Bindu Paul: Project 2

Section 0. 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.

Section 1. Statistical Test

1.1. 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. 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. 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. 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. I used the OLS using Statsmodels

2.2. 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.

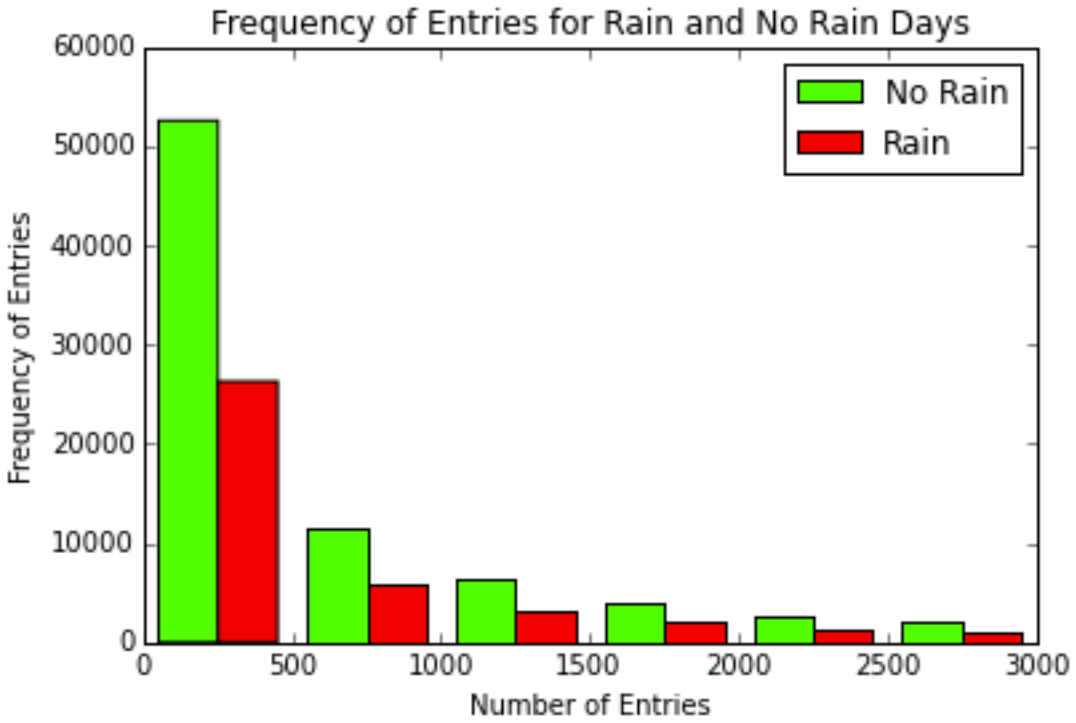
2.3. I used “Hour” with and without “rain” features because I wanted to analyze how much of an impact rain had on the R^2 value. I used the dummy variable because it increased my R^2 value drastically.

2.4. Hour = 59.483324; rain = 16.28

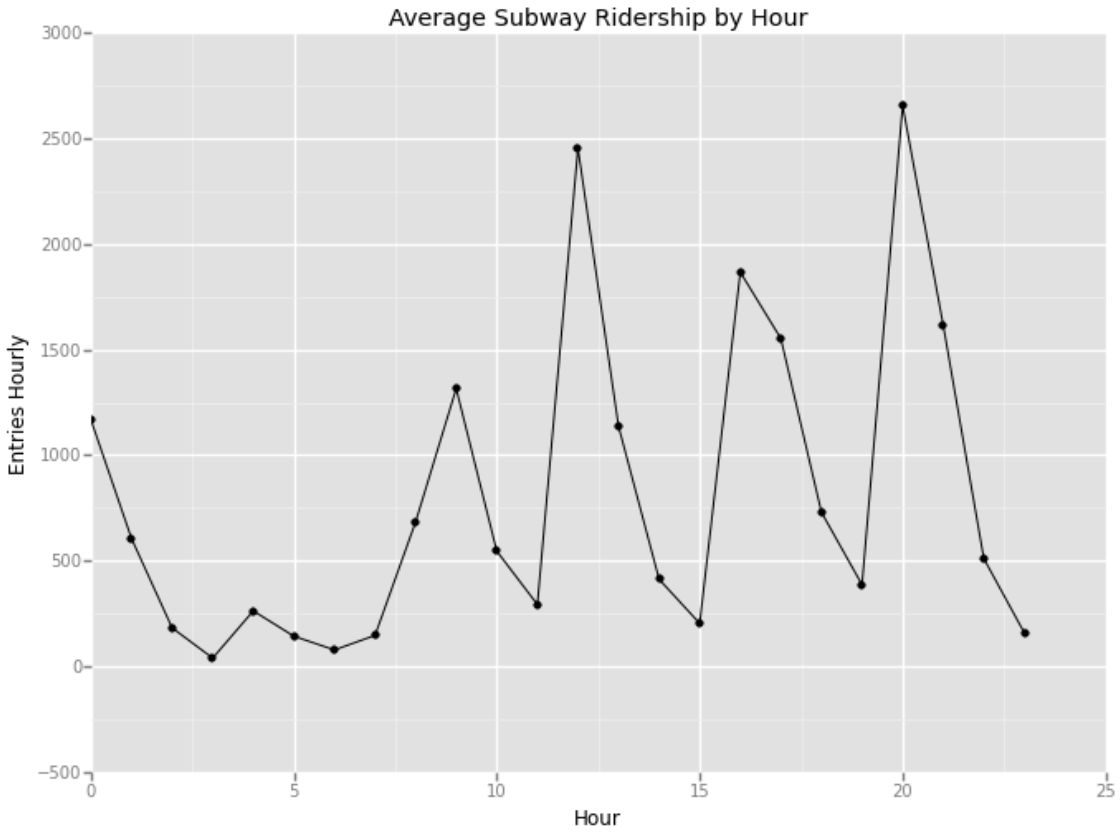
2.5. R^2 = 0.45756 with rain, R^2 = 0.45754 without rain

2.6. 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



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



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

Section 4. Conclusion

4.1. 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. 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. 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.