Project Title: Identifying the most effective mean of public transportation

Group Tag: W

Group Members: Jae Woong Ham & Andrew Thvedt

Please note that our group has obtained prior permission to work in a group of only two members from Professor Brambor due to our shared passion for this topic of interest.

Project Rationale

In the bustling city of New York City, people have a variety of choices of public transportation to choose from. However, everyone's priority and budget are different to get from point A to B given any point in time. For example, someone might be willing to pay more money than average to get somewhere fast, whereas another person values cost-efficient methods. As such, there is always a burden of choice of which form of transportation is ideal depending on the person's preferences in relation to its destination.

Although Google Maps and Ride Sharing applications currently do a fair job of giving real-time feedback on price and journey duration, this type of feature is often lacking for public forms of transportation. First, the TLC does not have a mobile application that tracks and provides real-time status of taxis. Although such mobile applications exist for trains and buses separately, they cannot be compared side by side to other forms of transportation. Secondly, Google Maps currently has a feature that does give estimates of public transportation, including buses and trains, and allows the user to compare them against other forms of transportation. However, their estimates of public transportation are often misleading and subject to errors. Lastly, even though TLC does not have an application, they have a highly detailed record of taxi and train services. Considering the factors mentioned above, this presents an excellent opportunity for exploratory data analysis to derive important insights on how the public transportation compares to each other and other forms of transportation in the heart of New York City.

Abstract

Our group is interested in finding out which mode of transportation is most beneficial in terms of either duration or price at a given time frame. As such, exploratory data analysis will be conducted at different levels with different datasets.

- Individual-level
  - Taxi: Duration of the trip and amount of fare will be assessed in relation to pick up and drop off 'zones', time of day, and day of the week. These zones are defined by TLC referring to specific regions within NYC. From this, our group will

    be able to identify which zones are the busiest and/or most expensive given the time and day of the week.
    - Subway: Since similar, highly detailed data is available for subway trips, this can serve as a comparison for taxi trips. We can leverage the same zones from the Taxi trips to compare subway activity, traffic, trip duration, and delays for roughly the same trip. Being able to compare this information can allow us to provide insights into historical trends and determine when and where certain modes of transportation are more efficient.
    - Bike: Bike will provide another transportation method for comparison. Similarly to taxi and subway data, detailed trip information allows us to analyze activity across zones and time. Combining all three sets of information will allow us to find which modes of transportation meet a customer's needs most often.
- Regional-level
    - Our group aims to compare different modes of transportation within five major areas in relation to geographically identified boroughs of New York City, which are the following: Manhattan, Brooklyn, Queens, The Bronx, and Staten Island
    - When comparing, our group will use universal metrics common to all three datasets available for buses, trains, and taxis, which are the duration of trip, fare, time and day of the trip, and number of trips given each of the five areas
    - Our group will visualize in aggregated amount for the following, but not limited to:
        - Trips per month (market share)
        - Unique vehicles/ drivers per month
        - Average duration of the trip
        - Average fare of the trip
- City-level
    - The same approach will be employed as regional-level, but everything will be aggregated together at city-level, which refers to New York City as a whole

Data

Our group will be using a total of three data sets. Each data set pertains to a specific mode of public transportation.

TLC Dataset
Our group will be using the monthly dataset for green and yellow taxis available on the official NYC government TLC website. As stated before, the variables of interest are the data and time when the meter was engaged and disengaged, trip distance in miles, the taxi zone in which the taximeter was engaged and disengaged, fare amount and total amount. The data contains complete trip information dating back to 2009.

MTA Subway Time Historical Data
We will also be using data from New York City Transit's feed of real-time train arrival estimates. This data collection began in 2014 and is ongoing. The data comes from the GTFS-realtime

feed and is in GTFS format. GTFS stands for General Transit Specification, an open data format for public transportation data created by Google. This provides extremely detailed data on each trip, and is comprehensive dating back six years.

City Bike Trip Data
Citi Bikes also provides detailed trip data for use within New York City. This information will provide another transportation alternative to taxi and subway data. As with the other data, this dataset contains highly detailed trip information, including trip duration, start time, end time, and beginning and ending longitude/latitude.

Visualizations

Maps: this will be our primary means of visualization, as it will be the most effective way of showing which areas of New York City at regional or city level are most congested or the highest fare or vice versa in relation to mode of transportation. In other words, it will be a powerful visual tool to quickly inform people which mode of transportation is most ideal.

Line charts: this will be a useful tool for quickly summarizing in the format of time-series trends for certain metrics, such as number of trips, growth in trips, market share of different modes of transportation, etc. Therefore, it will be a very informative tool in demonstrating either growth or fluctuation from month to month for single point metrics.

Bar charts: will be another effective method for comparing information between transportation methods, including trip duration, cost, wait time, etc.

Interactive charts: Many of these charts will be displayed with interactive options, using plotly. This can help the user further explore the data. This will be particularly effective when combined with heatmaps, line charts (showing use over time), and many other visualizations using transportation data.