P1: Analyzing the NYC Subway Dataset

QUESTION – Do more people ride the NYC Subway when it is raining vs not raining?

1 STATISTICAL TEST

The 2 sided - Welch t-test was used to compare the means of the two groups we were looking at: rainy days and non-rainy days. A two-tail test was used, as the question being addressed is looking for any significant difference whether positive or negative (ridership could be significantly higher or significantly lower on rainy days compared to non-rainy days).

This particular test was used because the variances of two distributions is unequal (they have unequal stds therefore unequal variances as well since variance = std^2) and also the underlying distribution does not need to be normal. This test is used to check if two populations have equal means based on the sample means (exactly what are trying to determine). Also the normal distribution assumption for a regular t-test can be relaxed under this Welch t-test.

Ho - Null Hypothesis: muRain - muNoRain = 0

There is no significant difference in ridership on rainy vs non-rainy days

Ha - Alternative Hypothesis: muRain - muNoRain != 0

There is a statistically significant difference in ridership on rainy vs non-rainy days

MEASURES	**VALUES**
rain group mean ridership	2028
norain group mean ridership	1845
difference between group means	182
p-value	4.64 e-07

On average, more people ride do ride the NYC subway on rainy days vs non-rainy days. The difference between the means of 182 is statistically significant; the p-value = 4.64 e-07 is less than alpha = 0.05 (5% significance level), which means we can reject the null hypothesis (that there is no difference between the rainy day and non-rainy day groups).

2 LINEAR REGRESSION

I used the OLS (Ordinary Least Squares) using Statsmodels to compute and produce predictions for ***ENTRIESn_hourly***.

These are the features used in the model I chose:

hour, weekday, rain, fog, tempi
dummy variables - ***station*** & ***UNIT***

hour was chosen because it was highly linearly correlated with ***ENTRIESn_hourly*** relative to the other features and intuitively because ridership would likely vary based on the time of day as more people would probably ride the subway to and from work from 7 to 9 am and 4 to 6 pm.

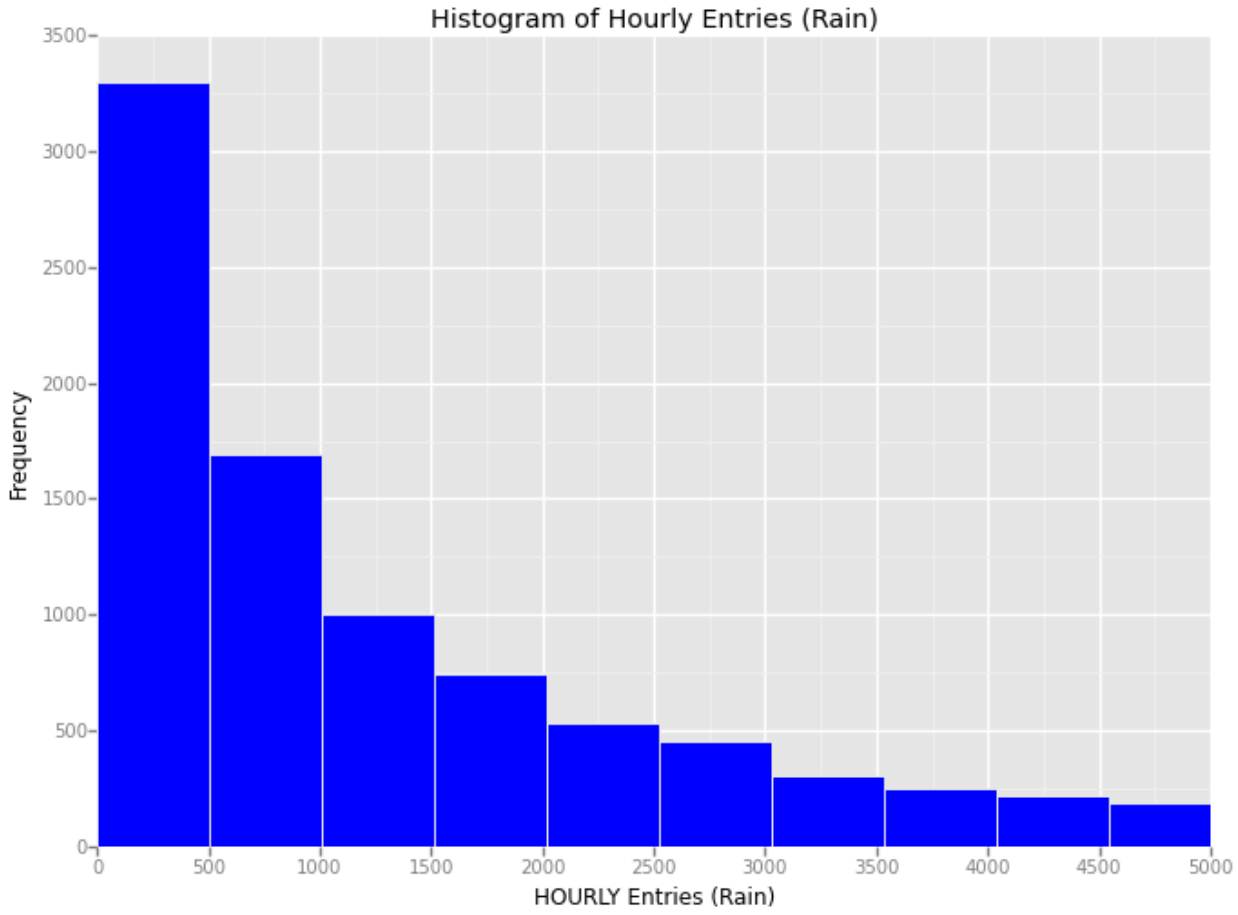
weekday was chosen because I would expect ridership to be higher on the weekdays when most people are commuting to work as opposed to the weekend when they are more likely to be at home.

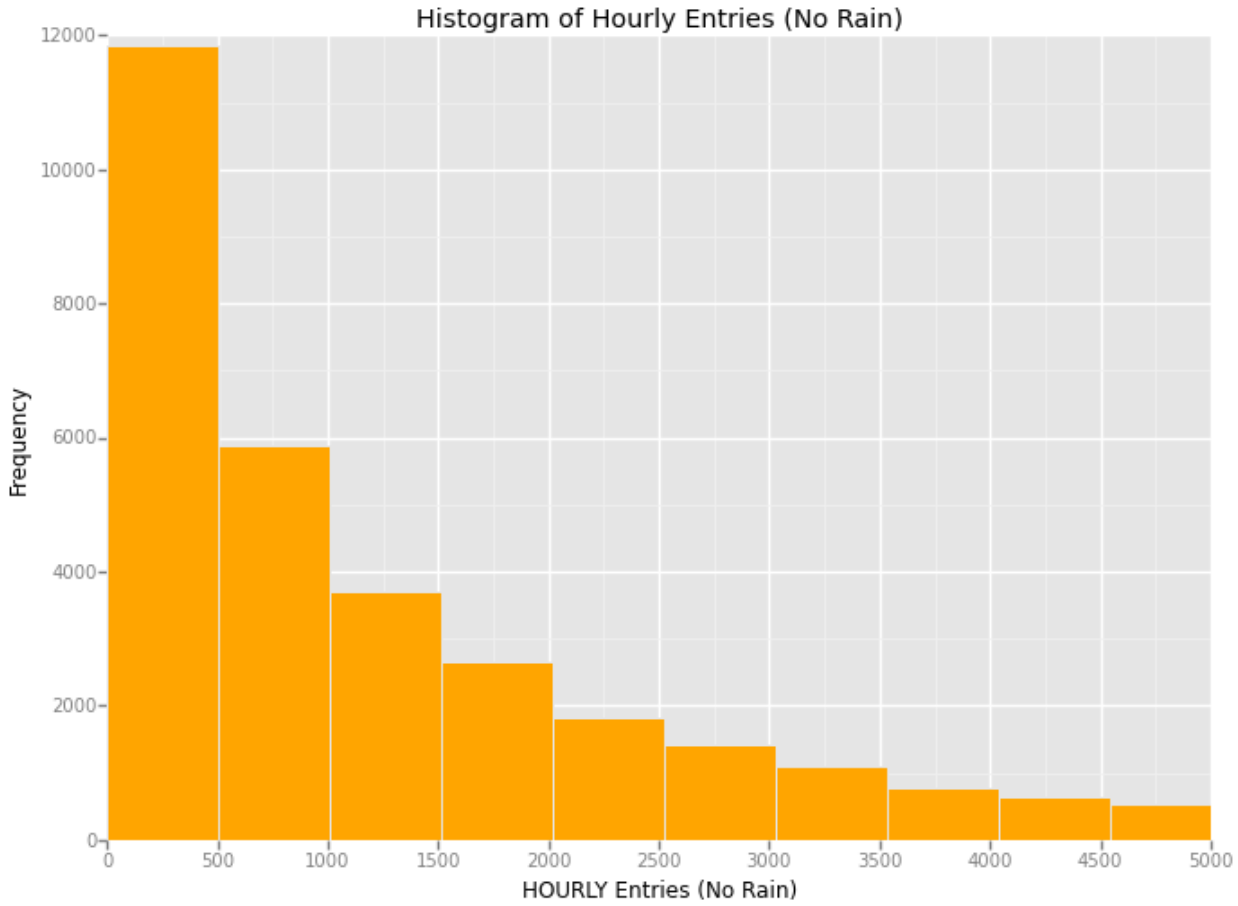
station and ***UNIT*** were transformed to dummy variables because larger/more trafficked stations and units would be expected to have more riders as opposed to smaller stations or units.

rain and ***fog*** were chosen under the assumption that New Yorkers ride the subway more during bad weather instead of walking outside. Same with ***tempi***, an assumption was made that people prefer to walk outside instead when the temperature was higher.

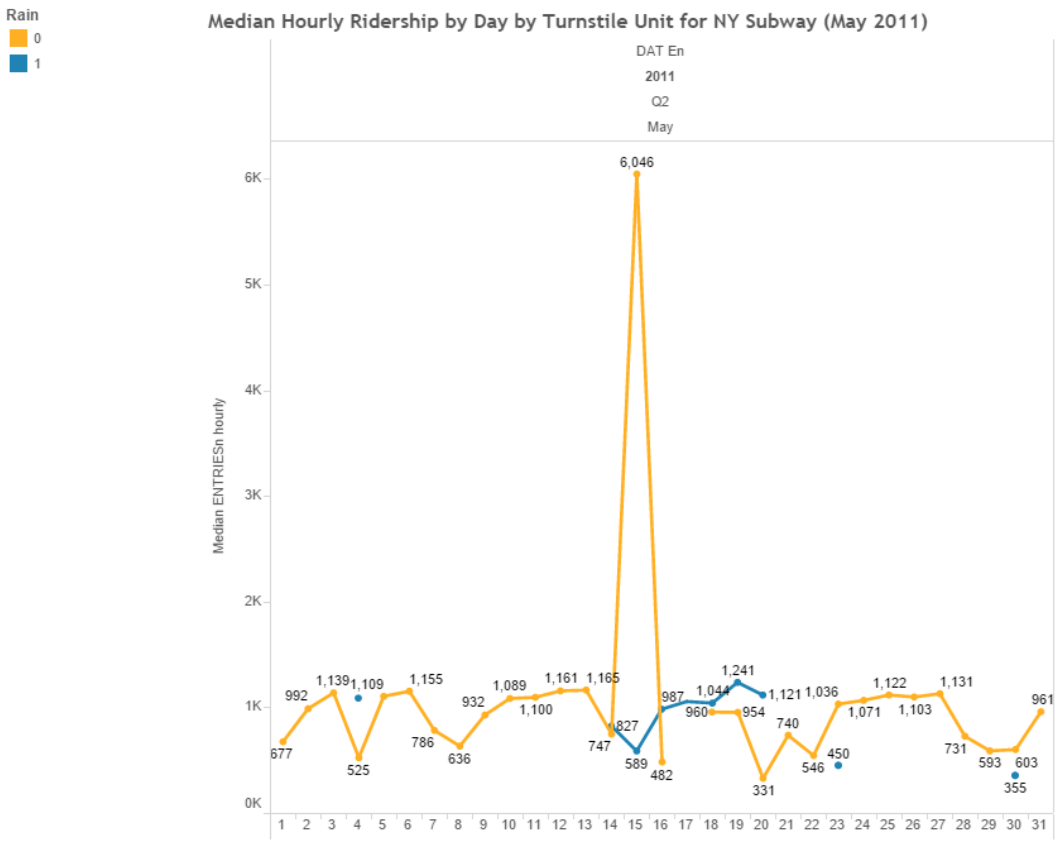
However, after computing R2(r-square) I found these three "weather" variables didn't increase the predictive power of the model as R2 did not improve when adding them to the model.

3 VISUALIZATION





From both histograms, it appears the shape of the distribution for rain any non-rainy days is nearly identical (long tail skewed to the right). The differences in height can be attributed to the fact that the majority of the data (over 3/4) was collected on non-rainy days. If we had an equal amount of data (15 rainy days in a month and 15 non-rainy days) then the height difference would be something to investigate. From the histograms, it doesn't appear there is any major difference between rainy and non-rainy days.



This second visualization was created using Tableau Public (which connects to Excel and csv files). It shows the daily median Entries on a Unit basis. The days that it rained are colored in blue. Some interesting insights from this visualization are:

- There are days when it rained consecutively from the 14th to the 20th and then there are 3 data points that could reflect the fact that it rained for a brief part of the day (1,109, 450, and 355).
- On the 15th of May, there was a huge spike in ridership, perhaps there was a concert or some other big event on that date that increased ridership more than usual. There was a Yankees vs Red Sox game on this day at Yankee stadium which could help explain this.
- Rainy days seem to follow the same general trend in terms of median ridership as typical non-rainy days. Median ridership doesn't increase past 1300 except for one outlier (6,046).

4 CONCLUSION

From my interpretation of the data, more people on average do ride the subway when it is raining vs when it is not raining.

However, although the average difference between rainy and non-rainy days is statistically significant as per the statistical Mann-Whitney t-test, for all practical purposes there is no large

difference in ridership on non-rainy vs rainy days. This also makes sense intuitively as most of the time people take the subway to save commuting time and money. It wouldn't make much sense to take a cab or walk if you usually take the subway because it is a sunny day(non-rainy). Most regular subway riders would not change their commuting habits due to the weather (except snow or extreme weather for example). It would be interesting to see what kind data is available in winter months (we were given only May's data). Also many people don't even check the weather forecast before leaving or planning their commute so they have already planned that they will be taking the subway.

5 REFLECTION

One of the major shortcomings of the dataset is it only provides data from a single month (May) in 2011. If the provided data was from different months or years, we could see if May was a typical month or perhaps there is some sort of seasonality (maybe ridership is higher in the winter as opposed to the summer when people are on vacation and the weather is sunny outside).

The model that was built might not perform very well on new test data because it is fitted very well to the training data that was provided. If we had data from other months and years there would be other time periods to test the model. Also it is difficult to generalize these results to the population (all NYC subway riders) as it is not a random sample of data that was taken (random sample would have had data from random days in the year).

REFERENCES

Welch's t-test
http://en.wikipedia.org/wiki/Welch%27s_t_test

r2, a measure of goodness-of-fit of linear regression
http://www.graphpad.com/guides/prism/6/curve-fitting/index.htm?r2_ameasureofgoodness_of_fitoflinearregression.htm

Piazza post -Plotting two separate histograms with ggplot
https://piazza.com/class/i23uptiifb6194?cid=109

May 15, 2011 in New York
http://scores.espn.go.com/mlb/boxscore?gameId=310515110