615 Final Project

Hsueh-Pin Liu

2022-12-17

Environment

First, I download the data from the MBTA website, even though the order is to focus on the data from November 2021 to October 2022, the data for October isn't completed, so I choose data from October 2021 to September 2022 for analysis.

EDA

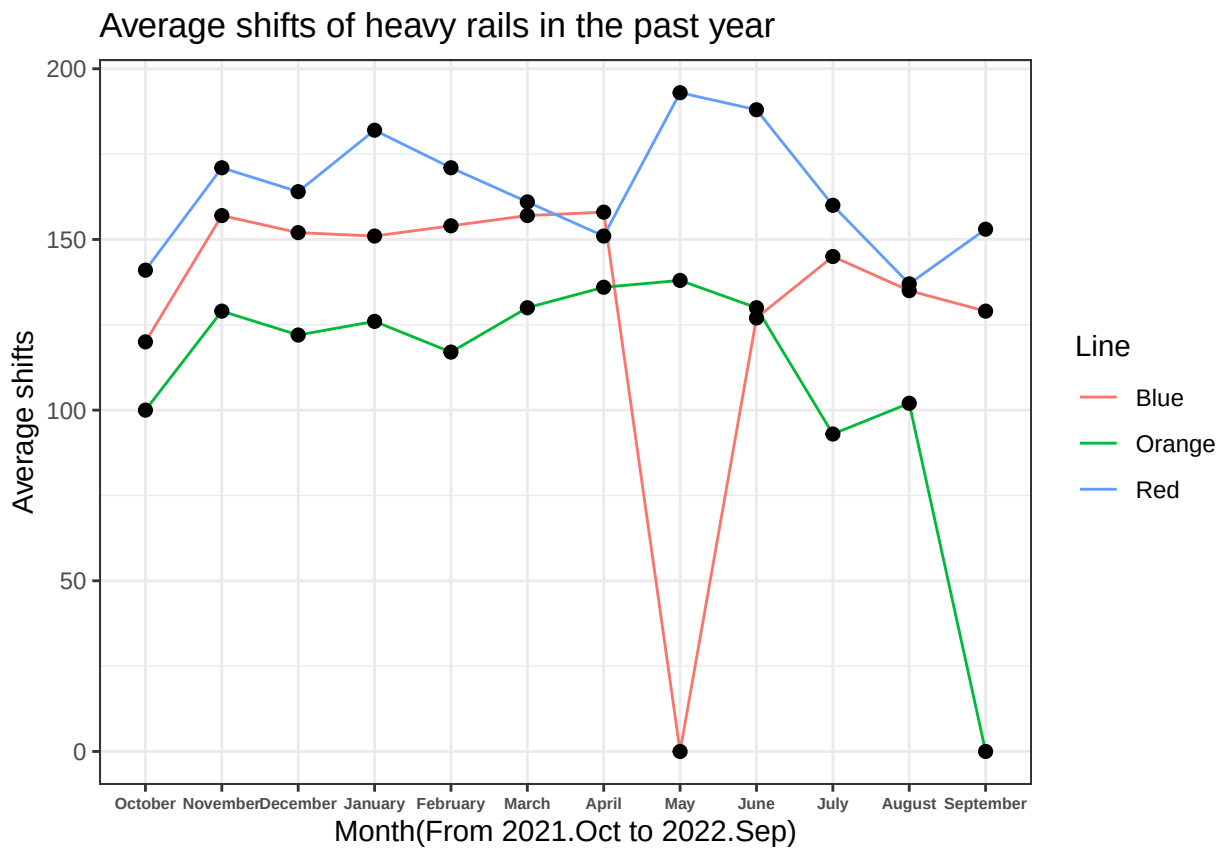
After I tidy the data, I focus on how many shifts are there in each month so I make plots below. I choose the first 7 days for each month to make the analysis.

```
stops <- rbind(dplyr::select(lr1,c(3,4)),dplyr::select(hr1,c(3,4)))
nrow(unique(stops))
```

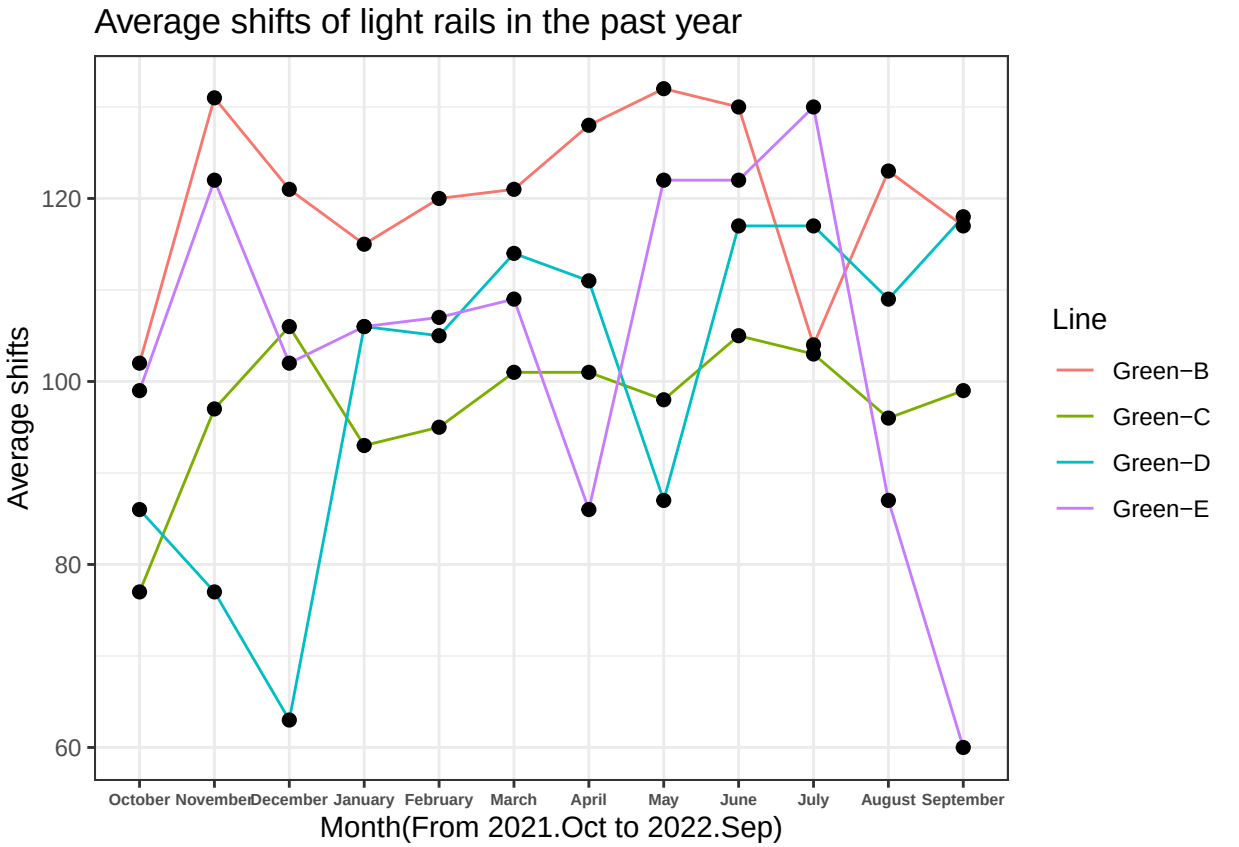
```
## [1] 285
```

So there are 285 stop_id, including same stops because some stops have two ids for two different directions. Then let's take a look at the plots.

```
hr <- rbind(hrOct,hrNov,hrDec,hrJan,hrFeb,hrMar,hrApr,hrMay,hrJun,hrJul,hrAug,hrSep)
hr <- filter(hr,from_stop_id==70061|from_stop_id==70038|from_stop_id==70036)
nhr <-  select(hr,c(1,2,4,6,9))
nhr <- unique(nhr)
result1 <- table(nhr$route_id,nhr$month)
result1 <- as.data.frame(result1)
names(result1) <- c("Line","Month","Times")
result1$Times <- round(result1$Times/7)
result1$Month <- factor(result1$Month,levels = c("October","November","December","January","February","
ggplot(data = result1, mapping = aes(x = Month, y = Times, group = Line)) +
geom_line(aes(color=Line))+
geom_point(color="Black", size=2)+
theme_bw()+
theme(axis.text.x = element_text(size=6,face="bold"))+
labs(x="Month(From 2021.Oct to 2022.Sep)",y="Average shifts",title="Average shifts of heavy rails in th
```



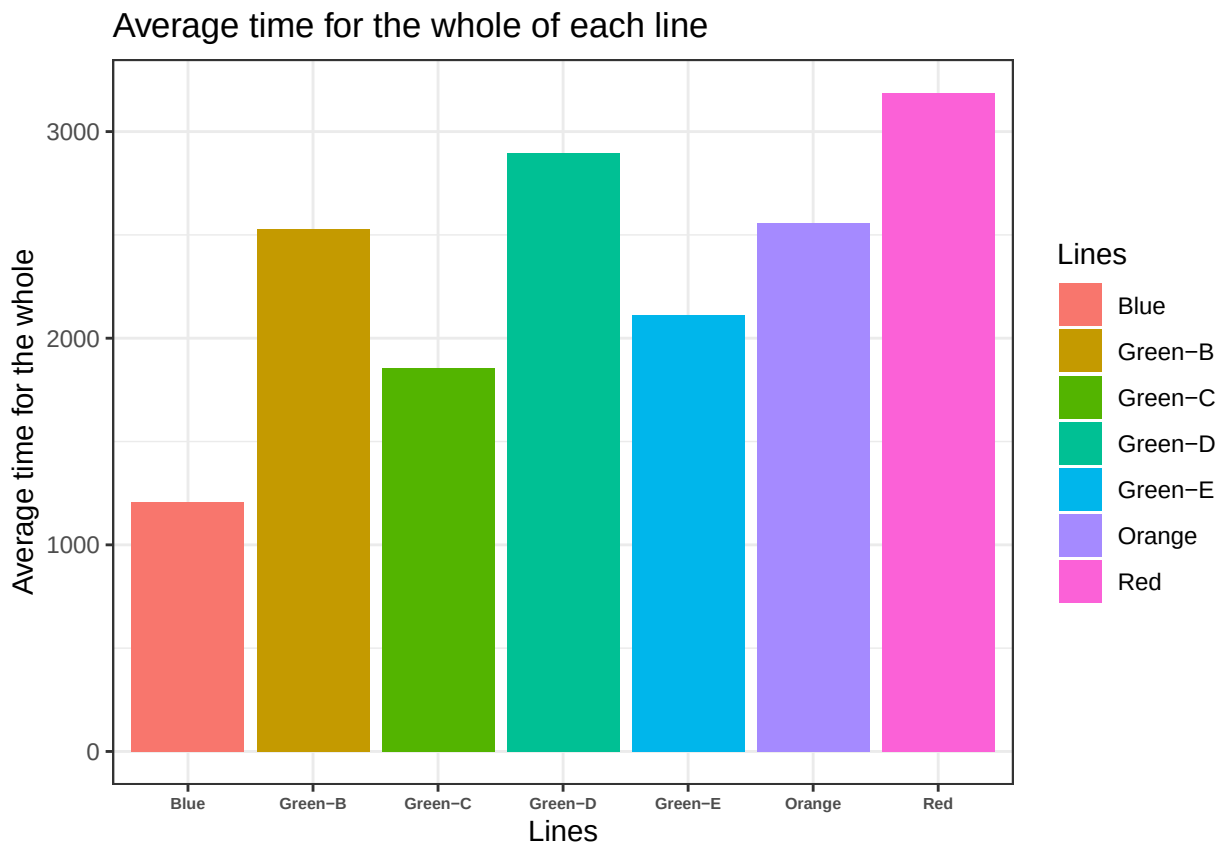

```
lr <- rbind(lrOct,lrNov,lrDec,lrJan,lrFeb,lrMar,lrApr,lrMay,lrJun,lrJul,lrAug,lrSep)
lr <- filter(lr,from_stop_id==70110|from_stop_id==70236|from_stop_id==70160|from_stop_id==70260)
nlr <-  select(lr,c(1,2,4,6,9))
nlr <- unique(nlr)
result2 <- table(nlr$route_id,nlr$month)
result2 <- as.data.frame(result2)
names(result2) <- c("Line","Month","Times")
result2$Times <- round(result2$Times/7)
result2$Month <- factor(result2$Month,levels = c("October","November","December","January","February","
ggplot(data = result2, mapping = aes(x = Month, y = Times, group = Line)) +
geom_line(aes(color=Line))+
geom_point(color="Black", size=2)+
theme_bw()+
theme(axis.text.x = element_text(size=6,face="bold"))+
labs(x="Month(From 2021.Oct to 2022.Sep)",y="Average shifts",title="Average shifts of light rails in th
```



Looking at these two plots, some of the points are equal to zero and some are relatively small. At first, I thought it was a data problem, but in fact, it's because sometimes the railways are out of service, so there's no data there. And we can see there are more shifts in November than most of the other months for all railways, and basically, the number of shifts is in a reasonable range of about 100 times a day.

Then let's take a look at plots of the mean time of each line that a subway can run the whole distance.

```
hr <- rbind(hrOct,hrNov,hrDec,hrJan,hrFeb,hrMar,hrApr,hrMay,hrJun,hrJul,hrAug,hrSep)
hr <- filter(hr,(from_stop_id==70061&to_stop_id==70105)|(from_stop_id==70038&to_stop_id==70060)|(from_s
result3 <- tapply(hr$travel_time_sec,hr$route_id,mean)
result3 <- as.data.frame(result3)
names(result3) <- c("Time")
lr <- rbind(lrOct,lrNov,lrDec,lrJan,lrFeb,lrMar,lrApr,lrMay,lrJun,lrJul,lrAug,lrSep)
lr <- filter(lr,(from_stop_id==70110&to_stop_id==70201)|(from_stop_id==70236&to_stop_id==70201)|(from_s
result4 <- tapply(lr$travel_time_sec,lr$route_id,mean)
result4 <- as.data.frame(result4)
names(result4) <- c("Time")
result3_4 <- rbind(result3,result4)
Lines <- c("Blue","Orange","Red","Green-B","Green-C","Green-D","Green-E")
ggplot(data = result3_4, mapping = aes(x =Lines, y = Time,fill=Lines))+
geom_bar(stat = 'identity')+theme_bw()+
theme(axis.text.x = element_text(size=6,face="bold"))+
labs(x="Lines",y="Average time for the whole",title="Average time for the whole of each line")
```



From the plot above, we can see that if we consider all the subways at the same speed, the red line is probably the longest and the blue line is the shortest. And for us BU students, the green-B line is also long enough so students from everywhere in Boston can arrive at school by taking the ride.

What I've done and not done

For the data on railways, I've made EDAs to analyze the data and better understand it, and for the shiny application, I've made a map that can choose where you are and where to go, but I haven't found a way to correctly calculate how much time spent using the data because there are so many lines and I can't solve the line-changing problem. Also, because of lacking data for buses and ferries, I can't calculate them as well. The shiny app sometimes can't open because of the huge data, but it can work in github.