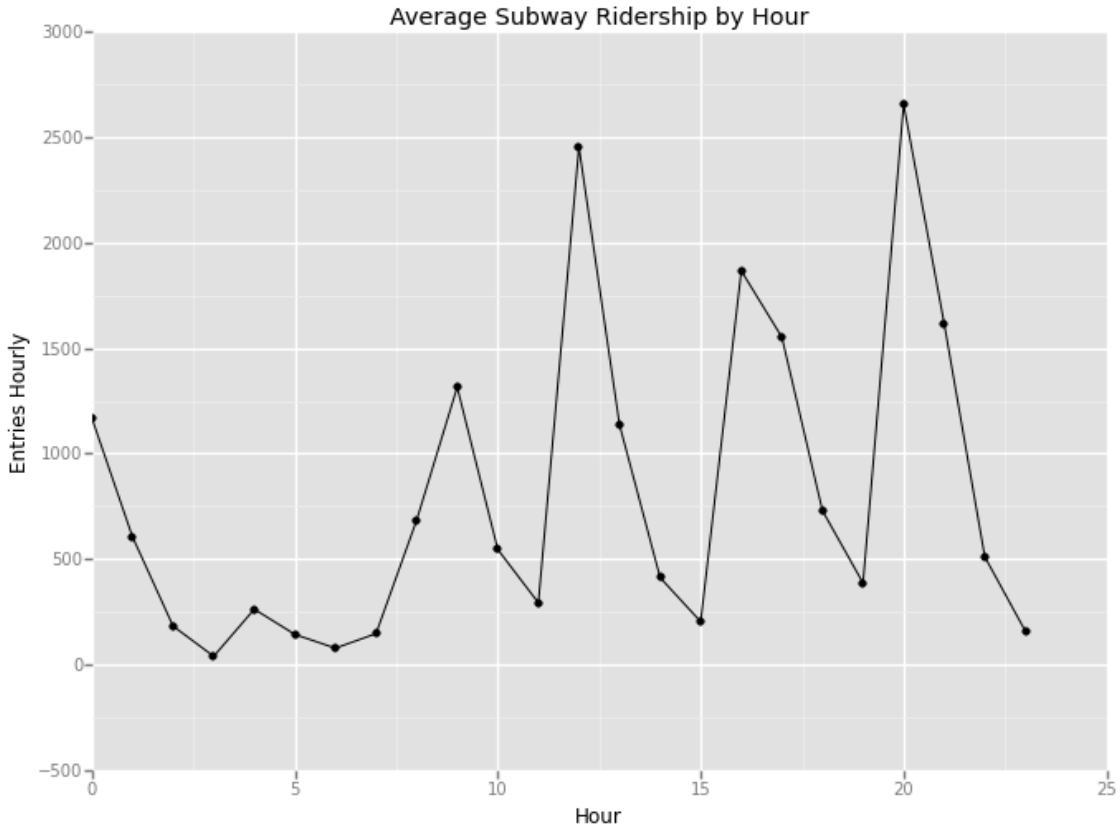
Please include two visualizations that show the relationships between two or more variables in the NYC subway data. Remember to add appropriate titles and axes labels to your plots. Also, please add a short description below each figure commenting on the key insights depicted in the figure.

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



The graph above shows the frequency of entries hourly is greater on days with no rain.

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).



The graph above shows average entries per hour and max entries at 12pm and 8pm.

Section 4. Conclusion

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?

From my analysis I conclude that more people ride the subway when it is raining. This can be seen when comparing the means of entries hourly for both rain (1105) and no rain (1090). Also in graph 3.1 there is a greater difference between the max and min frequencies of no rain compared to rain. The graph would be more accurate if both had equal counts and the frequency of each bin could be analyzed.

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.

I came to this conclusion because R^2 with rain (0.45756) is greater than without rain (0.45754). The R^2 shows the relationship of rain on entries hourly positively increasing

towards 1, which means rain does have a small effect. We can also analyze the means in 1.3 that shows Rain to be greater by 15 entries hourly. I also accept my null hypothesis that shows a greater chance for entries during the rain when equal samples are drawn.

Section 5. Reflection

5.1 Please discuss potential shortcomings of the methods of your analysis, including, Dataset, Analysis, such as the linear regression model or statistical test.

The dataset does not portray if any day was a special event day/holiday. For example, there could be days where many tourists were in NYC and used the subway system for no rain days, which would increase the entries. The data would be more precise if the population was isolated and we eliminated regular riders. The linear regression value is not the most precise which also makes the data less reliable. If the means of rain versus no rain had a greater difference then the data would be more accurate. It's hard to conclude if the results were due to chance or more people actually ride during rainy days.

References

http://www.ats.ucla.edu/stat/mult_pkg/faq/general/tail_tests.htm - research about one-tail and two-tailed tests.
https://discussions.udacity.com/t/problem-set-3-plotting-error-histogram-issue/22058/4 - discussion about pivot table error
http://blog.minitab.com/blog/adventures-in-statistics/how-to-interpret-regression-analysis-results-p-values-and-coefficients - information about regression lines
http://stats.stackexchange.com/questions/9573/t-test-for-non-normal-when-n50 - information about assumptions for t test.