The Association of Rain and Higher ridership in the New York Subway – Michael Wytock - *April 7, 2014*

Summary:

This study looks at the effect of rain on hourly subway ridership in New York City during May 2011 to assess the feasibility of performing a larger scale analysis of the impact of weather on subway ridership. Linear regression by gradient descent shows that rain is the most highly correlated weather variable with hourly ridership ($r^2 = 0.402$). The only predictor variable that was more highly correlated with hourly ridership was the hour of the day ($r^2 = 0.454$)

The distribution of mean hourly ridership during times of rain is significantly different from mean hourly ridership when it is not raining. When it is raining, an average of 1,115 riders enter the subway compared to 1,090 when it is not raining ($p = 0.02$). These results suggest that a follow-up study including yearlong or multi-year data could be fruitful in uncovering ridership trends. This could be done by using MapReduce to query multiple years of data to train a model through supervised learning. If the model reasonably approximates new data being generated by the subway as it grows, the findings of this study could inform maintenance schedules, optimize staffing, or enhance security

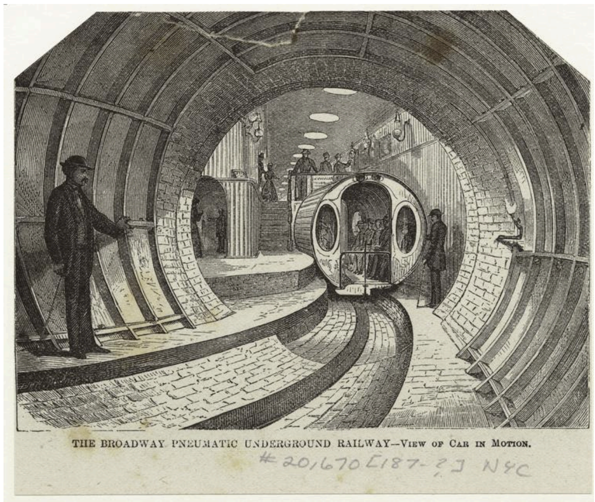
THE BROADWAY PNEUMATIC UNDERGROUND RAILWAY—View of Car in Motion.
# 20,670 [187-?] NYC

Figure 1. Sketch of the Beach Pneumatic Transit system that appeared in *Scientific American* (March 1870).

Background:

In March of 1870, an article entitled "The Pneumatic Tunnel Under Broadway" appeared in *Scientific American*. It detailed plans for constructing a new transit system underneath New York City using a "blowing engine" to propel carriages through the tunnel system using atmospheric pressure. A sketch of the design can be seen in **Figure 1**. Alfred Ely Beach, inventor and editor of *Scientific American*, had a prototype of the tunnel constructed in only 6 months without the public's knowledge.

Although infeasible for larger scale transit, many credit the pneumatic tunnel as a proof of concept for subterranean mass transit in New York City. Due in part to fears of the health effects of steam engine exhaust in the tunnels, the New York subway was put off until October 1904. Over the next 30 years, Interborough Rapid Transit and its predecessors developed most of the modern New York subway system.

One of the primary advantages of underground rapid transit then and now is reliability during poor weather. In fact, developers of the subway system used the blizzard of 1888 as an example of the benefits underground transport could provide. The blizzard crippled New York and brought the economy to a standstill. Today, over 1.6 billion riders per year rely on the subway for transit. Given that the average weekday ridership in 2013 was over 5.5 million, it would not be an exaggeration to state that New York's economy depends on the reliable service that the subway provides.

Many factors influence subway ridership, but this study seeks to understand what effect weather, and specifically rain, has on ridership. Understanding the impact of rain on subway ridership could inform, among other things, appropriate maintenance times, station staffing needs, security protocols, and preparation for extreme weather events. As a preliminary analysis, this study looks at New York subway ridership during May 2011 to determine the impact of rain on subway use. Based on the data, I suggest that rain is better correlated with hourly subway ridership than other weather-related factors and could potentially save the Metropolitan Transit Authority optimize service and spending.

Methods:

To assess the effect of weather on subway ridership, weather data during May 2011 from Weather Underground was combined with ridership data from the Metropolitan Transit Authority in a single pandas data frame. Because subway entries are logged cumulatively, hourly entries were obtained by taking the difference between each hourly ridership data point and the hourly ridership data immediately preceding it. Reformatting dates and adding a parameter to track the hour of day were also necessary. Finally, to obtain the data from all turnstile units, multiple files were concatenated from different stations across the city. These queries were done using the pandasql module to yield a data frame with 21 parameters and 131,951 entries.

SQL queries were also used to calculate descriptive statistics about the data. Examples of these calculations were the number of days in May 2011 when it rained, the average minimum temperature for rainy and clear days, and average hourly ridership during rainy and clear days.

Then, the normality of and differences between the hourly entry data for rainy and clear entry times were tested. The Shapiro-Wilk test, q-q plots, and histograms were used to test assumptions of normality for the hourly number of riders. A Mann-Whitney U-test was also performed to see if ridership during rainy and clear days followed the same distribution.

Finally, linear regression by gradient descent was performed to analyze rain, precipitation, hour of day, and temperature as features. To verify the assumption of homoscedasticity, residual plotting was performed. The correlation coefficient was used to compare the fit of different models. The feasibility of MapReduce on larger datasets was tested but was not necessary for the data given its small size.

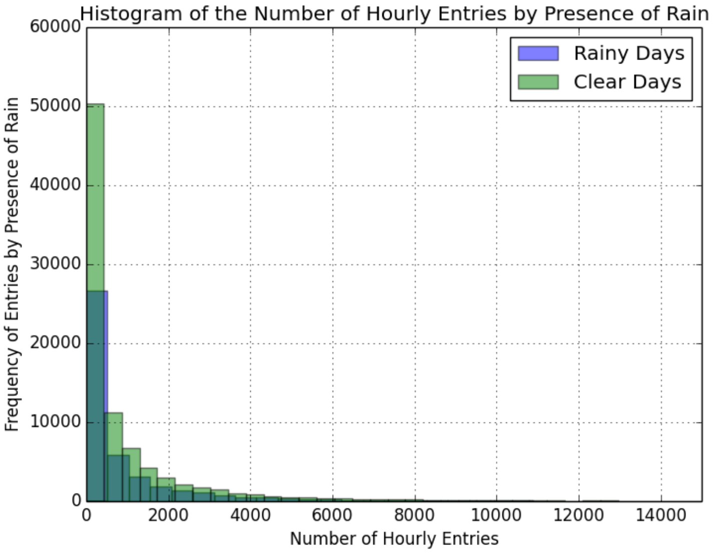

Figure 2. Histogram of the number of hourly subway entries and their frequency for both rainy and clear entry times.

All data was handled using pandas data frames and queried using the pandasql package. Plots were created using the ggplot2 and matplotlib packages in Python. Numpy and scipy packages were used for statistical analysis.

Results:

Descriptive Statistics

Out of the 30 days in May for which data was available, it rained on 10 days. The average number of hourly riders when it rained was 1,105, compared to 1,090 when it was clear. Hourly ridership did not follow a normal distribution as evidenced by the Shapiro-Wilk test (W = 0.595, p = 0.0) and histograms of the data (**Figure 2**). The distribution of entries during rainy hours was also statistically significantly different from the distribution of entries during clear hours, as evidenced by the Mann-Whitney U test (U = 1.92 * 10^9, p = 0.025).

Linear Regression by Gradient Descent

Linear Regression by gradient descent was used to assess which features correlated most highly with the number of hourly entries. The only feature that correlated more highly than the presence of rain and precipitation (r^2 = 0.402 for each) was the hour during which the entry data was collected (r^2 = 0.454). Other features with high correlation coefficients were minimum temperature, maximum temperature, and mean temperature (r^2 = 0.397, 0.397, 0.396, respectively).

The residuals of the gradient descent analysis were plotted for all models. All were approximately normally distributed. A sample plot of the residuals for the model including rain as its sole feature can be seen in **Figure 3.**

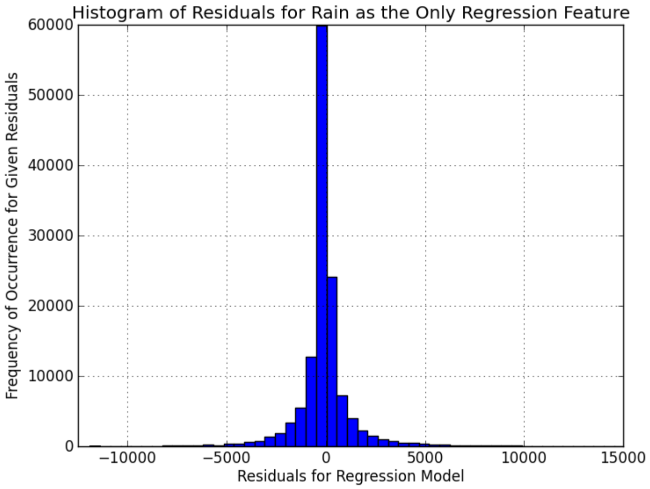

Figure 3. Sample histogram of the residuals for linear regression model using presence of rain as the only predictive feature of hourly subway ridership.

Analysis:

As evidenced by the Mann-Whitney U test, there is a significant difference in the ridership distribution when it is rainy as opposed to clear. Because hourly ridership is higher when it rains, there is a *suggested* relationship between rain and hourly ridership. The Mann-Whitney test was used because it is robust to violations of normality, which are present in this data as evidenced by the rejection of the null hypothesis of normality in the Shapiro-Wilk test and the histogram in **Figure 2**. To be clear, a causative relationship cannot be confirmed. Developing an argument for causation would require either a supervised learning hypothesis or a controlled study and would benefit from data collected over months when rain and temperature follow different patterns than in May.

Linear Regression models showed a positive relationship between rain and subway ridership. The only feature that was more highly correlated with ridership than rain was the hour of the day. The amount of rain does not seem to play a significant role in determining ridership. The correlation coefficient was the same for the model that included solely rainy versus clear weather as for the model that included exact precipitation quantities. The normally distributed residuals in **Figure 3** show that linear regression was an appropriate analysis in this situation. This is because a normal distribution shows that there are few high leverage outliers influencing the model fitting.

Gradient descent was used for the regression analysis because it is more efficient that methods such as Ordinary Least Squares for a model that may change given new data. Additionally its computational cost is lower than Ordinary Least Squares. There is a potential pitfall in using gradient descent that our cost function has located only a local minimum and not a global one. Once a model has been developed for a larger scale analysis, running other regression methods to verify that correlations are not substantially different would be advisable.

It is important to note that the hour during which the data were collected could have influenced the results since the hour of day was also found to be highly correlated to ridership relative to the other available features. Including both rain and the hour of day as features in the regression model resulted in a correlation coefficient higher that either feature individually ($r^2 = 0.460$), which suggests that each feature (hour of day and rain) is capturing different sources of variation. In future studies, care should be taken to ensure that trends due to the spikes in ridership at different times of day are not being interpreted as being due to weather events. **Figure 4** shows the ridership spikes at 00:00, 12:00, 20:00, and 21:00.

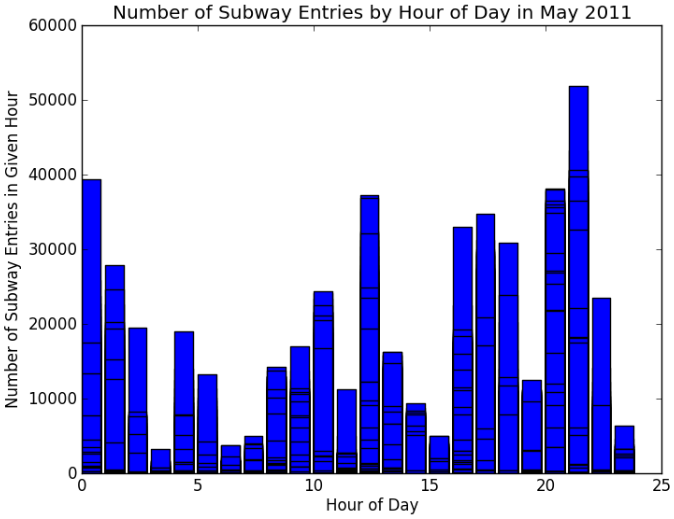

Figure 4 Average number of subway riders per hour of the day during May 2011.

While correlation coefficients do not imply causation, they do give a reliable indicator about which features would be most interesting to study in a year long or multi-year study. A couple of interesting follow-up studies could be done. First, yearly data or seasonal data could be analyzed to understand whether the correlations found during the month of May are similar during other months or seasons that have different weather patterns. Second, to avoid over- or under-fitting, we could use supervised learning to train our model on data from a given month or year and see how well the model performs on new data sets. This would provide better evidence for a causal relationship between the features in our model and subway ridership.

This analysis has multiple shortcomings, most of which would be eliminated by using a larger dataset. It is well known that weather patterns, especially rainfall and temperature vary over the course of a year and even year to year. It is also well-established that ridership on the New York subway has changed dramatically over the years. Record yearly ridership was achieved in 1949, and there have been many spikes and troughs in ridership since then. For these reasons, this study should be extended to include at least a whole year's worth of data before any actionable conclusions should be drawn. This would allow the model to take into account these changing weather patterns over the course of the year, and it would open the door to year-to-year analysis, which would put the model's predictive power to the test.

If the study were to be scaled up significantly, to the point where calculations could not be done on a single machine, MapReduce would be a necessary tool for analyzing the data. For example, if multiple yearly data sets were used, each machine could map hourly ridership values as a predictor variable of interest changes. Each machine could work in parallel on separate year ranges. Then, all values with the same key would be passed to a reducer that would output the results of a given query. Given modern computing power and disk space, the data would need to be significantly larger than the current dataset, approaching 5 terabytes rather than 15 megabytes.

Conclusion:

This pilot study suggests that weather, especially rain, is positively correlated with subway ridership in New York City. If this correlation holds up over a full year, I would recommend accounting for weather when planning subway maintenance and determining staffing. For example, planning maintenance during months when precipitation is low or infrequent could minimize impact on riders. Similarly, efficiency might be maximized if extra cars and station officers are used on days when it rains and fewer on clearer days.

Just as weather was a motivator for the creation of the subway back in the late 19th century, it should be considered as the subway evolves and expands during the 21st century. Subway planners should bear in mind the factors that cause increased ridership in order to ensure the durability and reliable use of the subway, on which the New York economy depends.