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 on the number of turnstile entries with and without rain. I used a two-tailed P value with a P-critical of .05. The null hypothesis is that given the two distributions of turnstile entries with or without rain, the distribution of both populations is the same. Therefore, random samples from each population have equal probability of either one exceeding the other.

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 appropriate for this dataset because the distribution of data is not normal. A two-tailed test is the best approach because ridership could either be greater or smaller during times of rain compared to a period of time without rain.

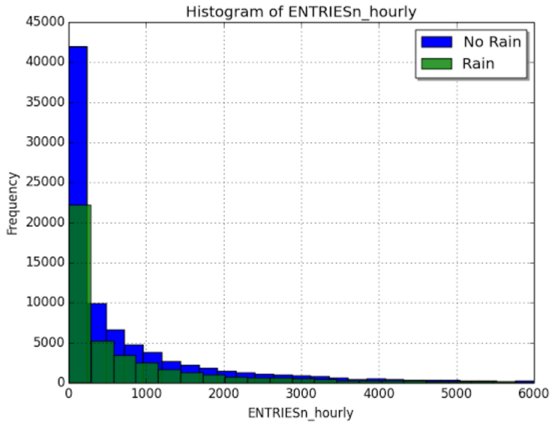

Figure 1: Mann Whitney U-Test appropriate due to positive skewed distribution

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.

The mean of entries during periods of rain was 1105.45. The mean of entries during periods without rain was 1090.00. The Mann-Whitney U score was 1924409167.0. The p-value for a two-tailed test was 0.049999825587, which is slightly under the p-critical of .05.

1.4. What is the significance and interpretation of these results?

Given that p is slightly under the p-crit of .05 for a two tailed test, we can reject the null hypothesis that the two distributions are the same.

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 descent.

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 the mean temperature, the amount of precipitation, and the dummy variables for turnstile units, day of week, and hour of day.

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.

The Mann-Whitney highlighted that rain likely played a factor in ridership. Through repeated tests and comparing R^2, I learned combinations of multiple weather variables often led to minimal R^2 changes. This likely meant they were collinear, and I settled on only using the amount of precipitation and mean temperature.

Turnstile units as a dummy variable were required since units in certain parts of the city are going to be significantly busier than other parts of the city. Intuition told me that time of day would greatly affect ridership, but I learned through trial and error that it was better to turn the hour of day into a dummy variable to achieve a better R^2.

During my visualization work, when I saw how greatly the day of week impacted ridership, I also converted this to a dummy variable.

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

Feature	Coefficient
precipitation	-19.13119233
meantempi	-45.80758203

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

R^2 = 0.514200027772

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?

Given that my R^2 value means the fit only explains ~51% of the overall variation of the data, it is probably not an ideal model.

My residuals initially appeared normally distributed with a mean of .03. This is typically indicative that I do not have some unaccounted aspect of the data skewing the prediction.

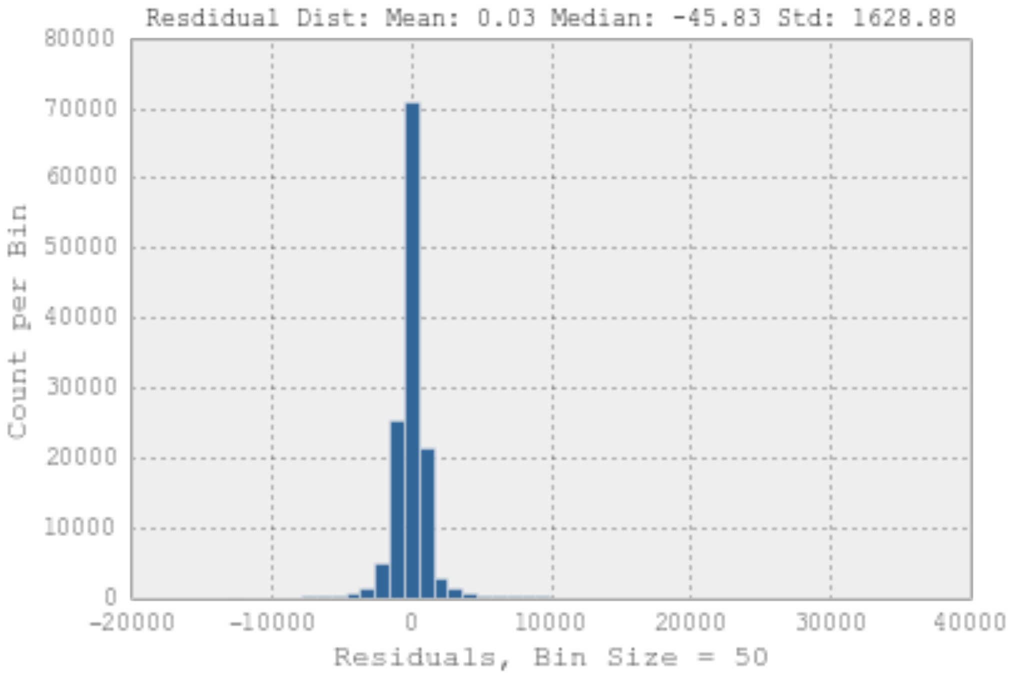

Figure 2: Residual histogram - normal or leptokurtic distribution?

However, a probability plot of the residuals highlights the long tails at the ends of the distribution. An ideal plot would be linear with a positive slope.

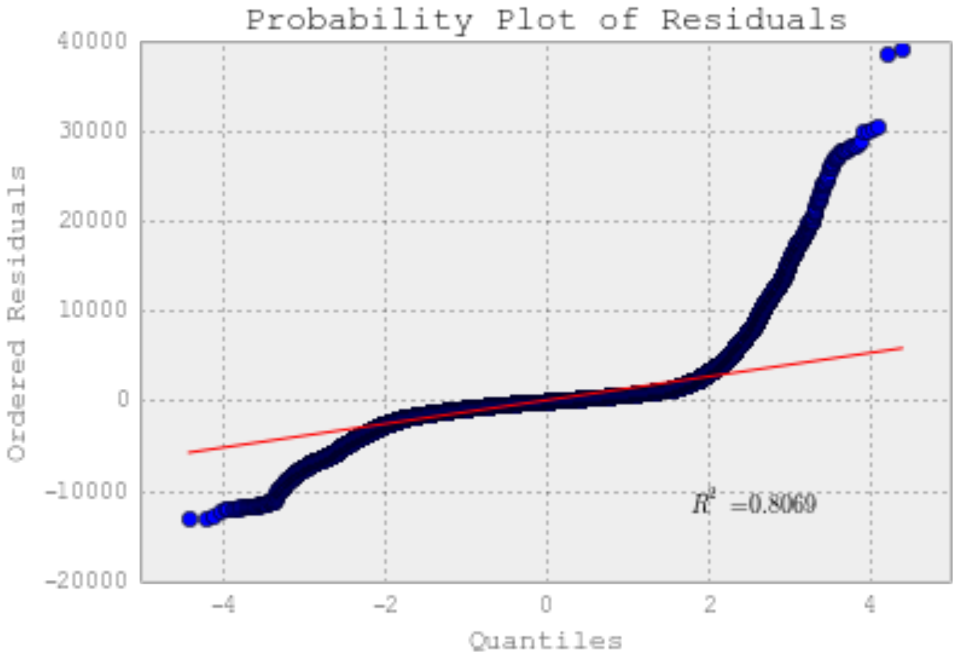

Figure 3: Probability plot of residuals with long tails

3. Visualization

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

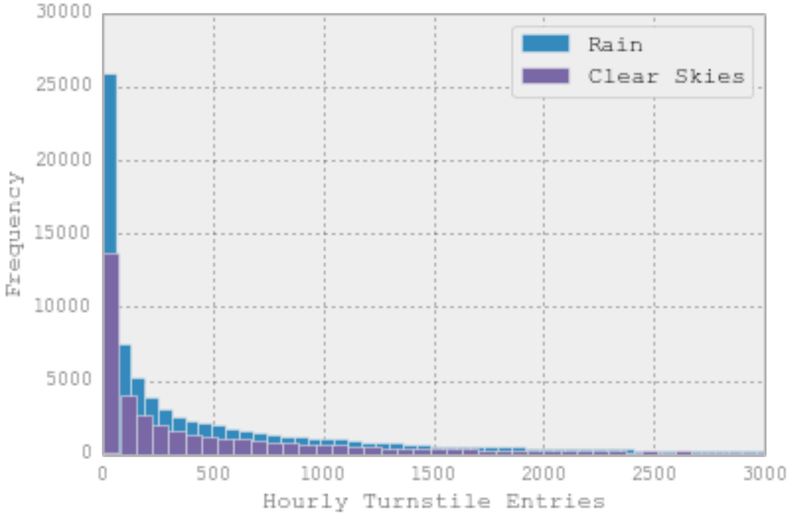

Figure 4: Histogram of hourly turnstile entries with or without rain.

Note: The x-axis of the histogram was capped at 3000 due to the extreme number of outliers for both cases. A boxplot below illustrates the number of outliers:

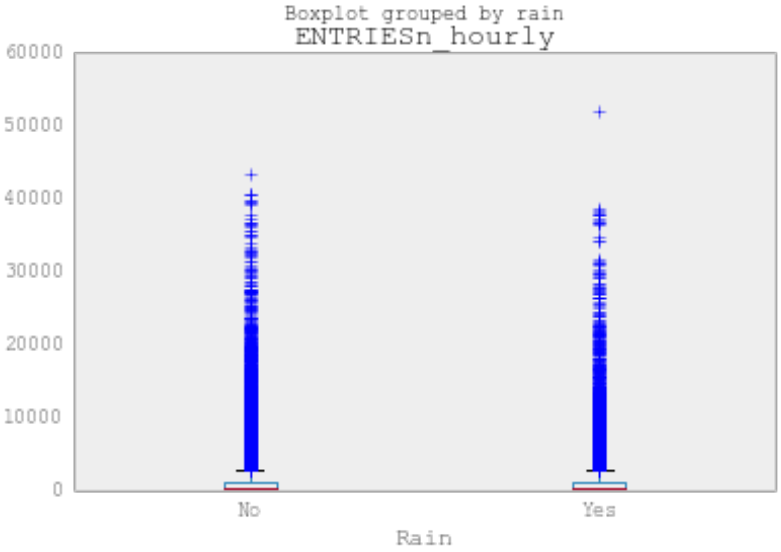

Figure 5: Boxplot illustrating significant outliers

3.2. One visualization can be more free form.

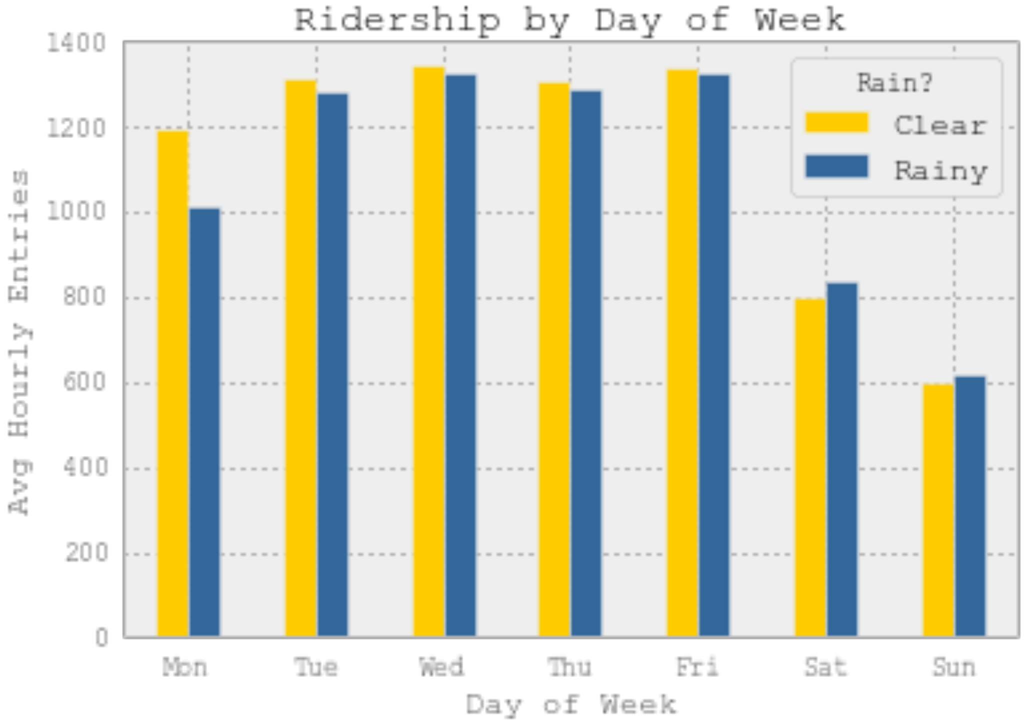

Figure 6: Average hourly turnstile entries by day of week and rain conditions reflects change on weekend.

Depicting average hourly turnstile entries by day of week and rainy weather, it appears there is a difference in ridership on weekdays compared to weekends.

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?

I believe in general, more people ride the NYC subway on weekdays when it is not raining, but weekend ridership will be higher on 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.

First, the extreme value of the Mann-Whitney U test, along with passing the p-critical, provides an indication that one sample set will be greater than the other. The mean of hourly turnstile entries on rainy days (1105.45) is greater than the mean of days without rain (1090). However, given the significant number of outliers in the histogram, this can skew means significantly. When looking at the median, hourly turnstile entries on rainy days (282) are still greater than on clear days (278). Initial conclusions would say that ridership is greater on rainy days.

However, my linear regression model shows that as precipitation increases, hourly turnstile entries go down. In addition, when plotting the median of hourly turnstile entries by day, it's clear that on busy weekdays, ridership is greater on clear days, rather than rainy days. On weekends, ridership does pickup on rainy days – especially when using the median to compare.

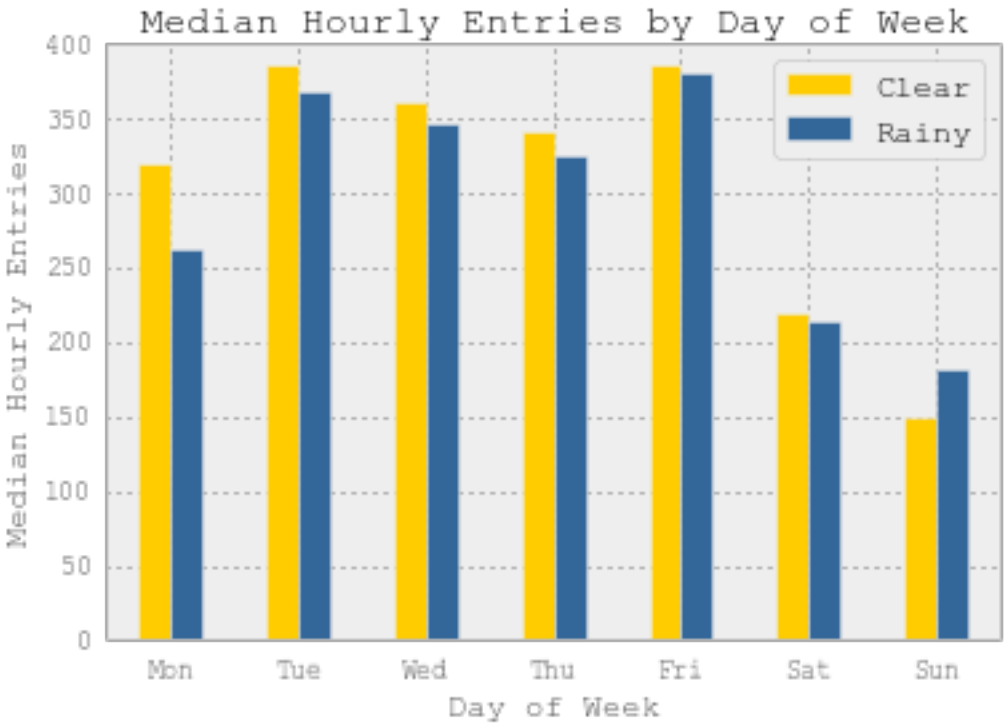

Figure 7: Median of hourly turnstile entries by day of week and rain conditions

5. Reflection
    5.1. Please discuss potential shortcomings of the methods of your analysis including dataset or analysis such as linear regression model or statistical test.

    Looking at a plot of linear regression residuals by dataset entry, there is a clear pattern that emerges. It appears the accuracy of the model is impacted by specific times of day – likely the morning rush hour. Additional reflection on what drives ridership early in the morning might produce better results.

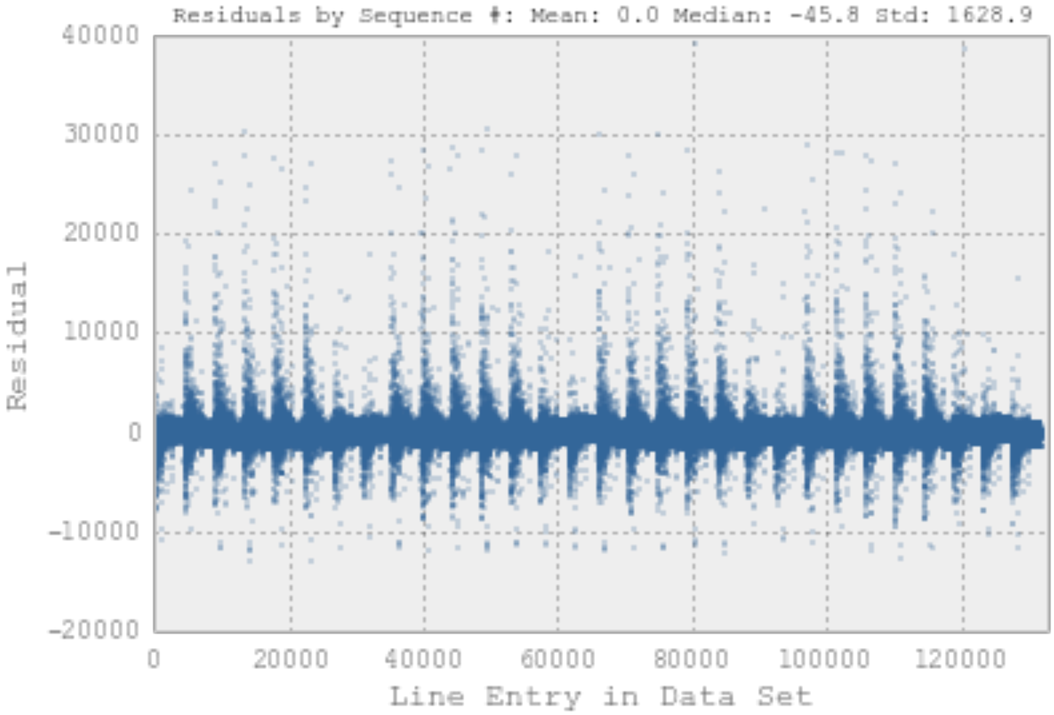

Figure 8 Residual values plotted by sequence number in data set demonstrate that the accuracy of model possibly impacted by specific times of day. (Data set is sequenced by time.)

In addition, the data set is only for the month of May during which average temperatures ranged from 55 to 78 degrees Fahrenheit. While that it typical for New York City in May, it certainly is not the typical weather pattern in December. Ridership on rainy days where the temperature hovers at the freezing point could vary dramatically compared to a typical day in May. Ideally, a data sampling for at least a full year would provide a better means to model ridership.

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

I was curious to see how ridership was impacted with the busiest stations. I selected the top quartile of entries per hour, and compared medians on rainy days versus cloud days. Rain on the weekends definitely drove ridership up during busy times.

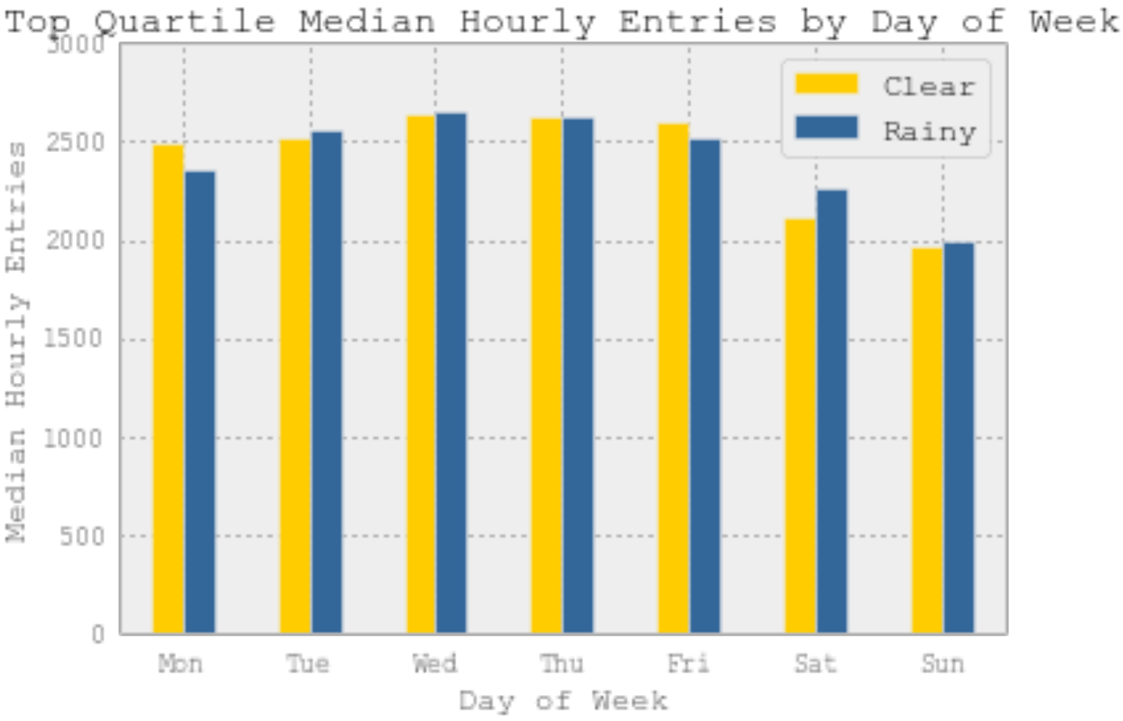

Figure 9 Top quartiles of busiest entries - comparing the median of hourly turnstile entries by day of week and weather. Busiest stations see increased ridership during the weekend during rainy conditions.

6. References used

6.1. Udacity Office Hours were an excellent source of additional information for me. They were especially helpful in understanding residuals and Mann-Whitney.

6.2. "Interpreting results: Mann-Whitney test" http://www.graphpad.com/guides/prism/6/statistics/index.htm?how_the_mann-whitney_test_works.htm

6.3. "Are the model residuals well behaved?" http://www.itl.nist.gov/div898/handbook/pri/section2/pri24.htm

6.4. "Working with Data Frames" by Greg Reda http://www.gregreda.com/2013/10/26/working-with-pandas-dataframes/

6.5. Ignatio's iPython Notebook presented during office hours. http://nbviewer.ipython.org/url/www.alma.cl/~itoledo/Presentation1.ipynb