MA615 Final Project

Yin Xu

2022-12-10

MBTA Data EDA

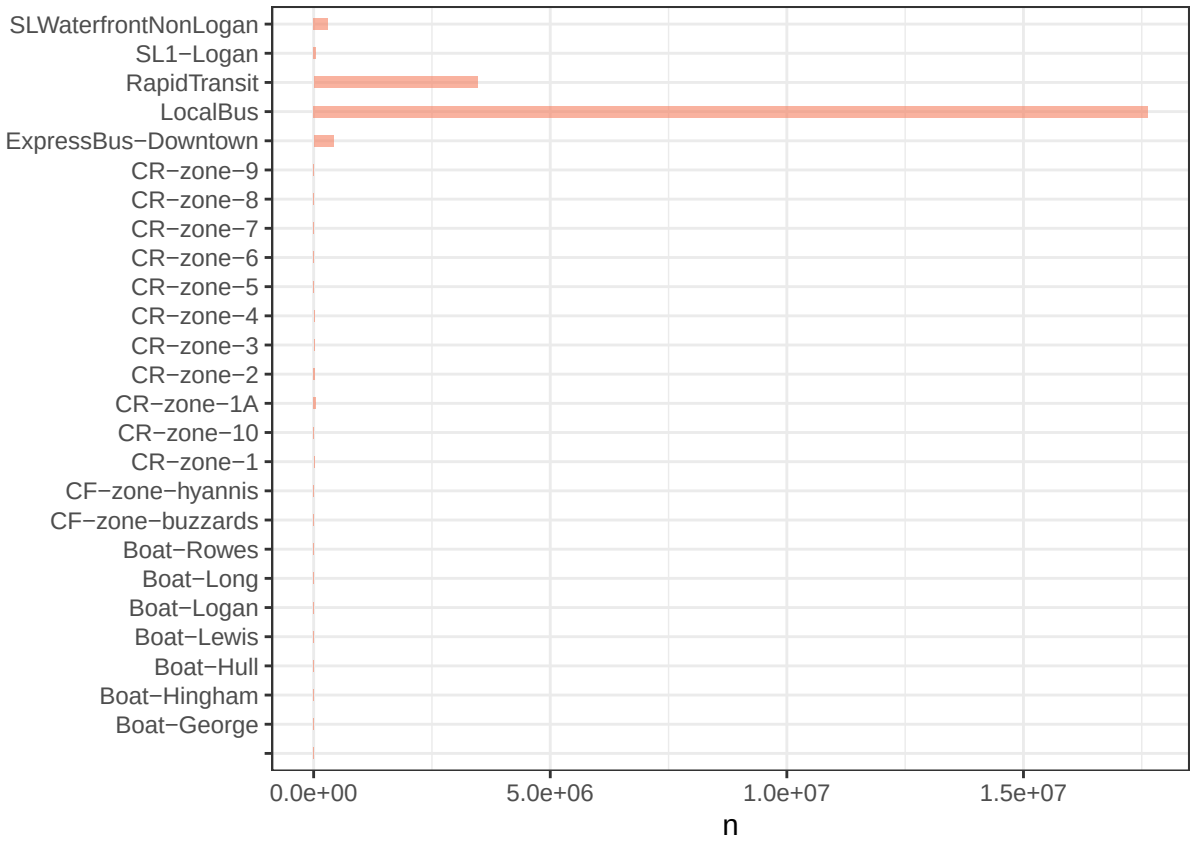
This report includes the data from MBTA in last 12 months with all categories of transit (https://cdn.mbta.com/archive/archived_feeds.txt), and I picked one week from each month. The dataset is large, to run code faster and drop the useless data, I will do the data cleaning first.

In this report, I'm gonna graph the distribution of the `stops`. Besides, I found that typicality is worthy of analysis. According to the reference(https://github.com/mbta/gtfs-documentation/blob/master/reference/gtfs.md#calendar_attributestxt), in this dataset, current valid values are: 0 (or empty): Not defined; 1: Typical service with perhaps minor modifications; 2: Extra service supplements typical schedules; 3: Reduced holiday service is provided by typical Saturday or Sunday schedule; 4: Major changes in service due to a planned disruption, such as construction; 5: Major reductions in service for weather events or other atypical situations.

Read Data and Data Cleaning

Count stops

```
newdata_df %>%
  ggplot( aes(x=zone_id, y=n)) +
    geom_bar(stat="identity", fill="#f68060", alpha=.6, width=.4) +
    coord_flip() +
    xlab("") +
    theme_bw()
```



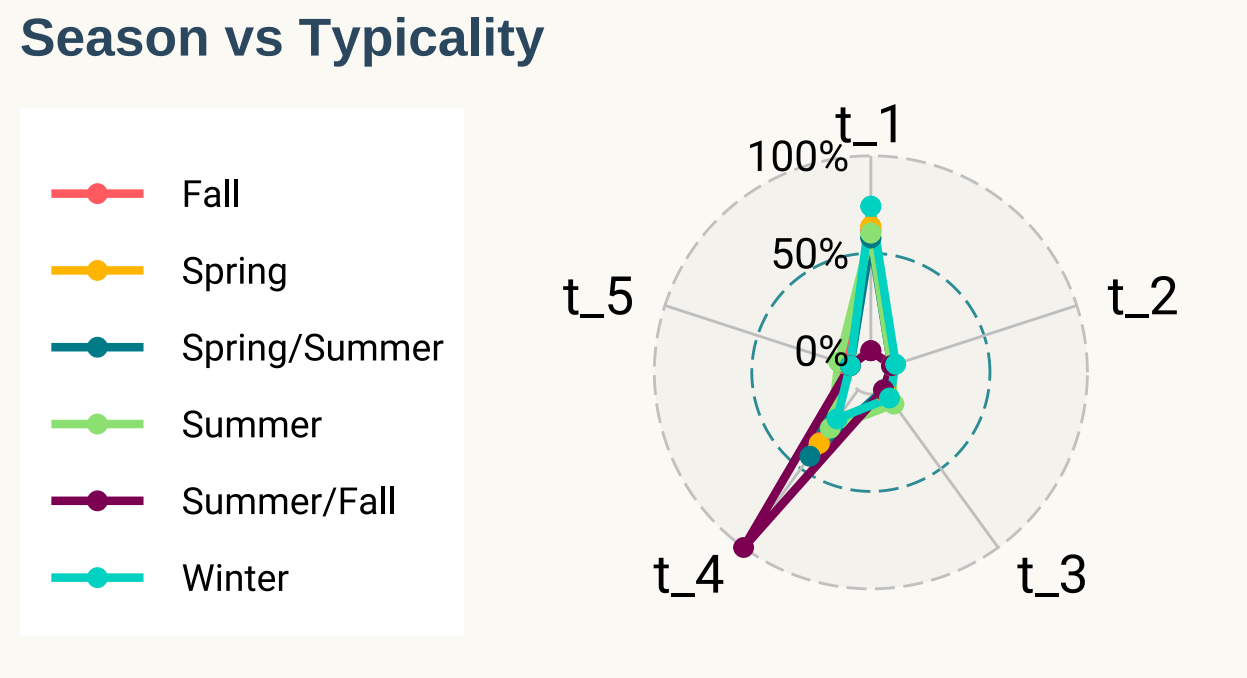
The graph shows that bus has the most stops.

Season vs Typicality

```
#Clean Data
new_cadata <- ca_data %>%
  count(rating_description, service_schedule_typicality, sort = TRUE) %>%
  pivot_wider(names_from = service_schedule_typicality, values_from = n)
new_cadata[is.na(new_cadata)] <- 0

new_cadata <- new_cadata %>%
  mutate(t_1 = `1`/(`1`+`2`+`3`+`4`+`5`),
         t_2 = `2`/(`1`+`2`+`3`+`4`+`5`),
         t_3 = `3`/(`1`+`2`+`3`+`4`+`5`),
         t_4 = `4`/(`1`+`2`+`3`+`4`+`5`),
         t_5 = `5`/(`1`+`2`+`3`+`4`+`5`))
new_cadata <- new_cadata[,c(1,7:11)]
```

```
library(ggradar)
library(palmerpenguins)
library(tidyverse)
library(scales)
library(showtext)

font_add_google("Roboto", "roboto")
```

```
showtext_auto()

#plot1
#code citation: gg gallery
plot_1 <- new_cadata %>%
  ggradar(
    font.radar = "roboto",
    grid.label.size = 7,
    axis.label.size = 7,
    group.point.size = 3
  )+
  labs(title = "Season vs Typicality") +
  theme(
      plot.background = element_rect(fill = "#fbf9f4", color = "#fbf9f4"),
      panel.background = element_rect(fill = "#fbf9f4", color = "#fbf9f4"),
      plot.title.position = "plot",
      plot.title = element_text(
        family = "lobstertwo",
        size = 20,
        face = "bold",
        color = "#2a475e"
    )
  )
plot_1
```



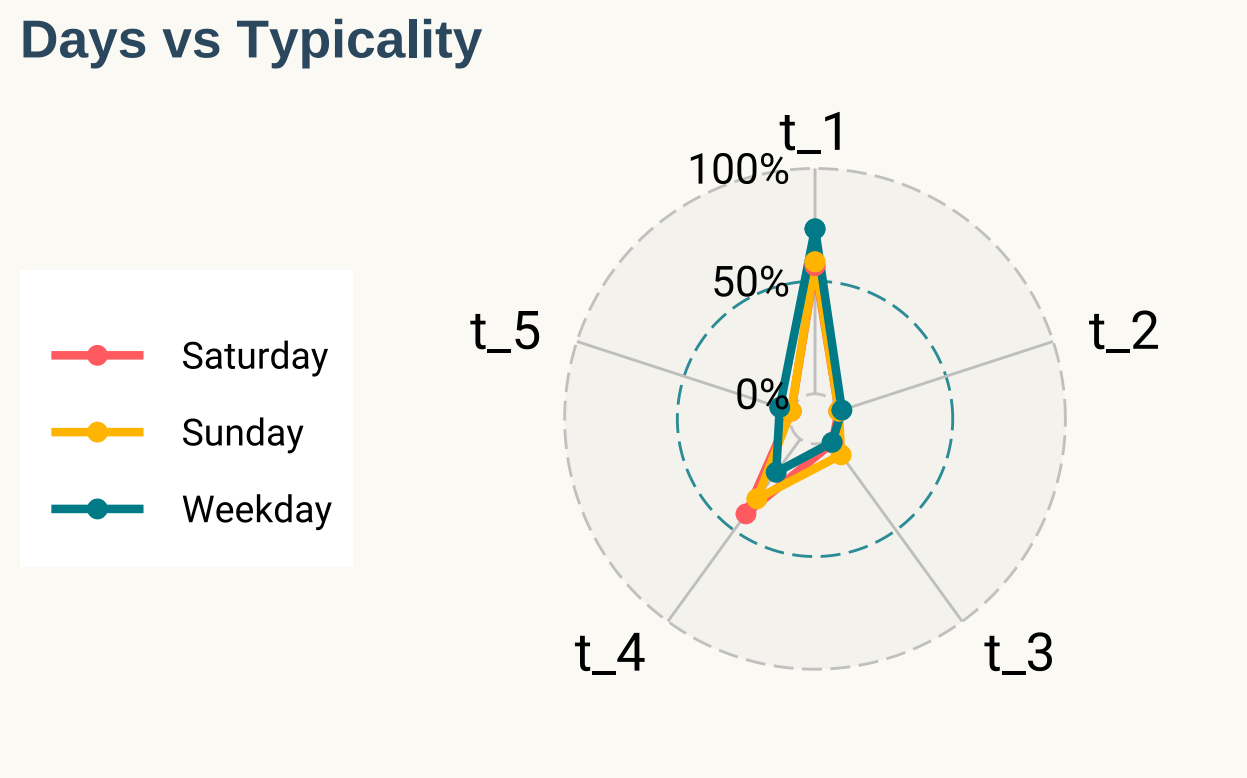
According to this graph, we can know that in most seasons, the most typical service reported is typical

service with perhaps minor modifications, but in summer/fall, major changes in service due to a planned disruption becomes the most, such as construction. Therefore, it can be concluded as MBTA generally tend to be constructed in summer/fall.

Days vs Typicality

```
#Clean Data
new_data_days <- ca_data %>%
  count(service_schedule_type, service_schedule_typicality, sort = TRUE) %>%
  pivot_wider(names_from = service_schedule_typicality, values_from = n)
new_data_days[is.na(new_data_days)] <- 0

new_data_days <- new_data_days %>%
  mutate(t_1 = `1`/(`1`+`2`+`3`+`4`+`5`),
         t_2 = `2`/(`1`+`2`+`3`+`4`+`5`),
         t_3 = `3`/(`1`+`2`+`3`+`4`+`5`),
         t_4 = `4`/(`1`+`2`+`3`+`4`+`5`),
         t_5 = `5`/(`1`+`2`+`3`+`4`+`5`))
new_data_days <- new_data_days[,c(1,7:11)]
```

```
plot_2 <- new_data_days %>%
  ggradar(
    font.radar = "roboto",
    grid.label.size = 7,
    axis.label.size = 7,
    group.point.size = 3) +
  labs(title = "Days vs Typicality") +
  theme(
      plot.background = element_rect(fill = "#fbf9f4", color = "#fbf9f4"),
      panel.background = element_rect(fill = "#fbf9f4", color = "#fbf9f4"),
      plot.title.position = "plot",
      plot.title = element_text(
        family = "lobstertwo",
        size = 20,
        face = "bold",
        color = "#2a475e"
    )
  )

plot_2
```



This graph shows that weekdays or weekends do not influence the typical service that much, weekdays have the less type4(Major changes in service due to a planned disruption, such as construction), MBTA might consider that the construction cannot influence the traffic in weekdays, so this kind of issue might be moved to saturdays and sundays.