Effects of Weather on NYC Subway Ridership

Introduction

The first time I went to New York City, I was captivated by the sheer liveliness of the place. It was marvellous walking along the streets, seeing a blend of modern skyscrapers as well as old architectural buildings. The only downside is my first impression of the New York City subway; it's dark, old and with walls full of graffiti. However, after taking a second glance, I realized that it's a well-used subway, that in spite of its appearance, it delivers over a billion rides in a year. In fact in 2013, the subway delivered 1.71 billion rides, averaging approximately 5.5 million rides on weekdays, about 3.2 million rides on Saturdays, and about 2.6 million rides on Sundays. [1] Wow, that is huge! I wonder if subway ridership is the same between sunny days and days with rain or fog. Do weather influence subway ridership?

New York City is one of the most populous urban agglomerations in the world and the most densely populated major city in the United States with approximately 55 million visitors per year[2]. Moreover, my personal impression of New York is that it is a fast-paced city, with many happenings in and around the city. In my opinion, rain or shine, people will use the subway as often as they need it, to go about their business in and around New York City. Hence, I think that subway ridership in sunny days should be the same in rainy or foggy days. Let's see if this assessment is true by doing a study on subway ridership in different weather types.

New York City Subway

The New York City Subway is a rapid transit system owned by the City of New York and leased to the New York Transit Authority, which is a subsidiary of the Metropolitan Transportation Authority (MTA). It is the busiest rapid transit rail system in the United States and Americas, as well as the seventh busiest rapid transit rail system in the world. It offers rail service 24 hours a day, 7 days per week. It is one of the largest public transportation systems in the world by number of stations, with 468 stations in operation. Overall, the system contains 232 miles (373 km) of routes, translating into 656 miles (1,056 km) of revenue track; and a total of 842 miles (1,355 km) including non-revenue trackage.[1]

The New York City Subway is one of the world's oldest public transit system and was known as the Interborough Rapid Transit Subway, or IRT. Even with elevated train lines springing up around the city, the need for an underground rapid transit railroad was obvious as a solution to street congestion and to assist development in outlying areas. On October 27, 1904, the first IRT subway line opened, in which the City Hall station stood as the showpiece of the new subway.[3]

Determining subway ridership helps to underline the importance of preparing the subway not only for good weather, but especially for bad weather, when subway operation is prone to service disruption. For example, if there is a forecast of rain, cleaning of drains, removing of trash and making repairs ensure that track areas are free and clear of any debris so that no clogged drains will occur, allowing for a build-up of water significant enough to stop service[4]. Furthermore, this study will highlight people's reliance on the subway for their transportation needs, thus making the needed push to upgrade the subway system in order to eliminate service disruption during bad weather.

---

[1] http://en.wikipedia.org/wiki/New_York_City_Subway

[2] http://en.wikipedia.org/wiki/New_York_City

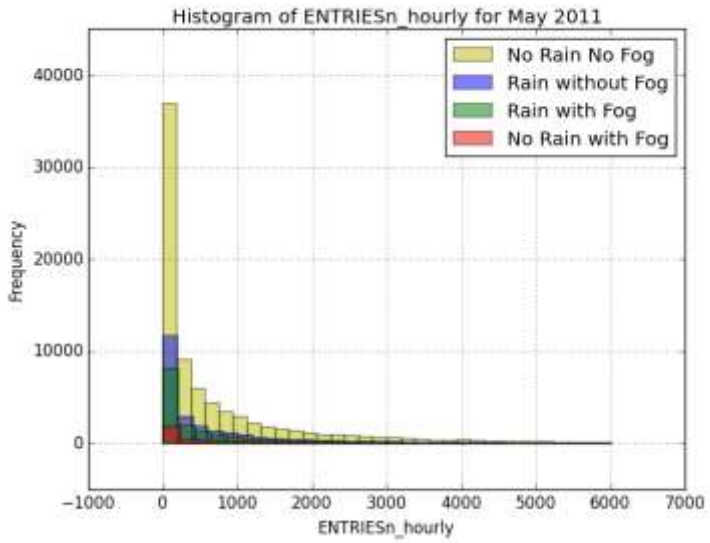
[3] http://www.nycsubway.org/wiki/The_First_Subway

[4] http://www.mta.info/news-rain-subway-new-york-city-transit/2014/07/03/nyc-transit-prepping-heavy-rain

Subway Ridership Data Comparison

For this study we have a sample size of 131, 951 subway records for the month of May 2011. Each record consists of the following information: total entries per hour, total exits per hour, date, hour, the unit id, fog, rain, precipitation and thunder indicators, and other weather measurements like pressure, dew, wind, and temperature. In this study, we will use the total entries per hour (ENTRIESn_hourly) field to determine ridership; and for weather types, we will use fog and rain fields. We will limit our comparison into 4 weather types: No Rain No Fog, Rain without Fog, Rain with Fog, and No Rain with Fog.

First, let's look at the histogram below for the data distribution. We can say that data is normally distributed if we see a bell curve in our histogram, meaning that 50% of the data can be found above or below the mean value.



As you can see from the above diagram, our data is not normally distributed; 44.78% percent of our data have ENTRIESn_hourly value less than 200. Also, 63.28% percent of our data belong to "No Rain No Fog".

Furthermore, here is the breakdown of sample sizes for each weather type and their medians:

Weather Type	Sample Size	Median
No Rain No Fog	83,499	273.0
Rain without Fog	26,403	286.0
Rain with Fog	17,701	278.0
No Rain with Fog	4,348	401.5

Based on the median values, it looks like more riders take the subway on "No Rain with Fog". Let's see if this is significant; let's test the distinctness of our sample data. Here, I've decided to use Mann Whitney U test[5] due to the following reasons: our data is not normally distributed, the samples is based on independent observations, the dependent variable, which is total entries per hour, is ordinal or continuous, and the data distribution of the samples are of the same shape[6]. This test returns a two-sided p value that indicates how distinct the samples for each weather type. When p value < 0.05, we can say that our data is distinct between the samples; otherwise, the distributions of the samples are similar.

[5] I've decided to use the Mann-Whitney U test from R, wilcox.test(x,y, exact=FALSE, correct=TRUE) function, as I find the python's scipy.stat.mannwhitneyu function problematic. This is because the mannwhitneyu test for "No Rain No Fog" vs "Rain without Fog" is giving me a p_value of nan.

[6] https://statistics.laerd.com/spss-tutorials/mann-whitney-u-test-using-spss-statistics.php

Here, I've decided to make "No Rain No Fog" as the base comparison, so there are only see 3 sets of test. This is because I am comparing the ridership in sunny days and other days with rain or fog. Below are the p values:

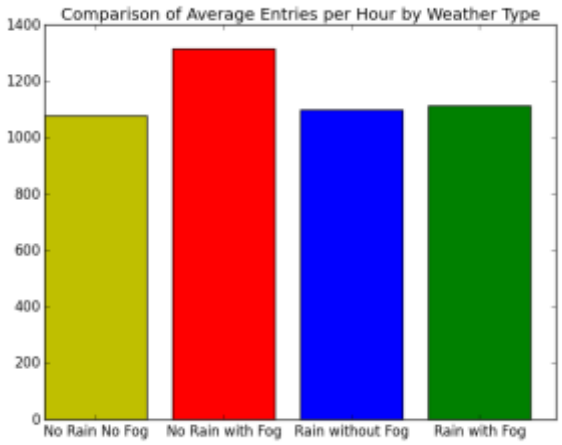
Sample X	Sample Y	P Value
No Rain No Fog	Rain with Fog	0.1234
No Rain No Fog	Rain without Fog	0.001732
No Rain No Fog	No Rain with Fog	2.2e-16

From the above results, we can say that the observed samples for "Rain without Fog" and "No Rain with Fog" are distinct compared with "No Rain No Fog". Furthermore, we can also say the median value for "No Rain with Fog" did not occur by chance and this suggest that more people do take the subway in this weather type. However, let's see if the average of the total entries per hour supports this. Using DataFrames groupby and sum function, below are the results we got.

Weather Type	ENTRIESn_hourly	total_hours	Average of Total Entries per Hour
No Rain No Fog	90,057,579	83,499	1,078.55
Rain without Fog	29,015,664	26,403	1,098.95
Rain with Fog	19,738,943	17,701	1,115.13
No Rain with Fog	5,720,141	4,348	1,315.58

Note that if we are dealing with larger sample sizes, in petabytes or even terabytes, we can use MapReduce framework because MapReduce process data faster. MapReduce has two components: Mapper and Reducer. The Mapper generates a list of key-value pairs, i.e. key being the weather type and value the ENTRIESn_hourly; while the Reducer massages the data in order to generate a summary, like counting of duplicate words or calculates for the average. In between the Mapper and Reducer, the MapReduce framework shuffles and groups the data accordingly so that when data is received by the Reducer, it is assured that it is ordered by the key. In actuality, MTA can benefit in using MapReduce because the amount of data that is analysed is in the region of billions per year.

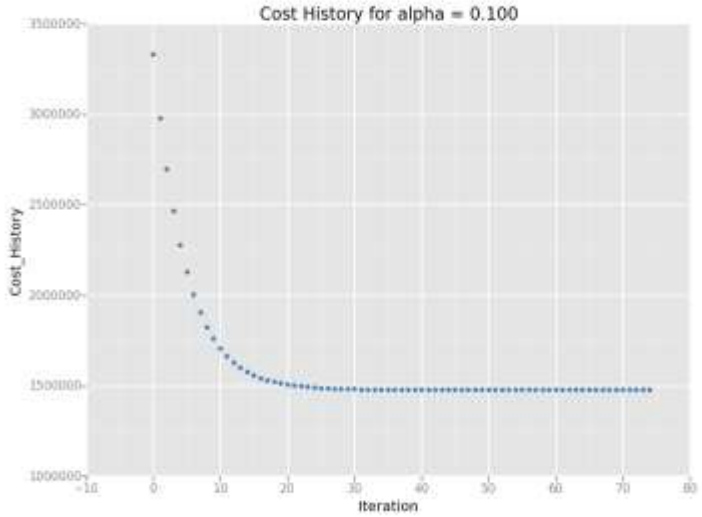
To effectively see the difference in the averages among the weather types, here is the bar chart.



The average value calculated for each indicates that there is no big difference in the ridership for each weather type. However, it is good to note that "No Rain with Fog" is higher by 18% compared to "No Rain No Fog" which has the least average.

Predicting ridership

Predicting data shows the likely trend of ridership; thus, running our model to the predicted data validates our findings and supports our conclusions. Hence, to confirm that there is really no big difference in ridership based on weather types, let's see if we will get the same result in the predicted data. There are 114,758 out of 131, 951 samples that fall within the standard deviation of 2337.01 for ENTRIESn_hourly. This means that our data points are not too scattered and we can use Linear Regression in predicting data. Linear Regression finds the best line that will fit a set of data points. A standard approach to Linear Regression is determining the cost function, the minimum error value that measures how good the line is. One method of finding the cost function is Gradient Descent, an iterative process of finding the minimum cost function by going opposite of the fastest increase. Another method is Ordinary Least Squares (OLS)[7], which minimizes the sum of squared distances between the actual data and the predicted data. However, I will use Gradient Descent as it has more extensions[8][9]. Looking at our data, there are several features that we would like to include in predicting the ENTRIESn_hourly value. These are: unit, hour, rain, fog, precipitation, mean pressure, mean dew, mean wind, and mean temperature. Gradient Descent uses a Learning Rate (alpha) or step size; how big a downward step we want to take so that we don't overshoot the bottom. Below is the cost history data point chart, using 75 iterations. As you can see, the cost is going down and is tapering to a minimum value at around the 40th iteration.



The objective of cost history to track the progress of the gradient update, which generates the final theta or parameters that can be applied to our linear equation for predicting the best line that will fit our data points. Here is the final theta, note that only the first 8 is shown and the unit dummy values[10] were left out.

```
final theta= [ -6.71026347e+01   6.56326494e+00   3.06183531e+01   4.64000839e+02
  -1.94536401e+02  -5.24965348e+01  -5.42589041e+01  -1.46188446e+01 … ]
```

Now, let's look at the data distribution of the predicted data.

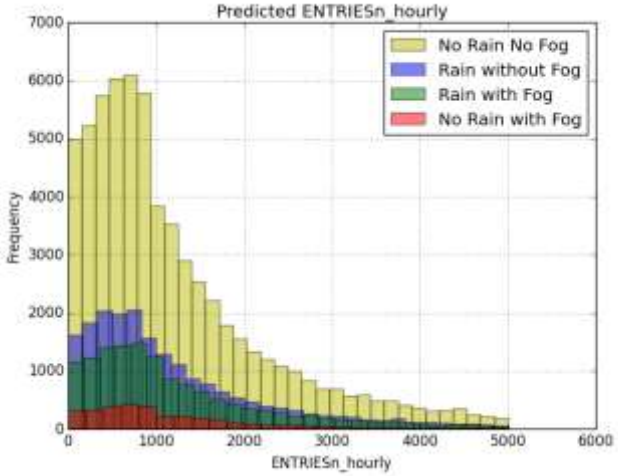
---

[7] http://en.wikipedia.org/wiki/Ordinary_least_squares

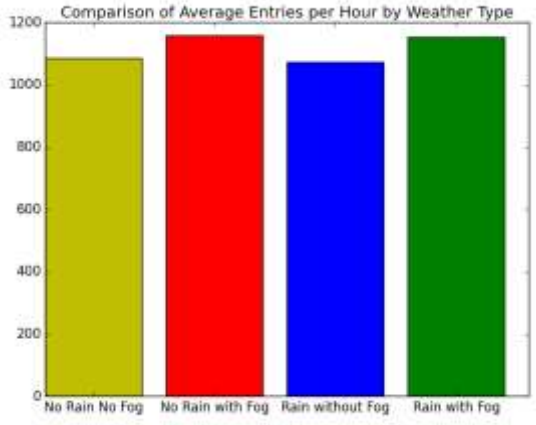
[8] http://en.wikipedia.org/wiki/Gradient_descent#Solution_of_a_non-linear_system

[9] http://www.statisticsviews.com/details/feature/5722691/Getting-to-the-Bottom-of-Regression-with-Gradient-Descent.html

[10] https://www.moresteam.com/whitepapers/download/dummy-variables.pdf



The coefficient of determination (R2), also called as r squared, tells us how well the predicted data fits into our model. The more R2 is close to 1, the better our predicted data is. The R2 value calculated for our predicted data is 0.458760[11], meaning that there is a 46% chance that our predicted data will hold and will fit future samples.



There is still no big difference in the average entries per hour for the predicted data. Also, "No Rain with Fog" still has the highest average at 1,156.86, seconded by "Rain with Fog" at 1,152.39.

Conclusion

Our test shows that there is no big dip or a significant increase in subway ridership among the different weather types that we studied. Therefore, different weather conditions do not influence subway ridership and subway service is equally important in sunny, rainy and foggy days. However, it is important to note that there is a slight increase in ridership when there is a fog. Hence, important preparations and repairs should be planned and implemented to ensure reliable subway service on all weather conditions, especially, when there is fog.

It should be noted that our study did not consider other weather conditions like snow, heat wave, cyclones and other extreme weather that may affect ridership data. These weather conditions were not reflected on our sample data. Thus, the conclusions that we made can only be applied to normal weather conditions.

[11] I also tried OLS using the same features in predicting data, and I got the same R2 result.