EDA of MBTA Transit Services

Hao He

2022-12-07

Introduction

In this report, the historical record of MBTA transit services from the MBTA blue book open data portal and the archive is analyzed to figure out any pattern (or distribution) on travel times for different category of MBTA transit given the same stop pairs and therefore provide insights to MBTA riders about whether the estimated time and routes are reliable.

For the time period of MBTA services, I picked the 14th - 21th days from each of the month from November 2021 to October 2022, as I think the schedule may remain unchanged at the most of time in the middle of the month.

Below are some terminologies that would be helpful to understand the data:

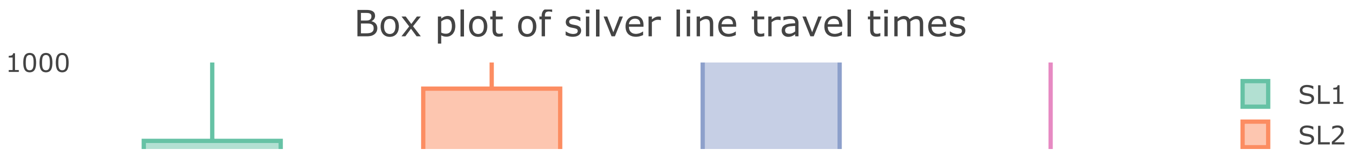
- `Stops` are associated with `trips` . A stop for any type of transit services is usually unchanged unless exception such as holiday occurs. Stops in the documentation could be numeric or combined with character, such as BNT-0000, or even complete in character, like Boat-Charlestown.
- `Trip` consists of several stops. Each trip refers to a single "run" of a vehicle along a particular route. In a route, a vehicle may not serve every one of its stops at all times. For example, on weekends it might skip several stops if there is a construction or a holiday event.

EDA

The data I used for examine travel times for bus and rapid transit are from MBTA blue book open data portal and the the data for the rest of modes including ferry and commuter rails are extracted from the archive GTFS feed of the developer site. I looked at several distributions of travel times/headways on each category of MBTA transit. Given the schedule in May had less changes basically (except for Memorial day), most of my analysis are produced for duration between May 14th and 21th.

Bus

Headway is the actual time between the trip and the previous trip at the stop, in seconds. This can be used for checking the reliability of bus service at a given stop for a trip. Since the headway can be only evaluated when using the headway standard, the missing values of headway caused by schedule standard are removed. I found silver line bus generally has more route stop pairs, thus, I narrow down to draw the distribution of the silver lines in May on for exploration. The stop at summer street has a extremely higher travel times compared to other stops.



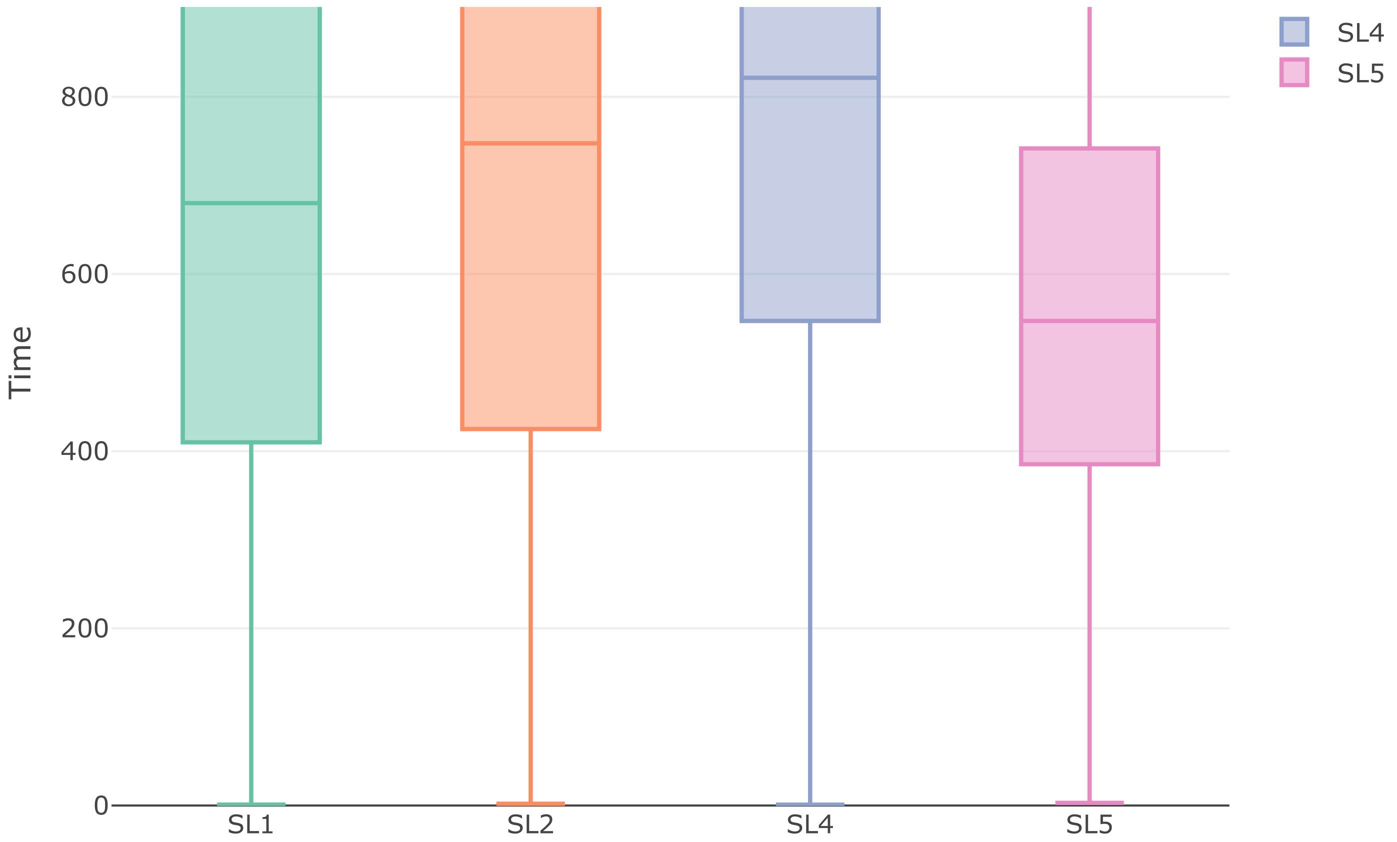


Figure 1. Silver Line bus travel times

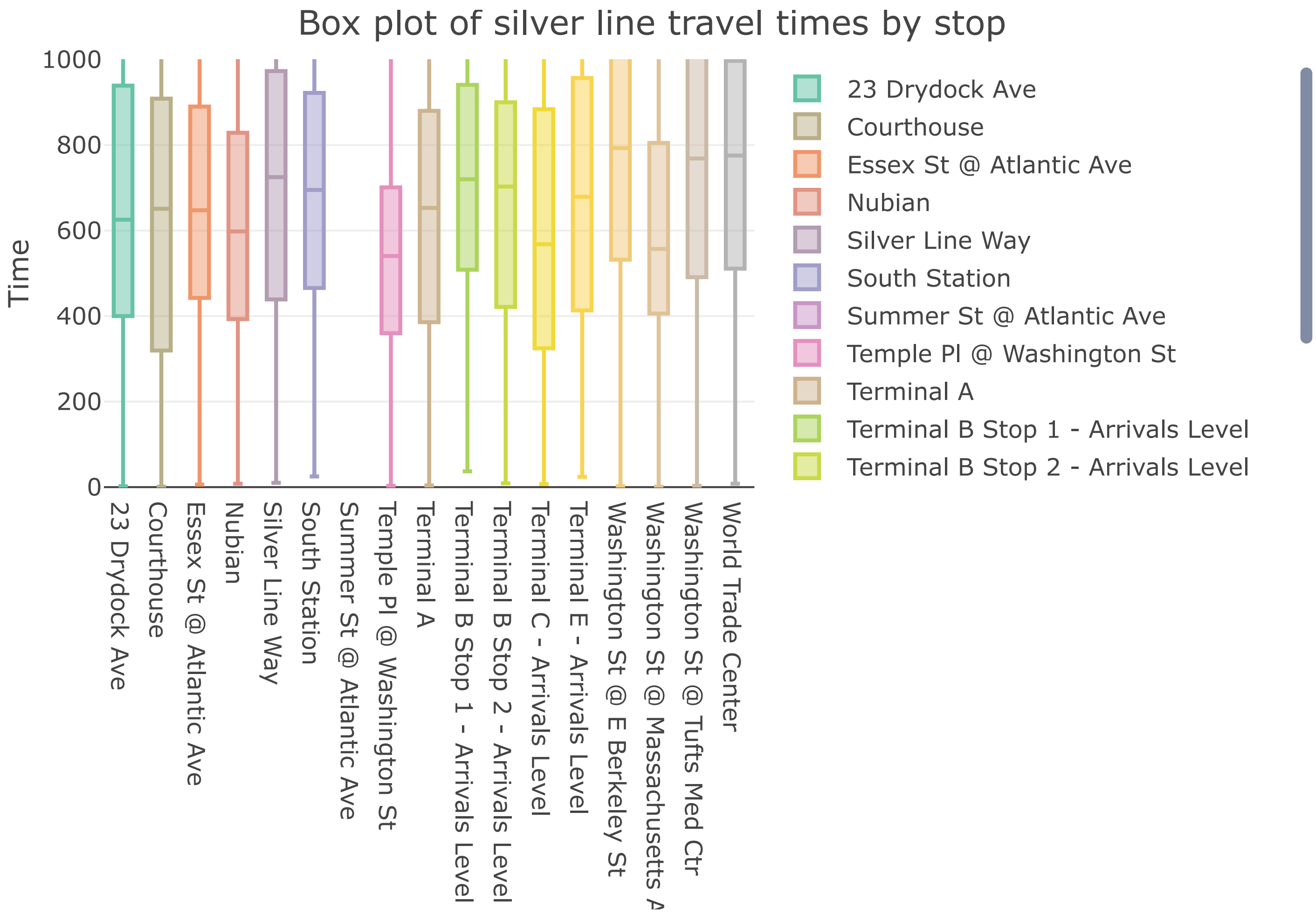


Figure 1. Silver Line bus travel times

Below I found that during my selected week, the exception on 20220521 causes the Saturday red line service towards Alewife be removed (1 means service added to schedule on this date, 2 means removed from schedule on this date). And why this has to be an exception is not clear, which would be good to for MBTA to show to riders otherwise this does not benefit a developer who wants to use this kind of information for prediction.

service_id	exception_date	holiday_name	exception_type
LRV22022-hlb22016-Sa-01	20220618		1
PRIV22022-hpj22017-Su-01	20220530	Memorial Day	1
RTL22022-hms22016-Sa-01	20220521		2
RTL22022-hms22017-Su-01	20220522		2
SpringWeekday	20220530	Memorial Day	2

Rapid transit

Below are the first 6 rows show the travel times for a given stop pairs along a route. Since the overlaid distribution of each line's travel times is a blur, I ploted them separately. It can seen that each line's distribution is right-skewed and the travel time in seconds is concentrated around 400 seconds or above except for Mattpan. This is not a surprise because Mattpan is actually part of the red line and it services much less stops than other subways for us to see an obvious trend. This plot also reflects that at the most of time other train seems less likely to running late compared to red line and Green-B as they have a higher frequency around higher travel times.

service_date	from_stop_id	to_stop_id	route_id	direction_id	start_time_sec	end_time_sec	travel_time_sec
2022-01-14	70016	70001	Orange	0	41936	42968	1032
2022-01-14	70016	70001	Orange	0	21363	22420	1057
2022-01-14	70016	70001	Orange	0	22209	23270	1061
2022-01-14	70016	70001	Orange	0	20580	21625	1045
2022-01-14	70016	70001	Orange	0	48071	49147	1076
2022-01-14	70016	70001	Orange	0	48591	49654	1063

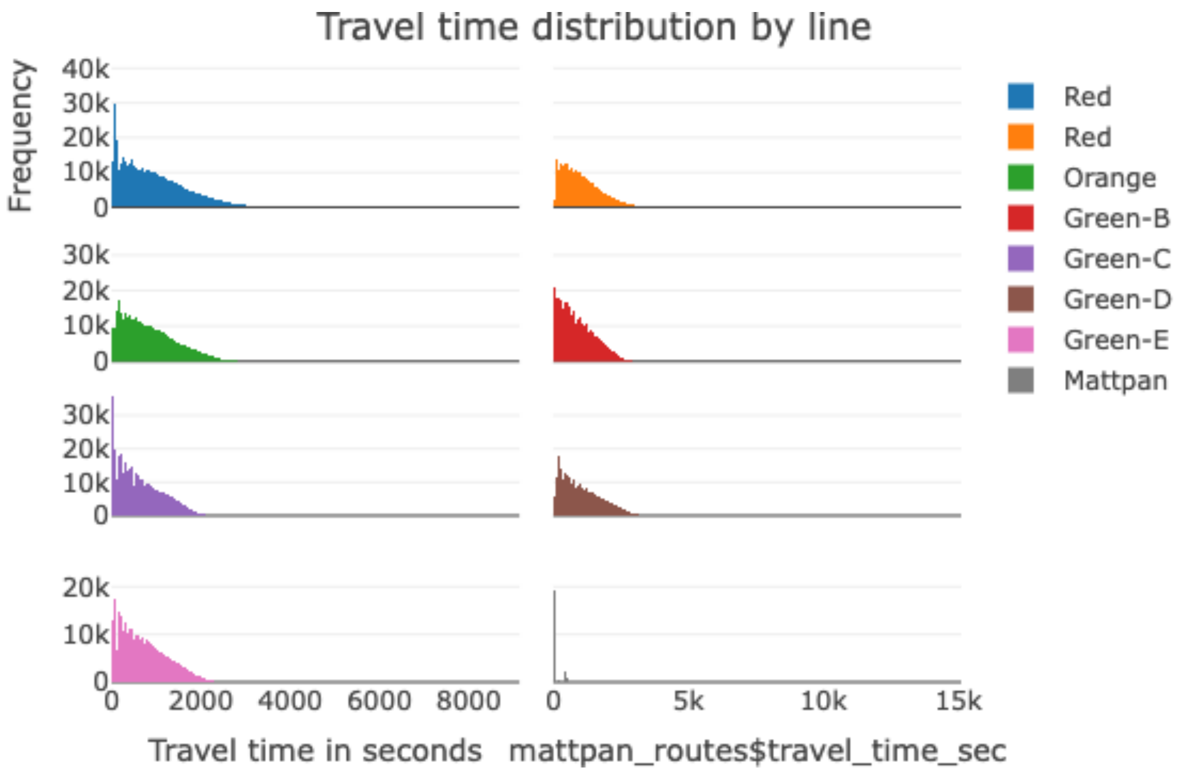


Figure 2. ravel time distribution by line

```
p_red_routes<- plot_ly(x=~red_routes$travel_time_sec, name = "Red", type = "histogram
") %>%
  layout( title = "Red Line Travel Times Distribution", xaxis=list(range=c(300,3500),
title=" travel time in seconds"), yaxis=list(title="frequency"))
p_red_routes

p_blue_routes<- plot_ly(x=~red_routes$travel_time_sec, name = "Red", type = "histogra
m") %>%
  layout( title = "Blue Line Travel Times Distribution", xaxis=list(range=c(300,3500)
,title= "travel time in seconds"), yaxis=list(title="frequency"))
p_blue_routes

p_orange_routes<- plot_ly(x=~orange_routes$travel_time_sec, name = "Orange", type = "
histogram") %>%
  layout(title = "Orange Line Travel Times Distribution", xaxis=list(range=c(300,3500
),title= "travel time in seconds"))
p_orange_routes

p_greenb_routes<- plot_ly(x=~greenb_routes$travel_time_sec, name = "Green-B", type =
"histogram") %>%layout(title = "Green-B Line Travel Times Distribution", xaxis=list(r
ange=c(300,3500),title= "travel time in seconds"))


p_greenc_routes<- plot_ly(x=~greenc_routes$travel_time_sec, name = "Green-C", type =
"histogram") %>%layout(title = "Green-C Line Travel Times Distribution", xaxis=list(r
ange=c(300,3500),title= "travel time in seconds"))
p_greenc_routes


p_greend_routes<- plot_ly(x=~greend_routes$travel_time_sec, name = "Green-D", type =
"histogram")%>%layout(title = "Green-D Line Travel Times Distribution", xaxis=list(ra
nge=c(300,3500),title= "travel time in seconds"))
p_greend_routes


p_greene_routes<- plot_ly(x=~greene_routes$travel_time_sec, name = "Green-E", type =
"histogram") %>%layout(title = "Green-E Line Travel Times Distribution", xaxis=list(r
ange=c(300,3500),title= "travel time in seconds"))
p_greene_routes


p_mattpan_routes<- plot_ly(x=~mattpan_routes$travel_time_sec, name = "Mattpan", type
= "histogram") %>%layout(title = "Mattpan Line Travel Times Distribution", xaxis=list
(range=c(300,3500),title= "travel time in seconds"))
p_mattpan_routes
```

To better understanding whether the travel times affected by the number of route operating for each subway line, I also made a bar chart to display the number of routes in MBTA by subway lines. Green lines generally has more routes (except for E) and red line is the subway line with the second highest number of routes.

Number of routes in MBTA by subway lines

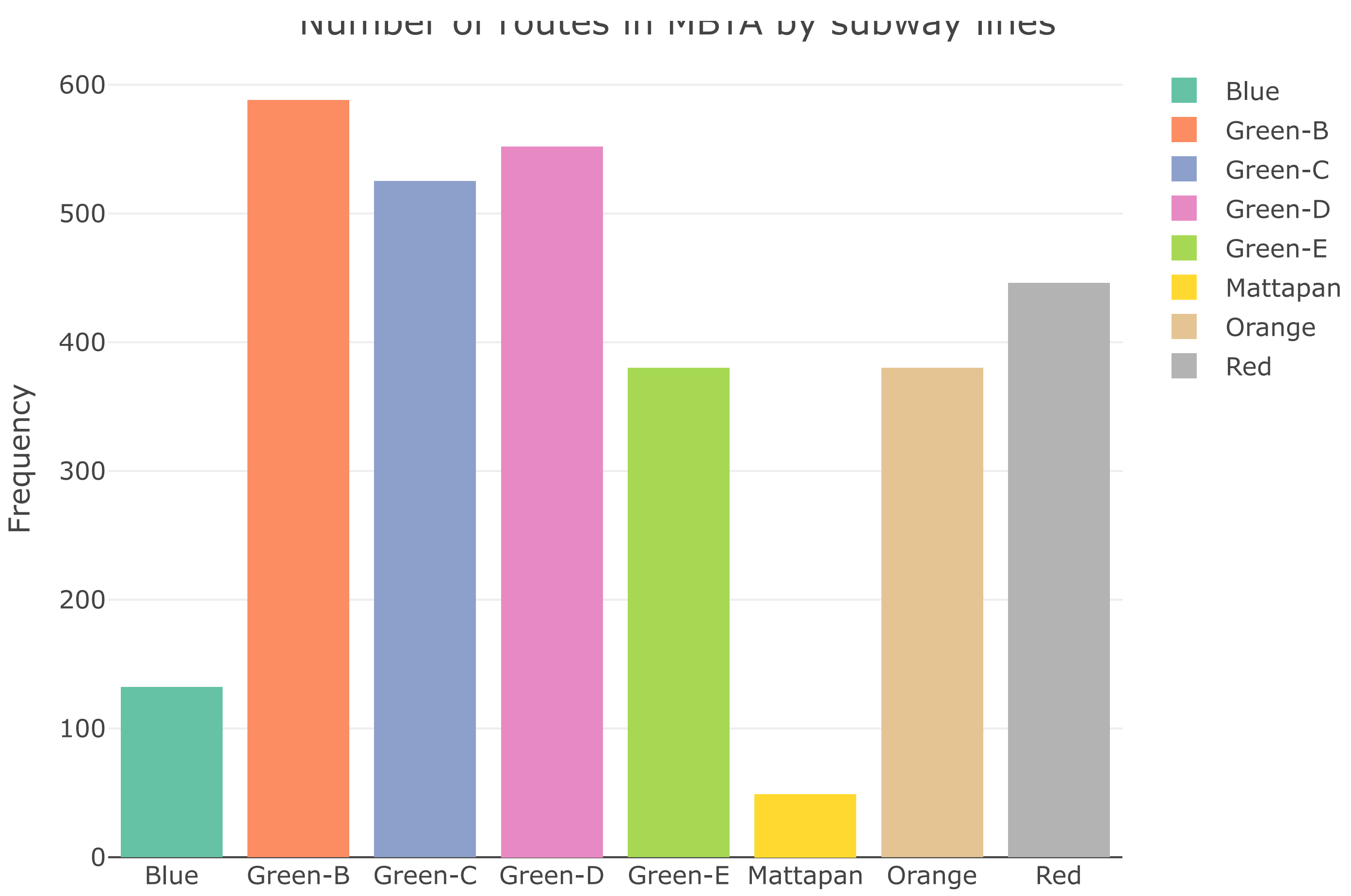
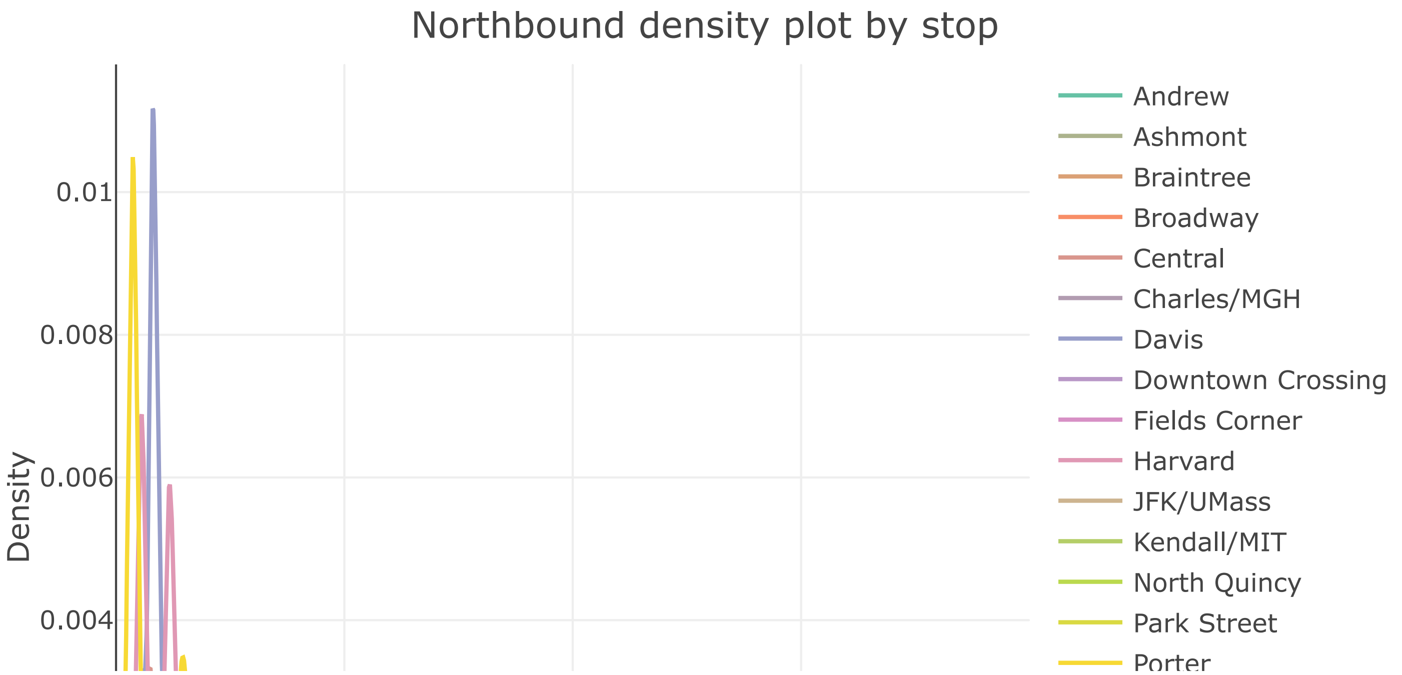


Figure 3. Number of routes in MBTA by subway lines

Red Line and Orange Line

I used to live in Quincy for 3 years and the travel time from North Quincy station is usually pretty long from my experience, here I focused more on analyze Red Line's travel times and its performance compared to orange line because both of them are either northbound or southbound. Based on the density plots of red line, the stop around Quincy area overall has a higher probability of having a higher travel time regardless of the direction of the red line when compared to stops like Harvard.

```
####RED LINE
# match stop_lon and stop_lat
red_routes$from_stop_id <- as.integer(red_routes$from_stop_id)
red_routes$to_stop_id <- as.integer(red_routes$to_stop_id)

# match departure stop_id
red_travel_times<- merge(red_routes, stop_coor, by.x = "from_stop_id", by.y = "stop_id") %>% rename(c(dept_lat = stop_lat,dept_lon = stop_lon, dept_name = stop_name))

# match arrival stop_id
red_travel_times <- merge(x = red_travel_times, y = stop_coor, by.x = "to_stop_id", by.y = "stop_id") %>% rename(c(arrv_lat = stop_lat,arrv_lon = stop_lon,arrv_name = stop_name))

# create a density plot at each stop
# reference: https://plotly.com/r/line-charts/
# Northbound
red_to_north<- red_travel_times %>% filter(direction_id ==1)
red_dept_dens_n<- with(red_to_north, tapply(travel_time_sec, INDEX = dept_name, density)) # now this is a nested list contains density travel times grouped by each stop

# use double bracket to unlist and store the associated density into a dataframe
red_to_north_df<- data.frame(
  x = unlist(lapply(red_dept_dens_n, "[[", "x")),
  y = unlist(lapply(red_dept_dens_n, "[[", "y")),
  dept_name = rep(names(red_dept_dens_n), each = length(red_dept_dens_n[[1]]$x))
)

# Northbound density plot
plot_ly(red_to_north_df, x = ~x, y = ~y, color = ~dept_name)%>% add_lines() %>%
  layout(title = "Northbound density plot by stop", yaxis=list(title="Density"), xaxis=list(title="Northbound Travel Times",range = c(0,4000)))
```



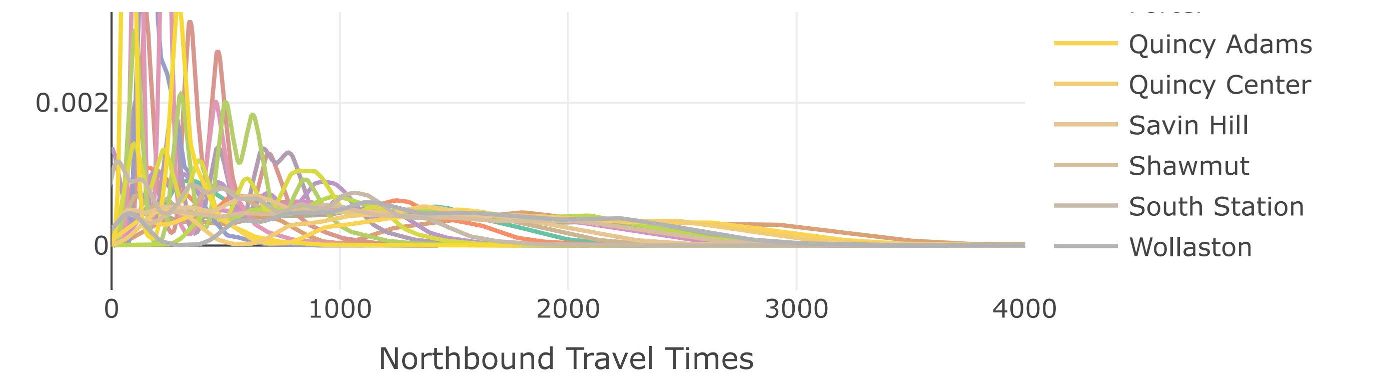
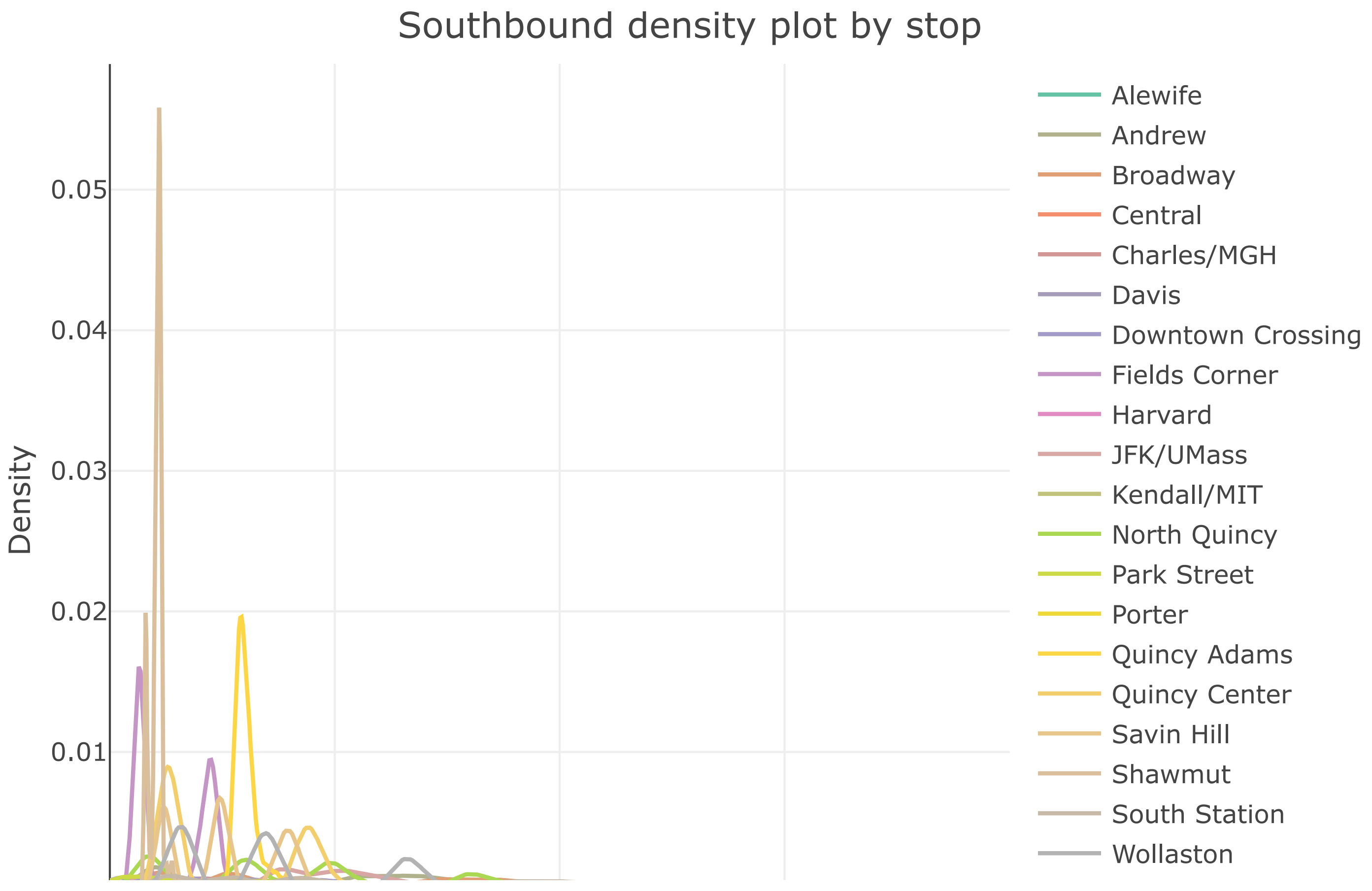


Figure 4. Red line travel time density plot

```
# Southbound
red_to_south <- red_travel_times %>% filter(direction_id == 0)
red_dept_dens_s<- with(red_to_south, tapply(travel_time_sec, INDEX = dept_name, densi
ty))
red_to_south_df<- data.frame(
  x = unlist(lapply(red_dept_dens_s, "[[", "x")),
  y = unlist(lapply(red_dept_dens_s, "[[", "y")),
  dept_name = rep(names(red_dept_dens_s), each = length(red_dept_dens_s[[1]]$x))
)

# Southbound density plot
plot_ly(red_to_south_df, x = ~x, y = ~y, color = ~dept_name)%>% add_lines() %>%
  layout(title = "Southbound density plot by stop", yaxis=list(title="Density"), xaxi
s=list(title="SouthBound Travel Times",range = c(0,2000)))
```



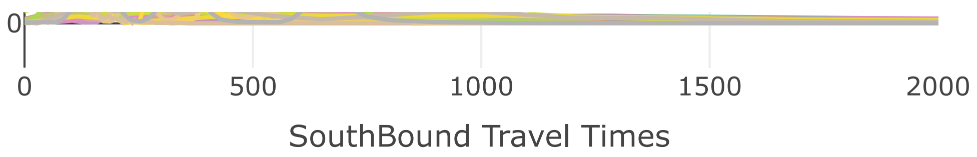
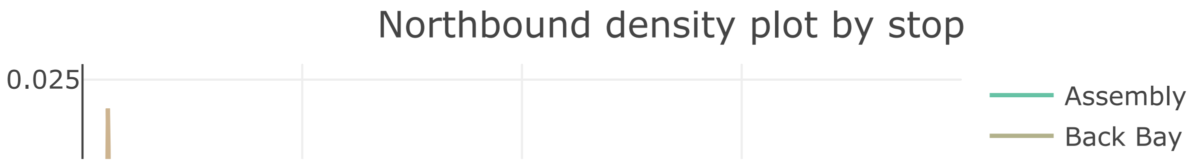


Figure 4. Red line travel time density plot

Below is the travel times density plot for orange line. According to the density plots, the higher concentration in travel times at a given stop is more related to its direction (depends on the southbound or northbound) which is more reasonable than the red line.

```
##### Orange line
# match stop_lon and stop_lat
orange_routes$from_stop_id <- as.integer(orange_routes$from_stop_id)
orange_routes$to_stop_id <- as.integer(orange_routes$to_stop_id)

# match departure stop_id
orange_travel_times<- merge(orange_routes, stop_coor, by.x = "from_stop_id", by.y = "
stop_id") %>% rename(c(dept_lat = stop_lat,dept_lon = stop_lon, dept_name = stop_name
))

# match arrival stop_id
orange_travel_times <- merge(x = orange_travel_times, y = stop_coor, by.x = "to_stop_
id", by.y = "stop_id") %>% rename(c(arrv_lat = stop_lat,arrv_lon = stop_lon,arrv_name
= stop_name))

# create a density plot at each stop
# reference: https://plotly.com/r/line-charts/
# Northbound
orange_to_north<- orange_travel_times %>% filter(direction_id ==1)
orange_dept_dens_n<- with(orange_to_north, tapply(travel_time_sec, INDEX = dept_name,
density)) # now this is a nested list contains density travel times grouped by each s
top

# use double bracket to unlist and store the associated density into a dataframe
orange_to_north_df<- data.frame(
  x = unlist(lapply(orange_dept_dens_n, "[[", "x")),
  y = unlist(lapply(orange_dept_dens_n, "[[", "y")),
  dept_name = rep(names(orange_dept_dens_n), each = length(orange_dept_dens_n[[1]]$x)
)
)

# Northbound density plot
plot_ly(orange_to_north_df, x = ~x, y = ~y, color = ~dept_name)%>% add_lines() %>%
  layout(title = "Northbound density plot by stop", yaxis=list(title="Density"), xaxi
s=list(title="Northbound Travel Times",range = c(0,4000)))
```



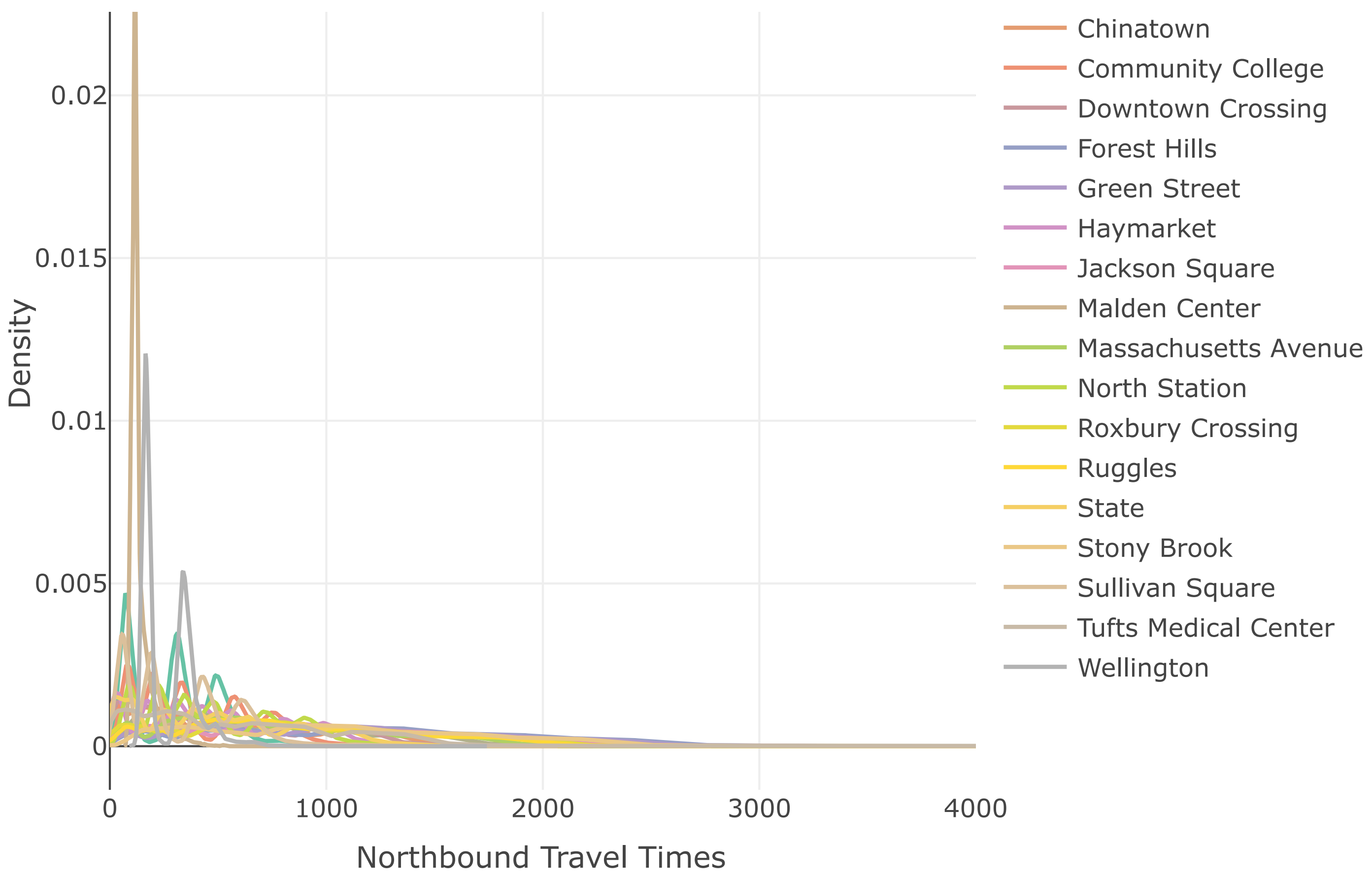
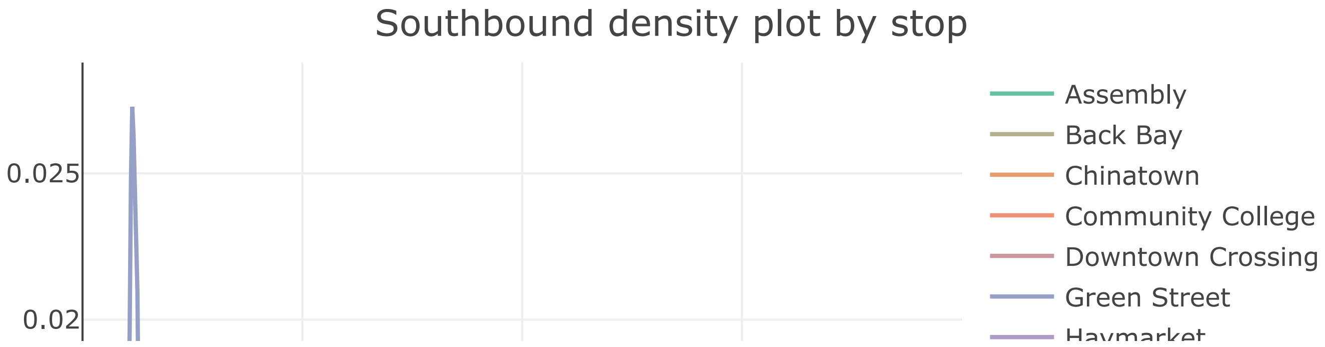


Figure 5. Orange line travel time density plot

```
# Southbound
orange_to_south <- orange_travel_times %>% filter(direction_id == 0)
orange_dept_dens_s<- with(orange_to_south, tapply(travel_time_sec, INDEX = dept_name,
density))
orange_to_south_df<- data.frame(
  x = unlist(lapply(orange_dept_dens_s, "[[", "x")),
  y = unlist(lapply(orange_dept_dens_s, "[[", "y")),
  dept_name = rep(names(orange_dept_dens_s), each = length(orange_dept_dens_s[[1]]$x)
)
)

# Southbound density plot
plot_ly(orange_to_south_df, x = ~x, y = ~y, color = ~dept_name)%>% add_lines() %>%
  layout(title = "Southbound density plot by stop", yaxis=list(title="Density"), xaxi
s=list(title="SouthBound Travel Times",range = c(0,2000)))
```



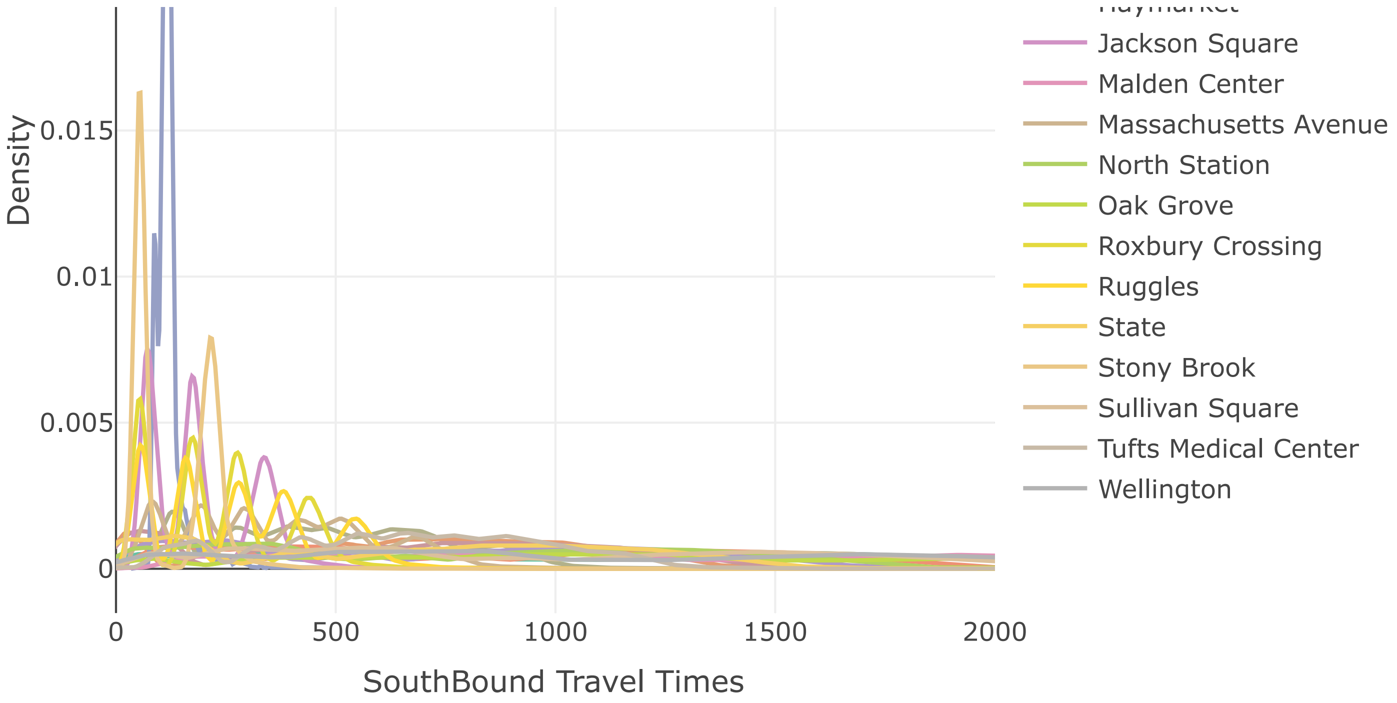


Figure 5. Orange line travel time density plot

Conlusion

Below is a summary of findings on the travel times and headways for bus and subway:

- Bus: Among all silver lines, SL5 has a lower headways than other silver line buses. So it is more likely for SL5 to
- Rapid transit: Red line has a longer travel time than orange line. This may be attributed to red line usually have longer distances between stops.

Since I don't have a consolidated dataset for commuter rail and ferry. I tried to took the schedule in may to represent a normal schedule for these two category of MBTA transit services.

This report has several limitations including limited time, incomplete data and the computation capability issue. And the way I preprocessing data is also affecting the evaluation of MBTA services. Besides, without the weather information, I tried to compare the travel times in September and that in May, these aren't too much changes, so I guess this is also a shortcoming of this analysis. The confidence interval should be considered if sampling is completed or a model is fitted for prediction. For future work, I can continue digging deeper into each month's MBTA services if I have more information about the dwell time for vehicles. I can combine this with headyways and travel times to see if there are a better measurement to compared with benchmark scheduled time. With more information about how the route may change by holidays and weather, I may have a better estimate on if the train/bus will be on time.