615 Final Project

Jingyu Liang

2022-12-14

Aquiring the data and explore the data

```
####import the data, cleaning and exploration####
y21HRQ4 <- read.csv("HRTravelTimesQ4_21.csv")

y21LRQ4 <- read.csv("LRTravelTimesQ4_21.csv")

HRQ1 <- read.csv("2022-Q1_HRTravelTimes.csv")

LRQ1 <- read.csv("2022-Q1_LRTravelTimes.csv")

HRQ2 <- read.csv("2022-Q2_HRTravelTimes.csv")

LRQ2 <- read.csv("2022-Q2_LRTravelTimes.csv")

HRQ3 <- read.csv("2022-Q3_HRTravelTimes.csv")

LRQ3 <- read.csv("2022-Q3_LRTravelTimes.csv")

# service_date is Nov 2021 to Sep 2022
date4 = unique(y21LRQ4$service_date) # Oct to Dec
date1 = unique(HRQ1$service_date) # Jan to Mar
date2 = unique(LRQ2$service_date) # Apr to Jun
date3 = unique(LRQ3$service_date) # Jul to Sep

# HR and LR's route_id are different
HRroute = unique(HRQ3$route_id) # orange, blue, red
LRroute = unique(LRQ3$route_id) # Green-BCDE, Mattapan

# delete October 2021 data
delete = grep("-10-", y21HRQ4$service_date)
y21HRQ4 %<>% filter(!row_number() %in% delete)

delete_lr = grep("-10-", y21LRQ4$service_date)
y21LRQ4 %<>% filter(!row_number() %in% delete_lr)


####pick a week randomly from each month (11 months)####

# pick two weeks from each of two months in 2021 (Nov to Dec)
```

```r
selected_week = data.frame()
for (i in 11:12){
  # select the week in HR (for blue, red, orange)
  HR_row = grep(paste("-", i, "-", sep = ""), y21HRQ4$service_date)
  HR <- y21HRQ4 %>% slice(HR_row) # a month's data of HR
  LR_row = grep(paste("-", i, "-", sep = ""), y21LRQ4$service_date)
  LR <- y21LRQ4 %>% slice(LR_row) # a month's data in LR

  days = unique(HR$service_date)
  start = sample(1:(length(days)-6),1)
  week = c(start:(start+6)) # finish selecting one week in this month

  # pick the data at the selected week day by day
  for (j in start:(start +6)){
    date_row  = grep(days[j], HR$service_date )
    date  <- HR %>% slice(date_row )
    selected_week = bind_rows(selected_week, date )

    date_row_LR = grep(days[j], LR$service_date )
    date_LR <- LR %>% slice(date_row_LR)
    selected_week = bind_rows(selected_week, date_LR)

  }

}

# pick 9 weeks randomly from each of 9 months in 2022 (Jan to Sep)
y2022 = bind_rows(HRQ1, LRQ1, HRQ2, LRQ2, HRQ3, LRQ3) # prepare the data of 2022
y2022_date = unique(y2022$service_date)

for (k in 1:9){
  # select the week in HR (for blue, red, orange)
  row = grep(paste("-0", k, "-", sep = ""), y2022$service_date)
  month <- y2022 %>% slice(row) # a month's data of HR & LR

  days22 = unique(month$service_date)
  start22 = sample(1:(length(days22)-6),1)
  week22 = c(start22:(start22+6)) # finish selecting one week in this month

  # pick the data at the selected week day by day
  for (m in start22:(start22 +6)){
    date_row22  = grep(days22[m], month$service_date )
    date22  <- month %>% slice(date_row22 )
    selected_week = bind_rows(selected_week, date22 )
  }

}
###########
```

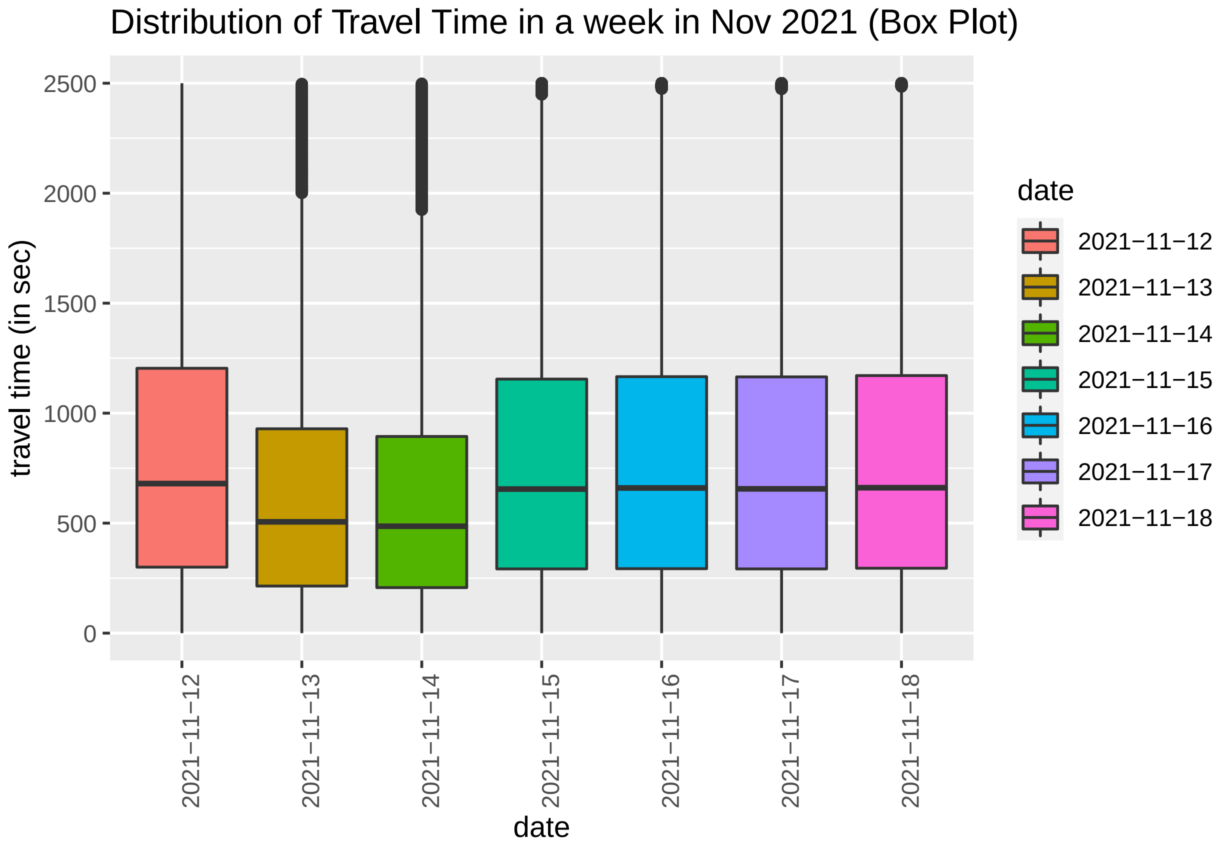
EDA

distribution plot of travel time in one week

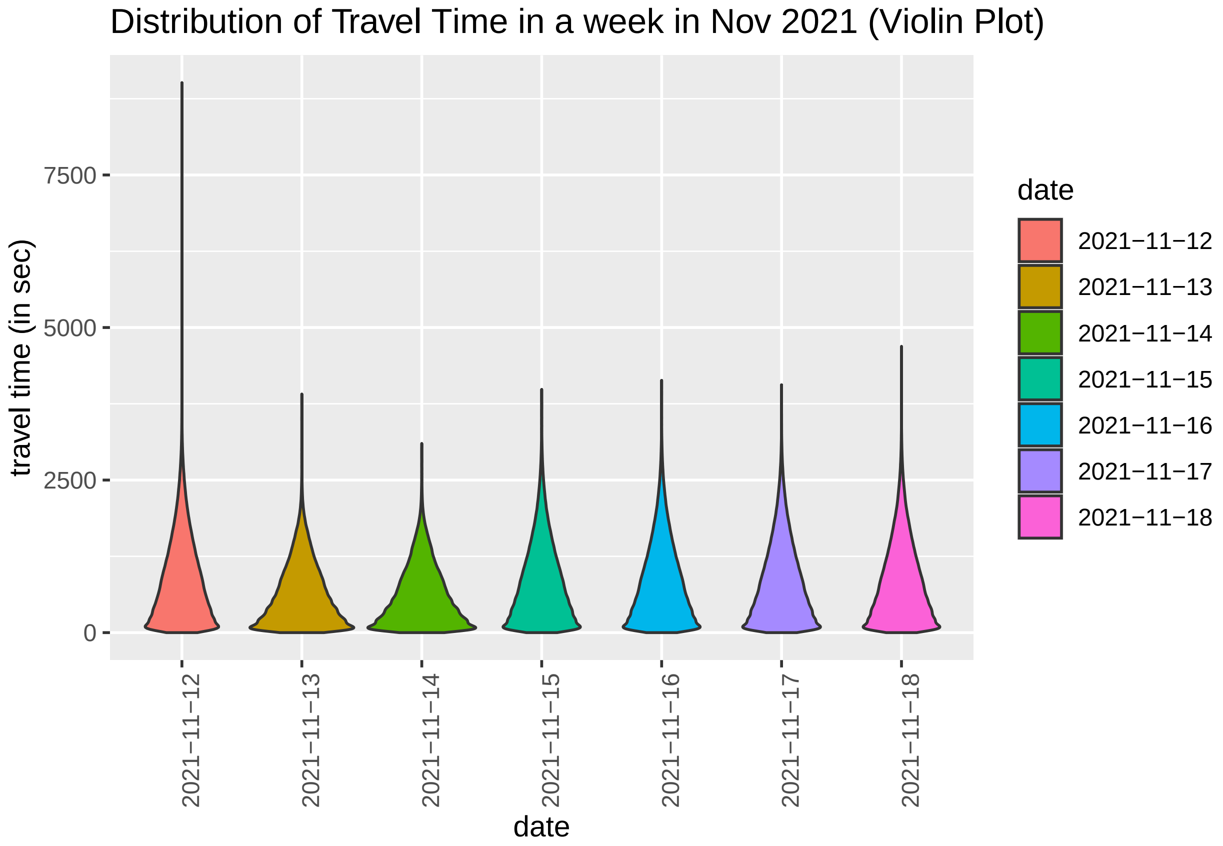
I use selected_week to do EDA. After plot different weeks, I found that the distributions of travel times between different weeks are similar, so I just show you one week's distribution plot. In the week in Nov 2021, the distributions of travel times are similar between different days in one week. In this week, the medians of travel time in each day are all around 600 seconds. From the violin chart and the distribution density plot, we can see in most cases, travel time is less than 2500 seconds.

```
# get the data of the week in Nov 2021, and the week in May 2022
Nov <- selected_week %>% slice(grep("2021-11-", selected_week$service_date))
May <- selected_week %>% slice(grep("2022-05-", selected_week$service_date))

# plot the distribution of the travel time in Nov 2021 week
# I remove most of the outliers from the plot.

ggplot(data = Nov, aes(x = service_date, y = travel_time_sec, fill = service_date )) +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  ylim(min =  0, max = 2500) +
  geom_boxplot()+
  labs(title = "Distribution of Travel Time in a week in Nov 2021 (Box Plot)")+
  xlab("date") + ylab("travel time (in sec)") +
  labs(fill = "date")
```

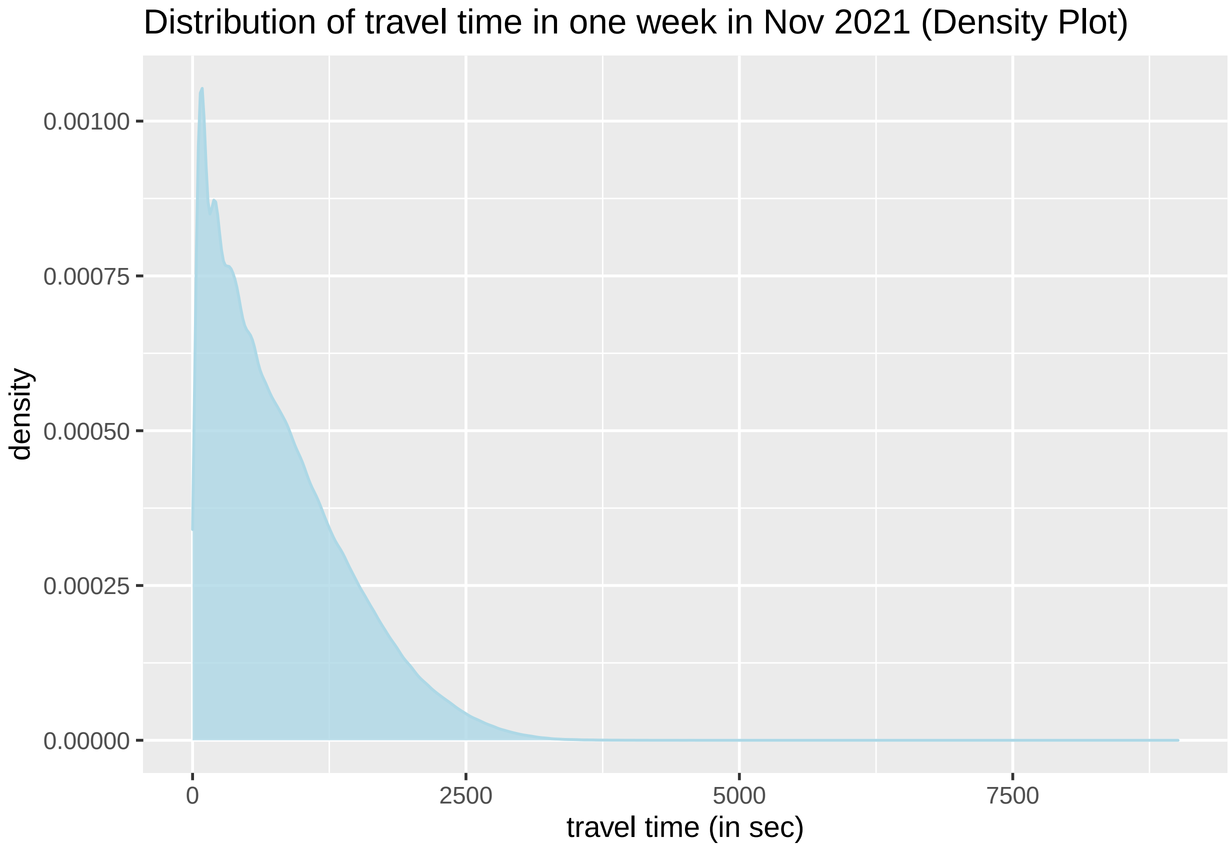
```
## Warning: Removed 26094 rows containing non-finite values (stat_boxplot).
```



```r
ggplot(data = Nov, aes(x = service_date, y = travel_time_sec, fill = service_date )) +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  geom_violin()+
  labs(title = "Distribution of Travel Time in a week in Nov 2021 (Violin Plot)")+
  xlab("date") + ylab("travel time (in sec)") +
  labs(fill = "date")
```

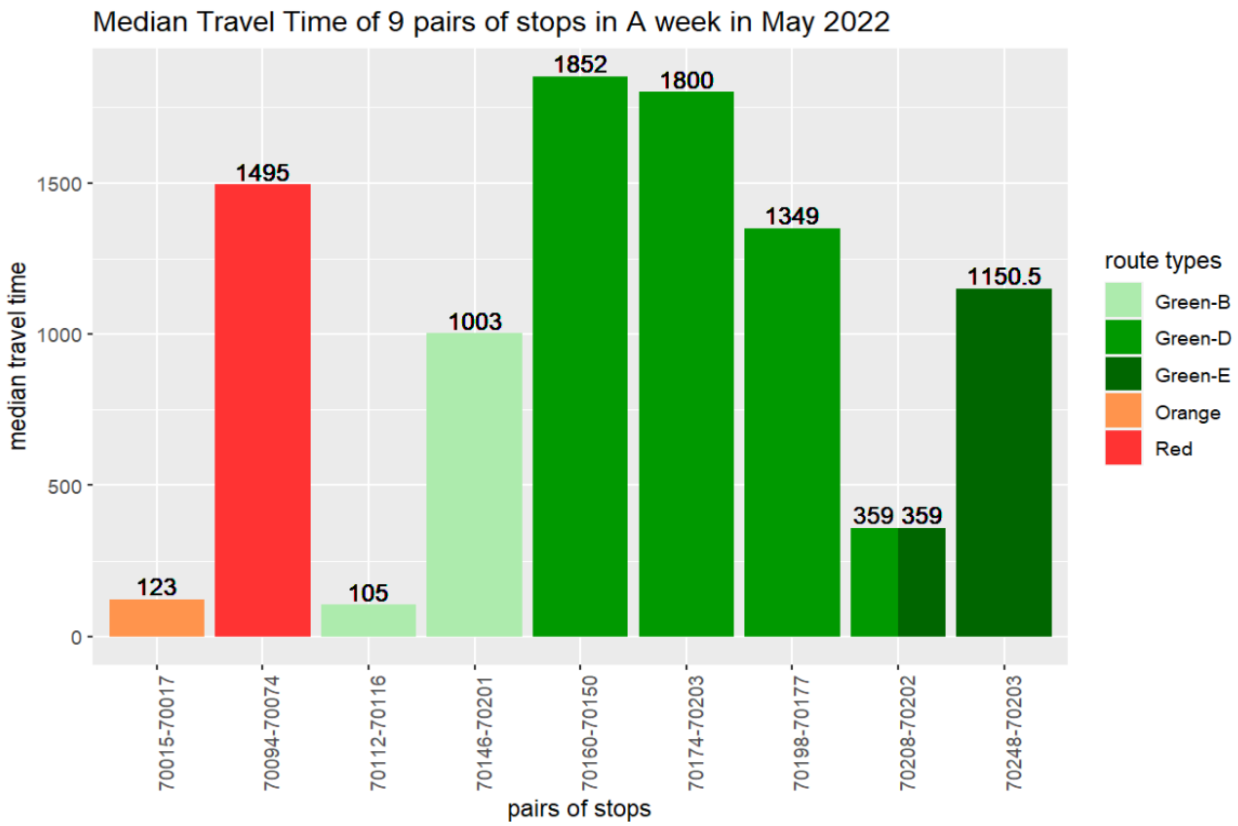


```r
ggplot(data = Nov, aes(x = travel_time_sec)) +
  geom_density(fill = "lightblue", color = "lightblue", alpha = 0.8) +
  xlab("travel time (in sec)") +
  labs(title = "Distribution of travel time in one week in Nov 2021 (Density Plot)")
```



# Median travel time of different pairs of stops in one week in May 2022 We use the data of the week in May 2022 to explore the data between different stops. In this week, the 5th and 6th (going from left to right in x-axis) pair of stops have the highest median travel time, and they are all Green-D line. This may represent that compared to other lines, the two stops are farther away from each other, or the speed of Green-D line is more slowly.

```
May$from_to_id <- str_c(May$from_stop_id, sep = "-", May$to_stop_id )
stoppairs_all = unique(May$from_to_id) # get all pairs of stops
set.seed(20)
pairs = sample(stoppairs_all, 9) # pick 9 pairs of stops at random

# Find the median of travel times of each stop pair and plot it
selected_pairs = data.frame()
median_pairs = c()
pairs_route = c()
for (i in 1:9){
  pair_rows = grep(pairs[i],May$from_to_id)
  getpair <- May %>% slice(pair_rows)

  median_pairs <- c(median_pairs, median(getpair$travel_time_sec))

  getpair$median_travel_time =  median(getpair$travel_time_sec)
  selected_pairs = bind_rows(selected_pairs, getpair)
}

ggplot(data = selected_pairs, aes( x = from_to_id, y = median_travel_time, fill = route_id) ) +
  geom_bar(position = 'dodge', stat = "identity") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))+
  scale_fill_manual(values = c("#adebad", #b
                               "#009900", #d
                               "#006600", #e
                               "#ff944d", #orange & red
                               "#ff3333" )) +
  geom_text(aes(label=median_travel_time),
            position=position_dodge(width=0.9), vjust=-0.25) +
  xlab("pairs of stops") + ylab("median travel time") +
  labs(title = "Median Travel Time of 9 pairs of stops in A week in May 2022")+
  labs(fill = "route types")
```



```
#b "#adebad",c "#33cc33",d "#009900",e "#006600",red "#ff3333",orange"#ff944d"
```