MA 615 Final Project_EDA

Jiahao Liu

2022-12-07

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(fmsb)
```

I choose data from November 2021 to October 2022 to complete this EDA report.

Calendar

```
calendar_11_21 <- read.csv("C:/Users/Jiahao Liu/Desktop/calendar_attributes0.txt")
calendar_12_21 <- read.csv("C:/Users/Jiahao Liu/Desktop/calendar_attributes1.txt")
calendar_01_22 <- read.csv("C:/Users/Jiahao Liu/Desktop/calendar_attributes2.txt")
calendar_02_22 <- read.csv("C:/Users/Jiahao Liu/Desktop/calendar_attributes3.txt")
calendar_03_22_1 <- read.csv("C:/Users/Jiahao Liu/Desktop/calendar_attributes4_1.txt")
calendar_03_22_2 <- read.csv("C:/Users/Jiahao Liu/Desktop/calendar_attributes4_2.txt")
calendar_03_22_3 <- read.csv("C:/Users/Jiahao Liu/Desktop/calendar_attributes4_3.txt")
calendar_03_22 <- rbind(calendar_03_22_1, calendar_03_22_2, calendar_03_22_3)
calendar_04_22 <- read.csv("C:/Users/Jiahao Liu/Desktop/calendar_attributes5.txt")
calendar_05_22 <- read.csv("C:/Users/Jiahao Liu/Desktop/calendar_attributes6.txt")
calendar_06_22 <- read.csv("C:/Users/Jiahao Liu/Desktop/calendar_attributes7.txt")
calendar_07_22 <- read.csv("C:/Users/Jiahao Liu/Desktop/calendar_attributes8.txt")
calendar_08_22 <- read.csv("C:/Users/Jiahao Liu/Desktop/calendar_attributes9.txt")
calendar_09_22 <- read.csv("C:/Users/Jiahao Liu/Desktop/calendar_attributes10.txt")
calendar_10_22 <- read.csv("C:/Users/Jiahao Liu/Desktop/calendar_attributes11.txt")

calendar <- rbind(calendar_11_21,calendar_12_21,calendar_01_22,calendar_02_22,calendar_03_22,calendar_0
```

```
count(calendar,service_schedule_typicality)
```

```
##   service_schedule_typicality   n
## 1                           1 878
## 2                           2   4
## 3                           3  33
## 4                           4 472
## 5                           5  17
```

```
data <- data.frame(
  category=c("Typicality-1", "Typicality-2", "Typicality-3", "Typicality-4", "Typicality-5"),
  count=c(878,4,33,472,17)
)

# Compute percentages
data$fraction <- data$count / sum(data$count)

# Compute the cumulative percentages (top of each rectangle)
data$ymax <- cumsum(data$fraction)

# Compute the bottom of each rectangle
data$ymin <- c(0, head(data$ymax, n=-1))

# Compute label position
data$labelPosition <- (data$ymax + data$ymin) / 2

# Compute a good label
data$label <- paste0(data$category, "\n value: ", data$count)

# Make the plot
ggplot(data, aes(ymax=ymax, ymin=ymin, xmax=4, xmin=3, fill=category)) +
  geom_rect() +
  geom_label( x=3.5, aes(y=labelPosition, label=label), size=2.5) +
  scale_fill_brewer(palette=4) +
  coord_polar(theta="y") +
  xlim(c(2, 4)) +
  theme_void() +
  theme(legend.position = "none")
```

```
#unique(calendar$rating_description)
calendar_rate <- data.frame(calendar$service_schedule_typicality,calendar$rating_description)

calendar_rate_1 <- calendar_rate %>%
  filter(calendar.rating_description == 'Fall')
#count(calendar_rate_1,calendar.service_schedule_typicality)

calendar_rate_2 <- calendar_rate %>%
  filter(calendar.rating_description == 'Spring')
#count(calendar_rate_2,calendar.service_schedule_typicality)

calendar_rate_3 <- calendar_rate %>%
  filter(calendar.rating_description == 'Summer')
```

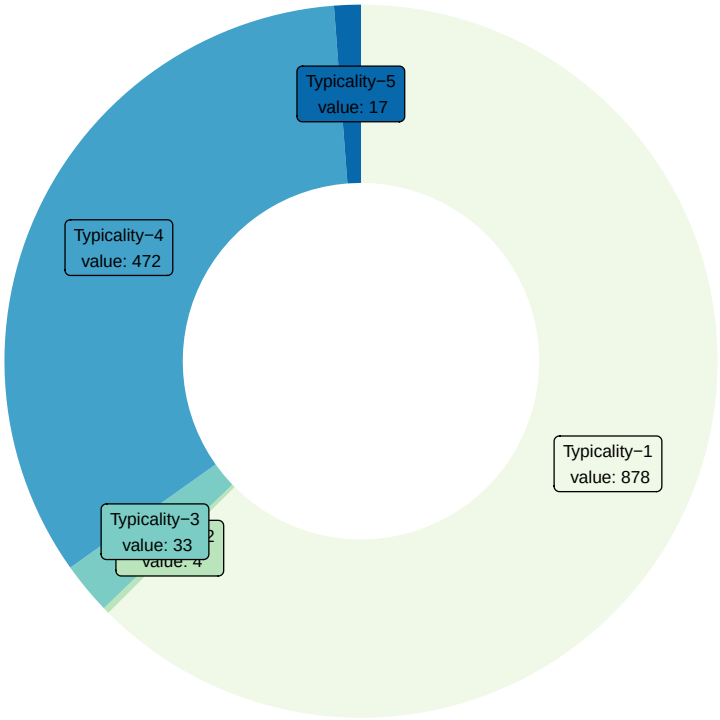
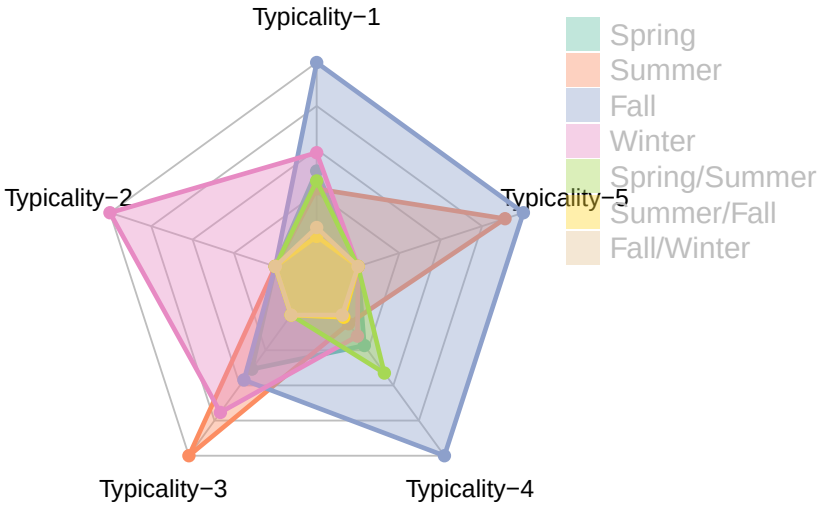


Figure 1: Typicality Distribution

```
#count(calendar_rate_3,calendar.service_schedule_typicality)

calendar_rate_4 <- calendar_rate %>%
  filter(calendar.rating_description == 'Winter')
#count(calendar_rate_4,calendar.service_schedule_typicality)

calendar_rate_5 <- calendar_rate %>%
  filter(calendar.rating_description == 'Summer/Fall')
#count(calendar_rate_5,calendar.service_schedule_typicality)

calendar_rate_6 <- calendar_rate %>%
  filter(calendar.rating_description == 'Fall/Winter')
#count(calendar_rate_6,calendar.service_schedule_typicality)

calendar_rate_7 <- calendar_rate %>%
  filter(calendar.rating_description == 'Spring/Summer')
#count(calendar_rate_7,calendar.service_schedule_typicality)

season_rate <- as.data.frame(matrix( c(132,96,351,169,112,0,18,0,0,0,4,0,0,0,5,13,6,9,0,0,0,58,28,210,45
colnames(season_rate) <- c("Typicality-1", "Typicality-2", "Typicality-3", "Typicality-4", "Typicality-5
rownames(season_rate) <- c("Spring", "Summer" , "Fall" , "Winter" , "Spring/Summer", "Summer/Fall", "Fal

season_rate
```

```
##               Typicality-1 Typicality-2 Typicality-3 Typicality-4 Typicality-5
## Spring                 132            0            5           58            0
## Summer                  96            0           13           28            8
## Fall                   351            0            6          210            9
## Winter                 169            4            9           45            0
## Spring/Summer          112            0            0           96            0
## Summer/Fall              0            0            0           19            0
## Fall/Winter             18            0            0           16            0
```

```
library(RColorBrewer)
coul <- brewer.pal(7, "Set2")
colors_border <- coul
library(scales)
colors_in <- alpha(coul,0.4)

radarchart(season_rate,axistype=0, maxmin=F,pcol = colors_border, pfcol=colors_in, plwd=2.5, plty=1,vlc
legend(x = 'topright', legend = rownames(season_rate), bty = "n", pch=15 , col=colors_in , text.col = "
```

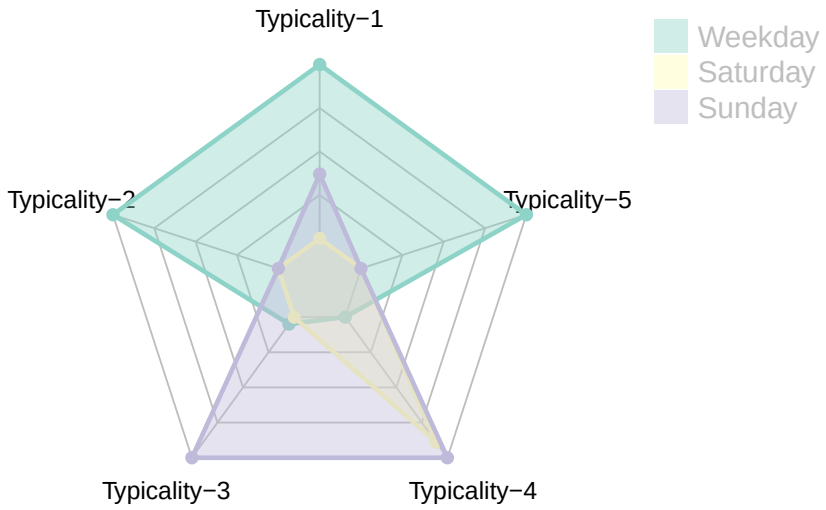


The figure shows that MBTA provide extra service in winter term, change holiday service to typical Saturday or Sunday schedule in Summer term, and in the fall term, MBTA usually change their service due to weather events or construciton.

```
week_rate <- data.frame(
  schedule = calendar$service_schedule_type,
  typicality = calendar$service_schedule_typicality)

week_rate_1 <- week_rate %>%
  filter(schedule == 'Weekday')
#count(week_rate_1,typicality)

week_rate_2 <- week_rate %>%
  filter(schedule == 'Saturday')
#count(week_rate_2,typicality)

week_rate_3 <- week_rate %>%
  filter(schedule == 'Sunday')
#count(week_rate_3,typicality)


week_rate_new <- as.data.frame(matrix( c(366,231,281,4,0,0,5,4,24,146,162,164,17,0,0) , ncol=5))
colnames(week_rate_new) <- c("Typicality-1", "Typicality-2", "Typicality-3", "Typicality-4", "Typicalit
rownames(week_rate_new) <- c("Weekday", "Saturday" , "Sunday")

week_rate_new
```

```
##          Typicality-1 Typicality-2 Typicality-3 Typicality-4 Typicality-5
## Weekday           366            4            5          146           17
## Saturday          231            0            4          162            0
## Sunday            281            0           24          164            0
```

```
library(RColorBrewer)
coul <- brewer.pal(3, "Set3")
colors_border <- coul
library(scales)
colors_in <- alpha(coul,0.4)

radarchart(week_rate_new,axistype=0, maxmin=F,pcol = colors_border, pfcol=colors_in, plwd=2.5, plty=1,v
legend(x = 'topright', legend = rownames(week_rate_new), bty = "n", pch=15 , col=colors_in , text.col =
```



The figure shows that MBTA often has planned disruption due to construction on Saturday and Sunday. For the weekday service, MBTA also provide extra service.

Route Patterns

```
route_11_21 <- read.csv("C:/Users/Jiahao Liu/Desktop/route_patterns0.txt")
route_12_21 <- read.csv("C:/Users/Jiahao Liu/Desktop/route_patterns1.txt")
route_01_22 <- read.csv("C:/Users/Jiahao Liu/Desktop/route_patterns2.txt")
route_02_22 <- read.csv("C:/Users/Jiahao Liu/Desktop/route_patterns3.txt")
```

```
route_03_22_1 <- read.csv("C:/Users/Jiahao Liu/Desktop/route_patterns4_1.txt")
route_03_22_2 <- read.csv("C:/Users/Jiahao Liu/Desktop/route_patterns4_2.txt")
route_03_22_3 <- read.csv("C:/Users/Jiahao Liu/Desktop/route_patterns4_3.txt")
route_03_22 <- rbind(route_03_22_1, route_03_22_2, route_03_22_3)
route_04_22 <- read.csv("C:/Users/Jiahao Liu/Desktop/route_patterns5.txt")
route_05_22 <- read.csv("C:/Users/Jiahao Liu/Desktop/route_patterns6.txt")
route_06_22 <- read.csv("C:/Users/Jiahao Liu/Desktop/route_patterns7.txt")
route_07_22 <- read.csv("C:/Users/Jiahao Liu/Desktop/route_patterns8.txt")
route_08_22 <- read.csv("C:/Users/Jiahao Liu/Desktop/route_patterns9.txt")
route_09_22 <- read.csv("C:/Users/Jiahao Liu/Desktop/route_patterns10.txt")
route_10_22 <- read.csv("C:/Users/Jiahao Liu/Desktop/route_patterns11.txt")

route <- rbind(route_11_21,route_12_21,route_01_22,route_02_22,route_03_22,route_04_22,route_05_22,rout
```

```
pattern_time <- route %>% filter(
  route_pattern_time_desc == 'Early mornings only'|route_pattern_time_desc == 'Weekday evenings only'|r
)



count(route,route_pattern_typicality)
```

```
##   route_pattern_typicality    n
## 1                        1 5604
## 2                        2 2408
## 3                        3 2895
## 4                        4  572
```

```
data <- data.frame(
  category=c("Typicality-1", "Typicality-2", "Typicality-3", "Typicality-4"),
  count=c(5604,2408,2895,572 )
)

# Compute percentages
data$fraction <- data$count / sum(data$count)

# Compute the cumulative percentages (top of each rectangle)
data$ymax <- cumsum(data$fraction)

# Compute the bottom of each rectangle
data$ymin <- c(0, head(data$ymax, n=-1))

# Compute label position
data$labelPosition <- (data$ymax + data$ymin) / 2

# Compute a good label
data$label <- paste0(data$category, "\n value: ", data$count)

# Make the plot
ggplot(data, aes(ymax=ymax, ymin=ymin, xmax=4, xmin=3, fill=category)) +
  geom_rect() +
  geom_label( x=3.5, aes(y=labelPosition, label=label), size=5) +
  scale_fill_brewer(palette=4) +
  coord_polar(theta="y") +
```

```
  xlim(c(2, 4)) +
  theme_void() +
  theme(legend.position = "none")
```

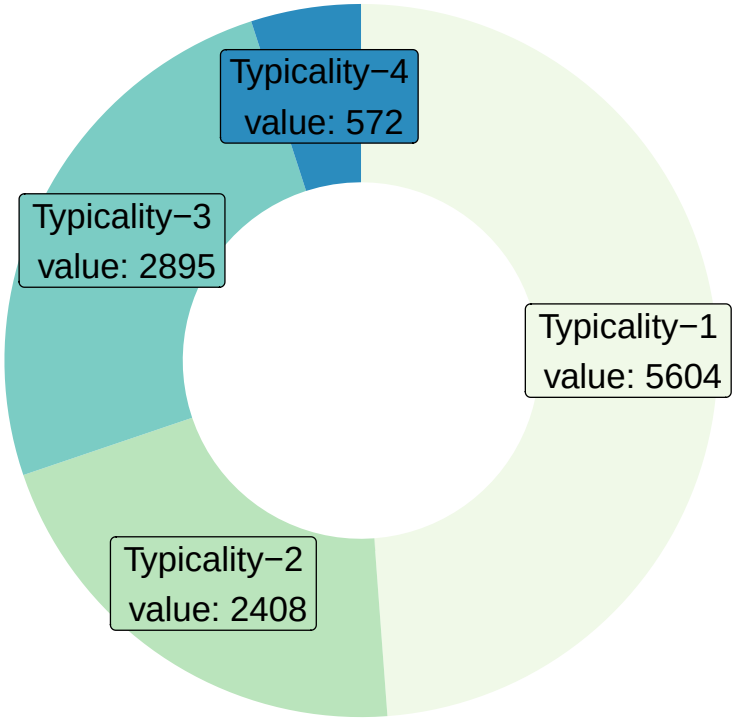


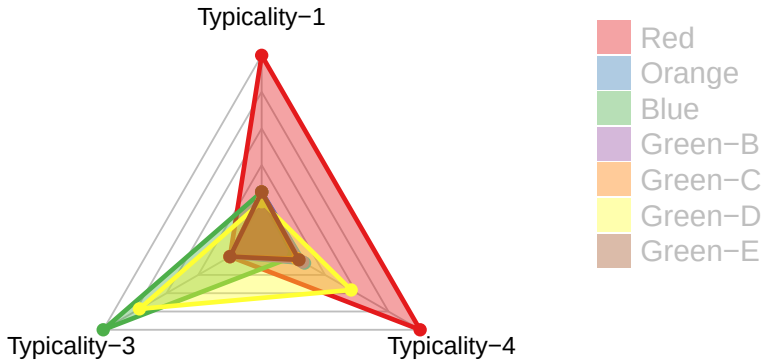
Figure 2: Typicality Distribution

```
#unique(route$route_id)
route_train <- route %>% filter(
  route_id == 'Red'|route_id == 'Orange'|route_id == "Blue"|route_id == "Green-B"|route_id == 'Green-C'
)

route_train_1 <- route_train %>%
  filter(route_id == 'Red')
#count(route_train_1,route_pattern_typicality)

route_train_2 <- route_train %>%
  filter(route_id == 'Orange')
#count(route_train_2,route_pattern_typicality)

route_train_3 <- route_train %>%
  filter(route_id == 'Blue')
#count(route_train_3,route_pattern_typicality)
```

```
route_train_4 <- route_train %>%
  filter(route_id == 'Green-B')
#count(route_train_4,route_pattern_typicality)

route_train_5 <- route_train %>%
  filter(route_id == 'Green-C')
#count(route_train_5,route_pattern_typicality)

route_train_6 <- route_train %>%
  filter(route_id == 'Green-D')
#count(route_train_6,route_pattern_typicality)

route_train_7 <- route_train %>%
  filter(route_id == 'Green-E')
#count(route_train_7,route_pattern_typicality)

route_train_new <- as.data.frame(matrix( c(56,28,28,28,28,26,28,0,0,14,0,0,10,0,148,24,12,22,16,74,18)
colnames(route_train_new) <- c("Typicality-1", "Typicality-3", "Typicality-4")
rownames(route_train_new) <- c("Red", "Orange" , "Blue", "Green-B", "Green-C", "Green-D", "Green-E")

route_train_new
```

```
##          Typicality-1 Typicality-3 Typicality-4
## Red                56            0          148
## Orange             28            0           24
## Blue               28           14           12
## Green-B            28            0           22
## Green-C            28            0           16
## Green-D            26           10           74
## Green-E            28            0           18
```

```
library(RColorBrewer)
coul <- brewer.pal(7, "Set1")
colors_border <- coul
library(scales)
colors_in <- alpha(coul,0.4)

radarchart(route_train_new,axistype=0, maxmin=F,pcol = colors_border, pfcol=colors_in, plwd=2.5, plty=1
legend(x = 'topright', legend = rownames(route_train_new), bty = "n", pch=15 , col=colors_in , text.col
```



This figure shows that red line diversions from normal service, such as planned detours, bus shuttles, or snow routes. Blue line and Green-D line contains special routing which only runs a handful of times per day.