MA 615 Final EDA

JingjianGao

2022-11-30

```
library(shiny)
library(dplyr)
library(ggplot2)
library(tidyverse)
```

Introduction

This EDA report is simply about how reliable is MBTA service. I chose to use the MBTA data from June 9th 2022 because June 8th is my birthday. However, the data from the archive is way too messy to work with. Thus, I downloaded another zip from MBTA website which contains 680 MB of data.

Data from Archive

```
MBTA <- read.csv("MBTA Data.csv") # The Data Archive

lines <- read.csv("lines.txt")
lines <- replace(lines,is.na(lines),0)

pathways <- read.csv("pathways.txt")
pathways <- replace(pathways,is.na(pathways),0)

route_patterns <- read.csv("route_patterns.txt")
route_patterns <- replace(route_patterns,is.na(route_patterns),0)

stop_times <- read.csv("stop_times.txt")
stop_times <- replace(stop_times,is.na(stop_times),0)

stops <- read.csv("stops.txt")
stops <- replace(stops,is.na(stops),0)

transfers <- read.csv("transfers.txt")
transfers <- replace(transfers,is.na(transfers),0)

routes <- read.csv("routes.txt")
routes <- replace(routes,is.na(routes),0)

# A lot of NA values, the data cannot be used for EDA report
```

Data from MBTA Travel Times 2022

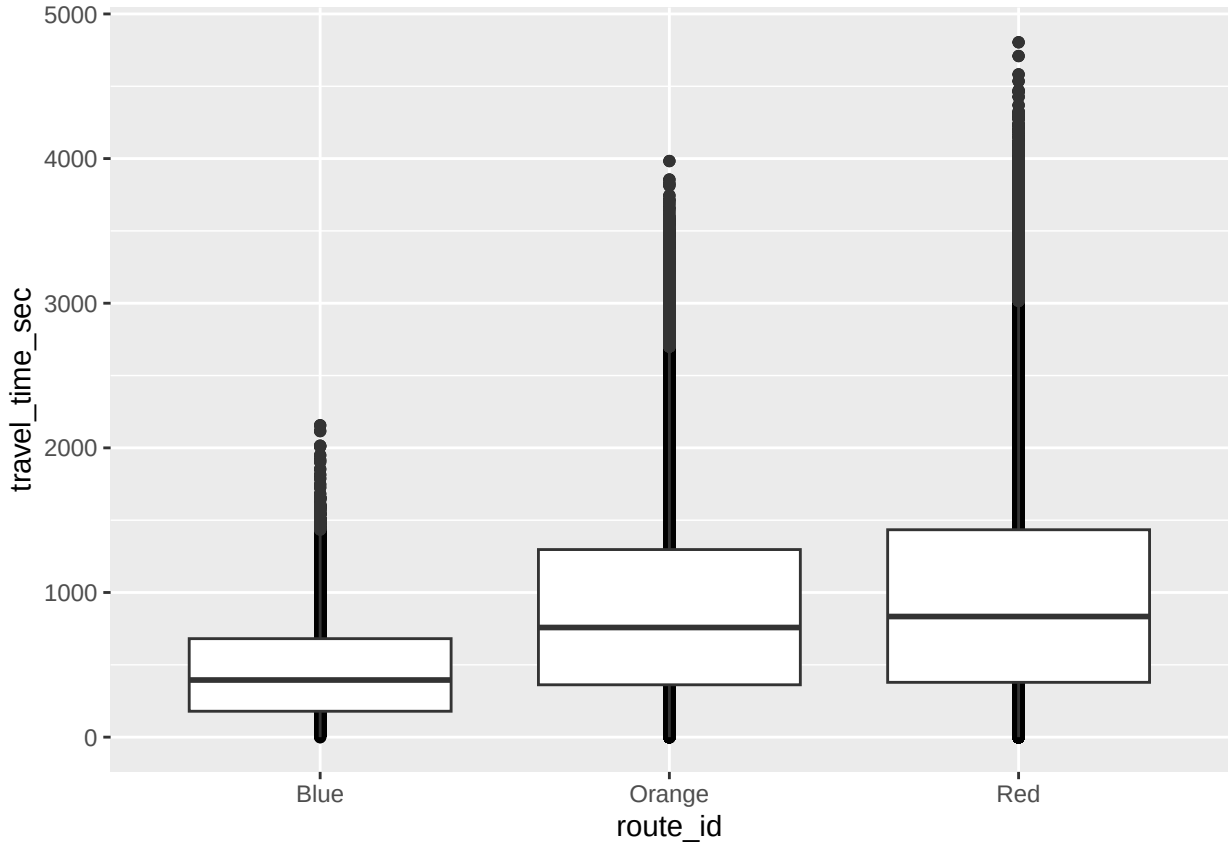
```
Q2 <- read.csv("2022-Q2_HRTravelTimes.csv")
week <- subset(Q2, service_date=="2022-04-01" | service_date=="2022-04-02" | service_date=="2022-04-03"
# 808587 observations from 04-01 to 04-07

orange <- subset(week, route_id =="Orange")
blue <- subset(week, route_id == "Blue")
red <- subset(week, route_id == "Red")
```

Methods

I am going to fit some regressions and plots in order to find associations between random variables. Simple graphs sometimes will produce better results and get the work done. I am going to compare Orange, Red, and Blue lines in terms of total travel time in seconds.

Travel time in seconds

```
ggplot(week, aes(x=route_id,y=travel_time_sec))+
  geom_point()+
  geom_boxplot()
```




```
mean(orange$travel_time_sec)
```

```
## [1] 877.826
```

```
mean(red$travel_time_sec)
```

```
## [1] 979.0948
```

```
mean(blue$travel_time_sec)
```

```
## [1] 447.2756
```

We can see that the general travel time between each stop for blue line is relatively a lot shorter than orange, and red line.

Conclusion

There is not a lot of information we can get from fitting graphs and regressions. The mean time between each stop for blue line is 447 seconds. The mean time between each stop for orange line is 878 seconds. The mean time between each stop for red line is 979 seconds. From the data of first week of April 2022, we can see that blue line has shorter time between each stop compared to orange and red lines, in general. I believe that the service is reliable. However, due to heavy snowing, the time may varies slightly.

Next step

It would be helpful if the data can include more info such as the general delay time. I believe that these data are easy to record and fit into the dataset.