Report of 615FinalProject

Lintong Li

2022-12-15

```
library(ggplot2)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.5.0
## v readr   2.1.3      v forcats 0.5.2
## v purrr   0.3.5
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
options(dplyr.summarise.inform = FALSE)
```

Clean data

```
#read files
Reliability <- read.csv("~/Desktop/MBTA_Bus_Commuter_Rail_Rapid_Transit_Reliability.csv")
#Split service date column
Re <- Reliability %>%  separate(service_date, c("service_date", "service_time"),
                                sep = ' ')
Re$service_date <- as.Date(Re$service_date)
#Fliter each week in every month from November 2021 to October 2022
ReNov <- Re %>% filter(service_date == "2021-11-22"|
                         service_date == "2021-11-23"| service_date == "2021-11-24"|
                         service_date == "2021-11-25"| service_date == "2021-11-26"|
ReDec <- Re %>% filter(service_date == "2021-12-20"| service_date == "2021-12-21"|
                         service_date == "2021-12-24"| service_date == "2021-12-25"|
ReJan <- Re %>% filter(service_date == "2022-01-24"| service_date == "2022-01-25"|
                         service_date == "2022-01-28"| service_date == "2022-01-29"|
ReFeb <- Re %>% filter(service_date == "2022-02-21"| service_date == "2022-02-22"|
                         service_date == "2022-02-25"|service_date == "2022-02-26"|
ReMar <- Re %>% filter(service_date == "2022-03-21"|service_date == "2022-03-22"|
                         service_date == "2022-03-25"| service_date == "2022-03-26"|
ReApr <- Re %>% filter(service_date == "2022-04-18"| service_date == "2022-04-19"|
                         service_date == "2022-04-22"| service_date == "2022-04-23"|
ReMay <- Re %>% filter(service_date == "2022-05-23"| service_date == "2022-05-24"|
                         service_date == "2022-05-27"| service_date == "2022-05-28"|
ReJun <- Re %>% filter(service_date == "2022-06-20"| service_date == "2022-06-21"|
```

```
                          service_date == "2022-06-24"| service_date == "2022-06-25"|
ReJul <- Re %>% filter(service_date == "2022-07-25"| service_date == "2022-07-26"|
                          service_date == "2022-07-29"| service_date == "2022-07-30"|
ReAug <- Re %>% filter(service_date == "2022-08-22"| service_date == "2022-08-23"|
                          service_date == "2022-08-26"| service_date == "2022-08-27"|
ReSep <- Re %>% filter(service_date == "2022-09-19"| service_date == "2022-09-20"|
                          service_date == "2022-09-23"|service_date == "2022-09-24"|
ReOct <- Re %>% filter(service_date == "2022-10-24"|service_date == "2022-10-25"|
                          service_date == "2022-10-26"| service_date == "2022-10-27"|
                          service_date == "2022-10-28"| service_date == "2022-10-29"|
```

```
#Create a new column called month
ReNov$month <- "21Nov"
ReDec$month <- "21Dec"
ReJan$month <- "22Jan"
ReFeb$month <- "22Feb"
ReMar$month <- "22Mar"
ReApr$month <- "22Apr"
ReMay$month <- "22May"
ReJun$month <- "22Jun"
ReJul$month <- "22Jul"
ReAug$month <- "22Aug"
ReSep$month <- "22Sep"
ReOct$month <- "22Oct"
```

```
#Bind multiple dataframe
Reoutput <- rbind(ReNov,ReDec,ReJan,ReFeb,ReMar,
                  ReApr,ReMay,ReJun,ReJul,ReAug,ReSep,ReOct)
```

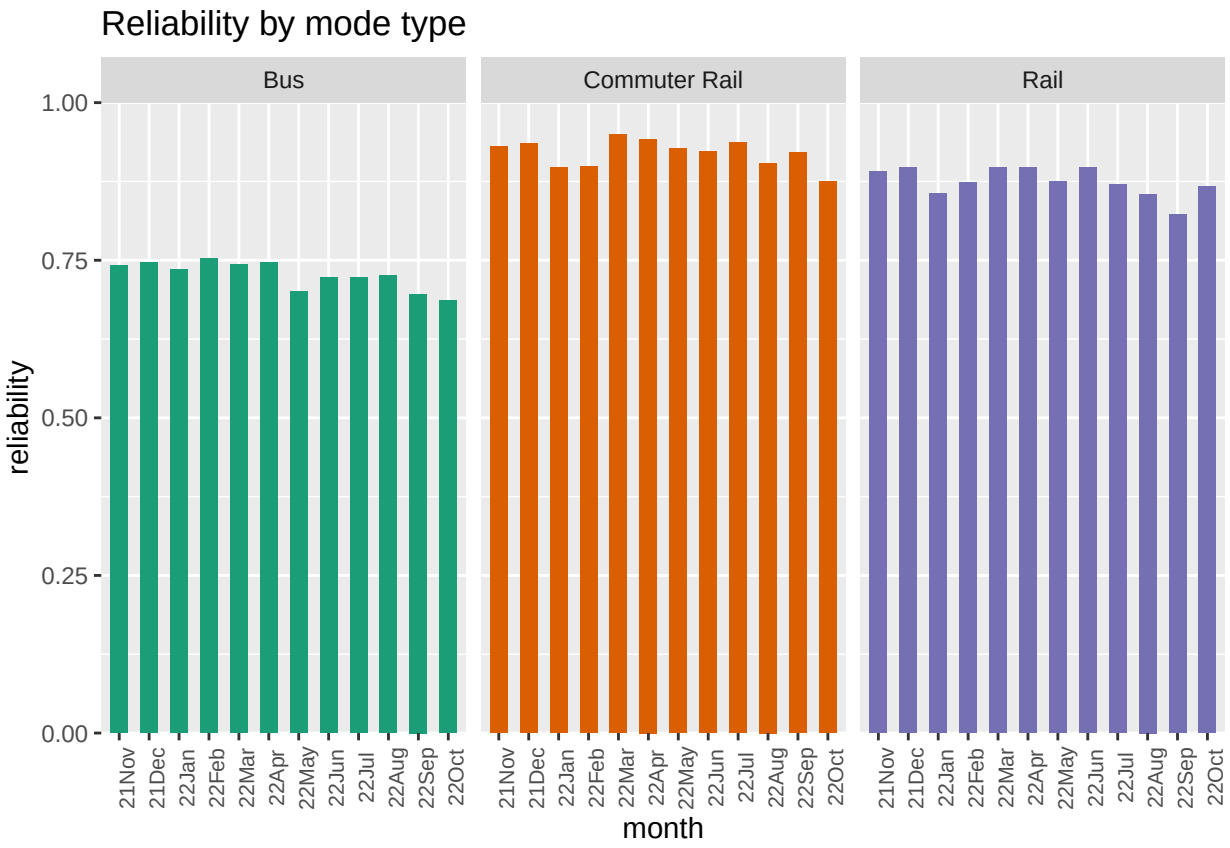
```
#Read route data
routes <- read.csv("~/Desktop/routes.txt",header = TRUE)
#Join Reliability data with route
ReJoin <- left_join(Reoutput, routes,
              by = c("gtfs_route_id" = "route_id"))
Re <-  select(ReJoin, c(1,3,6,7,8,11,12,13,20,22,23))
```

```
#Order data by month
order = c("21Nov", "21Dec","22Jan","22Feb","22Mar","22Apr","22May","22Jun",
          "22Jul","22Aug","22Sep","22Oct")
#Calculate Reliability in group
ReSum <- Re %>% group_by(mode_type,month) %>%
  summarise(across(c(otp_numerator,otp_denominator,cancelled_numerator), sum))  %>%
  mutate(reliability = otp_numerator/otp_denominator) %>%
  mutate(month = factor(month, levels = order))
```

EDA

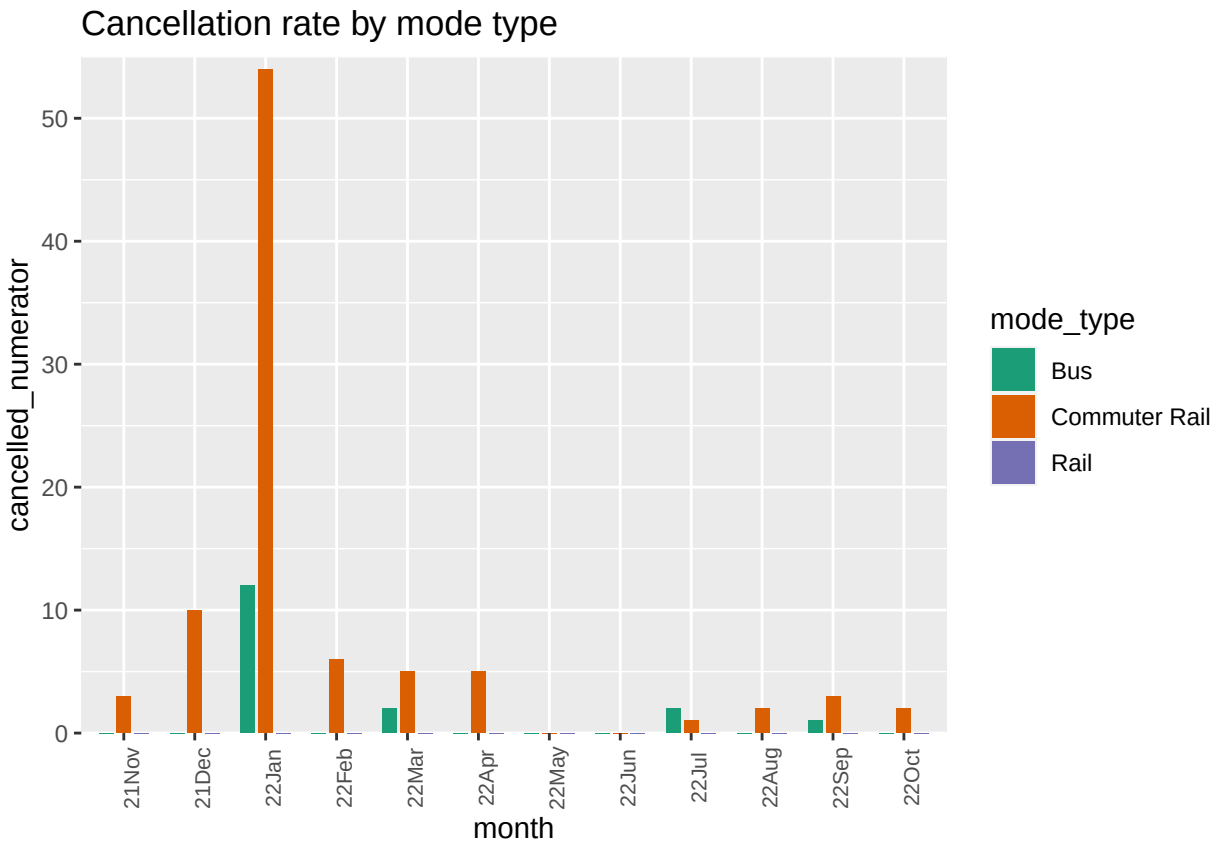
Bus & Commuter Rail & Rapid Transit

```
ggplot(ReSum, aes(x = month, y = reliability, fill = mode_type)) +
         geom_bar(position = position_dodge(0.75),width=0.6,stat = "identity")+
  coord_cartesian(ylim=c(0,1)) +
  scale_y_continuous(expand = c(0,0)) +
  scale_fill_brewer(palette = "Dark2")+
  theme(axis.text.x = element_text(angle=90,size=8))+
  guides(fill = "none") +
  labs(title = "Reliability by mode type") +
  facet_wrap(~ mode_type)
```



The plot shows that the reliability of Bus is the lowest, and the reliability of commuter tail is higher than rapid transit. There is no obvious difference according to month distribution.

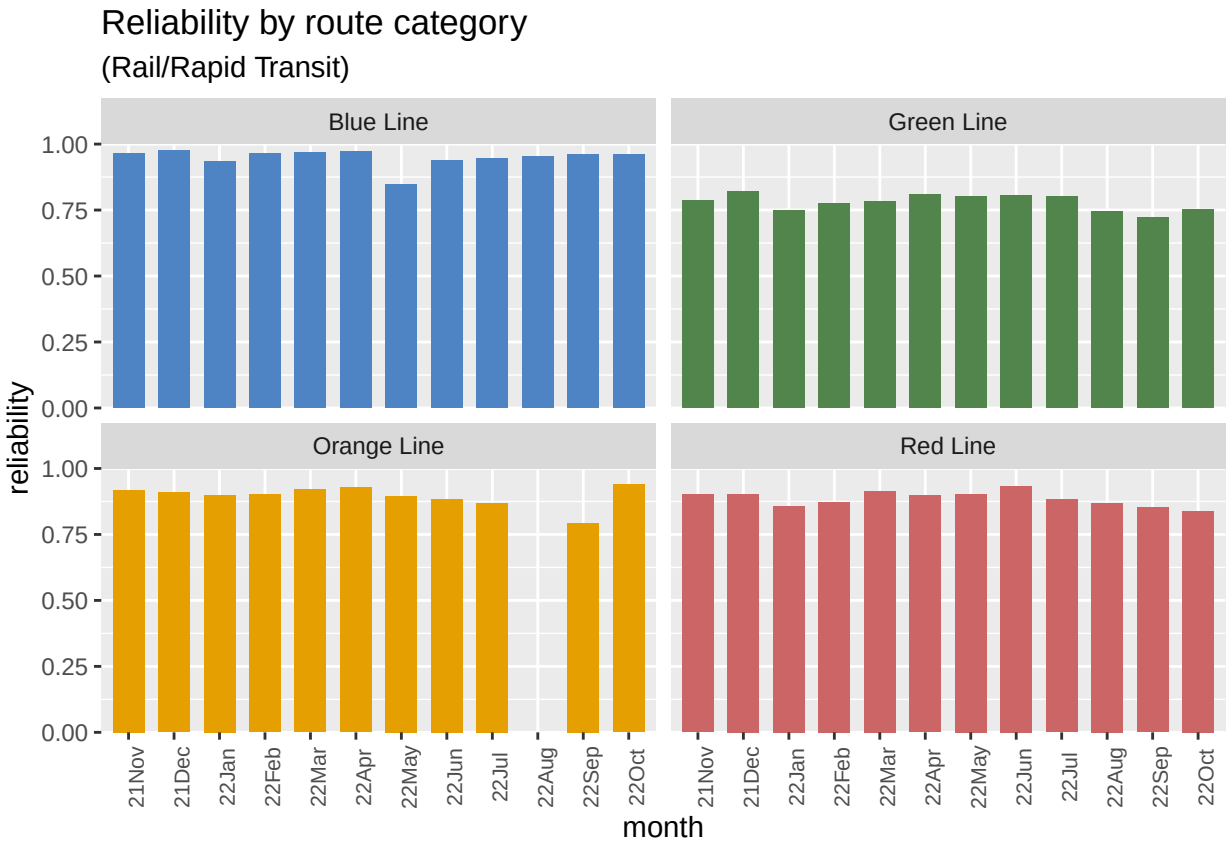
```
ggplot(ReSum, aes(x = month, y = cancelled_numerator, fill = mode_type)) +
         geom_bar(position = position_dodge(0.75),width=0.6,stat = "identity")+
  coord_cartesian(ylim=c(0,55)) +
  scale_y_continuous(expand = c(0,0)) +
  scale_fill_brewer(palette = "Dark2")+
  theme(axis.text.x = element_text(angle=90,size=8))+
  labs(title = "Cancellation rate by mode type")
```



The plot shows that the cancelled number of Commuter Rail system the highest and almost appears every month, so I think this system is not suitable. The highest number is in December 2021.

Rapid Transit

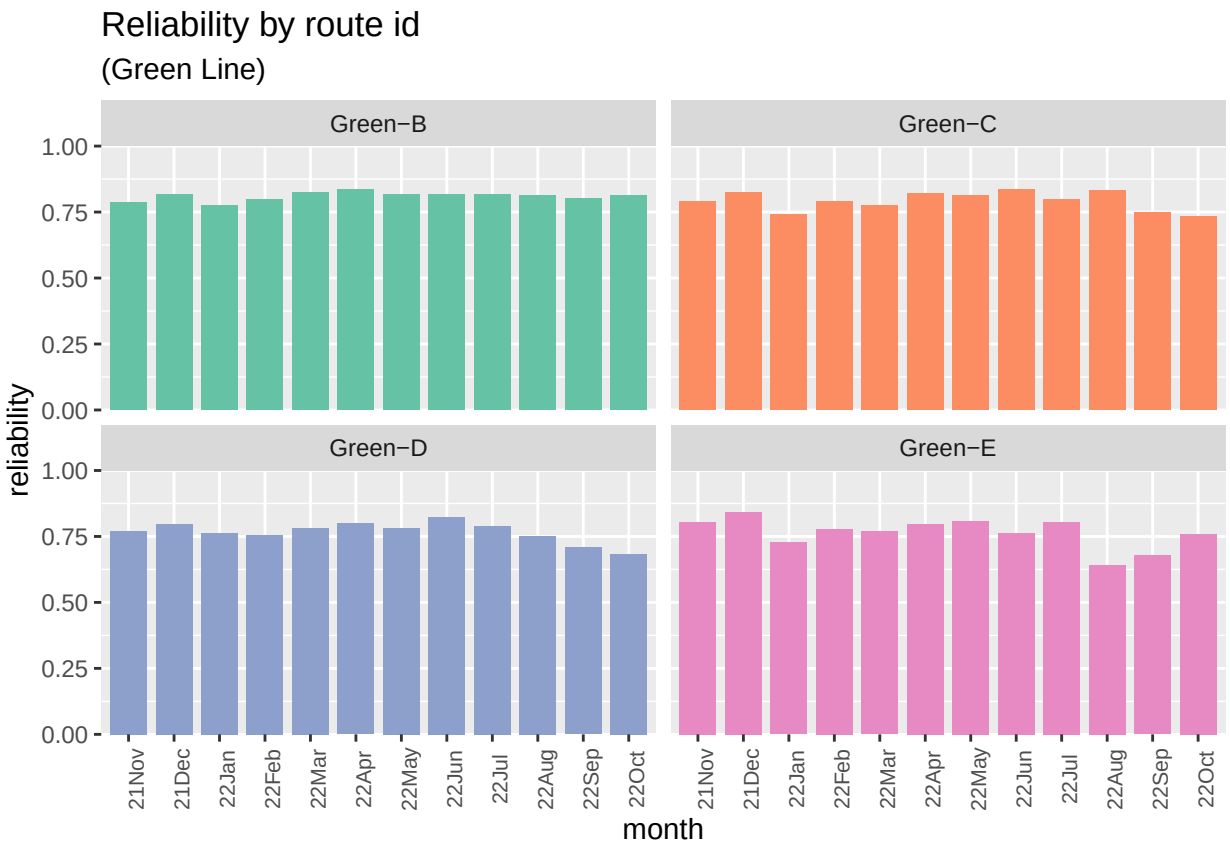
```
Re_rail <- Re %>% filter(mode_type=="Rail")
ReSum_rail <- Re_rail %>% group_by(route_category,month) %>%
  summarise(across(c(otp_numerator,otp_denominator,cancelled_numerator), sum))  %>%
  mutate(reliability = otp_numerator/otp_denominator) %>%
  mutate(month = factor(month, levels = order))
ggplot(ReSum_rail, aes(x = month, y = reliability, fill = route_category)) +
         geom_bar(width=0.7,stat = "identity")+
  coord_cartesian(ylim=c(0,1)) +
  scale_y_continuous(expand = c(0,0)) +
  scale_fill_manual(values=c("#4E84C4", "#52854C", "#E69F00","#CC6666"))+
  theme(axis.text.x = element_text(angle=90,size=8))+
  guides(fill = "none") +
  labs(title = "Reliability by route category",
  subtitle = "(Rail/Rapid Transit)") +
  facet_wrap(~ route_category)
```



The plot shows that the reliability of Blue Line is the highest, and Green Line is the lowest. Obviously Orange Line is missing data in August, so I assume it is out of service at this time.

Green Line

```
Re_railG <- Re %>% filter(route_category=="Green Line")
ReSum_railG <- Re_railG %>% group_by(gtfs_route_id,month) %>%
  summarise(across(c(otp_numerator,otp_denominator,cancelled_numerator), sum))  %>%
  mutate(reliability = otp_numerator/otp_denominator) %>%
  mutate(month = factor(month, levels = order))
ggplot(ReSum_railG, aes(x = month, y = reliability, fill = gtfs_route_id)) +
         geom_bar(width=0.8,stat = "identity")+
  coord_cartesian(ylim=c(0,1)) +
  scale_y_continuous(expand = c(0,0)) +
  scale_fill_brewer(palette="Set2") +
  theme(axis.text.x = element_text(angle=90,size=8))+
  guides(fill = "none") +
  labs(title = "Reliability by route id",
  subtitle = "(Green Line)") +
  facet_wrap(~gtfs_route_id)
```

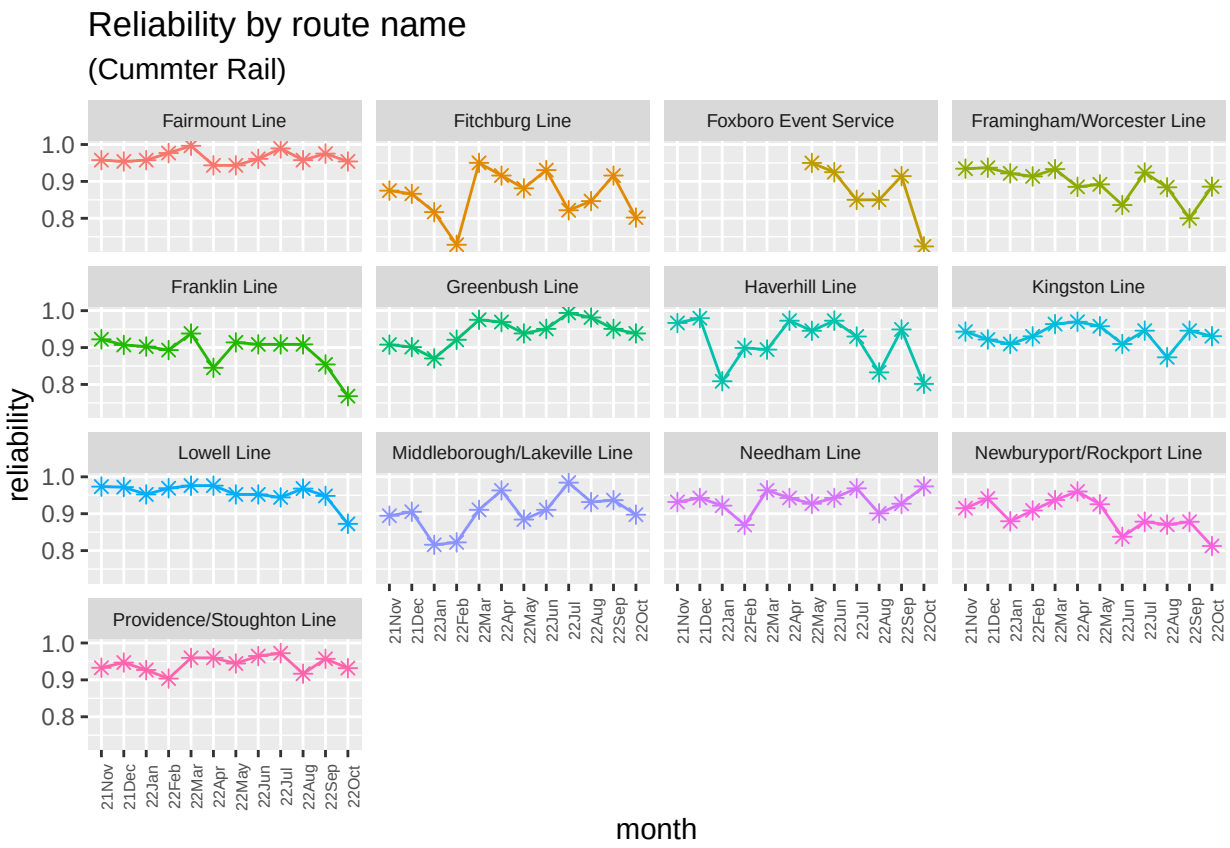


The plot shows that Green-B system is more stable, because the data is evenly distributed from month to month.

Commuter Rail

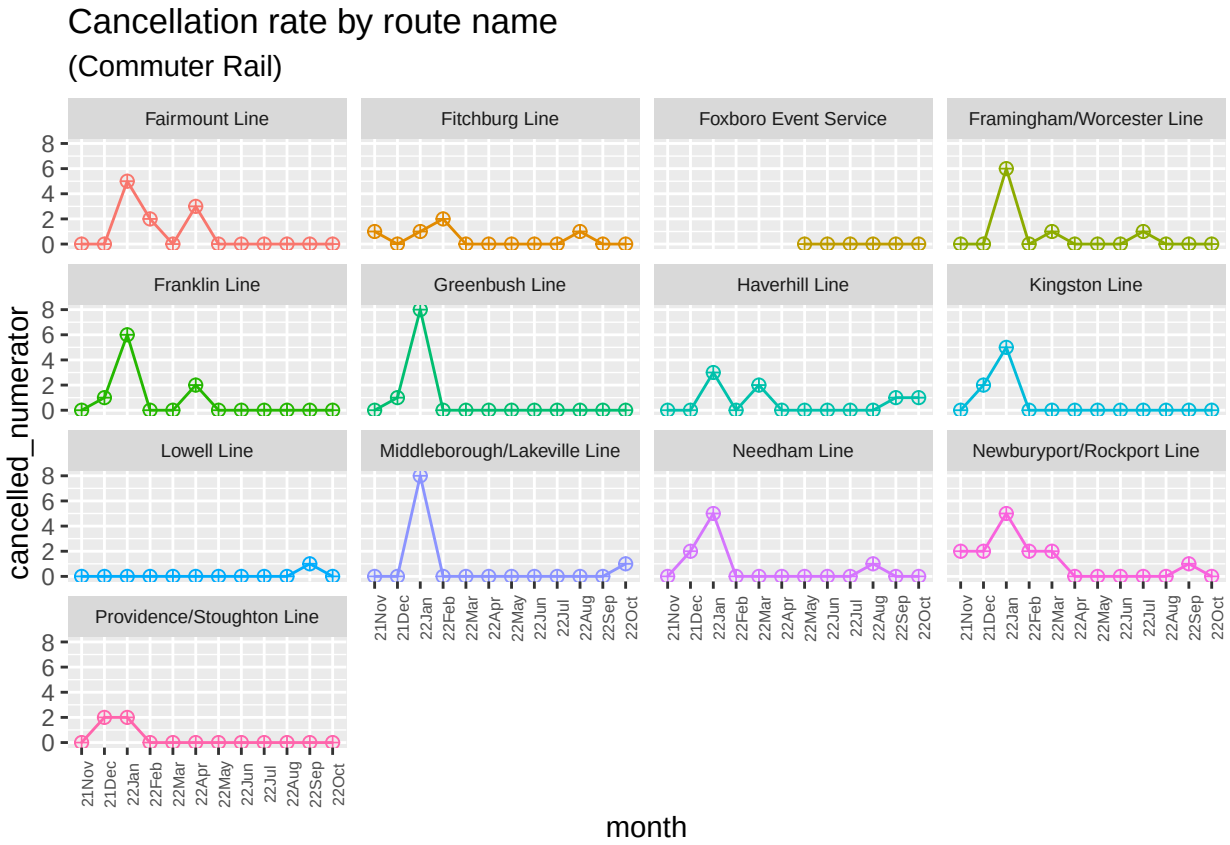
```
Re_commuter <- Re %>% filter(mode_type=="Commuter Rail")
ReSum_commuter <- Re_commuter %>% group_by(route_long_name,month) %>%
  summarise(across(c(otp_numerator,otp_denominator,cancelled_numerator), sum))  %>%
  mutate(reliability = otp_numerator/otp_denominator) %>%
  mutate(month = factor(month, levels = order))
ggplot(ReSum_commuter, aes(x = month, y = reliability, group = route_long_name))+
  geom_line(aes(color=route_long_name),linetype=1,size=0.5)+
  geom_point(aes(color=route_long_name),shape=8,size=2)+
  theme(axis.text.x = element_text(angle=90,size=6),
        strip.text.x = element_text(size = 7))+
  guides(color = "none") +
  labs(title = "Reliability by route name",
  subtitle = "(Cummter Rail)") +
  facet_wrap(~ route_long_name)
```

```
## Warning: Using 'size' aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use 'linewidth' instead.
```



The plot shows the reliability of each type of Commuter Rail has different trends and fluctuations from month to month, with most plummeting in January and February.

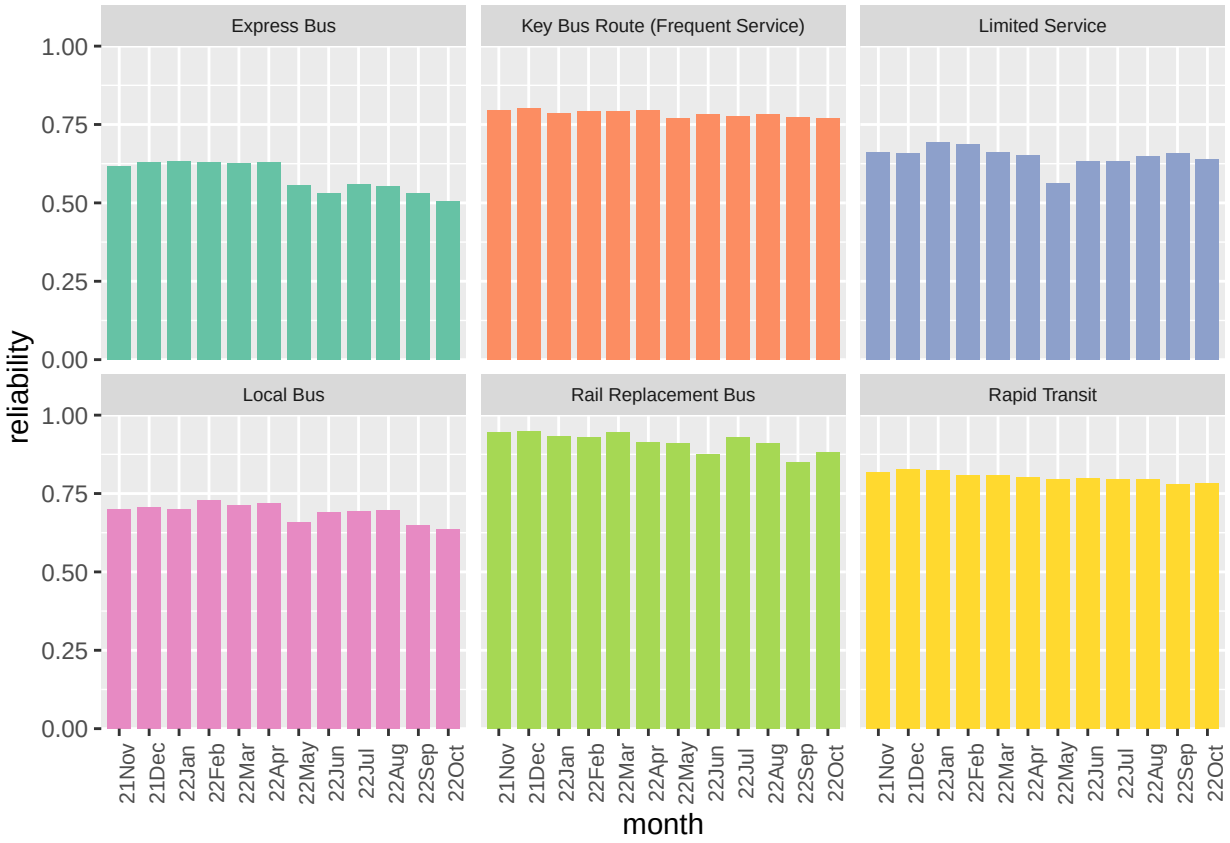
```
ggplot(ReSum_commuter, aes(x = month, y = cancelled_numerator,
                           group = route_long_name))+
  geom_line(aes(color=route_long_name),linetype=1,size=0.5)+
  geom_point(aes(color=route_long_name),shape=10,size=2)+
  theme(axis.text.x = element_text(angle=90,size=6),
        strip.text.x = element_text(size = 7))+
  guides(color = "none") +
  labs(title = "Cancellation rate by route name",
  subtitle = "(Commuter Rail)") +
  facet_wrap(~ route_long_name)
```



The plot shows the distribution cancelled number of each Commuter Rail are increasing sharply in January and February. I consider it is because of the cold weather, which prevents the normal operation.

Bus

```
Re_bus <- Re %>% filter(mode_type=="Bus")
ReSum_bus <- Re_bus %>% group_by(gtfs_route_desc,month) %>%
  summarise(across(c(otp_numerator,otp_denominator,cancelled_numerator), sum))  %>%
  mutate(reliability = otp_numerator/otp_denominator) %>%
  mutate(month = factor(month, levels = order))
ggplot(ReSum_bus, aes(x = month, y = reliability, fill = gtfs_route_desc)) +
         geom_bar(position = position_dodge(0.75),width=0.8,stat = "identity")+
  coord_cartesian(ylim=c(0,1)) +
  scale_y_continuous(expand = c(0,0)) +
  scale_fill_brewer(palette="Set2")+
  theme(axis.text.x = element_text(angle=90,size=8),
        strip.text.x = element_text(size = 7))+
  guides(fill = "none") +
  facet_wrap(~ gtfs_route_desc)
```

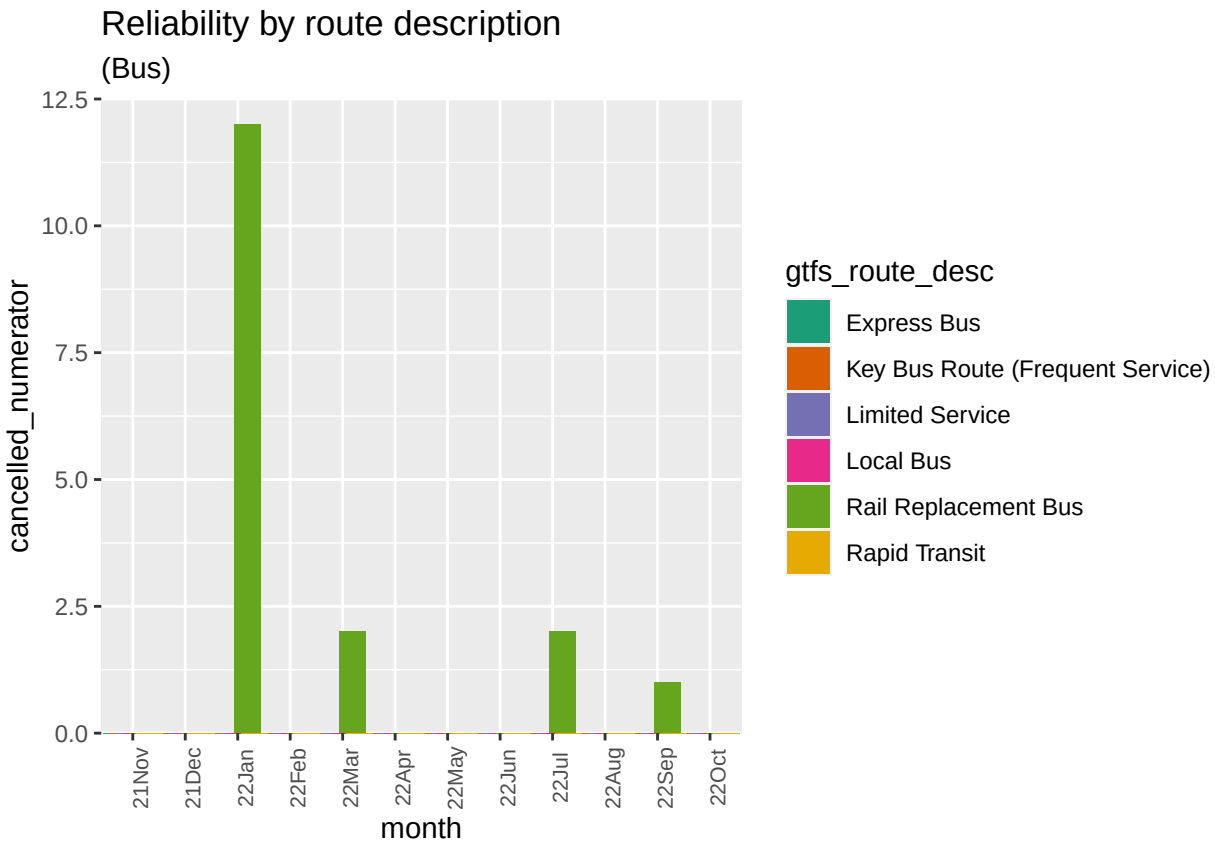


The plot shows the reliability of Rail Replacement Bus is the highest, and Express Bus and Limited Service is the lowest.

```
ggplot(ReSum_bus, aes(x = month, y = cancelled_numerator, fill = gtfs_route_desc)) +
         geom_bar(position = position_dodge(0.75), width=3, stat = "identity")+
  coord_cartesian(ylim=c(0,12.5)) +
  scale_y_continuous(expand = c(0,0)) +
  scale_fill_brewer(palette = "Dark2")+
```

```
  labs(title = "Reliability by route description",
  subtitle = "(Bus)") +
  theme(axis.text.x = element_text(angle=90,size=8))
```

```
## Warning: `position_dodge()` requires non-overlapping x intervals
```



The plot shows that the cancelled number of Rail Replacement Bus appears in four months, January 2022 is the highest.