Analyzing the NYC Subway Dataset

Christopher Velayo

August 9, 2015

Section 0. References

- https://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test
  - Used to determine critical U value
- http://stats.stackexchange.com/questions/70034/mann-whitney-u-test-critical-values-for-very-large-samples
  - Topic which led to using the wikipedia article to determine the critical U value

Section 1. Statistical Test

Statistical test

To analyze this data, I chose to use the Mann-Whitney U test. Based on an exploratory boxplot, days with rain appear to have a slightly higher median and upper quartile. Given this, I ran the test using a one-tail P value. To formally state the hypotheses: the null hypothesis is that the average number of subway entries on non-rainy days is equal to or greater than the average number of subway entries on rainy days; and the alternative hypothesis is theat the average number of subway entries on rainy days is greater than the average number of subway entries on non-rainy days.

According to Wikipedia, for large samples, the distribution of U approaches a normal distribution. Given the standard p-value of .05, and the one-tailed assumption, we would use the z-score to determine the critical U value. For this test the z-score equation is $z = (U - m_U)/\sigma_U$ with $m_U = (n_1 n_2)/2$ and $\sigma_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$. For this dataset, $n_1 = 0$ and $n_2 = 0$. With these values, the critical U value would be 160199770.

Why this test?

Creating histograms of the number of subway entries show that whether or not the day is rainy, the distribution is not normal. Not having a normal distribution means that parametric tests such as the t-test would not be valid as parametric tests assume a normal distribution. The Mann-Whitney U test on the other is a non-parametric test which makes no assumptions about the distribution of the sample.

Results

Results	Values
Mean number of entries during rainy days	2028.2
Mean number of entries during non-rainy days	1845.5
p-value	0.00000274107
U value	153635121

Interpretation

Since the computed p-value is well below the alpha level of .05, we reject the null hypothesis and accept the alternative hypothesis that the number of subway entries on rainy days is significantly greater than the number of subway entries on non-rainy days.

Section 2. Linear Regression

Approach

I used the lm function of R to create the model and calculate the coefficients of theta. As the dataset was small enough, it was computationally feasible to calculate a closed-form model. Had the dataset been larger, I would have sought to use a gradient descent approach to be more efficient computationally.

Features

These were the features I used in the model: rain, day_week, hour, and UNIT. As all of these were catergorical data, their values were converted to dummy variables for purposes of the regression. Unsing dummy variables had for all categorical data led to a significant increase in the R-squared value compared to treating them as numerical data.

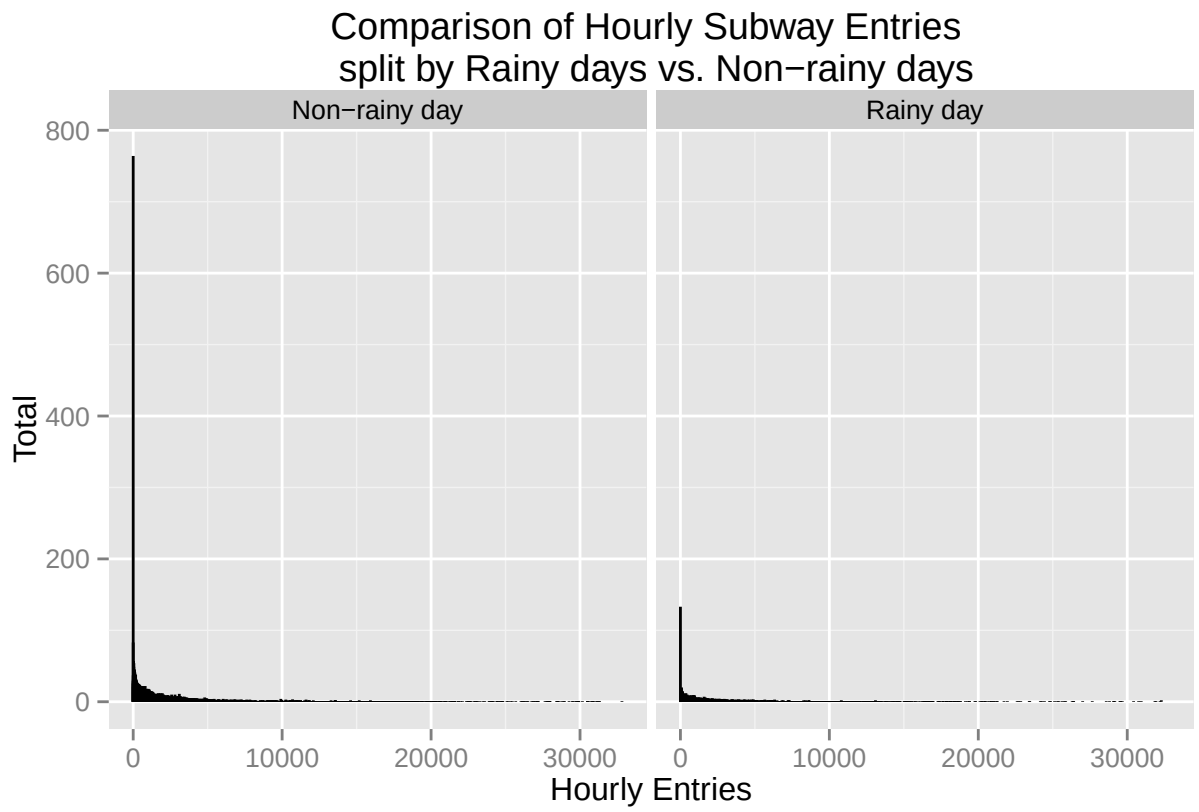
Feature selection criteria

As the purpose of this analysis was to explore the impact of rain on subway riders, several variables were considered to be potential indicators of rain. I explored the effect of using the following variables to indicate rain: rain, conds, precipi, and meanprecipi. None of these variables individually had a very high R-squared. This led me to explore other possible variables. I added day_week, hour and UNIT as I thought that the location of the subway station and the time of day and the day of the week might impact when people decided to use the subway. While other variables were considered, these three showed the biggest impact on the R-squared value. I then tried combining the rain features with these three features. This final selection presented the best balance between an increased R-squared value and simplicity.

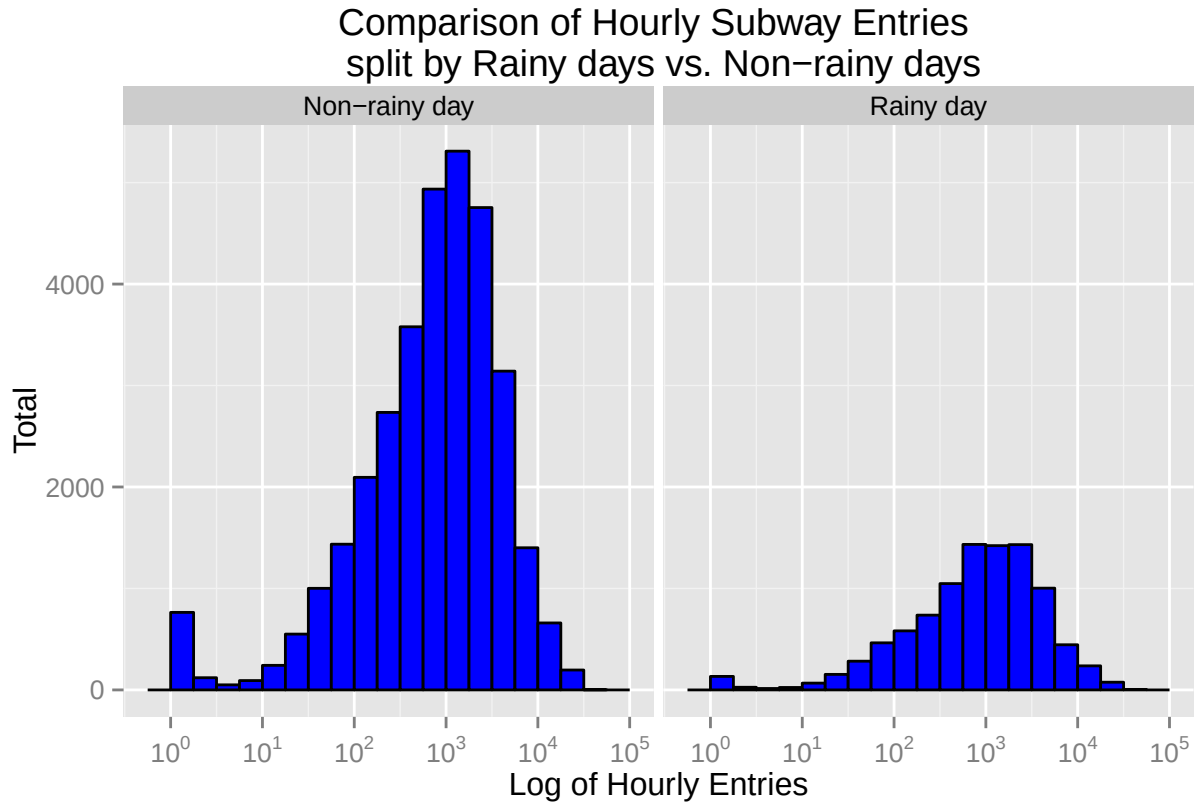
Results and Interpretation

My final model resulted in an R-squared value of 0.5440012. What this means is that this model explains a bit more than half of the variation. Considering that we are attempting to predict human behavior, this R-squared value suggests that this model is probably either appropriate, especially considering that a model with every feature has a similar R-squared value.

Section 3. Visualization

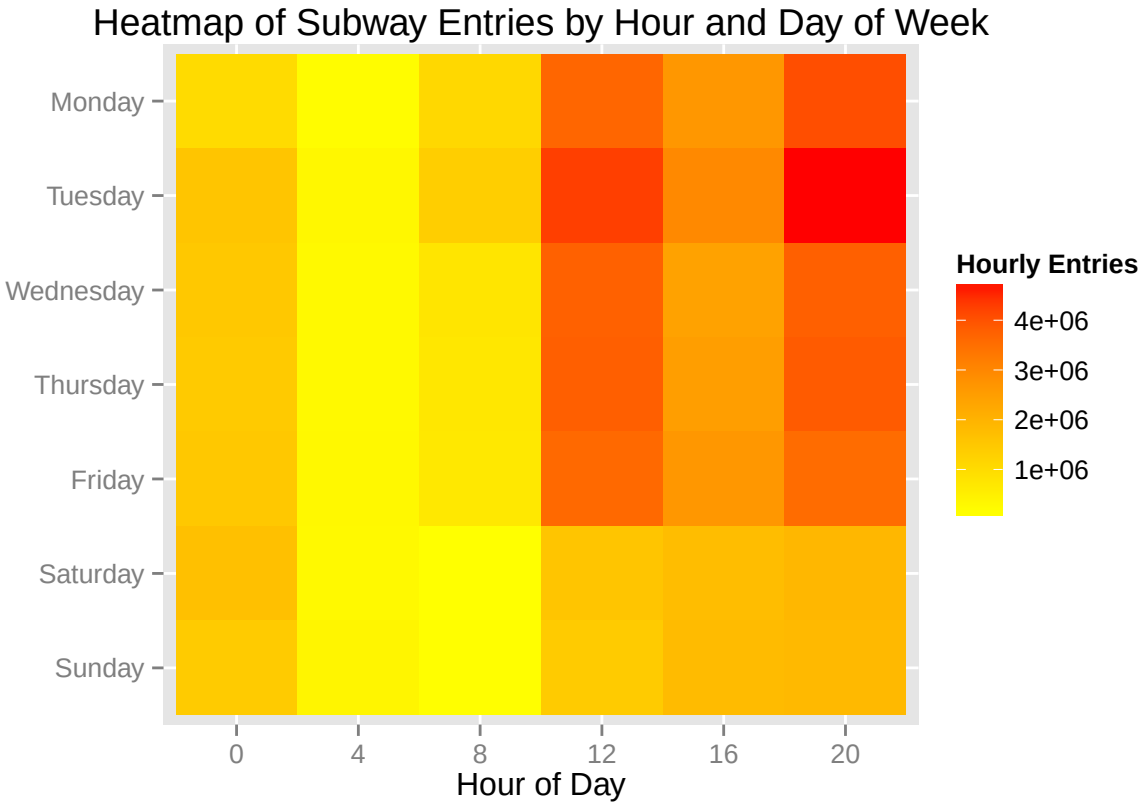


Looking at the plots above, one can see that at first glance, the two distributions appear to have fairly similar shapes. Of course, since there are many more entries marked as non-rainy compared to rainy entries, the scales are very different. At the same time, due to the fact that both distributions are long-tailed rather than normally distributed, outliers seem to be expanding the scale of the graph in such a way that makes it difficult to really see the detail of the bulk of the data.

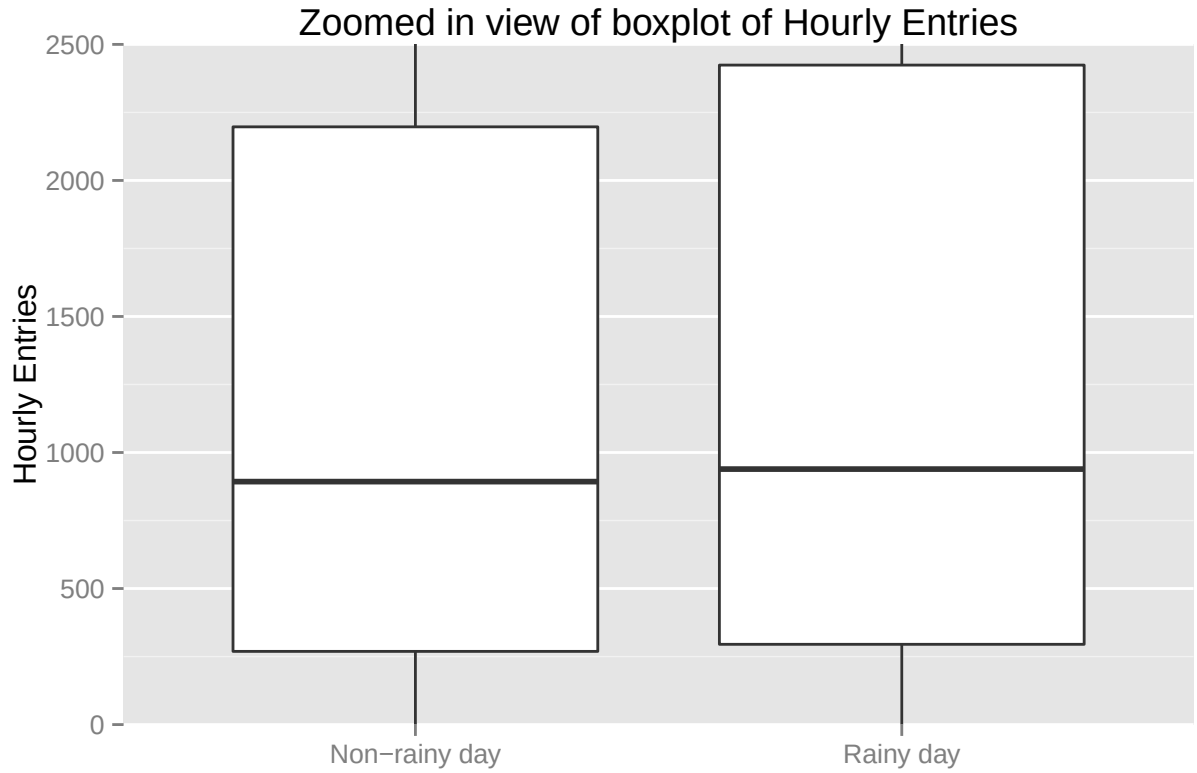


One way of comparing the distribution of the majority of the data and minimize the effect of outliers is to apply a transformation to the data. In this case, I have applied a log_10 transformation to the ENTRIES_nhourly data after adding one to each data point. Adding one prevents the errors that would arise from trying to get the log of 0, and as we are using log_10, the data distribution will not be affected.

Here we see that the bulk of the data shows similar distributions for rainy and non-rainy days, although oddly enough, the rainy day distribution shows a truncated peak. This would be something interesting to explore.



Here we see a distribution of subway entries plotted by the day of week and the hour of day. One can see that generally speaking, the subway is busier during the weekday afternoons, and that the 4AM hour is the time the subways seems to be the least used.



This is the exploratory boxplot referenced earlier. One can see that it is a small difference, but the median number of visits, and the top of the box on rainy days are higher than on non-rainy days. The plot had to be zoomed in because the outliers expanded the scale significantly enough that it was hard to tell any difference. The zoom method was chosen because it did not affect the characteristics of the boxplot the way other zoom methods would have.

Section 4. Conclusion

The answer to the question of whether or not subway ridership increases during rainy days is somewhat complicated based on the data. Based on the Mann-Whitney U test, there is in fact a statistically significant difference in means between ridership on rainy days compared to non-rainy days. Also, because a one-sided test was used, we can also say that ridership during rainy days is statistically higher than on non-rainy days. This is supported by the regression model. The dummy variable for rainy days has a positive coefficient, supporting an increase in ridership on non-rainy days. However, this support is somewhat qualified as the p-value of this coefficient was above .10, meaning the coefficient may not actually be different from 0.

On the other hand, the graphs suggest that the difference in ridership may not be practically significant. The difference between the two medians is only 46. Furthermore, when looking at the regression model, ridership seems to determined far more by the other coefficients rather than by the rainy day variable.

Overall, it appears that all things being equal, more people ride the subway when it is a rainy day than not. However, there are many more factors which could be used to predict ridership.

Section 5. Reflection

One limitation of this study has to do with the fact that the rain variable is a binary variable applying to the whole day. The fact that the variable is binary could overstate the potential impact of the rain. This hypothesis is supported by the fact that when looking at the precipi variable, the maximum value was .3 inches, not a significant amount. The effect of rain might have been better accounted for by a variable tracking the total continous amount of precipitation. After all, someone might not choose to ride the subway if they believe the rain will be minimal, whereas several constant hours of rain would be expected to have more people riding the subway.

One other limitation of the analysis relates to the nature of trying to assess and predict human decisions. Even using all the independent variables in the analysis only accounted for about 60% of the variability.

Finally, the fact that entries and exits were only reported every four hours ends up masking weather related effects. Had the measurement been reported every hour, this may have been more sensitive to brief patterns, and more variation may have been visible.