Analyzing the NYC Subway Dataset

Section 0. References

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

Section 1. Statistical Test

1.1: The statistical test used to analyze the NYC subway data is the Mann-Whitney U test. This is a test of the null hypothesis that two samples come from the same population against the alternative hypothesis that they do not. A two-tailed test was used. The p-critical value is .05. My p-value was 0.0498(.0249 * 2 for two tailed test)

1.2: This statistical test is applicable to the dataset because the data is not normally distributed as show in the histograms produced in the exercise. The Mann-Whitney U test is a non-parametric test and makes no assumption about distribution.

1.3: The results of the from the Mann-Whitney U test are as follows.
with_rain_mean = 1105.4464
without_rain_mean = 1090.2788
P-value = .0499

1.4: Since our p-value is less than the p-critical value, we can reject the null hypothesis at the 5% significance level. Our p-value represents the probability that the difference observed was due merely to chance. This result was close, however, and will take further analysis to prove the hypothesis that on average there are more subway riders during the rain.

Section 2. Linear regression

2.1: The approach I used to compute my regression model was A. OLS using Statsmodels.

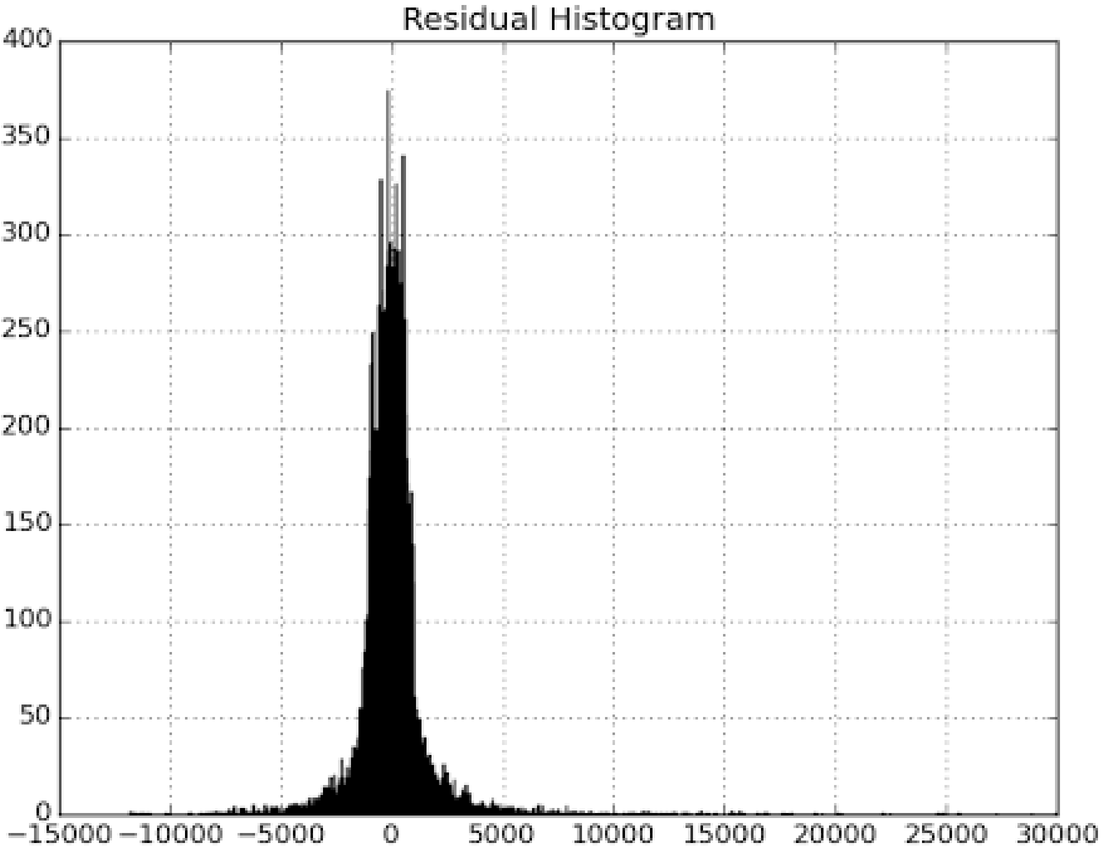
2.2: The features used in my model were as follows including a default dummy variable for unit.
Rain – Indicator (0 or 1) if rain occurred within the calendar day at the location.
Hour – Hour of the timestamp from TIMEn. Truncated rather than rounded.
Meantempi – Daily average of tempi for the location.
Unit – Remote unit that collects turnstile information.

2.3: I selected the prior features for the following reasons.
Rain – I figured people are less likely to walk when it is raining outside.
Hour – I figured there are certain hours pertaining to the work day when people ride the subway.
Meantempi – I figured when it is a really hot day people will be less likely to walk.
Unit – I figured some stations are busier than others and this would help predict ridership.

2.4: The parameters are as follows.
Rain – Coefficient of 43.46
Hour - Coefficient of 65.34
Meantempi - Coefficient of -10.71

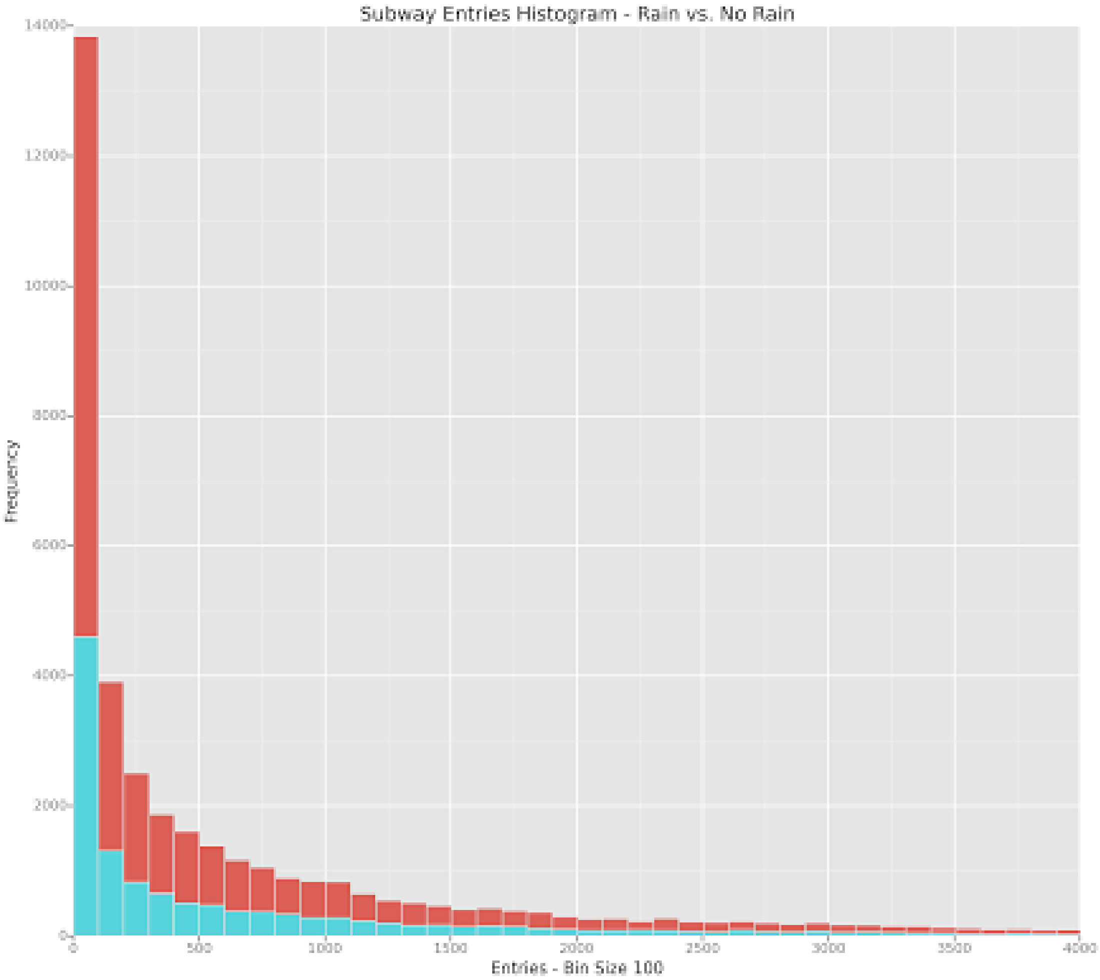
2.5: My models coefficient of determination is .4792

2.6: The R^2 of .4792 means that 47.92% of the variance in the dependent variable is explained by my model. Given this low R^2 value, I do not believe this model is appropriate to predict ridership for the dataset with a high level of certainty. Further analysis of the residuals can be found in the graph below. As seen in the graph there are long tails which suggest there are some very large residuals. This adds further doubt to the validity of our linear model. Spikes can also be seen visually suggesting a lack of normality and inappropriateness of a linear model.



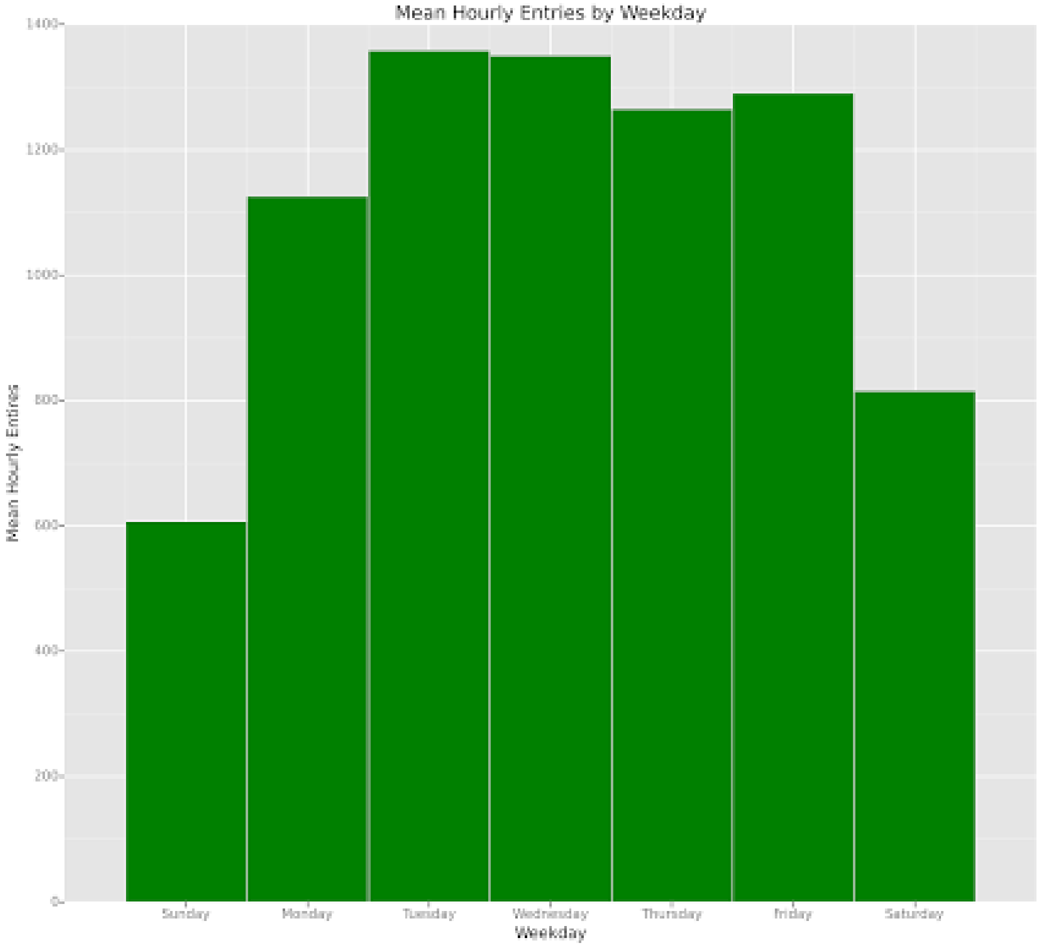
Section 3. Visualization

3.1:



This histogram represents the volume of ridership on the x-axis in bins of 100 and the frequency of occurrence on the y-axis. It is stacked to show the occurrences for rain(blue) vs. no rain(red). A key insight is that the data is not normally distributed. The majority of the records in our data consist of low volume ridership. There is also far more records for non-rainy days then there are for rainy days.

3.2:



This bar graph represents day of the week on the x-axis and average number of hourly entries on the y-axis. A key insight is that there tends to be more ridership on the weekdays then on the weekends. This makes sense intuitively since there will be more traveling for work on the weekdays.

Section 4. Conclusion

4.1: From my analysis and interpretation of the data, more people ride the NYC subway when it is raining than when it is not raining. This makes sense intuitively because when it is raining people will be less likely to walk for fear of getting wet. This leaves taxies and the subway. It is my understanding that taxies are harder to flag down when it is raining, so that leaves subway as the preferred method of transportation.

4.2: Firstly, my statistical Mann-Whitney U test allowed me to reject the null hypothesis that both data sets are statistically the same at the .05 significance level. This means that I am 95% confident that there is a difference between the two data sets. My p-value was very close to the critical p-value, however, and this alone is not enough proof to justify the conclusion that there are more subway riders when it rains. Although, it is a great starting point to further explore what seems intuitively correct though.

Secondly, when applying my linear regression I found that the coefficient for rain was a positive value. This means that there is a positive relationship on average between rain and the amount of people who ride the NYC subway. This, in combination with my Mann-Whitney U test, are the foundations of my conclusion.

Section 5. Reflection

5.1: The first thing that comes to my mind about the data is that it is all during the month of May, which is known to be one of the milder weather periods of the year. This can further be seen with the complete lack of thunderstorms in the data. In my opinion, data that spans a full year's weather cycle would be more useful. It is possible that during large thunderstorms the impact of rain on ridership would be more dramatic than during the light rain that happens in the month of May. Another issue might be the number of people who jump gates and are thus not reflected in the data. This is probably too small to have an impact though.

A shortcoming of my analysis is that it doesn't sufficiently isolate the rain variable's impact on ridership. I feel it would have been more useful to isolate one subway station ID during one time of the day and then compare ridership with rain vs. ridership without rain. As is, it is possible that rain only happened at times when ridership would have been high anyways. This could lead to false conclusions.