HRT EDA

Xu Luo

2022-12-17

#Read data In this EDA report, I download the MBTA Travel Times data. However, due to the size of data set(around 3.3Gb), I only randomly pick a week from each months from Nov.2021 - Oct.2022 for HRT data. I didn't included LRT and Bus Data since I think this data is enough for the main part of the project.

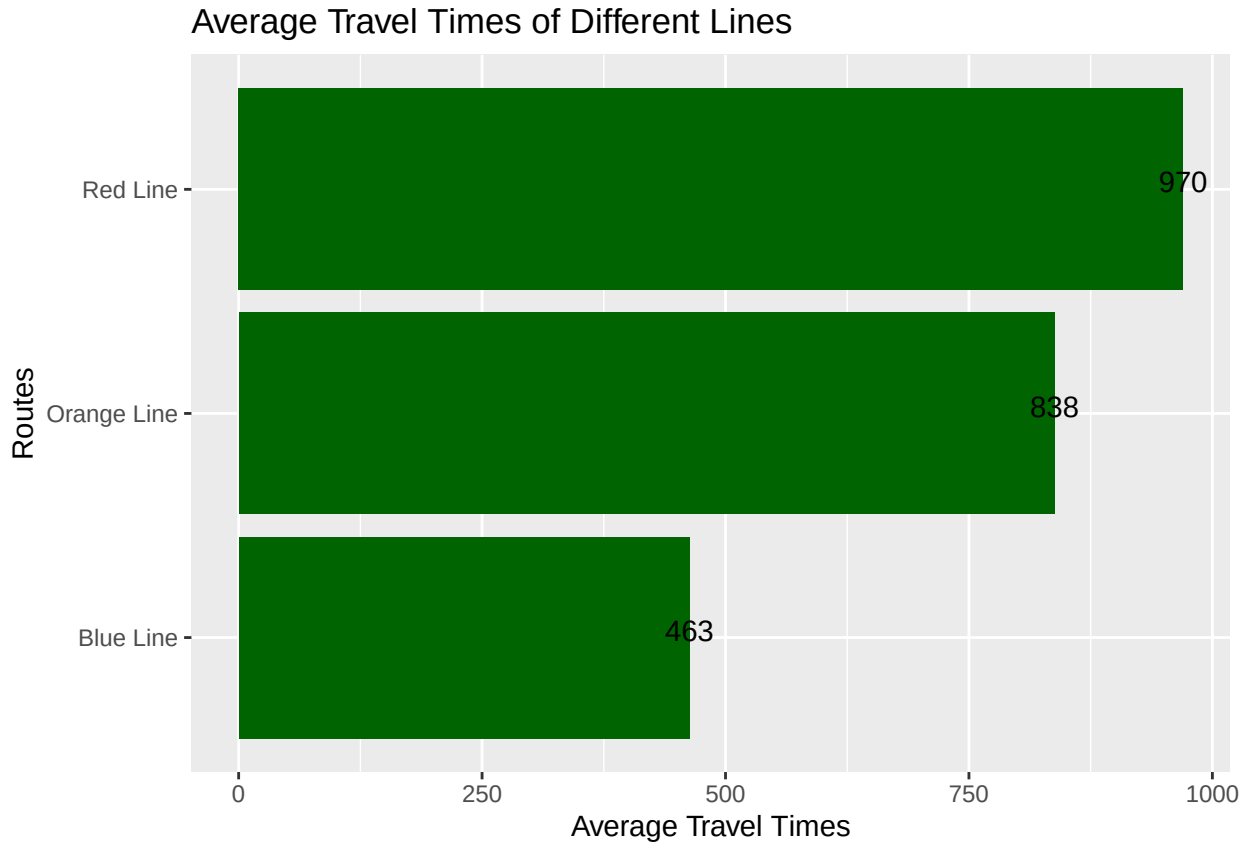
For the data cleaning process, I wrote functions to randomly select week from the original data set, and I output a selected data set call"HRT_year". I put the code in into another rmd file, because of the large data size .

#EDA In this HRT_year data, I subset 3 data set from Orange, Blue and Red Line, and assign each row a corresponding weekdays and months name

```
## [1] "English_United States.1252"
```

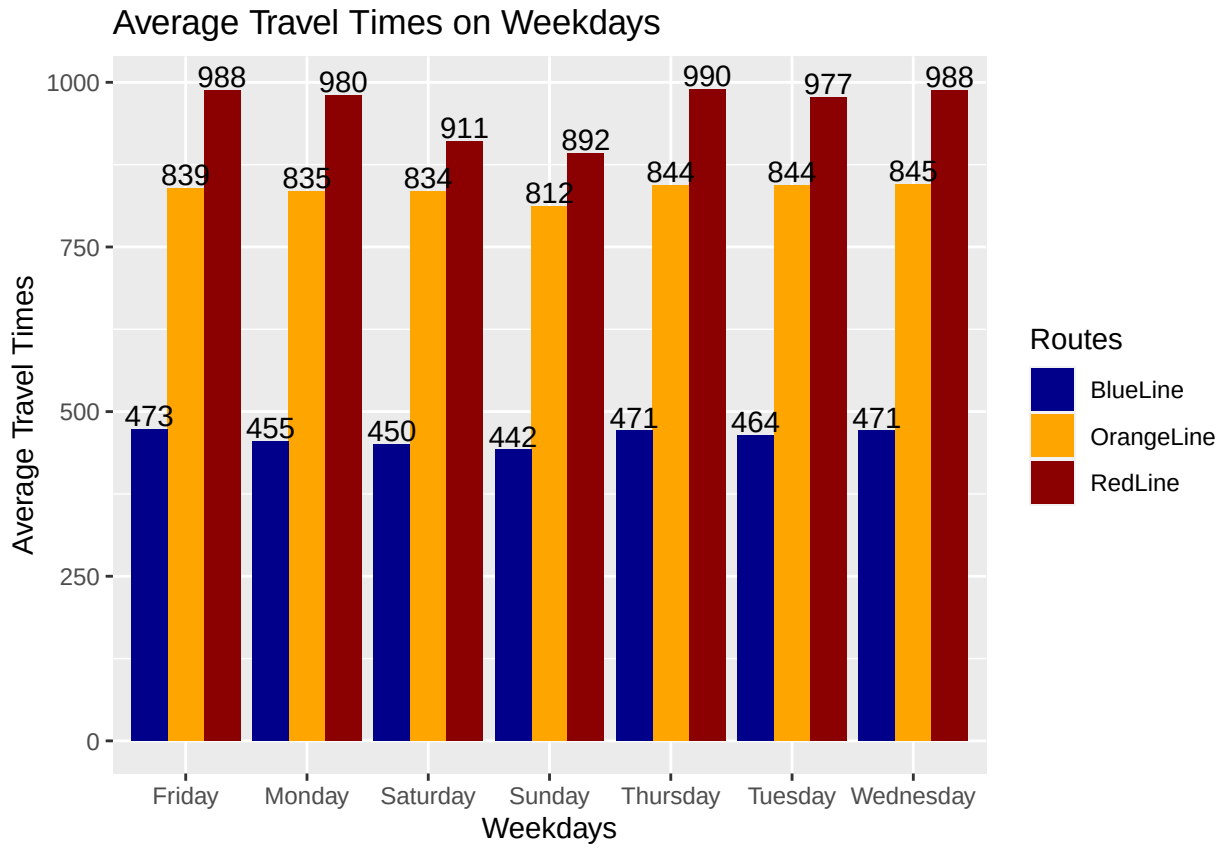
Plot1 average travel time of Orange, Blue, Red Line

In this plot, we can know that the red line has longest average travel time among all 3 lines



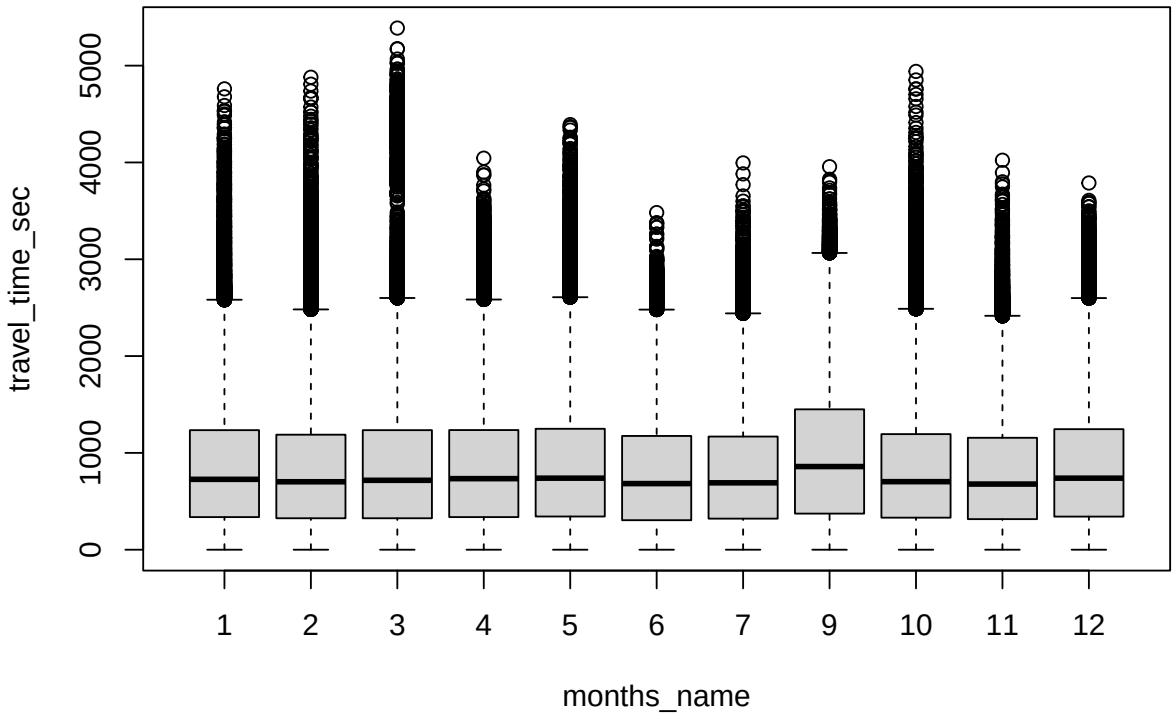
Plot 2 Average Travel time of Orange, Blue, and Red Line on different weekdays:

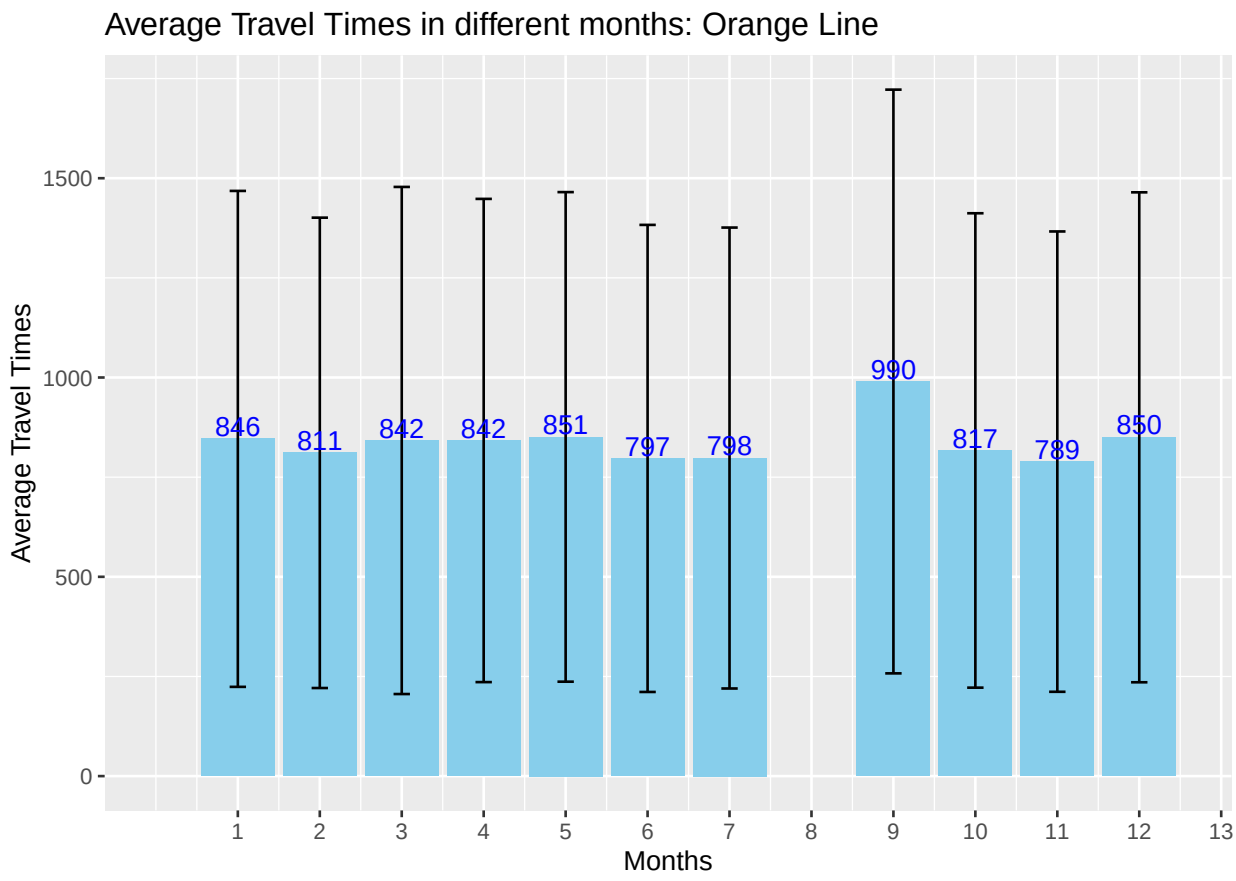
In this plot, I compared the average travel time of 3 lines on different weekdays. The Average time will be shorten in Red Line during the weekend, while other lines remain the same.



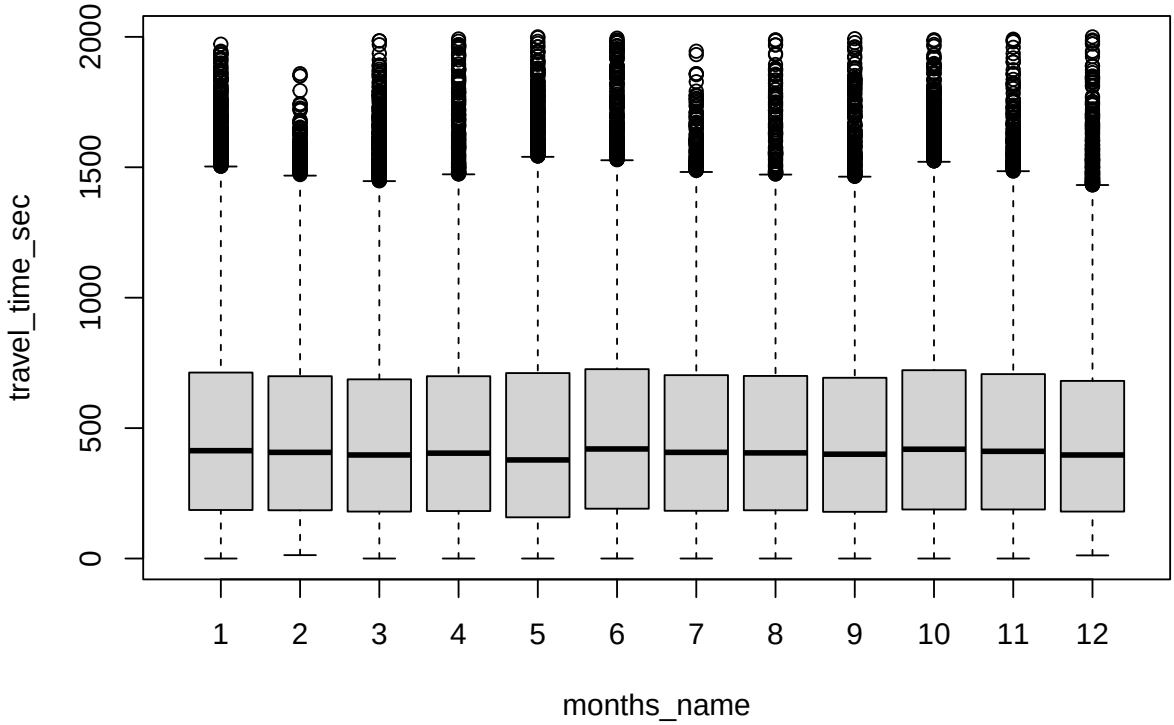
Plot 3: Boxplot and barchart of avg travel time in each month of Orange Line

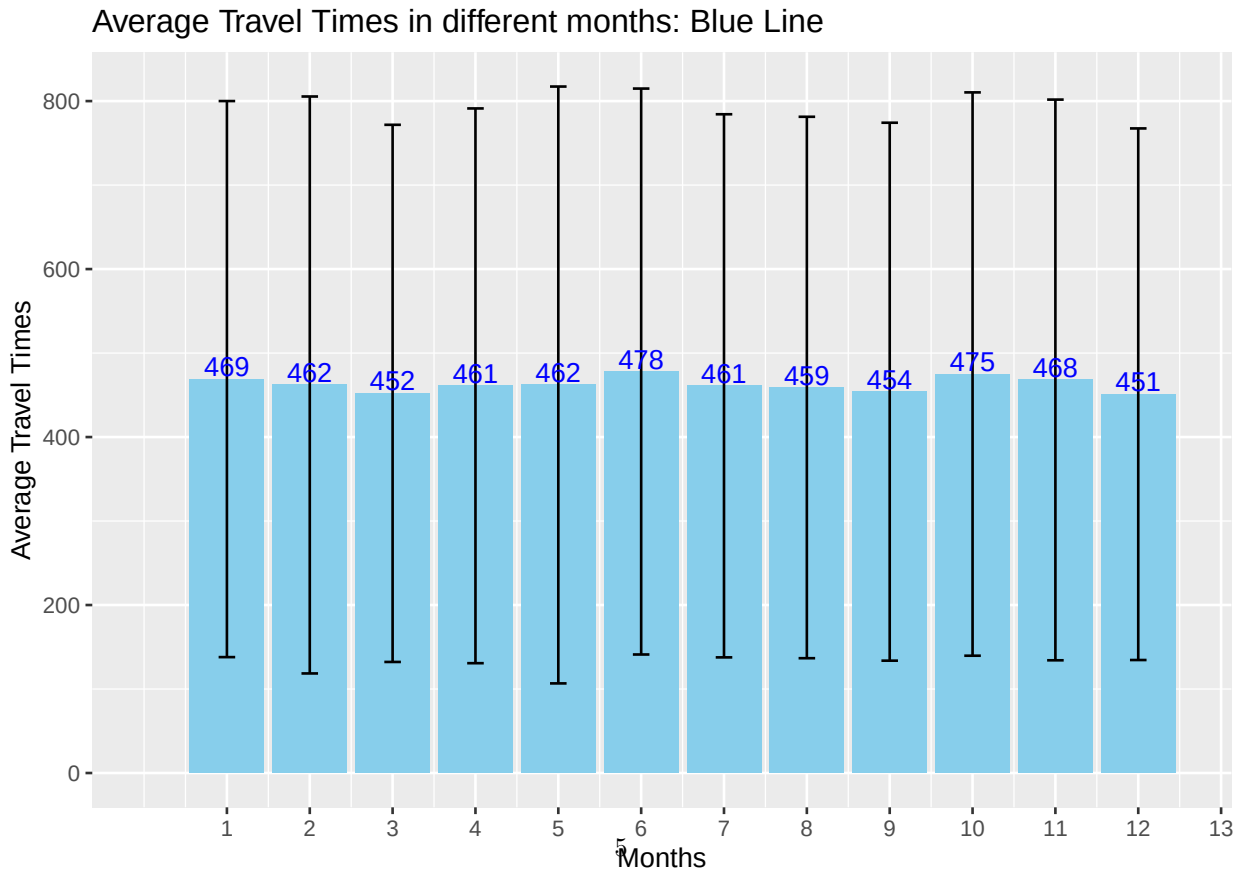
In the following plots, I compare the avg travel time in each month of each line. The travel time is stable in whole year. There is a lack of data in Orange Line since Orange Line Broke down on August.2022


travel_time_sec
0
1000
2000
3000
4000
5000
1
2
3
4
5
6
7
9
10
11
12
months_name

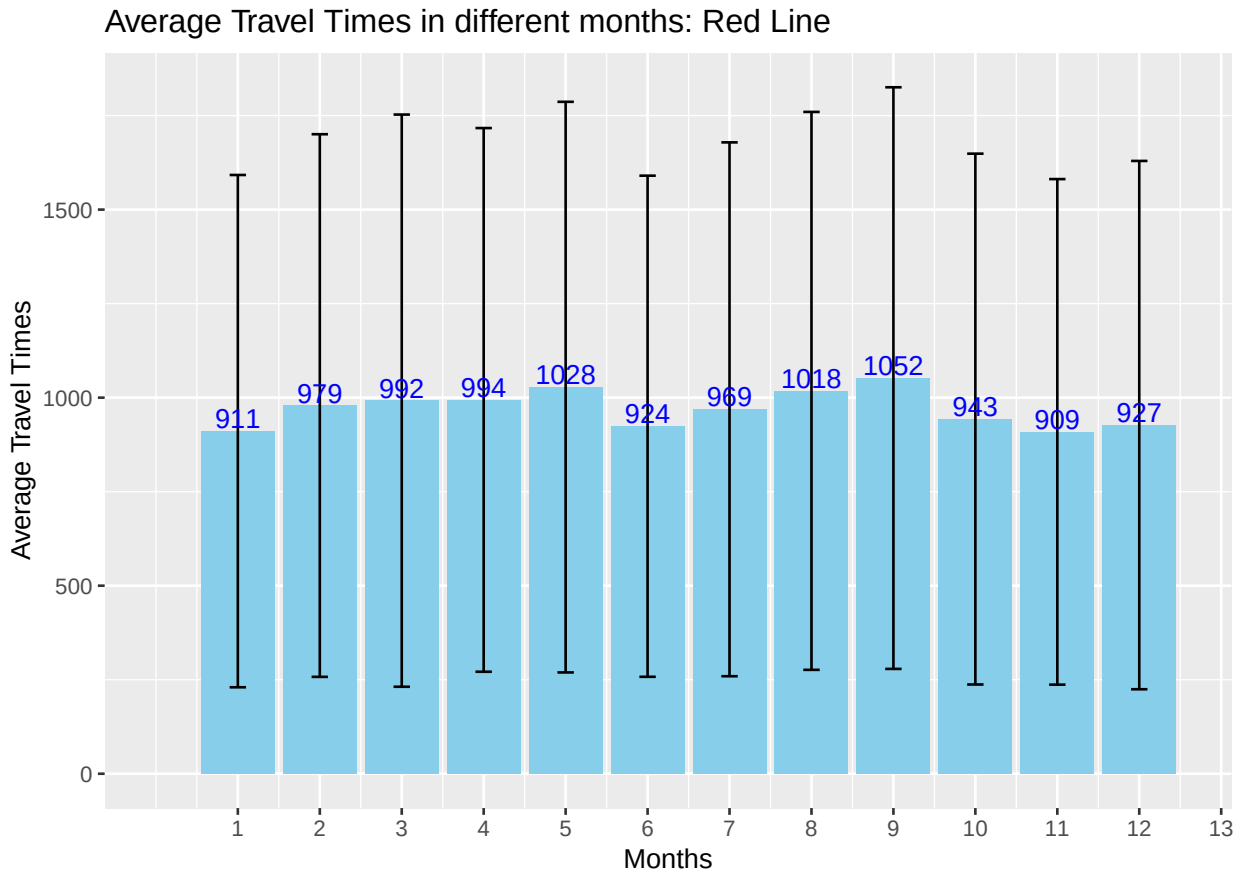

Average Travel Times in different months: Orange Line
Average Travel Times
1500
1000
500
0
846
811
842
842
851
797
798
990
817
789
850
1
2
3
4
5
6
7
8
9
10
11
12
13
Months

Plot 3b Boxplot and barchart of avg travel time in each month of Blue Line

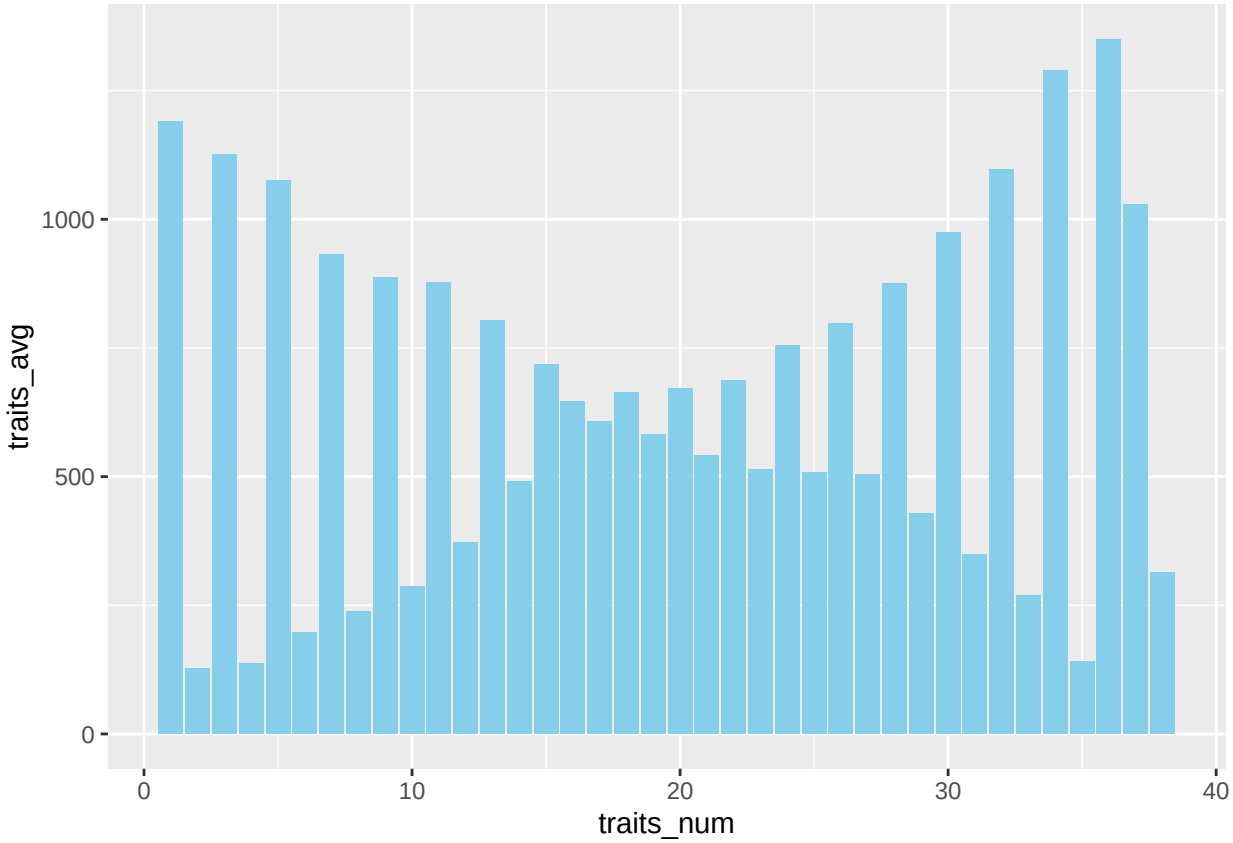




Plot 3b Boxplot and barchart of avg travel time in each month of Blue Line



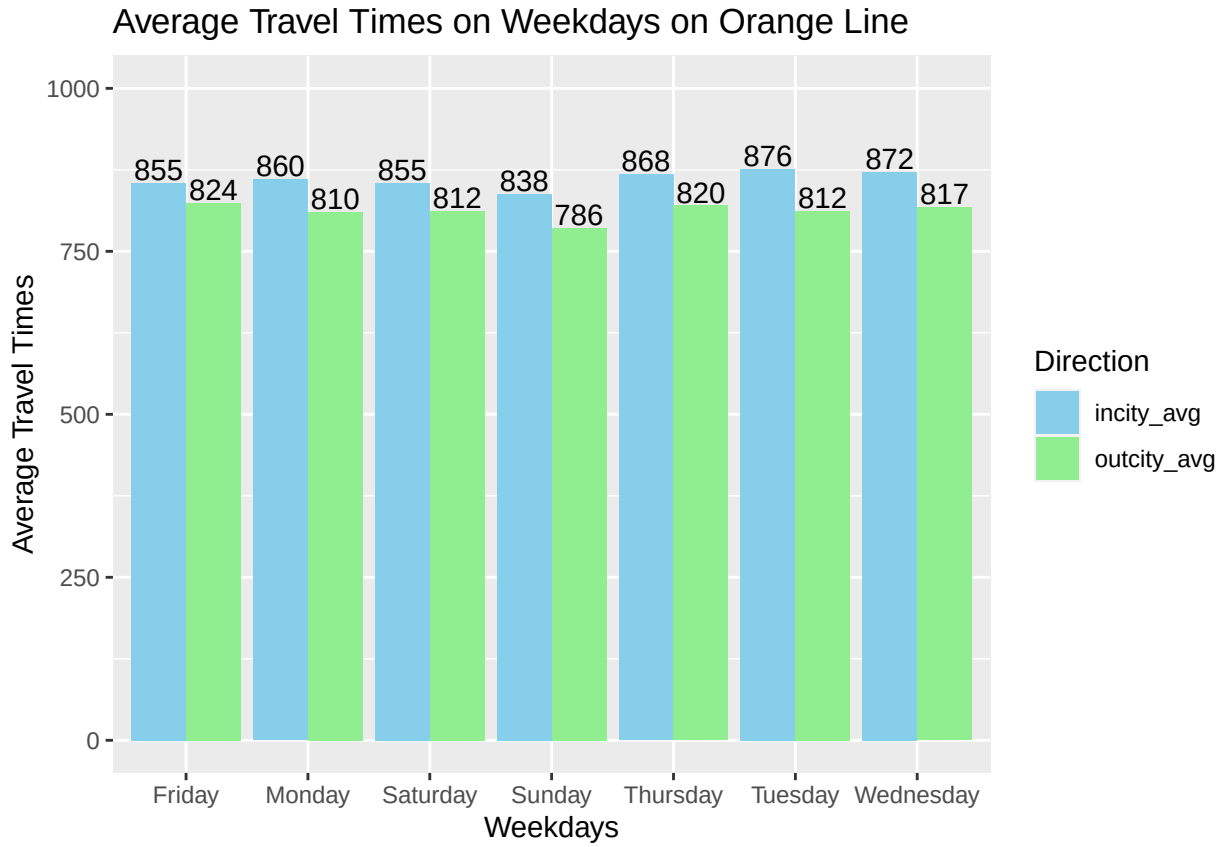
Plot 4: The average travel time of differnt traits



Whether the dircction could affect the travel times of Orange Line:

In the following plot, I compare the avg travel time of different line by different directions. The Orange Line costs more time when coming in city center, while the RED Line costs more when outbound of the city. The Blue line costs the same time when coming in and out of the city:

direction_avg	direction
863	Inbound
813	Outbound



Whether the dircction could affect the travel times of Blue Line:

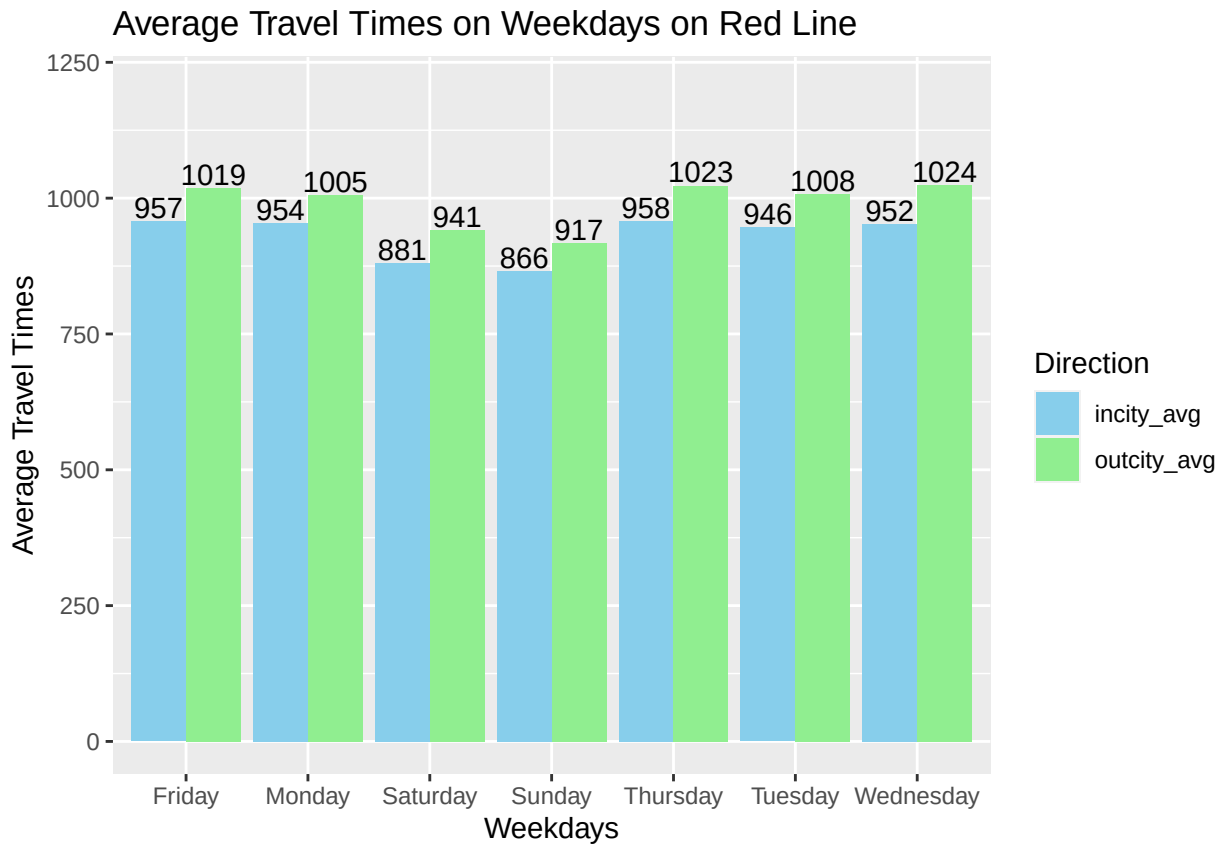
direction_avg	direction
460	Inbound
465	Outbound

wk_name	incity_avg	outcity_avg
Friday	470	477
Monday	451	458
Saturday	449	451
Sunday	441	443
Thursday	468	474
Tuesday	461	467
Wednesday	470	473

Whether the dircction could affect the travel times of Blue Line:

direction_avg	direction
940	Inbound
1000	Outbound

wk_name	incity_avg	outcity_avg
Friday	957	1019
Monday	954	1005
Saturday	881	941
Sunday	866	917
Thursday	958	1023
Tuesday	946	1008
Wednesday	952	1024



In the shiny app, since I didn't use leadlet as a method to create the map, I didn't achieve to create an interactive map. The size and the position of the map are also need to adjust.

Overall, I spent too much time on cleaning the large size data(which is almost undoable with my PC), and just left not enough time to create the shiny app.In the future study, I will be better in organizing my work and make a detailed schdule in advance.