Analyzing NYC Subway Data

by Frank Corrigan

Overview

The New York City subway system is the largest rapid transit system in the world by number of stations, with 468 stations in operation. By annual ridership, the New York City Subway is the busiest rapid transit rail system in the United States and in the Americas, as well as the seventh busiest rapid transit rail system in the world[1]. New York City visitors using these underground trains will probably see chaos in motion and fail to realize the systematic patterns that occur in this network.

Wikipedia lists the six busiest subway stations in the world as Beijing, Soel, Shanghai, Moscow, Tokyo, and Guangzhou

This analysis is an attempt to better understand how riders use the NYC subway and the intrinsic patterns generated by it. In order to do so we use a one month long dataset (May 2011) containing roughly 42,000 subway entry observations by station unit and time along with corresponding weather metrics for each observation.

The original dataset contains 27 variables. Of those 27, this report focuses on:

- *Turnstile (str)*
- *Date (str)*
- *Time (str)*
- *Entries (int)*
- *Day of Week (int)*
- *IsWeekday? (int)*
- *Rain? (int)*

The goal of this analysis is two-fold. The first objective is to deduce whether or not the occurrence of rain has an impact on ridership. The second aim is to build a linear regression model to predict number of subway entries. The analysis will employ a mix of statistical testing, data modeling, and visualization techniques in order to achieve those goals.

[1] Wikipedia, June 12, 2015. https://en.wikipedia.org/wiki/New_York_City_Subway.

Section 0. References

1. Dropbox link to data
2. Amazon AWS link to data dictionary
3. StackOverflow link for determining and removing outliers
4. StackOverflow link for plotting line chart in ggplot2
5. Slideshare link for interpreting and reporting Mann Whitney U-test results
6. IDRE link for conducting and interpreting test statistics
7. UCLA paper researching factors influencing transit ridership in San Francisco

Other ongoing resources include the Udacity learning platform and community as well as Open Intro Statistics textbook

Section 1. Statistical Test

This section focuses on performing statistical testing. Firstly, we compare distributions of the number of hourly subway entries when it is raining versus when it is not raining. Second, we compare distributions of the number of hourly subway entries on weekends versus weekdays. The goal of this sections is to find statistically significant differences between data groupings. The data used for each test is lightly cleaned -- outliers, defined as hourly entries more than 3 standard deviations away from the mean, are removed -- and the number of observations is 41,688.

Difference in Subway Entries When it Rains

After grouping the data by rain and non-rain observations and plotting histograms for each, it is evident that both sets are highly positively skewed (see Visualizations). The Shapiro Wilk test confirms this with p-values of <0.001 and <0.001 respectively. Due to the skewed nature of these datasets, the medians are calculated as the measure of central tendency. Accordingly, it appears that more people ride the subway when it rains than when it does not rain. To test that assumption, we use the Mann Whitney U-test.

The Shapiro-Wilk test tests the null hypothesis that the data was drawn from a normal distribution.

- SciPy.org

The null hypothesis for the Mann Whitney U-test states that the median hourly entries for when it rains is

equal to the median hourly entries for when it does not rain. This is a two-tailed test since it just checks whether or not the distributions are equal. The test was run 100 times on a subset sample of 4,999 observations and the results averaged. The Mann Whitney test indicated that subway entries were not statistically different for rain observations (Mdn = 893) than for non-rain observations (Mdn = 850), U = 2,291,921, p = 0.2692. Note that this test was conducted with scipy.stats in python which reports a one-tailed p-value of 0.1346 and was doubled to obtain the 2-tailed p-value.

The Mann-Whitney U test is used to compare differences between two independent groups when the dependent variable is either ordinal or continuous, but not normally distributed.

- Laerd Statistics

At an alpha level (p-critical value) of 0.025 (2-tailed test), we retain the null hypothesis that the distributions are equal and conclude that more people do not necessarily ride the subway when it rains than when it does not rain.

Difference in Subway Entries on the Weekend

Similar to above, both weekend observations (days that occur on weekend) and weekday observations (days that occur during the week) have an evident positive skew in number of hourly subway entries. The Shapiro Wilk test, again, confirms this with p-values of <0.001 and <0.001 respectively. The median values indicate that fewer people ride the subway on weekends than on weekdays. We use a Mann Whitney U-test to confirm this assumption.

The null hypothesis for the Mann Whitney U-test states that the median hourly entries on weekends is equal to the median hourly entries on weekdays. This is a two-tailed test since as it just checks whether or not the distributions are equal. The same sampling method was used as above with rain/non-rain testing. The Mann Whitney test indicated that subway entries are statistically different for weekend observations (Mdn = 700) than for weekday observations (Mdn = 1,031), U = 2,590,210, p = <0.001.

We conclude that while there is no statistically significant difference between rain and non rain observations, there is indeed a statistical difference between weekend and weekday entries.

At an alpha level (p-critical value) of 0.025 (2-tailed test), we reject the null hypothesis that the distributions are equal and conclude that less people ride the subway on the weekend days than on weekday days.

Section 2. Linear Regression

In order to achieve the second goal of this analysis, which is to predict hourly subway entries, we build a multivariate linear regression model. The final model was built using OLS with the Statsmodels python package, selected over gradient descent, for it's finer degree of accuracy.

Several manipulations were done on the data prior to performing the regression. First, outliers, defined as hourly entries three standard deviations outside the mean, were removed. Second, an 'isWeekend' variable was created which grouped Friday, Saturday, and Sundays as weekend days and Monday through Thursday as week days. Third, two time-related variables were created -- 'offset_entries_week' and 'offset_entries_day' – which are subway entries for a particular station unit and hour one week prior and one day prior respectively. Lastly, all rows that contain null values were removed.

Note that the original 'weekday' variable grouped Monday through Friday as week days. The recalculated set groups Monday through Thursday as week days.

The dataset used to perform the regression contained 31,608 observations. The dependent variable is hourly entries and the features used are 'isWeekend', 'hour', 'offset_entries_week', 'offset_entries_day', and 'UNIT'. Unit is represented as a dummy variable. No weather related variables were used as predictors.

Reasons Behind Feature Selection

Although, intuitively, weather seems like a significant indicator of subway ridership, data aggregation methods, plotting, and statistical testing proved the contrary. In no way did any of the weather variables significantly improve the 'goodness of fit measure' r-squared. Furthermore, less variables seemed like a winning strategy.

"When you have two competing theories that make exactly the same predictions, the simpler one is the better."
- Occam's Razor

Several quick plots of entries aggregated into different time buckets, such as by date or hour, revealed significant differences between weekday ridership and weekend ridership (see Visualizations) and between morning ridership and evening ridership. There is a noticeable difference between ridership on Monday

through Thursday and ridership on Friday, Saturday, Sunday. It is also visually evident that afternoons have higher ridership rates than the early mornings. Hence, ‘isWeekend’ and ‘hour’ are used as features.

The entries by day plot also reveals a strong week over week pattern. In an attempt to capture this pattern as a predictor variable, the model uses entries from one week prior and entries from the day prior. As we see in the model parameters, these features have substantial predictive power. Lastly, and perhaps most importantly, is the station unit. The difference in r-squared between a model that uses station unit as a predictor and one that did not was very large (approximately 0.40 versus 0.74).

Below is a screenshot of model results. The model parameters, as shown to the right, show that entries for the day or week prior for a particular station unit and time holds a great influence over the model. Higher entries before predicts higher entries now. While the hour coefficient is positive, the isWeekend coefficient is negative. This means that if the observation being predicted is a weekend, it will have a lower prediction than if it were a weekday.

Model intercept & coefficients:

Intercept = 1,546
isWeekend (x1) = -231
hour (x2) = 318
offset_entries_week (x3) = 446
offset_entries_day (x4) = 760

The r-squared value for the model is 0.745 and the adjusted r-squared (adjusted for # of observations) is 0.743. This means the nearly 75% of the variation in hourly entries can be explained by these variables. Depending upon the usage of this data model, we would argue that this linear model to predict ridership is appropriate for the data.

Dep. Variable:	y	R-squared:	0.745
Model:	OLS	Adj. R-squared:	0.743
Method:	Least Squares	F-statistic:	377.3
Date:	Tue, 16 Jun 2015	Prob (F-statistic):	0.00
Time:	16:42:52	Log-Likelihood:	-2.6142e+05
No. Observations:	31608	AIC:	5.233e+05
Df Residuals:	31364	BIC:	5.254e+05
Df Model:	243		

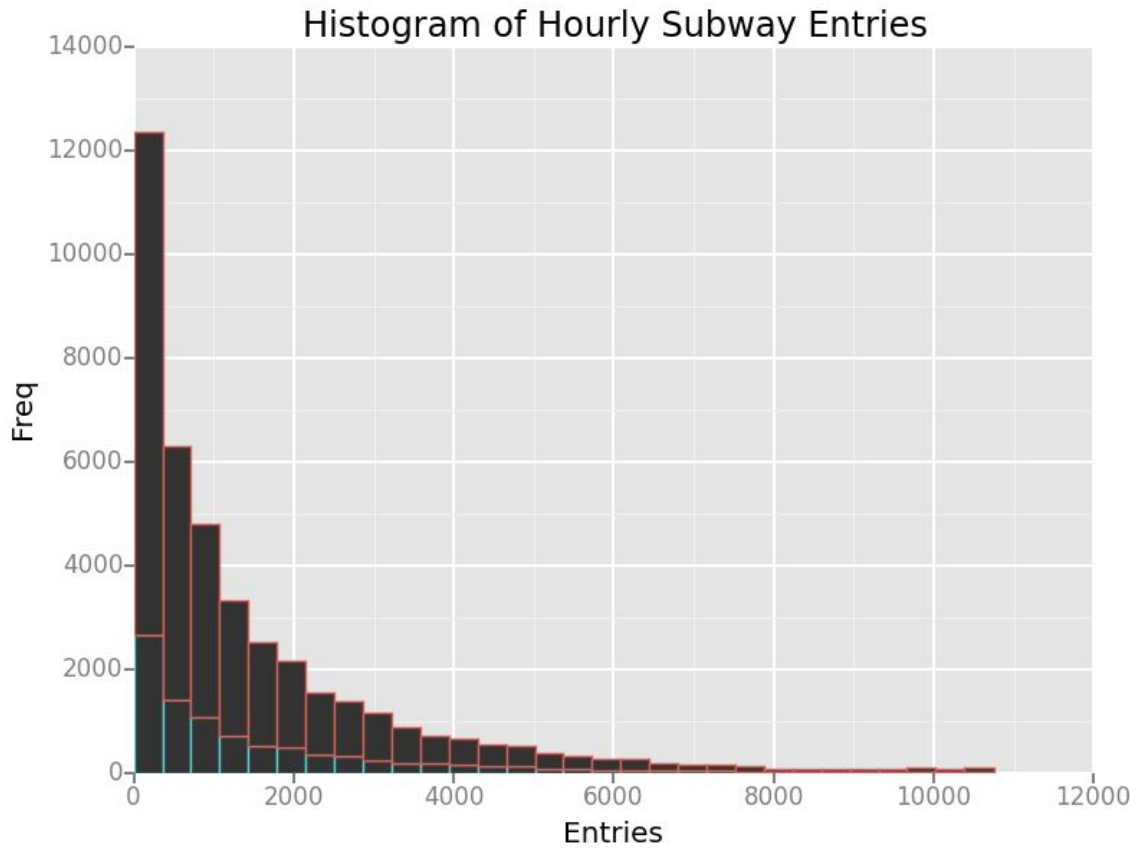
	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
const	1547.1382	7.380	209.630	0.000	1532.672	1561.604
x1	-226.2656	5.375	-42.100	0.000	-236.800	-215.731
x2	316.7676	5.900	53.688	0.000	305.203	328.332
x3	446.9512	6.184	72.277	0.000	434.831	459.072
x4	760.9712	6.282	121.143	0.000	748.659	773.283

It is worth noting again that the unit variable -- which is the particular subway turnstile -- was included as a dummy variable and there are 240 different units in the model which are not shown in the results here.

Section 3. Visualization

The following graphs attempt to visually capture the reasoning behind why and how the regression model was built.

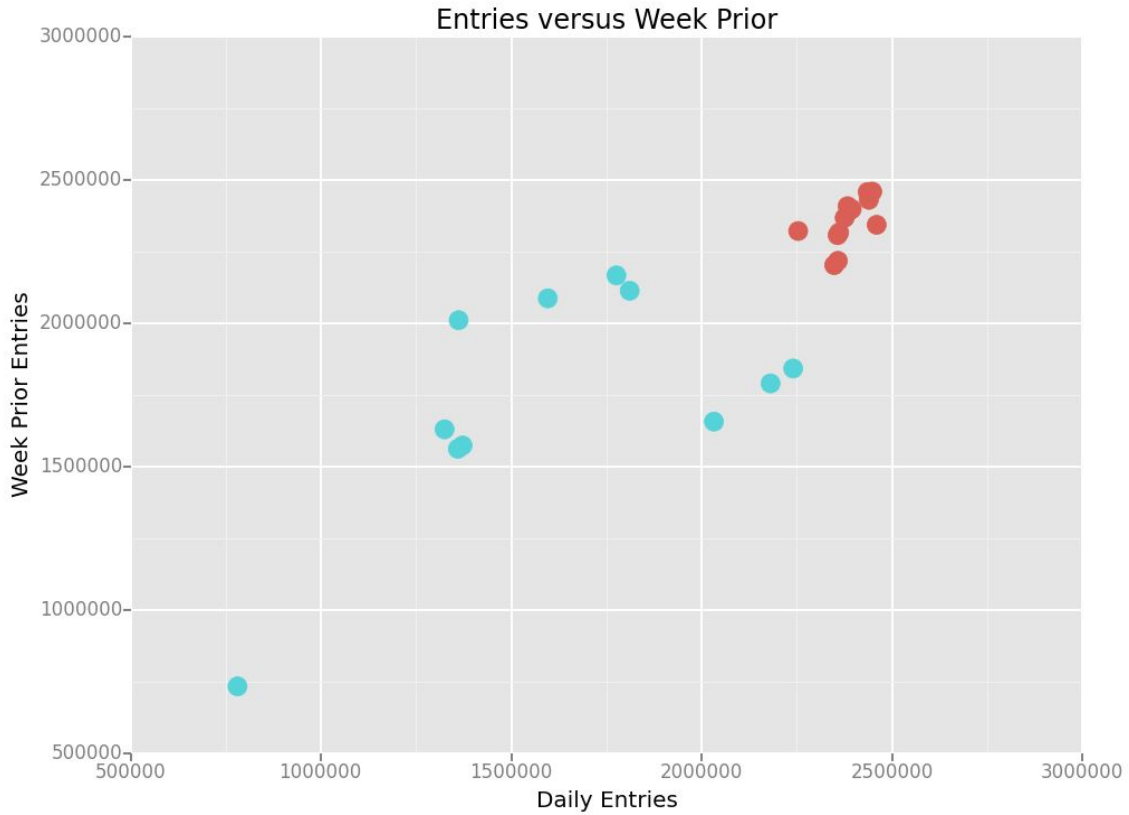
In the histogram below, the red bars represent subway entries during non rain observations while the green bars represent entries during rain observations. The high level of positive skew indicates that many turnstiles during a particular hour chunk saw no riders enter.



The histogram here compares the distribution of entries when there is rain (blue) versus when there is not rain (red). Both distributions are positively skewed.

It is easy to see that there are outliers present and that there are considerably more observations when it is not raining than when it is. Unfortunately, this was not a useful predictor in our model.

The regression model above heavily relies on past observations as a predictor for future observations. Indeed, this plot reveals the habitual nature of subway ridership. There appears to be a strong linear positive relationship between ridership one week and the next (especially in the red dots which indicate week days).



This scatter-plot tells a powerful story. The red dots represent weekday observations while the green dots represent weekend observations.

The red dots are more tightly knit, which shows a consistent pattern from week to week. However, the green dots are more spread out.

Overall, there appears to be a strong linear relationship which helps us to build a good fit regression model above.

Section 4. Conclusion

The goal of this analysis was twofold; to deduce whether or not the occurrence of rain has an impact on ridership and to build a linear regression model to predict number of subway entries.

According to the data, more people do not necessarily ride the subway when it is raining than when it is not raining. The Mann Whitney U-test concludes that there is not a statistically significant difference in these medians. Furthermore, when attempting to use rain (or even precipi for that matter) in the regression model, it added miniscule value to the r-squared, the coefficient was very small, and the p-value was large indicating that this was not a significant predictor in our model.

Rain does not significantly affect levels of subway ridership in NYC.

Past observations as well as day of the week are important predictors of future subway ridership.

It was clear that past ridership is a good indicator of future ridership. This was especially pronounced in the week over week scatterplot. The count and distribution of subway entries repeats itself particularly on weekdays. The regression results make it clear that past/future relationship as well as knowing whether it is a weekday or weekend held strongly to predict ridership.

Section 5. Reflection

While the model fit the data well and several strong patterns were identified, there is much room for improvement in this report. Utilizing a larger set of data that spans many months would be the easiest and fastest way to gain more insight about the subway system.

Regarding the statistical testing section, more diligence can be taken to test relationships between all variables to perhaps reveal a missing strong association. The r-squared in the regression model (0.74) is considerably far from explaining 100% of the variation in ridership which means there are some features of the system that we are not capturing. Lastly, it might be helpful to run alternative goodness of fit measures on the model to confirm the accuracy of the r-squared.