EDA Report for MA615 Final Report

Sutong Zhang

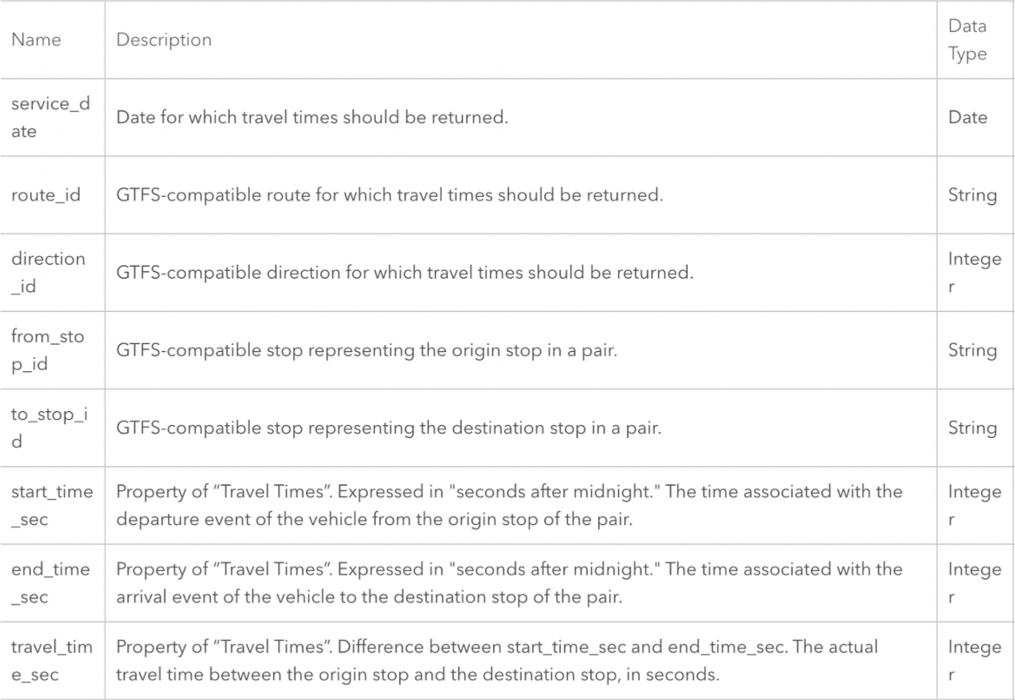
2022/12/15

Method

Data Processing

I found the dataset of 2022 MBTA Travel Times through the website of MBTA (https://mbta-massdot.opendata.arcgis.com/datasets/MassDOT::mbta-travel-times-2022/about). This dataset shows travel times between origin and destination pairs on a single line. It contains data up to the most recent completed quarter for 2022.

The following is the dictionary of the data:



Name	Description	Data Type
service_date	Date for which travel times should be returned.	Date
route_id	GTFS-compatible route for which travel times should be returned.	String
direction_id	GTFS-compatible direction for which travel times should be returned.	Integer
from_stop_id	GTFS-compatible stop representing the origin stop in a pair.	String
to_stop_id	GTFS-compatible stop representing the destination stop in a pair.	String
start_time_sec	Property of "Travel Times". Expressed in "seconds after midnight." The time associated with the departure event of the vehicle from the origin stop of the pair.	Integer
end_time_sec	Property of "Travel Times". Expressed in "seconds after midnight." The time associated with the arrival event of the vehicle to the destination stop of the pair.	Integer
travel_time_sec	Property of "Travel Times". Difference between start_time_sec and end_time_sec. The actual travel time between the origin stop and the destination stop, in seconds.	Integer

Through this dataset, I want to find out the relationship between travel time of the Rapid Transit in Boston and other factors such as Line, Date etc.

EDA

This dataset contains a lot of information, so I start to clean the data. First, I remove columns that unnecessary to me: **direction_id, from_stop_id, to_stop_id**. Theses variables only shows some location details that cannot be used in my report. Next, I delete **start_time_sec** and **end_time_sec**, which can be concluded information in the next column.

Secondly, I decide to merge the rows that share the same information in **service_date** and **routine_id**, and calculate the mean of the travel time as my vectors. Then since I decide to use the January 2022 data as my sample, I subset the data into service time that only in January. Also, I want to determine the weekdays and weekends' influence to the sample. So I create two more dataset to reach the goal.

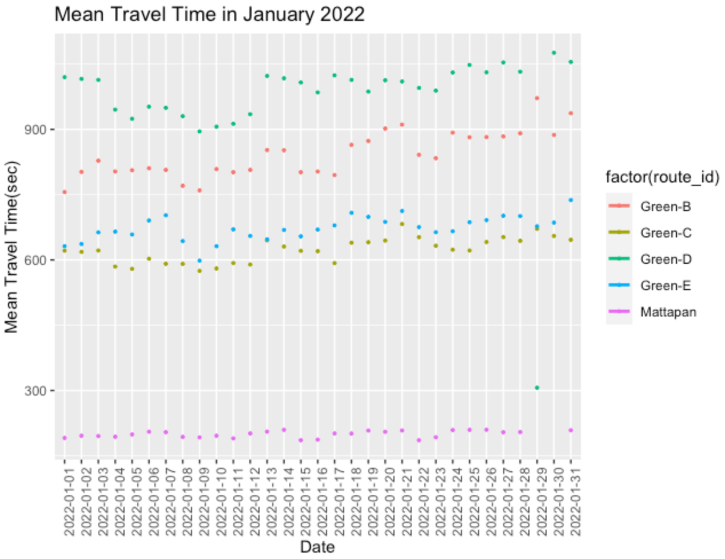


Figure 1: Mean Travel Time in January 2022

This plot shows the mean travel time in January 2022 by different lines in Boston. We can find that the Green-D line has the most Travel Time and the Mattapan line has the least travel time. Also, we can find that each lines has a stable mean of travel time through the month.

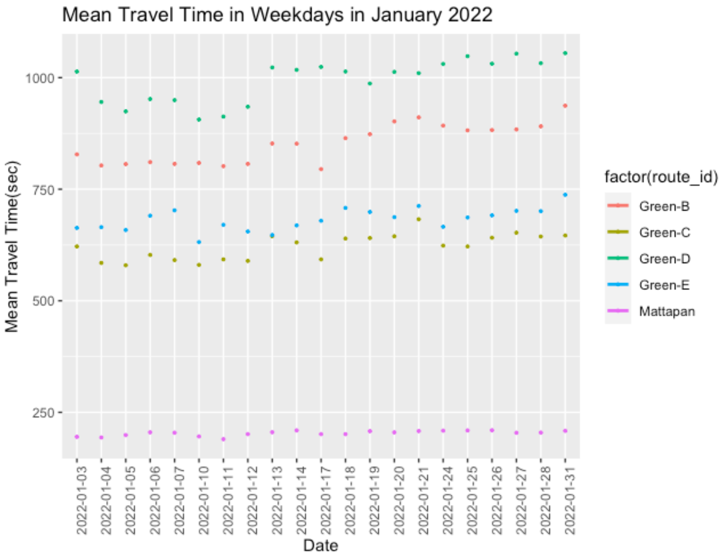


Figure 2 : Mean Travel Time in Weekdays in January 2022

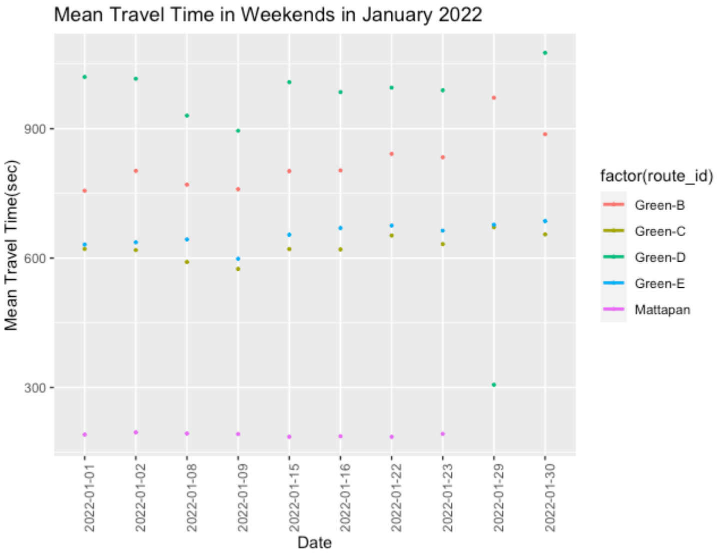


Figure 3: Mean Travel Time in Weekends in January 2022

These two plots show the difference in mean of travel time in weekdays and weekends in January 2022. Through the comparison, we can find that the weekdays have more travel time than weekends through the same distance.