MBTA project EDA

runze

2022-12-18

```
setwd("D:/615/final")
#load stops info (lat-lon)


LRQ4_21<-vroom("LRTravelTimesQ4_21.csv")
```

```
## Rows: 15678283 Columns: 8
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (1): route_id
## dbl  (6): from_stop_id, to_stop_id, direction_id, start_time_sec, end_time_s...
## date (1): service_date
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
LRQ1_22<-vroom("2022-Q1_LRTravelTimes.csv")
```

```
## Rows: 15879911 Columns: 8
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (1): route_id
## dbl  (6): from_stop_id, to_stop_id, direction_id, start_time_sec, end_time_s...
## date (1): service_date
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
LRQ2_22<-vroom("2022-Q2_LRTravelTimes.csv")
```

```
## Rows: 17167207 Columns: 8
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (1): route_id
## dbl  (6): from_stop_id, to_stop_id, direction_id, start_time_sec, end_time_s...
## date (1): service_date
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
LRQ3_22<-vroom("2022-Q3_LRTravelTimes.csv")
```

```
## Rows: 15789057 Columns: 8
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (1): route_id
## dbl  (6): from_stop_id, to_stop_id, direction_id, start_time_sec, end_time_s...
## date (1): service_date
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
T_2122<-rbind(LRQ4_21,LRQ1_22,LRQ2_22,LRQ3_22)

rm(LRQ4_21)
rm(LRQ1_22)
rm(LRQ2_22)
rm(LRQ3_22)


T_2122$service_date<-as.Date(T_2122$service_date)
T_2122$month<-month(T_2122$service_date)
T_2122$week<-week(T_2122$service_date)
T_2122$weekdays<-weekdays(T_2122$service_date,abbreviate = T)

#Randomly choose a week in these 12 months
sample_weeks<-c()
set.seed(726)
for (i in 1:12) {
  weeks_in_month<-levels(as.factor(T_2122$week[T_2122$month==i]))
  sample_weeks<-c(sample_weeks,sample(weeks_in_month[2:(length(weeks_in_month)-1)],1)) # To make sure p
}


smp_T<-T_2122[T_2122$week %in% as.numeric(sample_weeks),]

rm(T_2122)

smp_T$trip<-paste0(paste0("From ",smp_T$from_stop_id),paste0(" to ",smp_T$to_stop_id))# Add a variable

smp_T$trip<-as.factor(smp_T$trip)
```

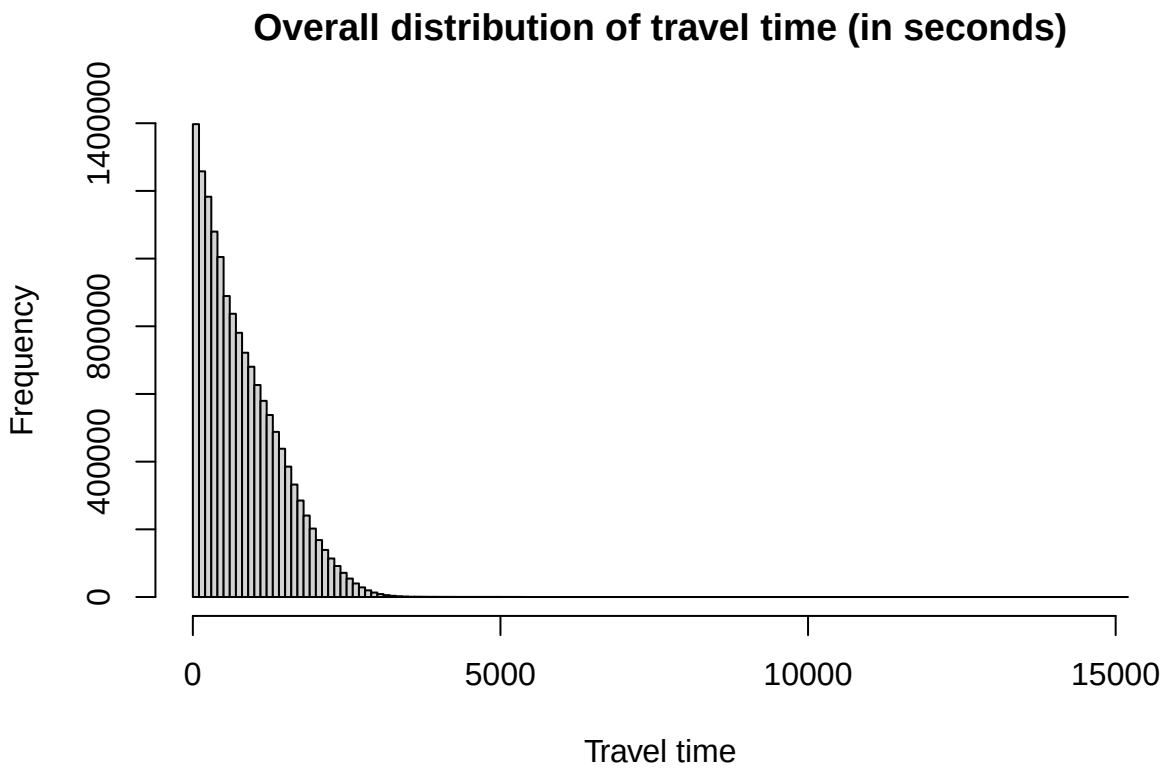
The light rail data EDA

Check the overall distribution of travel times

```
summary(smp_T$travel_time_sec)
```

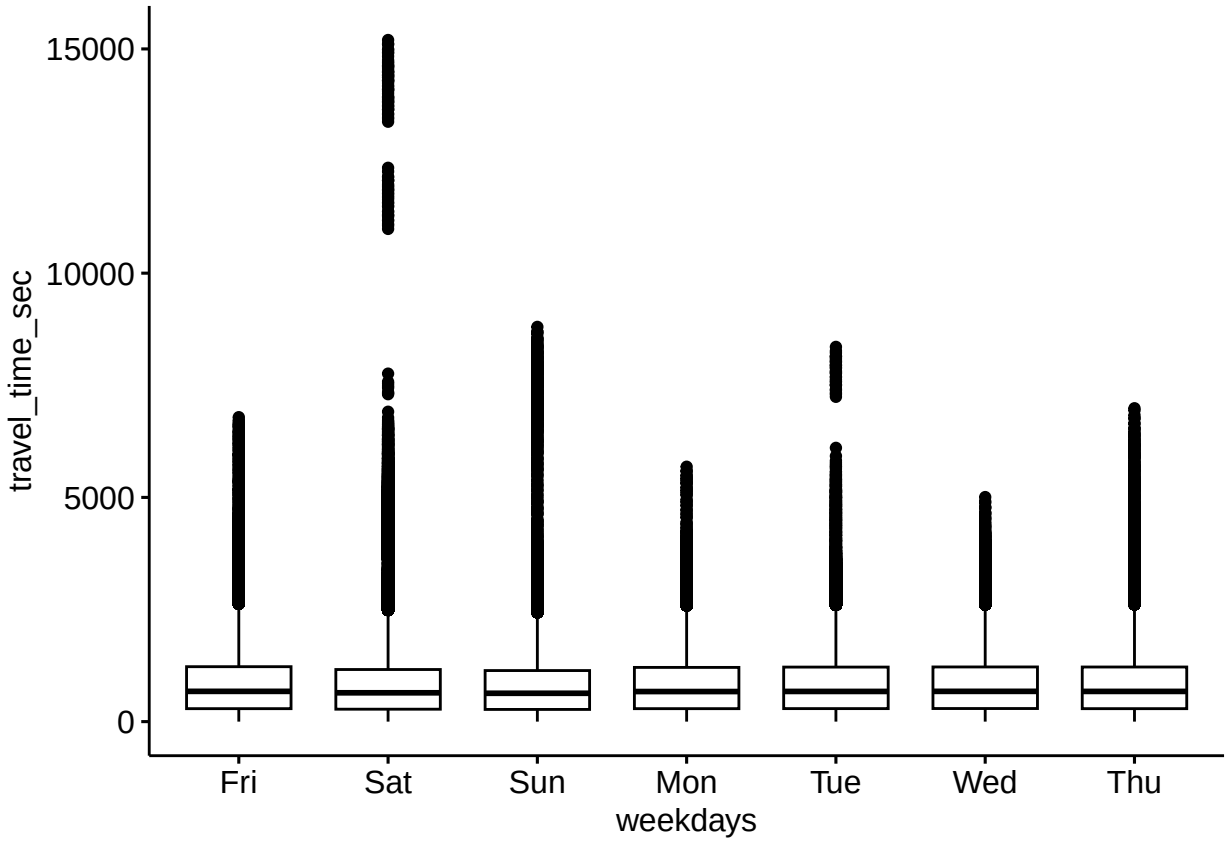
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   287.0   665.0   805.9  1201.0 15199.0
```

```
hist(smp_T$travel_time_sec,breaks = 120,main = "Overall distribution of travel time (in seconds)",xlab =
```



Check the distribution of travel times for each weekdays

```
ggboxplot(smp_T,x = "weekdays",y="travel_time_sec")
```



We can check many kinds of trips are there

```
trip<-levels(smp_T$trip)

length(trip)
```

```
## [1] 2023
```

There are total 2067 kinds of trips for light rails in boston

We can check which trip has the highest frequency

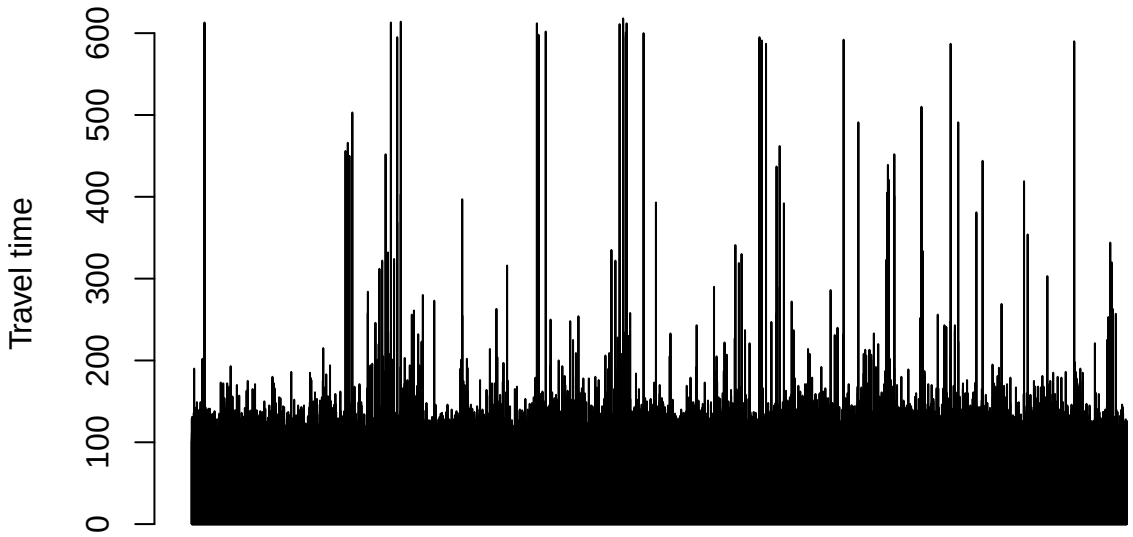
```
table(smp_T$trip)[table(smp_T$trip)==max(table(smp_T$trip))]
```

```
## From 70157 to 70155
##               39928
```

The most busy trip is from station 70159 to station 70147, which has a frequency of 41675

Then check how this trip's travel times are distributed

```
barplot(smp_T$travel_time_sec[smp_T$trip=="From 70159 to 70157"],ylab = "Travel time")
```



We can see that for most cases, this trip finishes within 200 seconds, but for some extream situation, the travel time can up to 600 seconds.