P1: Analyzing the NYC Subway Dataset

Is the average NYC Subway Ridership statistically different on rainy days vs non-rainy days?

1 STATISTICAL TEST

The Mann-Whitney U 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).

The Mann-Whitney test is appropriate in this case because the underlying distributions are not normally distributed (both the rainy and non-rainy distributions have a long tailed versus a bell curved shape). This is referred to as a non-parametric test, a statistical test that does not assume our data is drawn from any particular underlying probability distribution (such as a normal distribution).

Concretely, in this case we are using this test to determine if the two populations (rainy and non-rainy) have equal means based on the sample means computed from the provided data.

Null Hypothesis: There is **no** statistically significant difference in average ridership on rainy vs non-rainy days. The two populations are the same.

Alternative Hypothesis: There **is** a statistically significant difference in average 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	5.48 e-06

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 = 5.48 e-06 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 I considered to use in my final model:

UNIT was transformed to dummy variables because more trafficked units (or in heavily populated areas) would be expected to have more riders as opposed to less trafficked units (or units in less populated areas). Station was not needed because it is already strongly correlated with UNIT (if a station has more riders all the units in that station would also have more riders and vice versa).

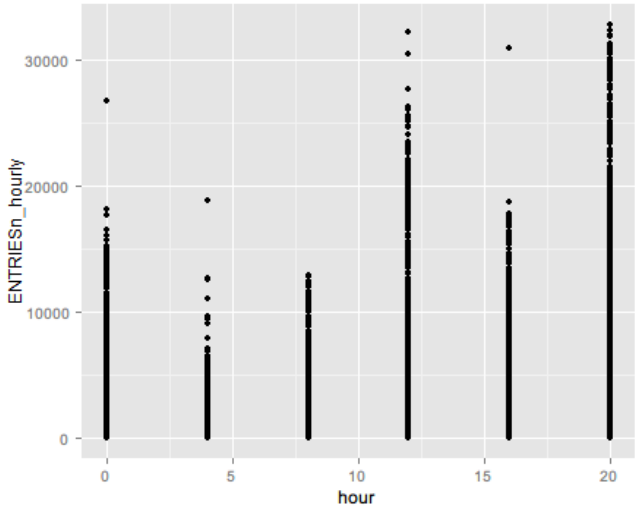
hour was chosen because it was highly correlated with ***ENTRIESn_hourly*** compared with other features (correlation coefficient, r, was 0.28 - 2nd highest correlation coefficient after ***EXITSn_hourly*** = 0.64) and intuitively because ridership would likely vary based on the time of day as more people would probably ride the subway to and from work from 7 to 9 am and 4 to 6 pm (for example).

weekday was initially 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.

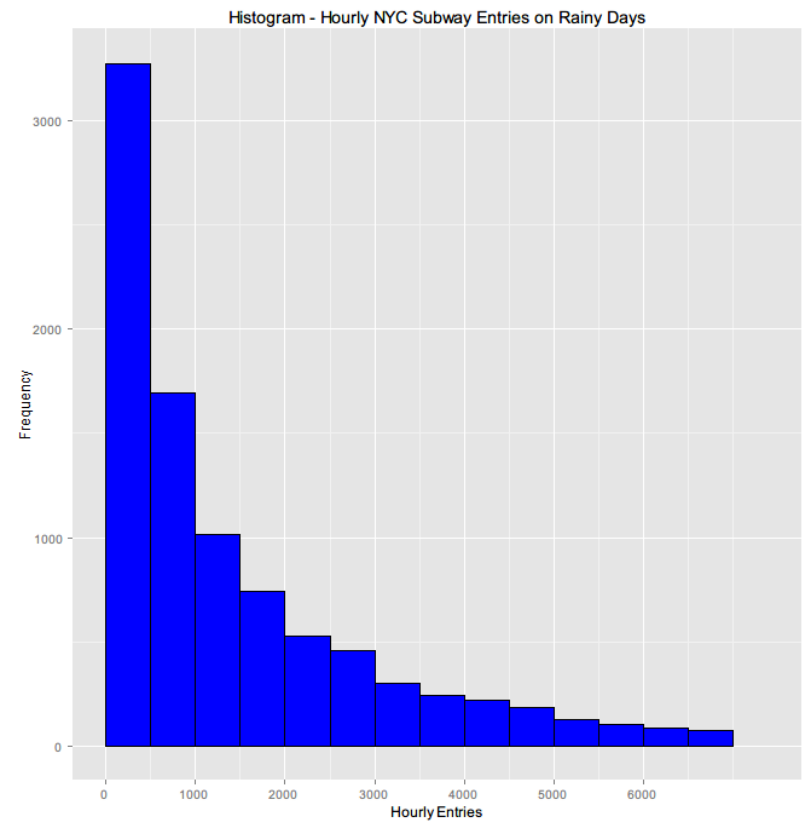
rain and ***fog*** were initially chosen under the assumption that New Yorkers ride the subway more during bad weather instead of walking outside. Same with ***tempi***, an assumption was made that people prefer to walk outside instead when the temperature was higher.

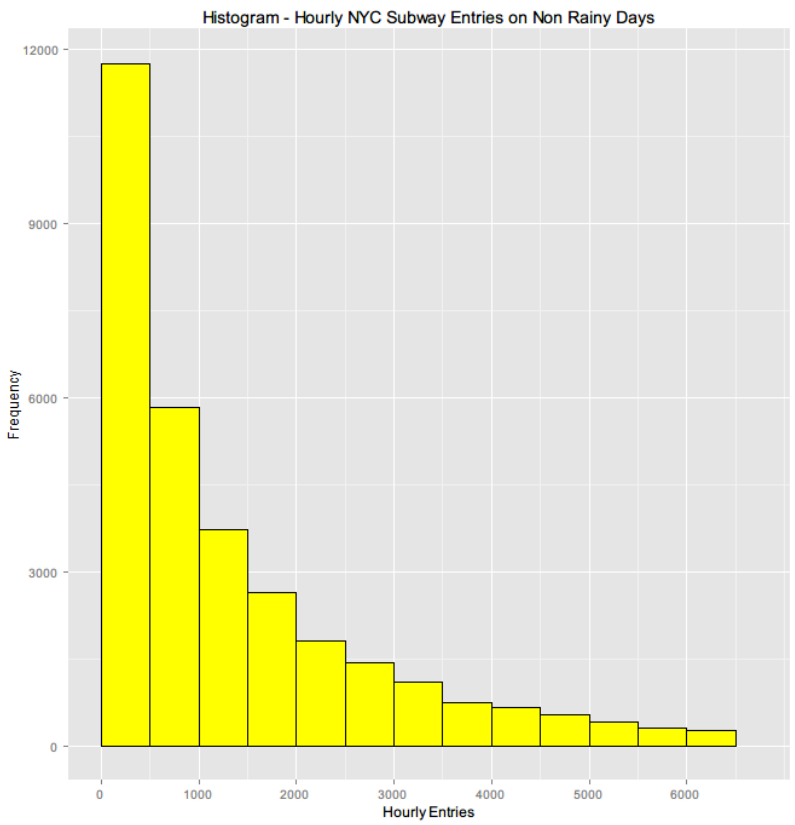
However, after computing R^2 I found these three "weather" variables did not help increase the predictive power of the model as R^2 did not improve when adding them to the model. Adding weekday alone or combined with hour also did not have a higher R^2.

The final value of R^2 was **0.61754**. The chosen features were **UNIT** (dummy variables) and **hour**. Just under 62% of the variation of **ENTRIESn_hourly** can be explained by this model with these features. This does not represent a strong linear relationship between the dependent variables and the independent variable; the linear model is not appropriate for this data set. This is further supported by the fact that hour of day and ridership do not have a linear relationship as shown below.



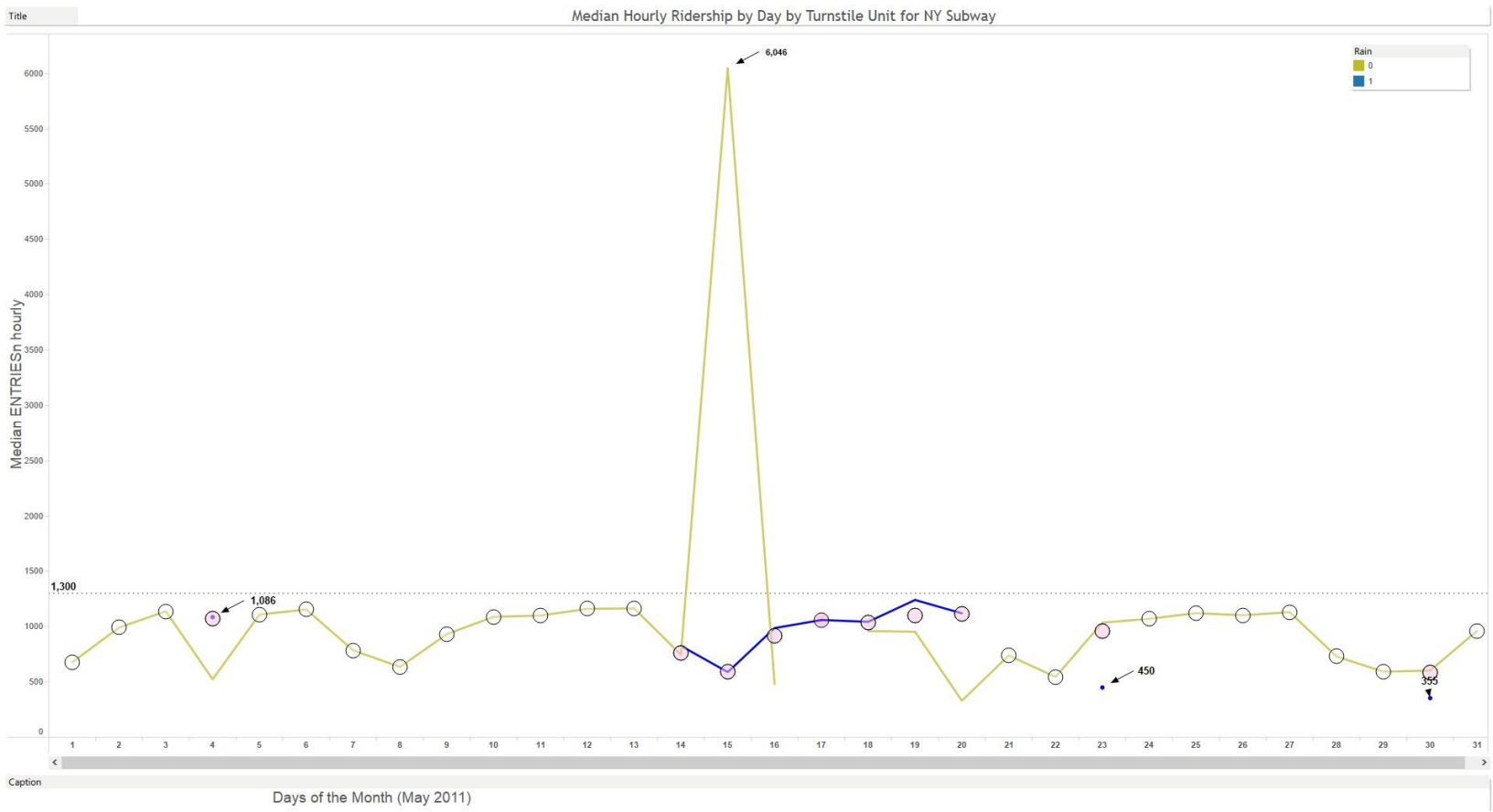
3 VISUALIZATIONS





From both histograms (created in R using ggplot2), it appears the shape of the distribution for rainy and 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.

*Note for both histograms, the 95th quartile was chosen as the cutoff point for the upper limit of the x-axis (just beyond 6000). This is enough data to show the shape of the distribution.



This second visualization was created using Tableau. It shows the daily **median** entries on a unit basis. The days that it rained are blue and that it didn't rain are yellow. The circles show the median of the original dataset (rainy and non-rainy data combined). 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 only (1,086, 450, and 355).
- In the middle of the graph, you can see that there was a huge spike in ridership (6,046); perhaps there was some big event on that date that increased ridership more than usual (a possible confounding variable). I actually found there was a baseball against the Boston Red Sox on May 15th, 2011 at Yankee stadium that could explain this outlier.
- 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) mentioned previously.
- The circles (original dataset with no segregation rainy/non-rainy separation) show the robustness of using the median as the statistic as opposed to the mean as it is not effected by outliers like the mean is.

4 CONCLUSION

From my interpretation of the data, more people on average do ride the subway in New York when it is raining vs when it is not raining. This is supported by the Mann-Whitney U Test results which show that the 182 difference in averages is statistically significant. However, the difference is not enough to say there is any practical difference or enough to predict higher or lower ridership based on the presence of rain. Supporting this conclusion, it is quite telling that adding rain to the predictive model did not increase R^2. If rain did have a large impact on ridership, we would have expected it to definitely be part of the final model that was chosen. Finally, the comparative histograms show the distributions are nearly identical in shape except for the y-axis height (frequency) which can be attributed to the fact that there were less rainy days than non-rainy days in May 2011.

5 REFLECTION

One of the major shortcomings of the dataset is it only provides data from a single month (May 2011). If the provided data was from different months and 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). Other confounding variables such as major sports events, concerts, and subway closures are not mentioned which could also affect the ridership of the subway.

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 and especially the dummy variables that were created from UNIT. If we had additional data from other months and years there would be other time periods to test the model to see if the model was more robust. Also it is difficult to generalize these results to the entire population (all NYC subway riders on any day of the year) as it is not a random sample of data that was taken at different times of the year.

REFERENCES

Piazza post - Problems with Mann-Whitney U test - no p-values returned by improved dataset
https://piazza.com/class/i23uptiifb6194?cid=517

Mann Whitney U test
http://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test

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