How Data Science can help MTA improve the NYC Subway

The New York City Subway is the largest rapid transit system in the United States, serving 1.7 billion rides per year. The system covers over 230 miles of track, with 460+ stations in operation. The Metropolitan Transportation Authority (MTA) has leased the NYC Subway from the City of New York.

The data scientist's job is to analyze the turnstile data collected by the MTA to determine key factors that affect ridership. If certain factors, such as weather and time of day, are found to play a statistically significant role, planned changes can be made by the MTA to improve its service and thus benefit riders.

Staffing levels at stations, frequency and mix of local and express trains on certain routes, number of cars on a train can all be adjusted based on weather forecast, hour, or day of the week.

From a longer-term perspective, the MTA can make strategic decisions for lasting improvements. Stations found to be rarely used at all times can be shut down. New stations can be built between adjacent stops that encounter very high commuter traffic, to relieve pressure on those stations. More retail spaces can be built at stations that show an upward trend in ridership, to increase revenue for the MTA.

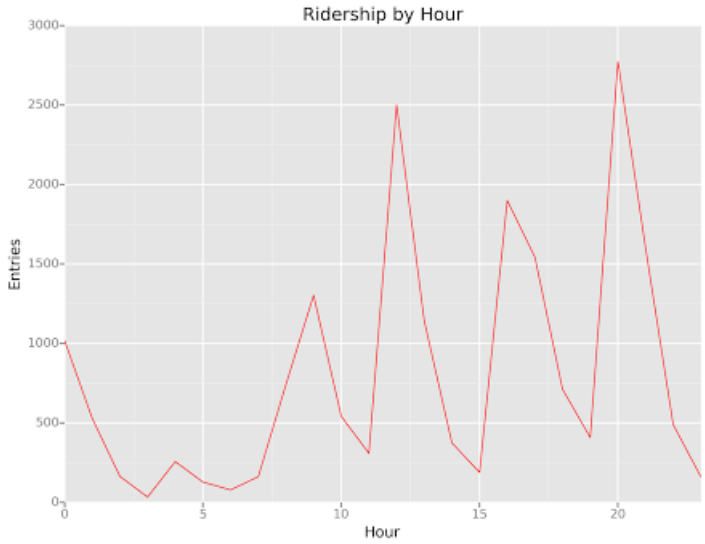
Commuters can benefit as well from a high-level view of the turnstile data. At peak hours, they may want to avoid crowded stations, and instead use less busy ones nearby. At wee hours of the day, female riders may want to avoid deserted stations, for safety reasons.

Determing key factors affecting ridership

The turnstile data provided by the MTA contains information about the number of commuters that entered and exited each station by date and hour for the entire month of May 2011. Also included are weather conditions for each hour of the day, such as temperature, incidence of rain or fog or precipitation, etc.

Only commuter entries were used to count ridership for data analysis, exits were ignored.

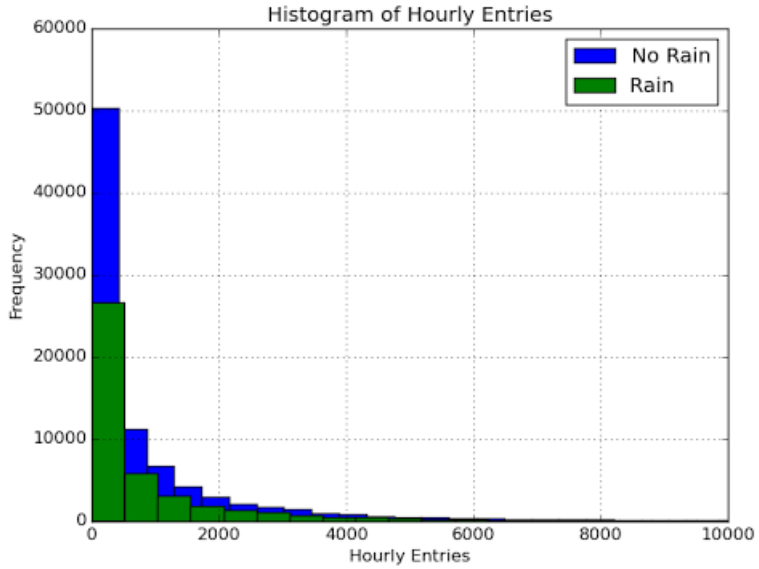
Common knowledge suggests that time of day would be an important factor affecting ridership. The graph below plots the average entries at each hour of the day. The spikes and falls at expected times visually confirm that hour of the day may well be a factor.



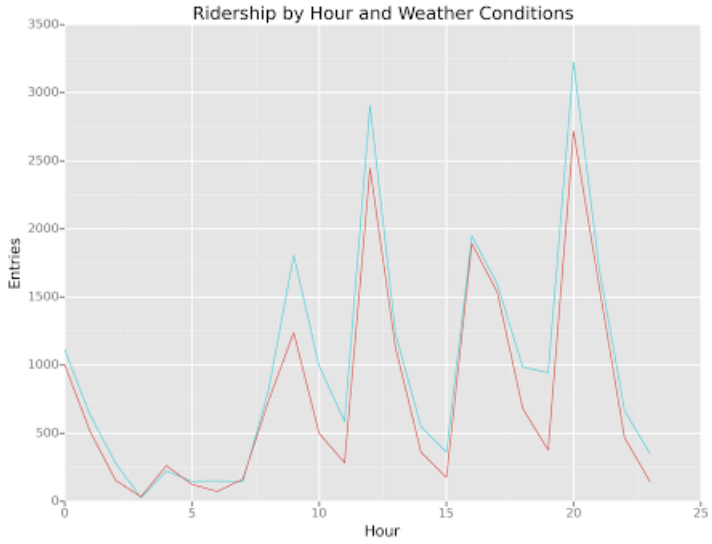
From the different weather conditions provided, rain could be a factor. Certain tests were run to determine its significance before being used in a model that can predict ridership based on provided factors.

If the average number of riders in “rain” and “no rain” are normally distributed, then Welch’s t test can be used to confirm the hypothesis that incidence of rain affects ridership. If the data is not normally distributed, the Mann-Whitney U test can be run instead.

The graph below shows that hourly ridership based on incidence of rain is not normally distributed, hence requiring the Mann-Whitney U test.



The Mann-Whitney U test on “rain” and “no rain” data returned a p-value of 0.02, which proves that the number of riders during rain is not the same as when there is no rain. The graph below visually confirms that ridership during rain (red line) is different than when there is no rain (green line).



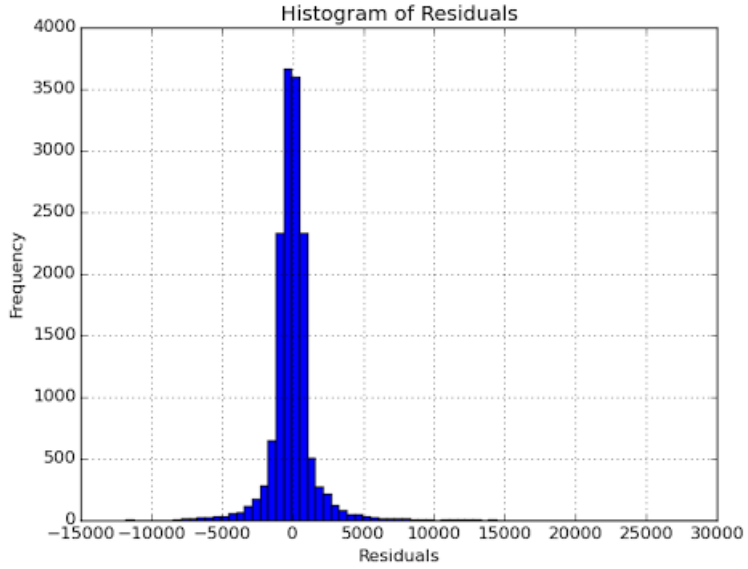
So, the incidence of rain may likely be a factor in predicting ridership. In winter months precipitation may affect ridership, so was also used as a factor alongside rain.

Building a model to predict ridership

A linear regression with gradient descent model can take several factors to predict an outcome such as ridership. One was built using hour, rain, precipitation, and mean temperature to predict riders at each hour of the day.

The model was then evaluated by comparing the predictions against actual outcomes. R^2 - a statistical measure that determines how well a model predicts outcomes – was calculated, and returned a value of 0.46.

The graph below plots the difference between actual outcomes and predictions made by the model, i.e., errors or residuals. The normally distributed histogram and the R^2 value tell us that the model is appropriate, and the chosen factors are significant in affecting ridership.



The MTA can indeed use weather forecasts about rain, snow, and temperature, as well as hour of the day to make decisions about its service. Noon and 8 pm seem to be peak hours. The frequency of express trains, the number of cars on each train, and staffing levels during those hours can be scaled up to handle the additional commuter traffic.

Commuters seem to use the subway less frequently during rains, possibly preferring to use bus services instead. If that hypothesis can be proven, then MTA could make adjustments to its bus services during inclement weather.

Analyzing Very Large Data Sets (Big Data)

The MTA turnstile data covered just one month. However, effective analysis often requires data that spans a much longer time frame.

Analyzing very large data sets is not always possible on one single computer. It requires distributed systems that bind together many computers in answering questions such as “Which was the busiest hour across all stations from 2001 to 2010?”.

One such distributed system is MapReduce, which uses many mappers and one or more reducers to analyze very large data sets.

A data set is first divided across multiple computers, and then mappers on each of the computers work in parallel to generally pare or filter the data. The output of each mapper is sent to one or more reducers that summarize it to answer the question posed.

The MapReduce ecosystem includes Hadoop Distributed File System (HDFS) for storing data, as well as programs like Pig and Hive for data analysis.

Author: Bhavin V. Choksi

Project: Udacity’s Intro to Data Science