mbta_EDA

Jiun Lee

2022-12-17

Introduction

This report is to check the the accuracy of MBTA's departure and arrival time. Using the dataset of travel times between origin and destination pairs in 2021, we will verify if MBTA is reliable enough.

Data Cleaning

Since it's a large dataset, we will subset the data for 2021-12-01 to 2021-12-07. Also, there are weird travel times under 10 seconds, which has to be removed.

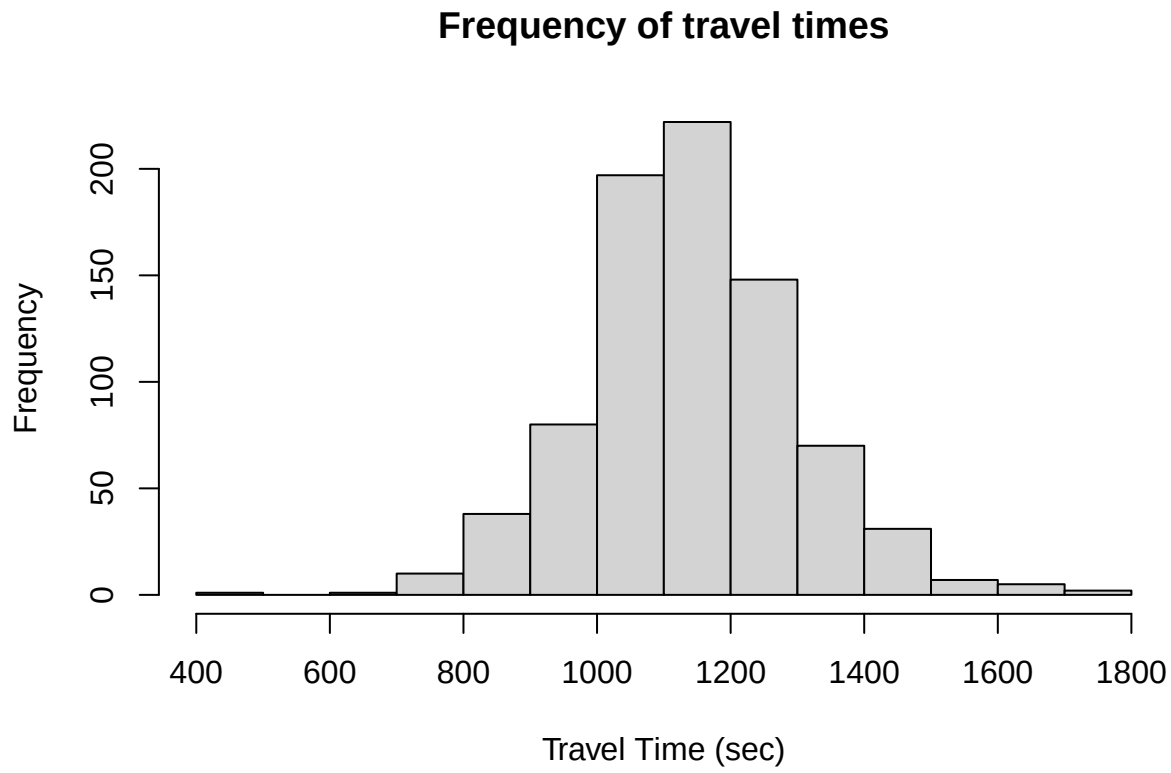
```
## Choose only 2021-12-01 ~ 2021-12-07
df <- data %>%
    mutate(service_date = lubridate::ymd(service_date)) %>%
    filter(service_date >= as.Date("2021-12-01") & service_date <= as.Date("2021-12-07"))

r <- which(df$travel_time_sec < 10)  # remove weird rows. Travel time cannot be under 10 seconds.
df <- df[-r, ]

# remove unnecessary columns rename columns
df <- df %>%
    select(-direction_id) %>%
    rename(date = service_date, from = from_stop_id, to = to_stop_id, line = route_id,
        traveltime = travel_time_sec, starttime = start_time_sec, endtime = end_time_sec) %>%
    group_by(from, to, line)
head(df)
```

```
## # A tibble: 6 x 7
## # Groups:   from, to, line [1]
##   date        from     to line    starttime endtime traveltime
##   <date>     <int>  <int> <chr>       <int>   <int>      <int>
## 1 2021-12-01 70112 170136 Green-B     57039   58000        961
## 2 2021-12-01 70112 170136 Green-B     58155   59215       1060
## 3 2021-12-01 70112 170136 Green-B     57757   59115       1358
## 4 2021-12-01 70112 170136 Green-B     66710   67678        968
## 5 2021-12-01 70112 170136 Green-B     67271   68332       1061
## 6 2021-12-01 70112 170136 Green-B     66006   66920        914
```

```
# make the subset of stops and route with randomly chosen
# from_stop_id,to_stop_id,route_id.
sub <- df %>%
    filter(from == 70110 & to == 170136 & line == "Green-B")
sub <- as.data.frame(sub)
```

With this subset, we will perform several visualizations.

Graphs



Histogram

The histogram is the most basic visualization you can use for EDA. As you can see, the distribution looks like a normal distribution.

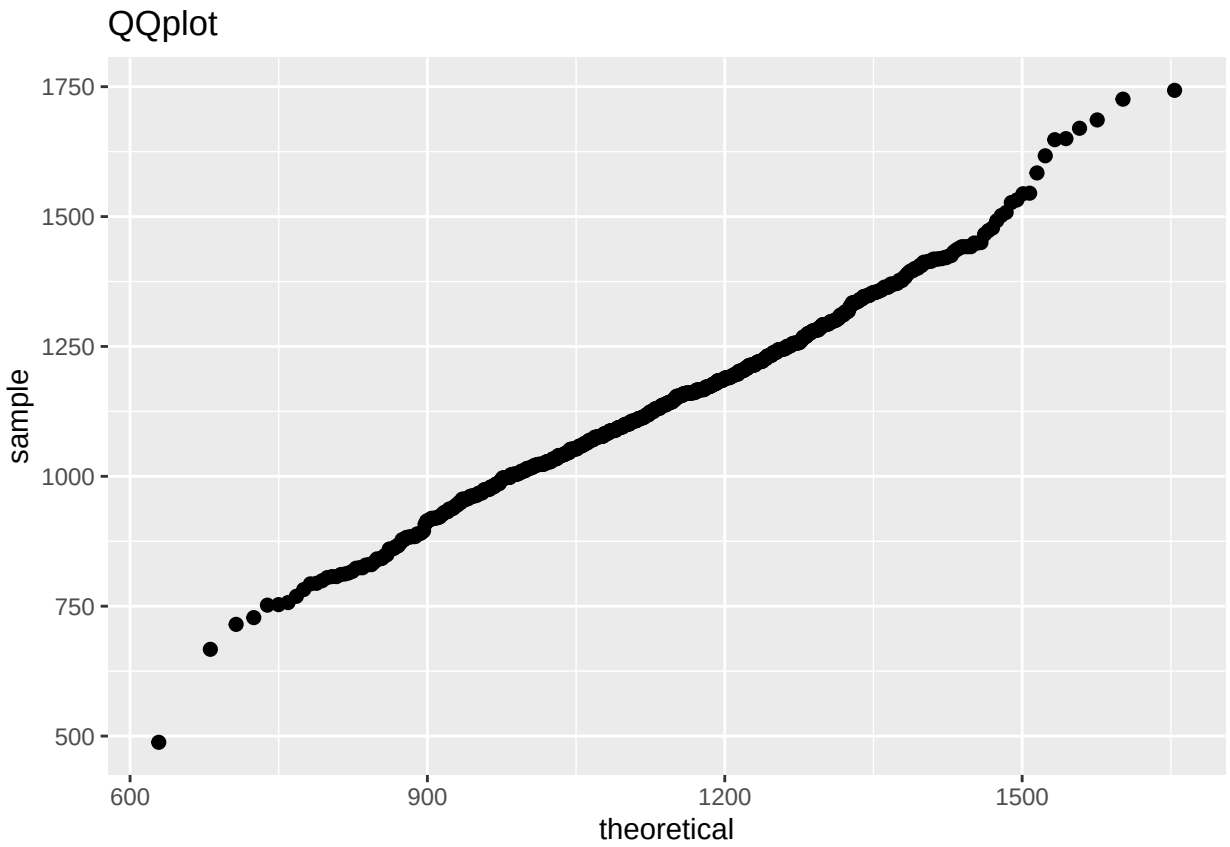
Normal probability plot

Let's see a normal probability plot to check the normal distribution.

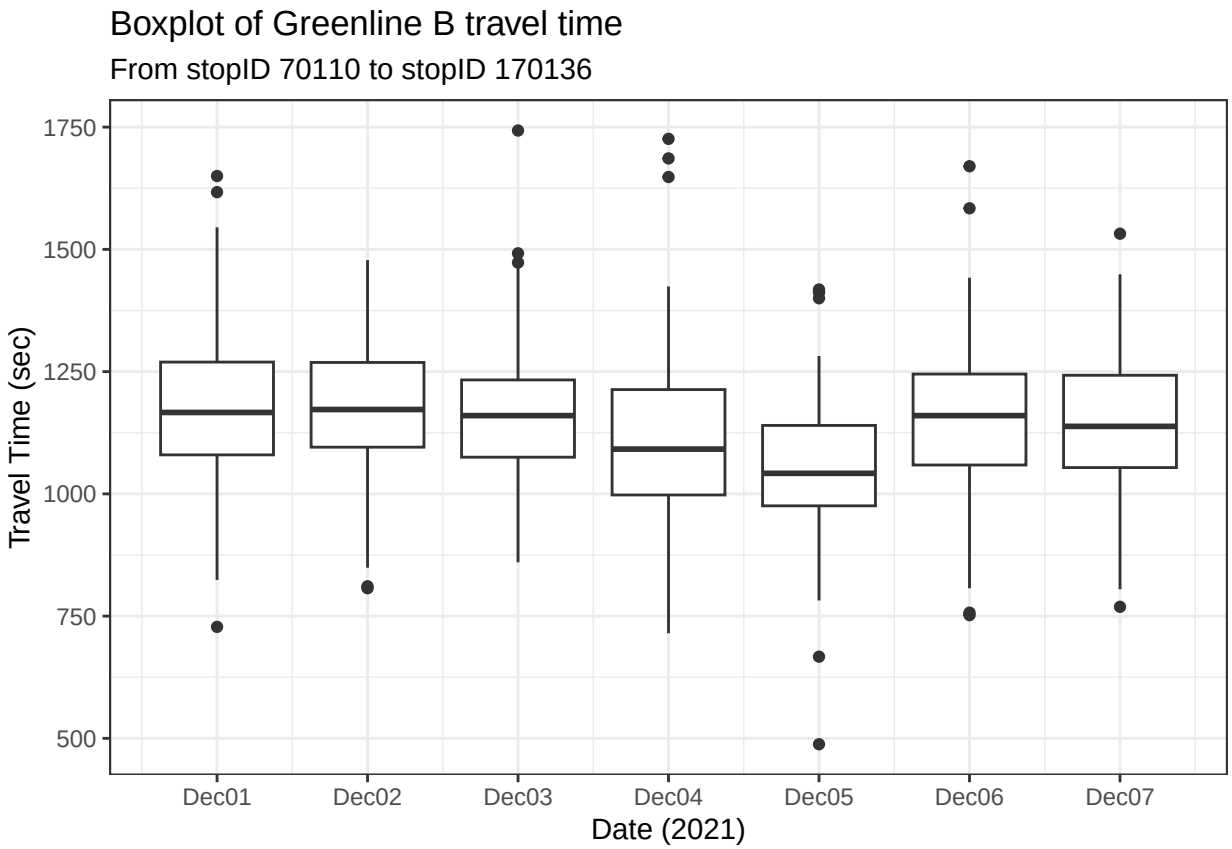
```
## Loading required package: ggplot2
```

```
##
## Attaching package: 'qqplotr'
```

```
## The following objects are masked from 'package:ggplot2':
##
##     stat_qq_line, StatQqLine
```



The graph is a straight line with a slope of 1, so we can confirm that the dataset is normally distributed.

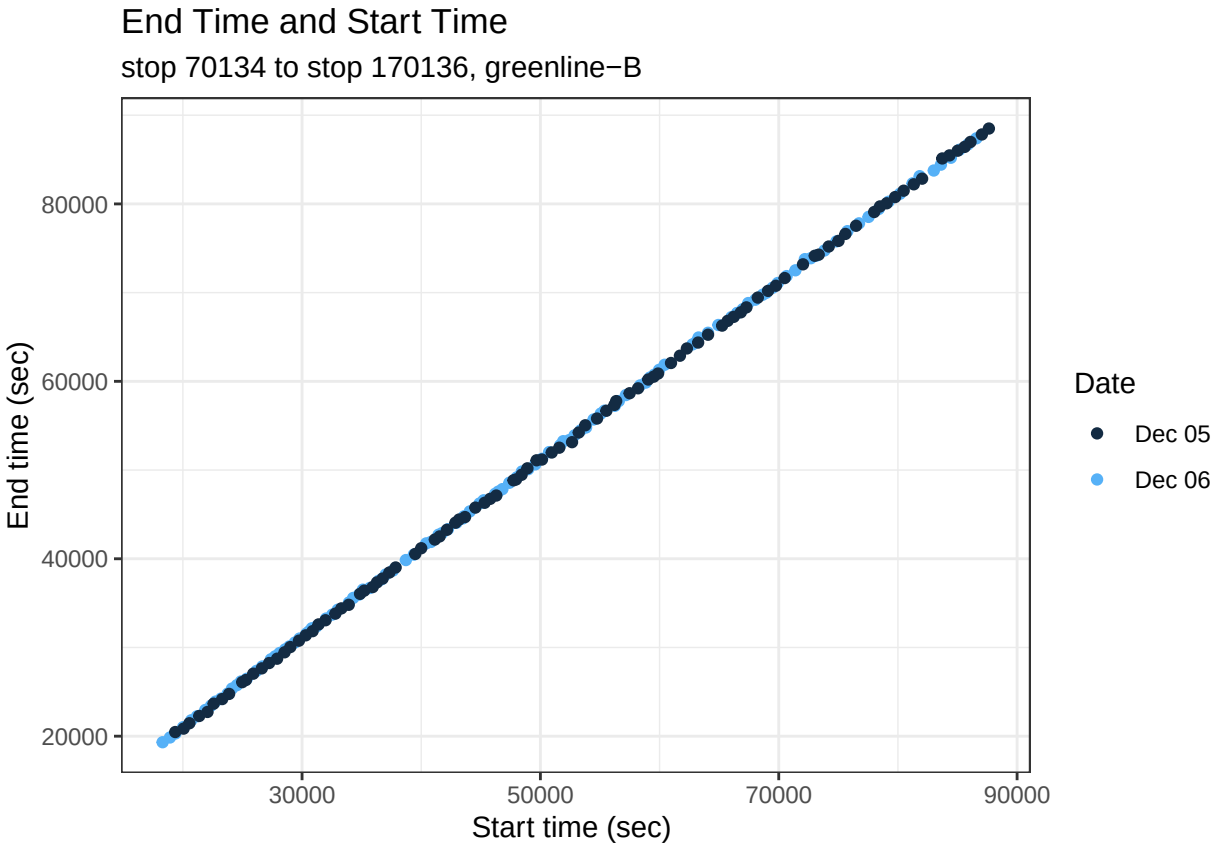


Box Plot

Let's see the box plot of travel time.
As the median of Dec05 is significantly lower than other days' medians, we can verify the travel time of Dec 05 generally is faster than usual. This is probably because Dec 05 is Sunday that has less passengers.

Scatter plot

A scatter plot will be useful to see the difference between Monday and Sunday.



Start and end time points of each day are mostly overlapped, gathering into one single line. It means the start and end times are steady, which means MBTA is dependable.