Analyzing the NYC Subway Dataset

1. Statistical test

The Mann-Whitney U-test was used to compare the mean values of the two groups that we considered: rainy days and not rainy days. A test with two tails was used, since the question in question seeks any significant difference, positive or negative (the number of passengers may be significantly higher or significantly lower on rainy days compared to rainy days). In this case, the Mann-Whitney test is appropriate, since base distributions are usually not distributed (both rain and non-rain have the shape of a long tail, rather than a bell-shaped shape). This is called a non-parametric test, a statistical test that does not assume that our data is taken from any particular basic probability distribution (such as the normal distribution).

Specifically, in this case, we use this test to determine whether the two populations (rainy and not rainy) have equal averages based on sample averages calculated from the data provided.

Zero hypothesis: there is no statistically significant difference in the average number of passengers in rainy and inclement weather. Two populations are the same.

Alternative hypothesis: there is a statistically significant difference in the average number of passengers in rainy and not rainy weather.

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 and more people ride the subway in New York on rainy and not rainy days. The difference between the average of 182 is statistically significant; p = 5.48 e-06 is less than alpha = 0.05 (5% significance level), which means that we can reject the null hypothesis (that there are no differences between groups of rainy and rainy days).

2. Linear regression

I used OLS (Ordinary Least Squares) using Statsmodels to calculate and generate forecasts for ENTRIESn_hourly.

Here are the features I decided to use in my final model:

UNIT has been transformed into dummy variables because it is expected that more units of trafficking (or in densely populated areas) will have more riders than units with fewer victims of trafficking (or units in less populated areas). The station was not needed, because it already strongly correlates with UNIT (if there are more racers at the station, all units at this station will also have more racers, and vice versa).

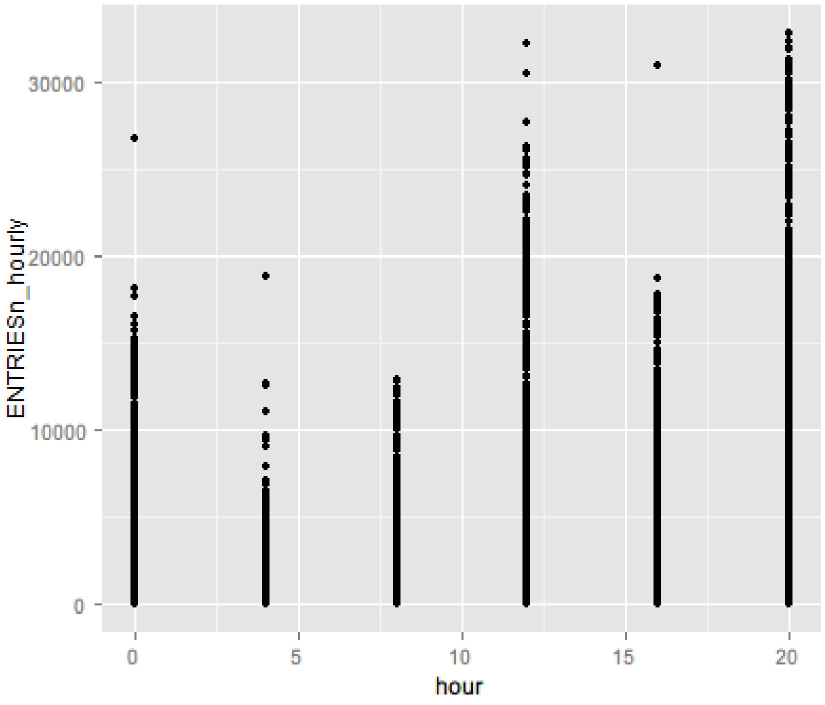
The hour was chosen because it was strongly correlated with ENTRIESn_hourly compared to other functions (the correlation coefficient r was 0.28 - the 2nd highest correlation coefficient after EXITSn_hourly = 0.64) and intuitively, because the passenger traffic is likely to be change depending on the time of day, as more people are likely to travel by metro to and from work from 7 to 9 and from 16 to 18 (for example).

The weekday was originally chosen because I expected the number of people traveling to be higher on weekdays when most people go to work and not on weekends when they are more likely to be at home.

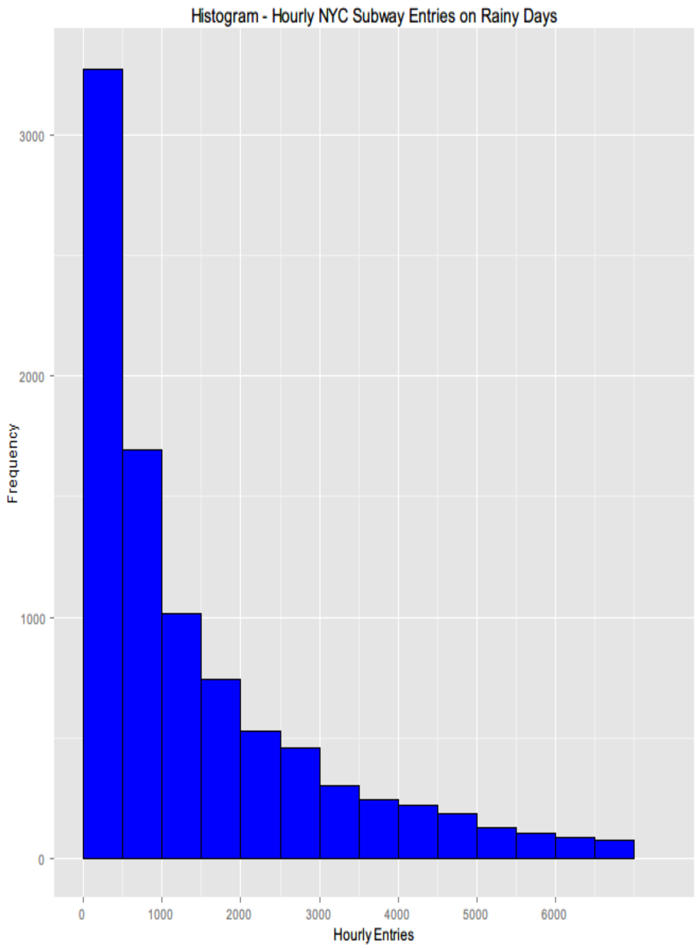
Rain and fog were originally chosen on the assumption that New Yorkers take the subway more in bad weather, rather than go outside. As with the pace, it was suggested that people prefer to walk outside when the temperature was higher.

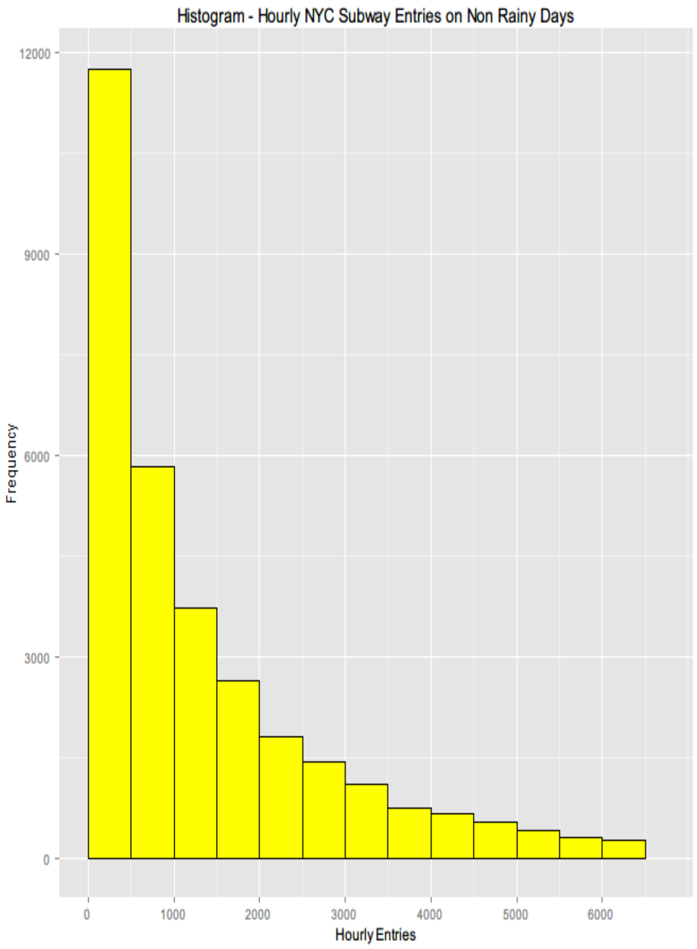
However, after R^2 was calculated, the variables did not help increase the predictable power of the models, since R^2 did not improve when they were added to the model. Add one day of the week.

The final R^2 value was 0.61754. The selected characteristics were UNIT (dummy variables) and hour. These features of this model can be explained by a little less than 62% of the ENTRIESn_hourly variation. This does not represent a strong linear relationship between the dependent variables and the independent variable; linear model is not suitable for this data set. This is also confirmed by the fact that the time of day and the number of passengers do not have a linear relationship, as shown below.



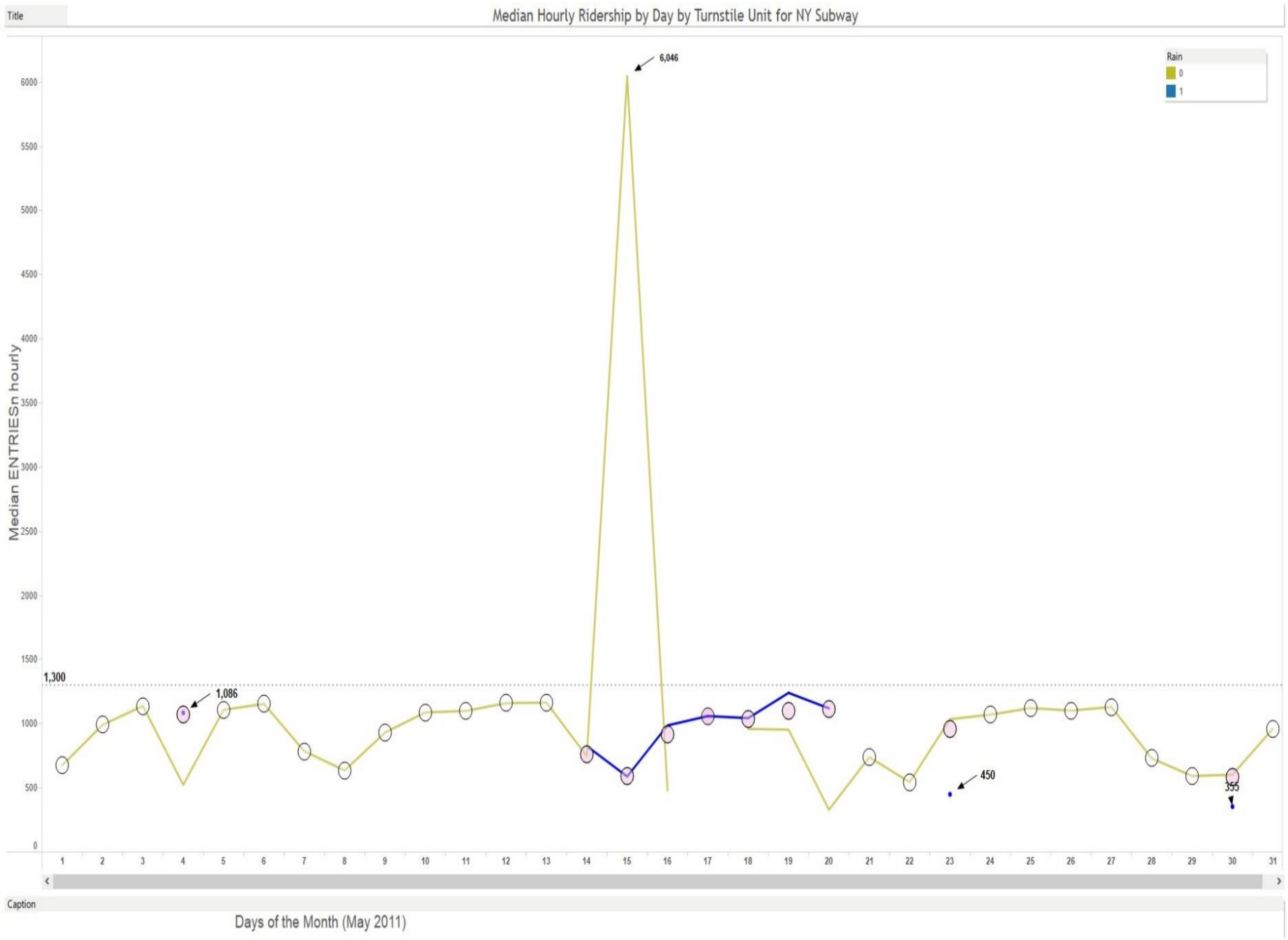
3. Visualizations





From both histograms (created in R using ggplot2) it can be seen that the form of distribution for rainy and rainy days is almost identical (the long tail is tilted to the right). Differences in height can be explained by the fact that most of the data (over 3/4) was collected on inclement days. If we had the same amount of data (15 rainy days in a month and 15 not rainy days), then the difference in height could be investigated. Judging by the histograms, there is not much difference between rainy and rainy days.[i]

Please note that for both histograms the 95th quartile was chosen as the cut-off point for the upper limit of the X axis (slightly more than 6000). This is enough data to show the distribution form.



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.

REFERENCES

https://www.dropbox.com/s/1lpoeh2w6px4diu/improved-data set.zip?dl=0
The zip file contains a CSV file used for this project and a PDF file describing variables

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