ANALYSIS OF RIDERSHIP BEHAVIOR ON NYC SUBWAY SYSTEM

Raja Shekar Kilaru & Sai Bhargav Komatireddy

Graduate Program in Operations Research and Industrial Engineering | The University of Texas at Austin

Background

New York City subway is operated by the New York City Transit Authority, which is a subsidiary of Metropolitan Transportation Authority (MTA). It was opened in 1904, and is one of the world's oldest public transit system, one of the world's most used metro systems, and the metro system with the most stations and most trackage. It offers services 24 hours per day and every day of the year. It has 469 stations and 846 miles of track length which makes it the system with highest number of stations and the longest.

By the number of riders, it is the busiest public transit system in the western world and seventh largest in the world. The transit systems that are busier than NYC subway system are only Beijing, Seoul, Shanghai, Moscow, Tokyo and Guangzhou. In 2015, the subway system delivered more than 1.76 billion rides with 5.7 million daily average rides on weekdays and 5.9 million rides on weekends combined. Ridership continues to increase and had recorded 6.1 million rides on a single on September 23rd 2014. The NYC subway system continues to serve as the most important mode of transport for the people living or working in NYC for their daily commute and running errands. Hence, we believe that it is important to understand the ridership behavior to strategize infrastructure improvements, plan business activities in and around the metro stations and other steps that are required to enhance the commuters experience.

Data Sources

Metropolitan Transit Authority (MTA) whose subsidiary operates the NYC subway system hosts tons of data on its website. One of such useful data for our project would be the turnstile data which the website updates on every Saturday for the preceding week. Turnstile data seemed to be best source of riders' numbers as it captures the entry of each rider entering the station. The important features in the data were the Unit identifier, Station Name, Date, Time, Entries and Exits. The data set lists the entries and exists by a cumulative number at every 4-hour intervals for each unit at a station. To study how the ridership behavior varies with different weather conditions, we scraped the weather data from website www.weatherunderground.com which hosts the historical weather data for every one-hour intervals. We selected the MTA data and weather data from May 2016 as it is a transitional month from rainy to summer season in NY. It has different weather conditions like rain, fog and sunny days.

Data Preprocessing

The input for weather underground website should be either the city name or zip code. As we are interested in knowing the weather conditions at each station, it is important that we obtain the zipcode for each station. To accomplish this, we used the geocoder module in Python to get the latitude/longitude coordinate information and zipcodes for each station. Also, the cumulative entries and exits are converted to absolute values. We use the extracted zip code as the input along with date to the weather website, to scrape the historical weather information by each hour for the specified date. Also, the key weather parameters were aggregated for 4-hour windows as the turnstile data is available for every 4 hours.

Analysis

Effect of time of the day and day of the week:

Our first focus was to study the variation in ridership by time of the day and day of the week. Hence, we decided to visualize how the number of entries and number of exists vary by every 4-hour window. We noticed that the ridership was high between 8:00 and 12:00 hours, and between 16:00 and 20:00 hours. This can be attributed to the daily work commute traffic. However, the high traffic between 20:00 and 24:00 hours may be giving an insight into late night work hours that the NYC business district is known for. (Fig 1)
Similarly, we noticed that the traffic was almost uniform on the weekdays and was considerably low on the weekends. This reinforces the understanding that the subway system was primarily used for traveling to and from workplace. (Fig 2)
However, to know which stations have higher traffic, we plotted a histogram for the top 30 stations and noticed that the stations with highest traffic are in the Manhattan area which further strengthens our belief of being the metro system being used for work commute. To further get a better view, we leveraged on the NYC map from the open street maps, transformed the coordinates of each station into pixel coordinates of the picture and plotted a scatter plot showing the average traffic at different times during the entire month of May. (Fig 3, Gif image)

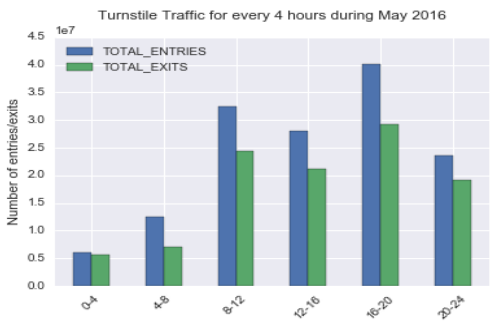


Fig 1

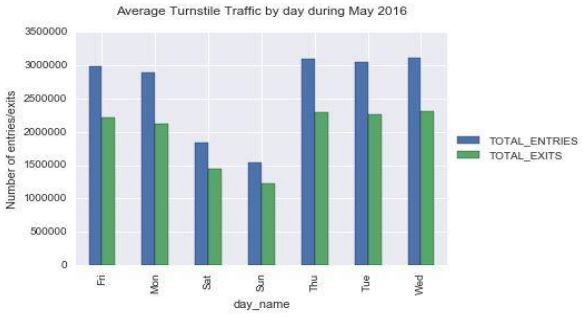
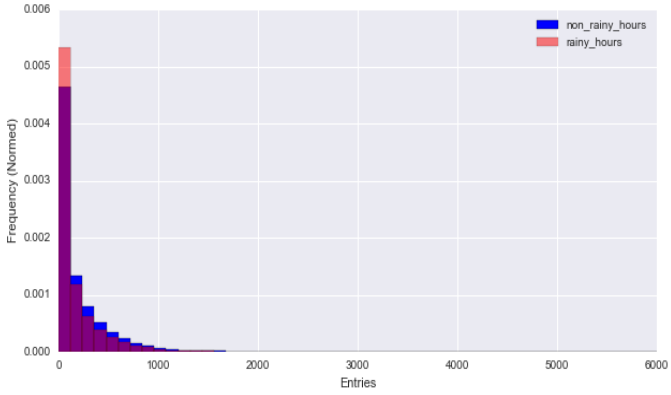


Fig 2

Effect of weather:
Further we studied the influence of rain and fog on the ridership. We plotted histograms to see the number of riders on rainy and non-rainy days and noticed that the median riders on rainy day was 54 and on non-rainy day was 88. Hence, we were of the opinion that rain decreases the riders. We plotted the similar graph for foggy and non-foggy days and observed that the riders on foggy day were more than on the non-foggy day (207 and 82 respectively).



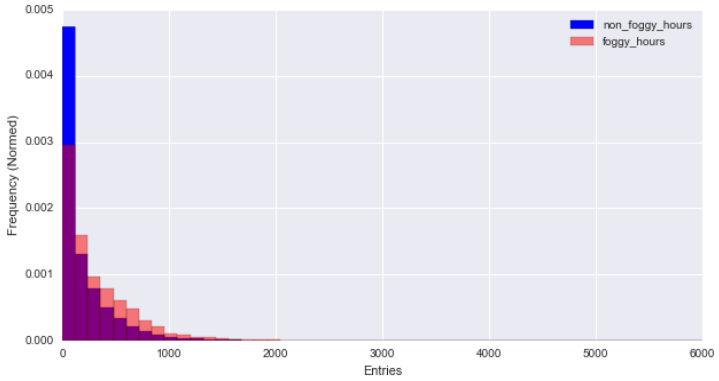


Fig 4 and Fig 5 (from top to bottom)

However, to further reinforce our belief about the effect of rain and fog on the ridership, we conducted a hypothesis test. As evident from the above figures (Fig 4 and Fig 5), the distribution were skewed to the right. Hence, we decided to conduct the Mann-Whitney hypothesis test which does not assume normality. Mann-Whitney test compares the medians of the two groups unlike the t-test which compares the means of the groups. From the test, we observed that the p-values were very low and hence we have statistical evidence to reject the null hypothesis at a lower significance level. Therefore, rain and fog will have an effect on ridership. It was also noticed that the effect size (which describes the strength of the phenomenon) of rain is lower than the effect size of both fog and day of the week. However, fog had the highest effect among the three factors.

	U-Statistic	**p-value (two tailed)**	**effect size**
Rain	2.8246E+10	0.0	0.0938
Fog	1247433644	1.05E-239	0.2581
Day-type	4.189E+10	0.0	0.1933

Effect size	**Effect**
< 0.1	Trivial
0.1-0.3	Small
0.3-0.5	Medium
> 0.5	Large

Regression Analysis:

We further performed regression analysis to learn how much of the total ridership entries is explained by the features we are aware of. To estimate the total entries, we used Day type (Weekday/Weekend), Time group, Average temperature, Average precipitation, Unit identifier as the independent variables.

We noticed that the R-squared value for the linear regression was 25.16%. However, it was noticed that residuals are not normally distributed. Hence, we fit ensemble models of Gradient Boosted regression and Random Forest Regression which resulted in R-squared values of 23.42% and 52.23% respectively.

Limitations

The Dataset considered was only for the month of May 2016. To reach solid conclusions, data entries should be distributed uniformly across different years and the 12 months of the year. We do not have an actual record of entries for each hour as the turnstile data contains only cumulative entries for every 4-hour interval. However, the weather data is available for each hour and hence, on aggregation there may be a bias associated.

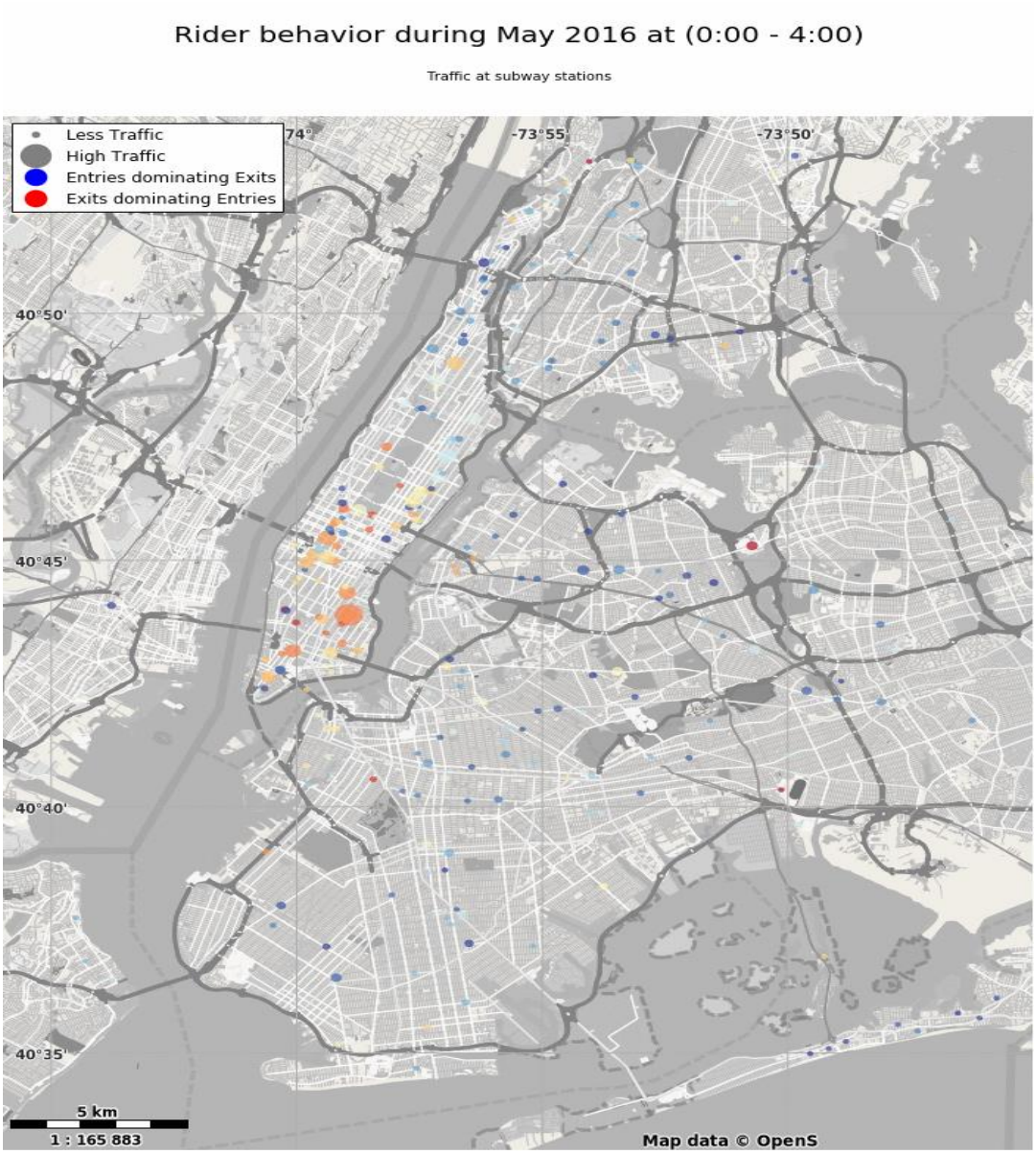


Fig 3