mbta_615final

Danya Zhang

2022-12-16

The MBTA, more commonly known as the "T" by locals, is the first subway system ever implemented in the United States. It was voted into law in 1964, serving the greater Boston area. Some time later, a consulting firm was hired to consolidate the various transit lines; it was then that the MBTA system became as we know it today with the color-coded lines.

In this report, we would like to examine the the reliability of the MBTA system using statistics and visualizations to see if departure and arrival times are accurate. Let's first read in the data. To make it easier, since the data is quite large, we will work with a subset of the data which represents a week of data.

Getting to know the data

Summary Statistics

Let's look at some summary statistics of the data.

```
## `summarise()` has grouped output by 'from_stop_id'. You can override using the
## `.groups` argument.
```

```
## # A tibble: 6 x 6
## # Groups:   from_stop_id [1]
##   from_stop_id to_stop_id   max   min  mean    sd
##          <int>      <int> <int> <int> <dbl> <dbl>
## 1        70110      70112   610    12  104.  45.2
## 2        70110      70114  1111     2  183.  66.3
## 3        70110      70116  1189    80  253.  71.4
## 4        70110      70120  1316   176  365.  80.9
## 5        70110      70124  1642   269  489. 100.
## 6        70110      70126  1767   324  597. 113.
```

There were some travel times that are clearly impossible, perhaps due to a data entry error. So we will delete any observations with less than a 10 second travel time. For the visualizations, we will further subset the data to a single from_stop_id, to_stop_id, direction_id, route_id, and direction_id (subset1).

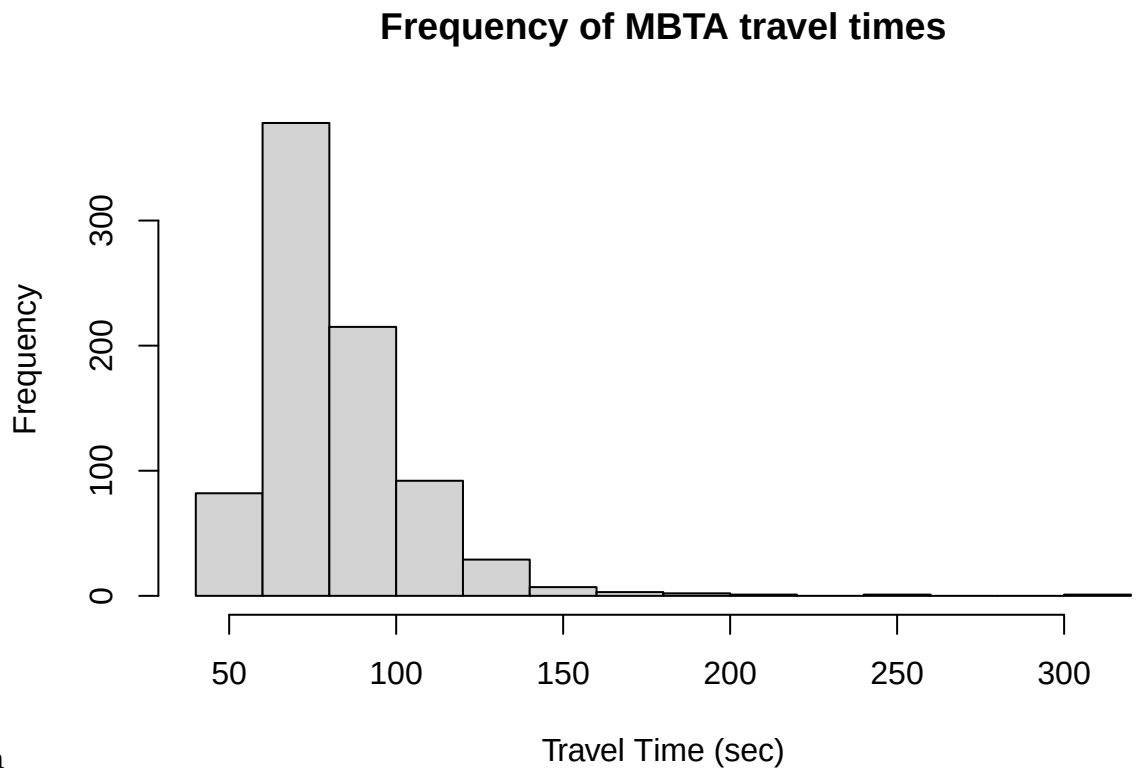
Data Cleaning

```
# some travel times that are clearly impossible. possible data entry
# error
short_ind <- data[which(data$travel_time_sec < 10), ]
remove_ind <- as.numeric(dimnames(short_ind)[[1]])
data <- data[-remove_ind, ]

# subsetting based on the same from_stop_id, to_stop_id, route_id,
```

```r
# direction_id
subset1 <- data %>%
    filter(from_stop_id == 70134 & to_stop_id == 170136 & route_id == "Green-B" &
        direction_id == 1)

# first week of 2022
subset1 <- as.data.frame(subset1)
```

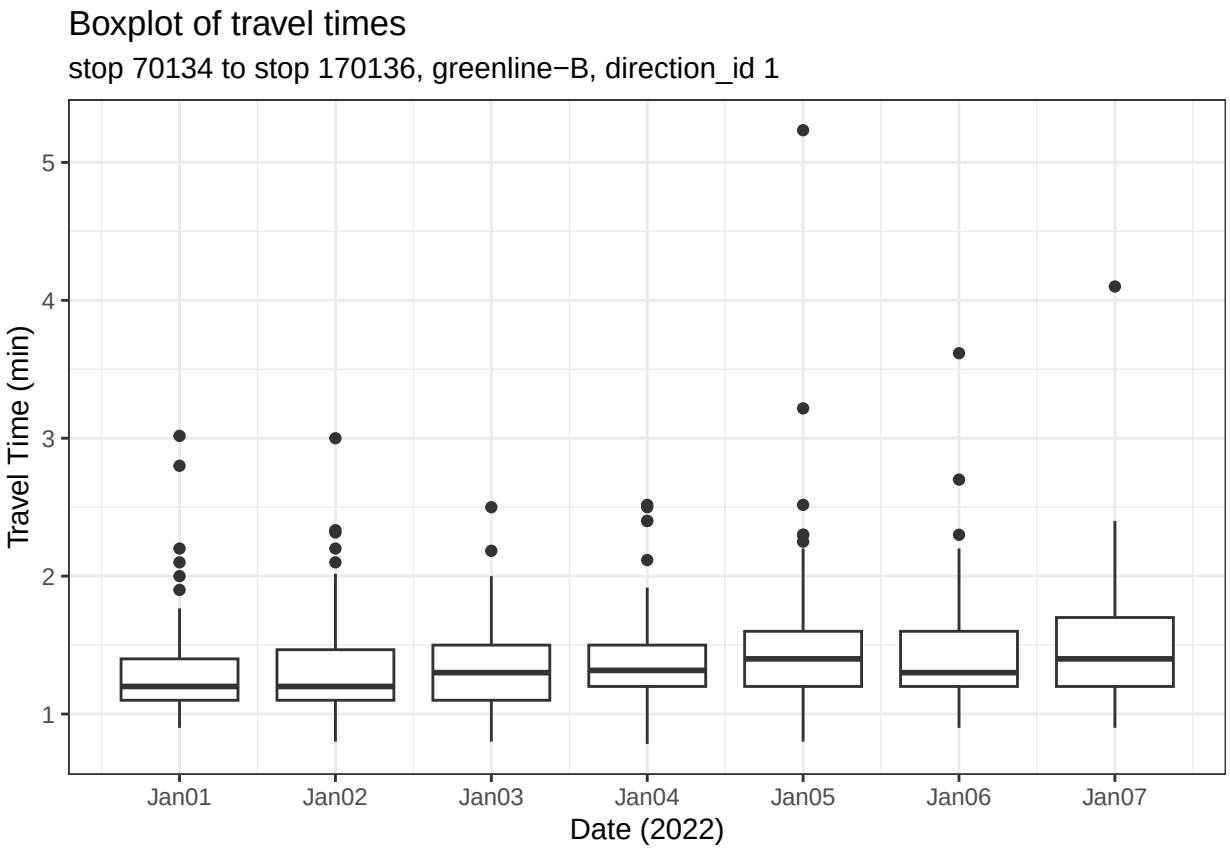
Now, let's visualize the data using some graphs.

Visualizations



Histogram

Aside from the very few outliers, the histogram is otherwise fairly bell-shaped. This suggest that travel times are roughly Gaussian and the MBTA travel times between the two stops are fairly quick.

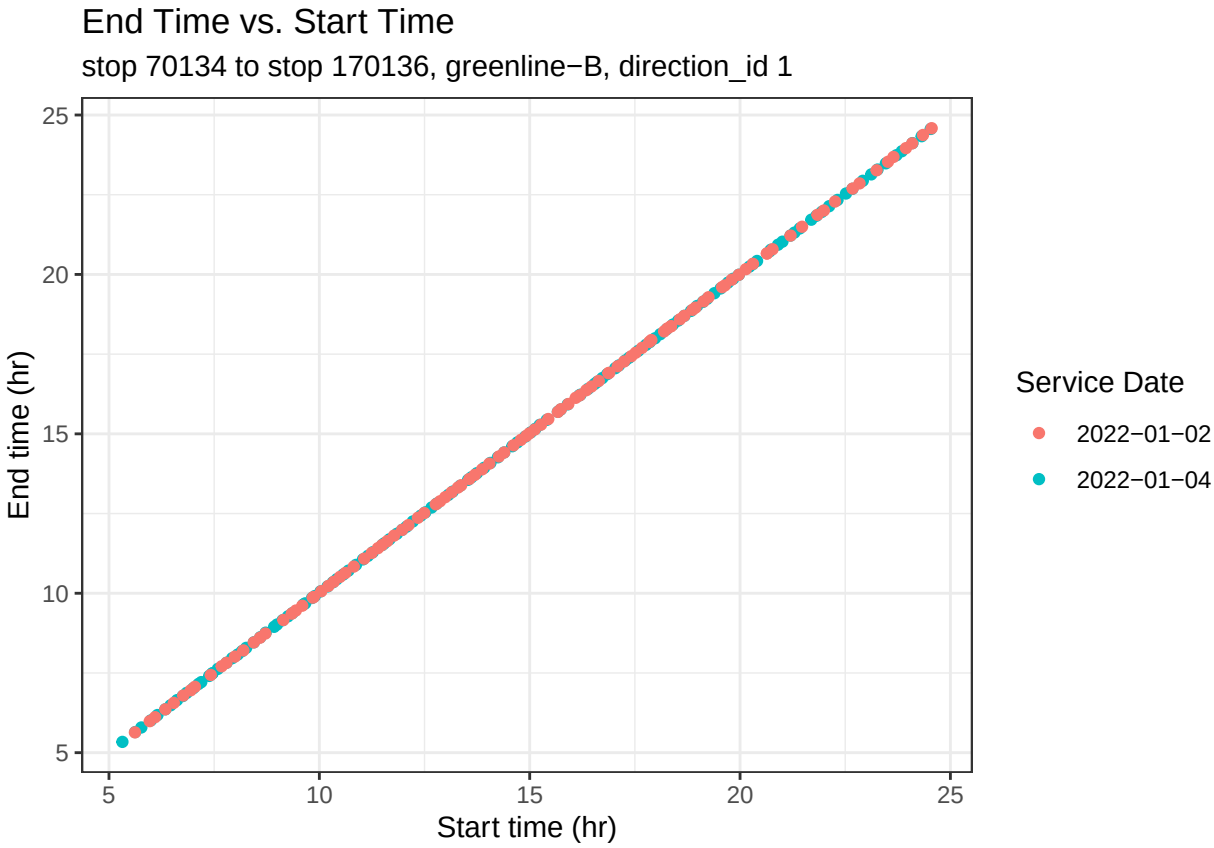


Boxplot

As the box plots are roughly on the same horizontal level as each other, there seems to be little to no variation. Ignoring the outliers, the medians, 25% quantile and 75% quantile are at roughly the same level. This is good as it means the Greenline-B is very consistent with travel times.

Scatter plot

Now, let's use a scatter plot to explore possible variation between two extremes. Here I have selected two very different days in terms of passenger traffic: Sunday and Tuesday. Let's see if there is a difference between weekdays and weekends.



Scatter plots appear linear and stacked on top of each other, which means start and end times are consistent between days, meaning the
MBTA is fairly reliable.

Hypotheses Tests

Moving on to the hypotheses tests, we'd like to use some statistical tests to see if MBTA travel times are reliable and consistent.

Anova test

Let's start with an ANOVA test. Our null and alternative hypotheses are as follows:

H_0 = there are no significant differences in travel times
H_1 = there are significant differences in travel times

```
##
## Attaching package: 'rstatix'
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

```
## ANOVA Table (type II tests)
##
##         Effect DFn     DFd       F        p p<.05      ges
## 1 service_date   6 1181636 103.925 2.1e-131     * 0.000527
```

The p-value for the anova test is very small this shows that we reject our null hypotheses, meaning that there

are significant differences in travel times. This might be surprising considering the graphs above did not show this. However, it is very important to know that p-values do not represent a definitive answer. Let's consider another type of test: the paired t-test.

Paired T-test

```
##      .y.                group1             group2                n1
##  Length:21          Length:21          Length:21          Min.   :159526
##  Class :character   Class :character   Class :character   1st Qu.:159526
##  Mode  :character   Mode  :character   Mode  :character   Median :162525
##                                                           Mean   :169709
##                                                           3rd Qu.:171806
##                                                           Max.   :191500
##        n2                p              p.signif             p.adj
##  Min.   :152899   Min.   :0.000000   Length:21          Min.   :0.0000
##  1st Qu.:152899   1st Qu.:0.000000   Class :character   1st Qu.:0.0000
##  Median :169313   Median :0.000000   Mode  :character   Median :0.0000
##  Mean   :167903   Mean   :0.069749                      Mean   :0.1100
##  3rd Qu.:174074   3rd Qu.:0.000483                      3rd Qu.:0.0101
##  Max.   :191500   Max.   :0.809000                      Max.   :1.0000
##  p.adj.signif
##  Length:21
##  Class :character
##  Mode  :character
##
##
##
```

```
## # A tibble: 21 x 9
##    .y.           group1 group2     n1     n2         p p.sig~1     p.adj p.adj~2
##  * <chr>         <chr>  <chr>   <int>  <int>     <dbl> <chr>       <dbl> <chr>
##  1 travel_time_~ 2022-~ 2022-~ 159526 162525 1.04e-  8 ****    2.19e-  7 ****
##  2 travel_time_~ 2022-~ 2022-~ 159526 191500 3.71e- 60 ****    7.78e- 59 ****
##  3 travel_time_~ 2022-~ 2022-~ 162525 191500 1.29e- 25 ****    2.71e- 24 ****
##  4 travel_time_~ 2022-~ 2022-~ 159526 171806 6.41e-  1 ns      1   e+  0 ns
##  5 travel_time_~ 2022-~ 2022-~ 162525 171806 3   e- 10 ****    6.3 e-  9 ****
##  6 travel_time_~ 2022-~ 2022-~ 191500 171806 3.99e- 66 ****    8.38e- 65 ****
##  7 travel_time_~ 2022-~ 2022-~ 159526 169313 2.46e-  9 ****    5.16e-  8 ****
##  8 travel_time_~ 2022-~ 2022-~ 162525 169313 3.82e- 32 ****    8.03e- 31 ****
##  9 travel_time_~ 2022-~ 2022-~ 191500 169313 1.14e-115 ****    2.4 e-114 ****
## 10 travel_time_~ 2022-~ 2022-~ 171806 169313 2.1 e-  8 ****    4.42e-  7 ****
## # ... with 11 more rows, and abbreviated variable names 1: p.signif,
## #   2: p.adj.signif
```