Mini-Project 1

Bravo Choi (5810288)

2024-04-20

Table of contents

1 Abstract

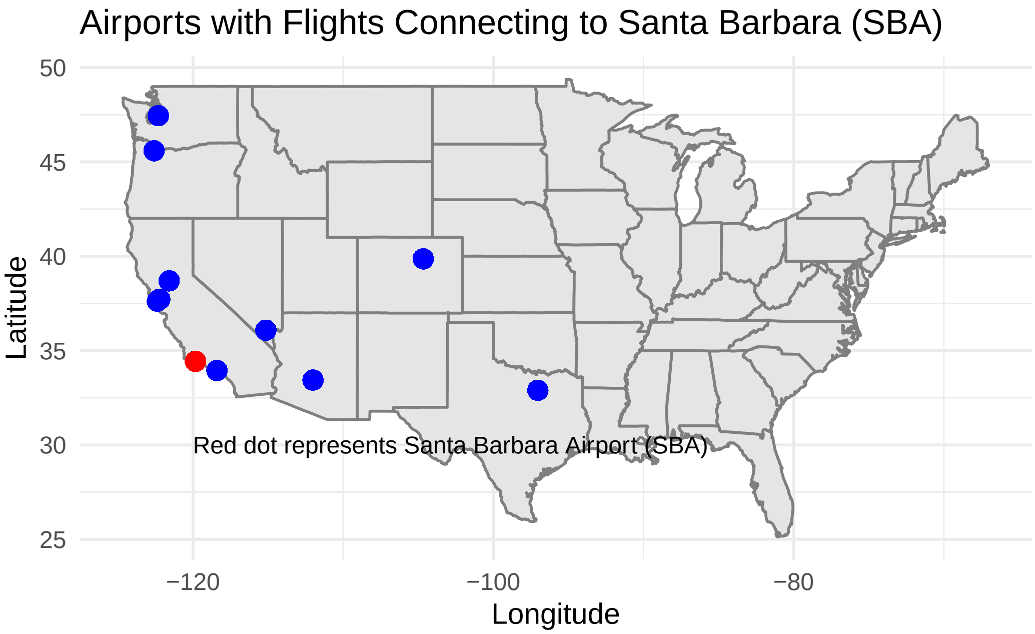
This report examines the difference between the number of flights landing in SBA and the number of flights leaving SBA, the distribution flight duration between the flights departing from and arriving in SBA, the association between the scheduled departure time and the length of delay, the distribution of departure times, the distribution of arrival times, and the months that have higher/lower average departure and arrival delays.

2 How many airlines are connected to Santa Barbara

We first used US map to see how many airlines are connected to Santa Barbara. From the map we can see that there are nine airlines connected Santa Barbara.

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.3     v readr     2.1.4
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.0     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.0
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to
```

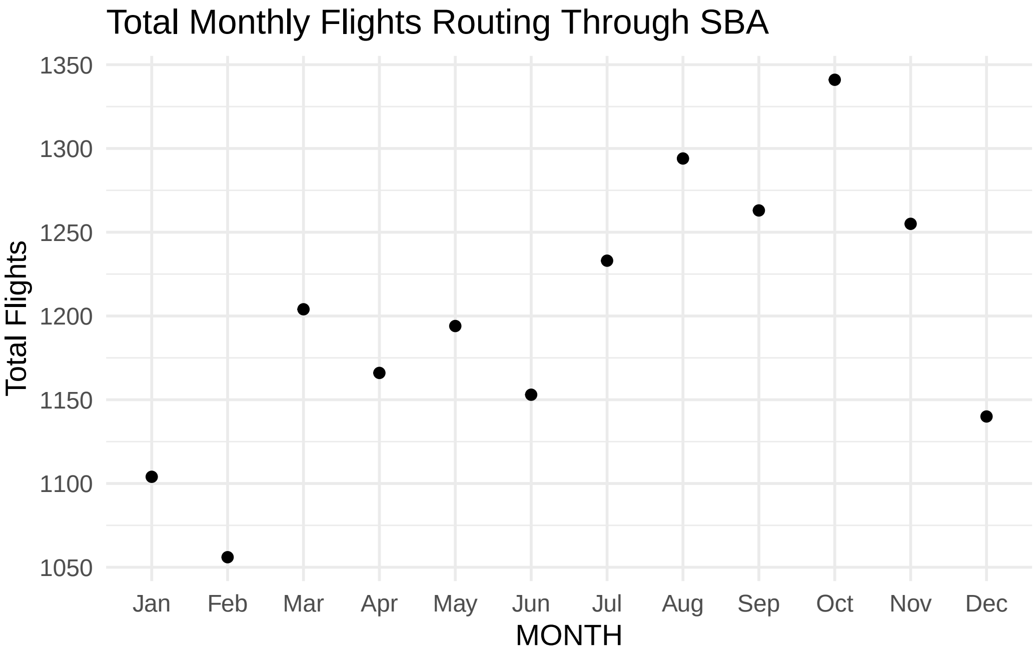
```
[1] "/home/jovyan/100-sp24/Mini_Projects/MP01"
```


Airports with Flights Connecting to Santa Barbara (SBA)
50
45
40
35
30
25
Latitude
Red dot represents Santa Barbara Airport (SBA)
−120
−100
−80
Longitude

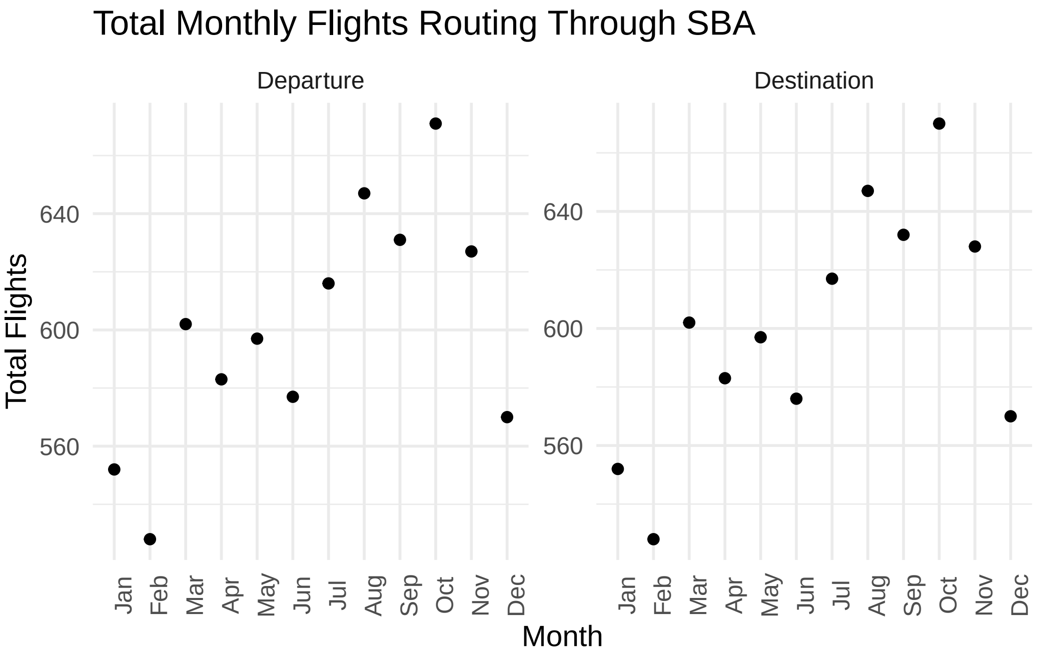
3 Investigating high and low seasons for airlines through Santa Barbara

Now let's identify the "high" and "low" seasons for travel to and from Santa Barbara. From the graph, we can see that the airline travel through Santa Barbara is more frequent around August to October. Comparing the graphs between arrival and departure, there is no much difference, indicating that the locals in Santa Barbara are equally likely to travel outside of the city as the visitors flowing into Santa Barbara.

```
`geom_line()`: Each group consists of only one observation.
i Do you need to adjust the group aesthetic?
```



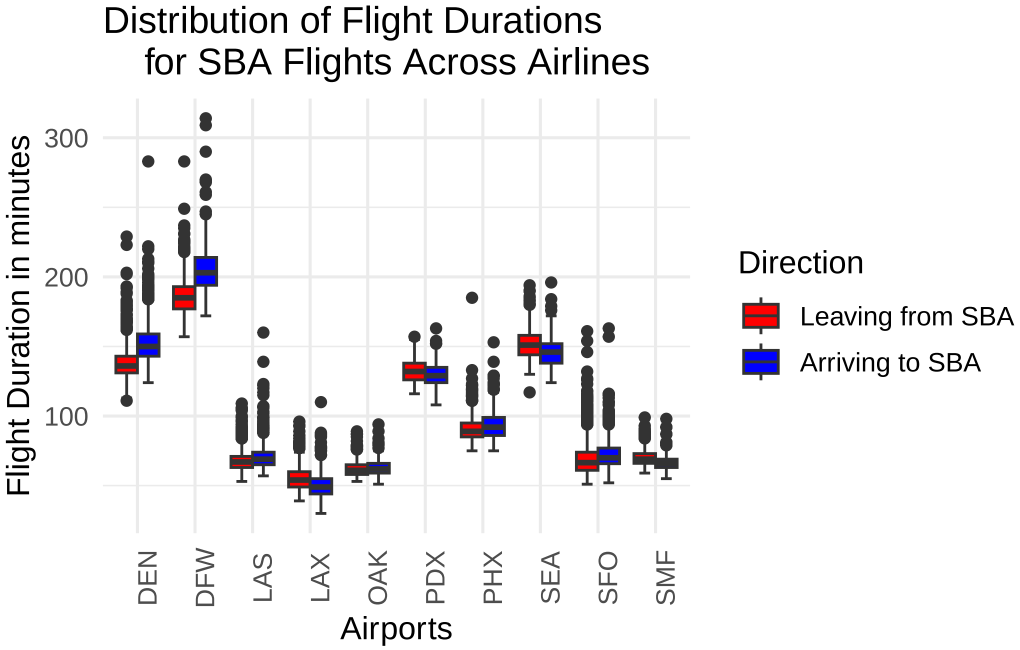
```
`summarise()` has grouped output by 'MONTH'. You can override using the
`.groups` argument.
`geom_line()`: Each group consists of only one observation. i Do you need to
adjust the group aesthetic?
`geom_line()`: Each group consists of only one observation. i Do you need to
adjust the group aesthetic?
```



4 Investigating any differences in the distribution of flight durations to and from SBA

We now see how ling each airline takes to travel to and from SBA. From the graph, we can see that on average, the flight duration between departing from and arriving to SBA is about the same across different airlines. We can also see that airline between SBA an DFW has the longest flight duration compared to others. I believe it's because DFW is located in Texas - that farthest state from California in our sample size.

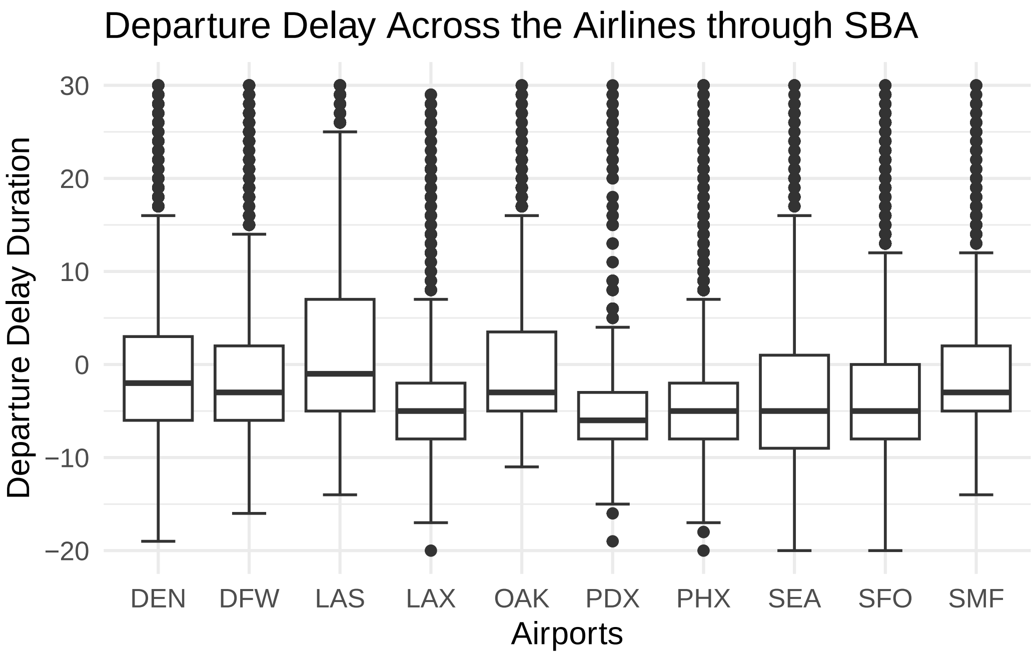
```
Warning: Removed 258 rows containing non-finite outside the scale range
(`stat_boxplot()`).
```



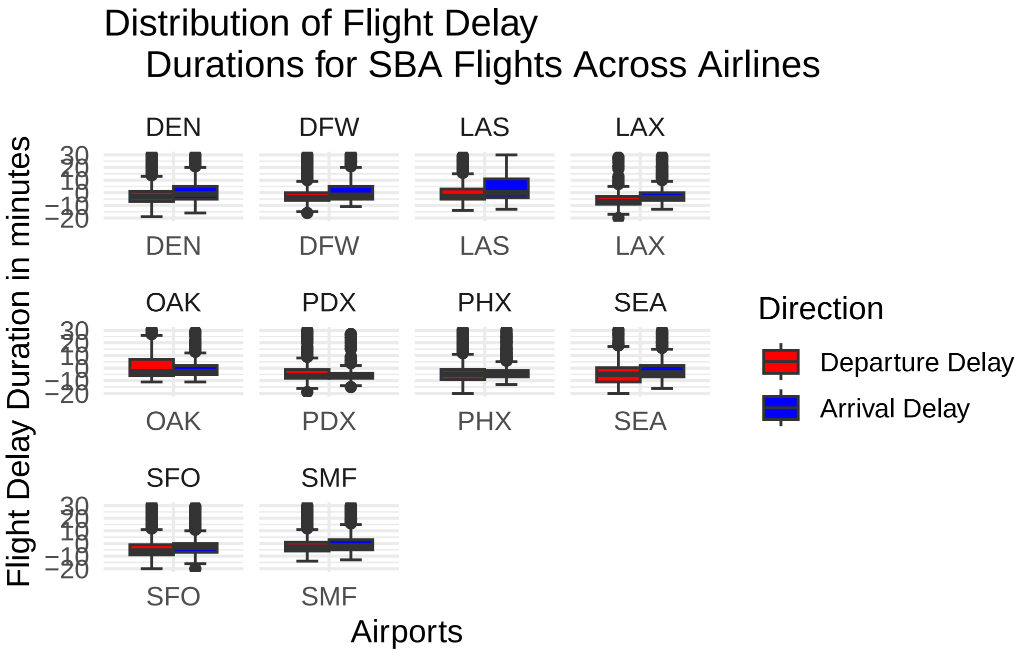
5 Investigating Delays

We now compare the departure - delay across all airlines connected SBA. From the graph, on average, all flights are able to take off earlier.In addition, we can see that LAS has the longest delay time.From the second graph, we can see that on average the arrival and departure delays are about the same across all airlines.

```
Warning: Removed 1918 rows containing non-finite outside the scale range
(`stat_boxplot()`).
```



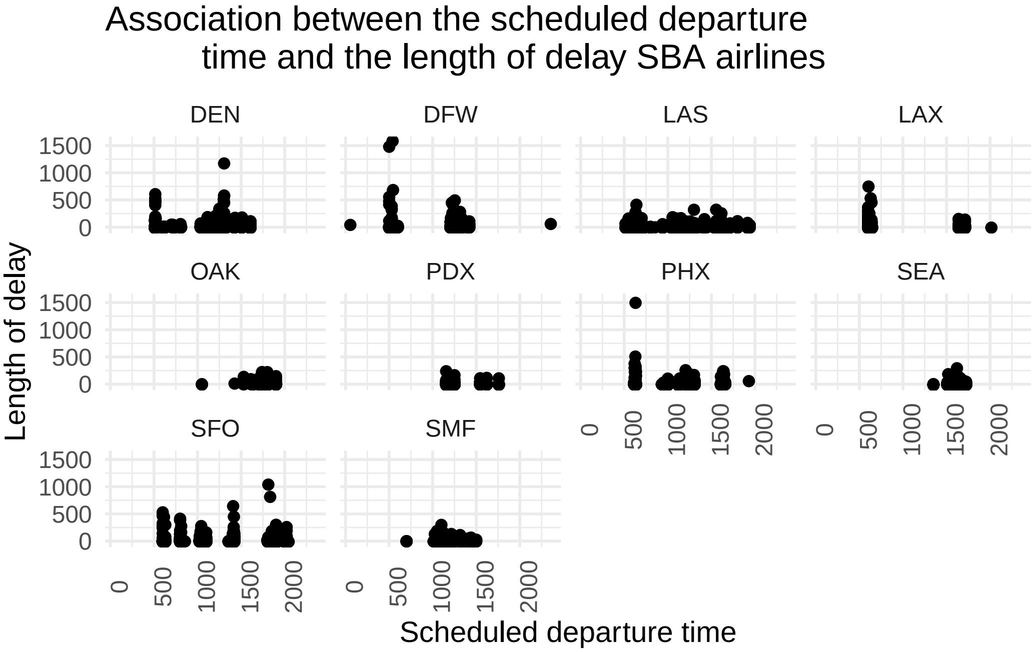
```
Warning: Removed 1918 rows containing non-finite outside the scale range
(`stat_boxplot()`).
```



6 Association between the scheduled departure time and the length of delay for flights departing from SBA

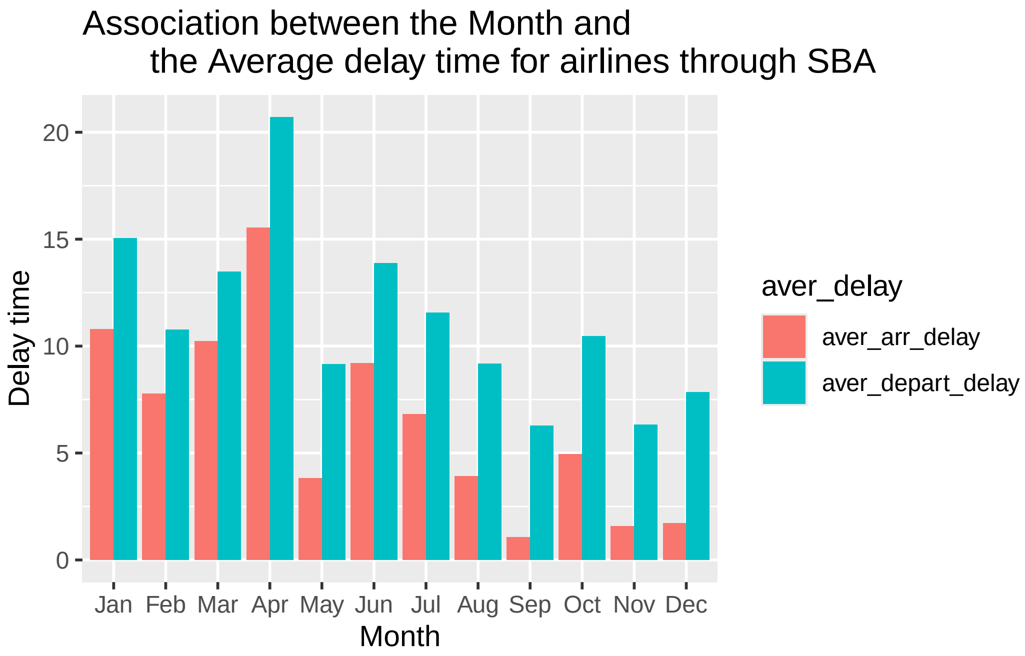
From the graph, we can see that for DFW, LAX, and PHX airlines, the length of delay for flights departing from SBA is at peak around 5 am in the morning. For other airlines, the length of delay is even across the day.

```
Warning: Removed 113 rows containing missing values or values outside the scale range
(`geom_point()`).
```



7 Which months have higher/lower average departure delays and arrival delays

We now investigate which month is the most busy season for airlines connected to SBA by looking for which month has the most delays. From the graph, we can see that most delays for both departure and arrival are clustered around March to April. There are less delays towards the end of the year.

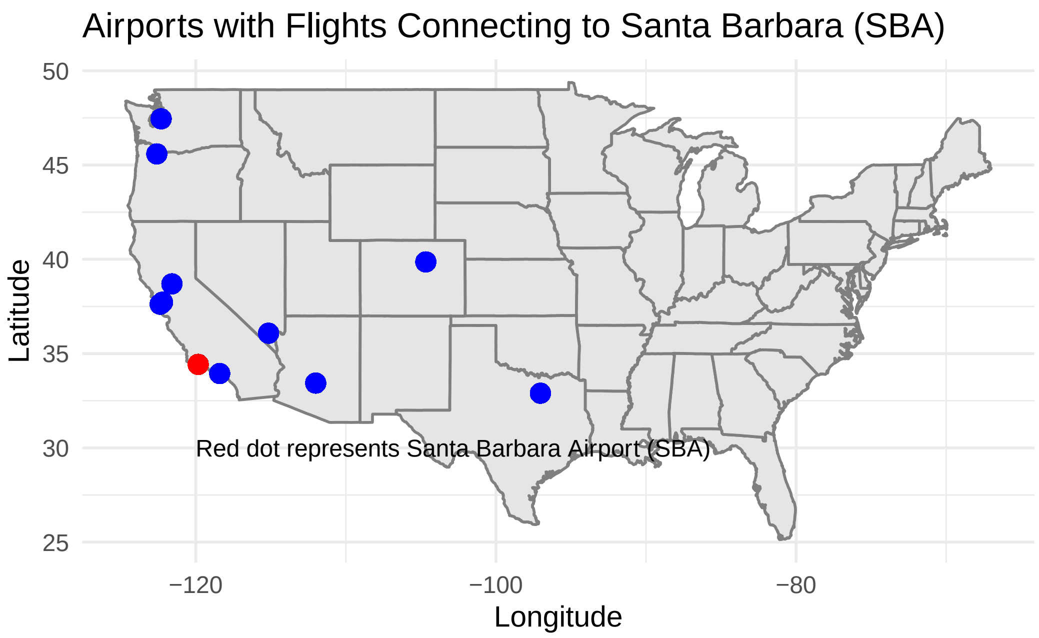


8 Appendix

```
library(tidyverse)
library(readxl)
library(dplyr)
rm(list=ls())

getwd() #get directory
```

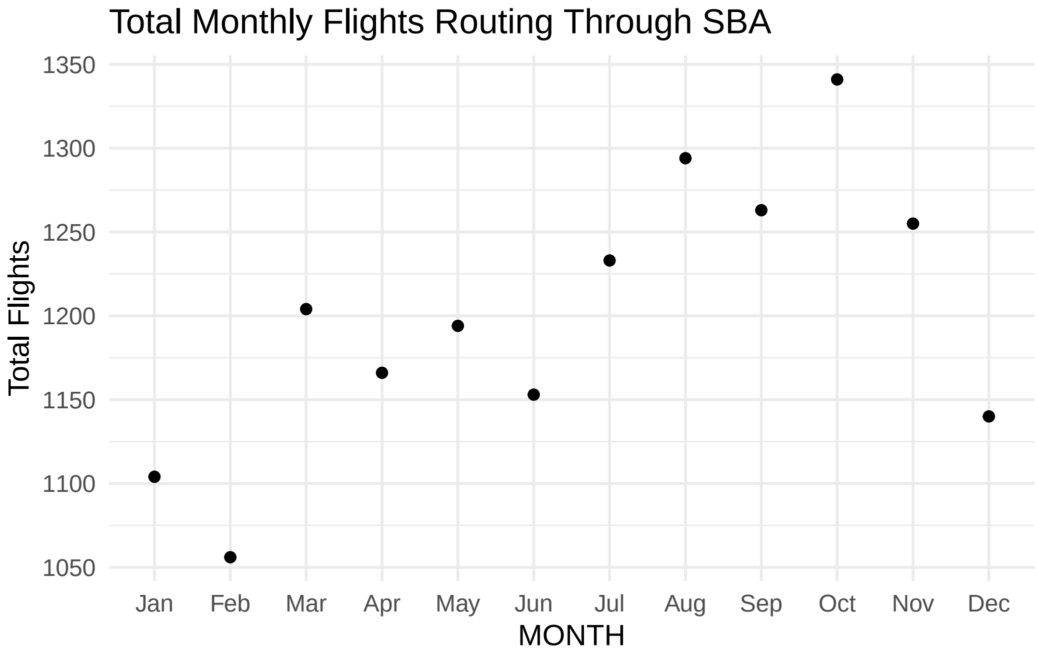
```
[1] "/home/jovyan/100-sp24/Mini_Projects/MP01"
```

```
# Combine all 12 months’ worth of data into a single data frame.
CA_Flights_Month <- list.files("/home/jovyan/100-sp24/Mini_Projects/MP01/data/", patte

CA_Flights_Month_df <- lapply(CA_Flights_Month, read.csv) #load files into data frame

combined_CA_Flights_Month <- bind_rows(CA_Flights_Month_df)

# Import Airport_Info
Airport_Info <- read.csv("data/Airport_Info.csv")

# Join additional 6 columns into Airport_Info
combined_CA_Flights_Month_Airpot_Info <- left_join(
  combined_CA_Flights_Month,
  Airport_Info,
  by = join_by (`ORIGIN`==`ARPT_ID`)
) %>%
  rename(
    "lat_origin" = x,
    "lon_origin" = y
  ) %>%
  left_join(
    Airport_Info,
    by = join_by(`DEST`==`ARPT_ID`)
  ) %>%
  rename(
    "lat_dest" = x,
    "lon_dest" = y
  )

# Convert month to characters
combined_CA_Flights_Month_Airpot_Info$MONTH <- month.abb[as.numeric(combined_CA_Flight


#Santa Barbara Airport Map

# How many connections
sba_connections <- combined_CA_Flights_Month_Airpot_Info %>%
  filter(ORIGIN == "SBA" | DEST == "SBA")

num_airports <- sba_connections %>%
  distinct(ORIGIN, DEST) %>%
  nrow()

airport_names <- sba_connections %>%
  distinct(ORIGIN, DEST) %>%
  pull(ORIGIN, DEST)


states <- map_data("state")

ggplot() +
  geom_polygon(data = states,
               aes(x = long, y = lat, group = group),
               fill = "grey90",
               colour = "grey50") +
  geom_point(data = sba_connections, aes(x = lat_origin, lon_origin), color = "blue",
  geom_point(data = sba_connections[sba_connections$ARPT_NAME.x == "SANTA BARBARA MUNI
  coord_quickmap() +  # Ensure equal aspect ratio
  theme_minimal() +  # Set plot theme
  labs(title = "Airports with Flights Connecting to Santa Barbara (SBA)", x = "Longitu
  annotate("text", x = -120, y = 30, label = "Red dot represents Santa Barbara Airport
```


Airports with Flights Connecting to Santa Barbara (SBA)
50
45
40
35
30
25
Latitude
Red dot represents Santa Barbara Airport (SBA)
−120
−100
−80
Longitude

```
# Exploring Flights
# reorder month column
month_order <- c("Jan", "Feb", "Mar", "Apr", "May", "Jun", "Jul", "Aug", "Sep", "Oct",
sba_connections$MONTH <- factor(sba_connections$MONTH, levels = month_order)
sba_connections <- sba_connections %>%
  select(sort(names(sba_connections)))

# pivot wider so every month has flights
sba_flight_counts <- sba_connections %>%
  group_by(MONTH) %>%
  summarise(total_flights = n())

# a line graph that visualizes the total number of monthly flights that route through
ggplot(sba_flight_counts, aes(x = MONTH, y = total_flights)) +
  geom_line() +
  geom_point() +
  labs(
    x = "MONTH",
    y = "Total Flights",
    title = "Total Monthly Flights Routing Through SBA"
  ) +
  theme_minimal()
```

```
`geom_line()`: Each group consists of only one observation.
i Do you need to adjust the group aesthetic?
```

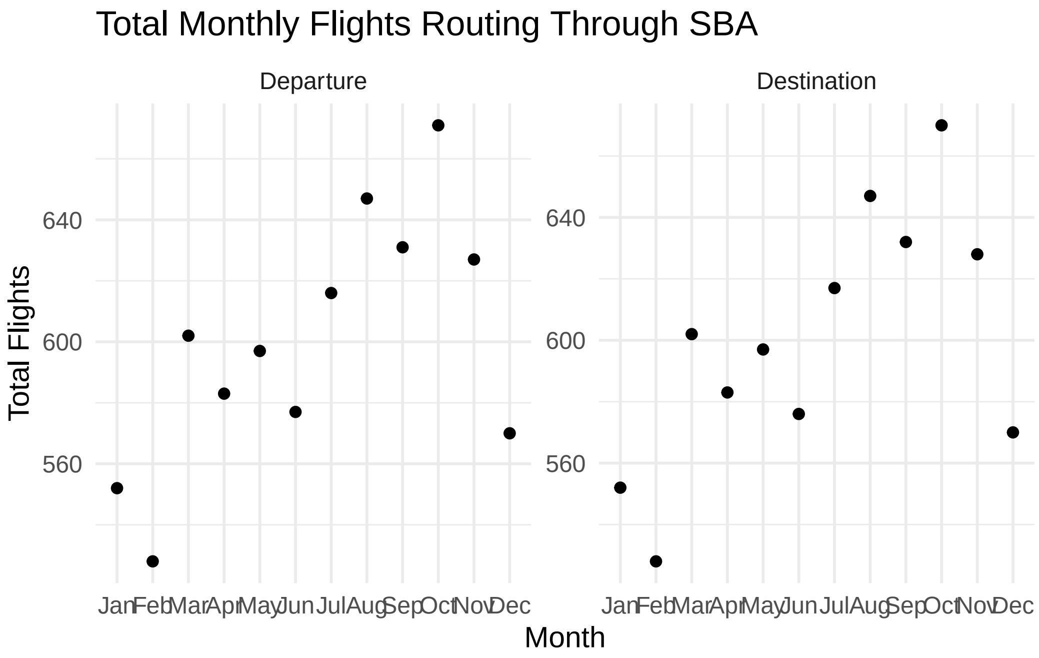


```
# Now, reproduce your graphic from the above step but facet based on whether the fligh
sba_flight_counts <- sba_connections %>%
  mutate(airport_type = ifelse(ORIGIN == "SBA", "Departure", "Destination")) %>%
  group_by(MONTH, airport_type) %>%
  summarise(total_flights = n())
```

```
`summarise()` has grouped output by 'MONTH'. You can override using the
`.groups` argument.
```

```
ggplot(sba_flight_counts, aes(x = MONTH, y = total_flights)) +
  geom_line() +
  geom_point() +
  labs(
    x = "Month",
    y = "Total Flights",
    title = "Total Monthly Flights Routing Through SBA",
  ) +
  facet_wrap(~airport_type, scales = "free_y") +
  theme_minimal()
```

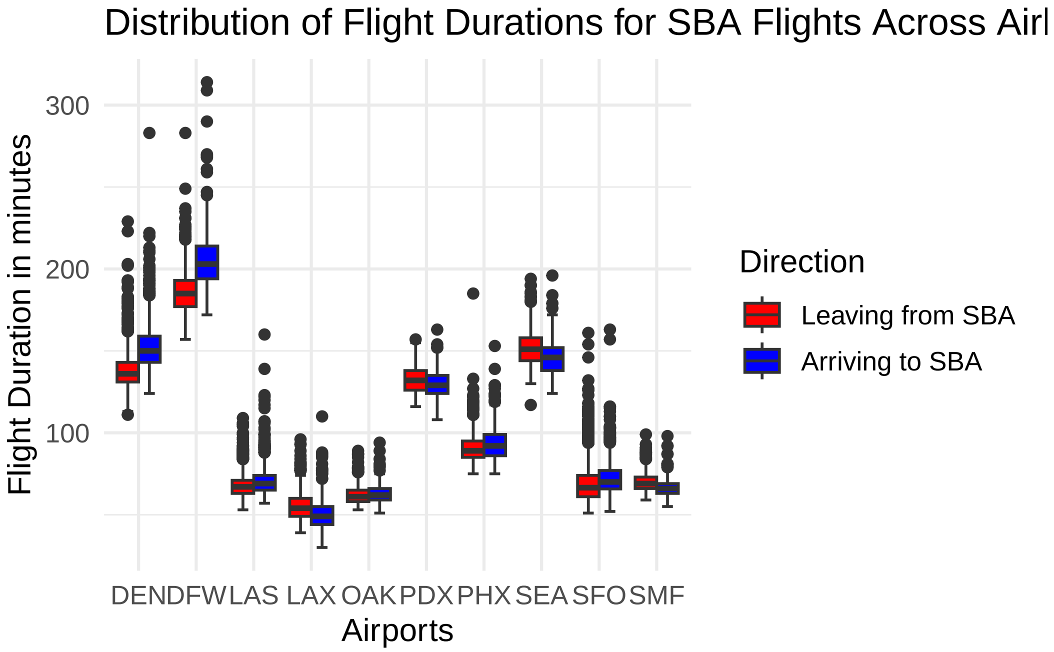
```
`geom_line()`: Each group consists of only one observation.
i Do you need to adjust the group aesthetic?
`geom_line()`: Each group consists of only one observation.
i Do you need to adjust the group aesthetic?
```



```
# Doubly-grouped side-by-side boxplot that displays the distribution of flight duratio

sba_connections_boxplot <- sba_connections %>%
  mutate(Direction = ifelse(ORIGIN == "SBA", "Departure", "Destination")) %>%
  mutate(Airports = ifelse(ORIGIN == "SBA", DEST, ORIGIN) )


#sba_connections_boxplot$ORIGIN <- factor(sba_connections_boxplot$ORIGIN, levels = sor

#sba_connections_boxplot <- sba_connections_boxplot %>%
  #select(`ACTUAL_ELAPSED_TIME`,`Airports`, `Direction`) %>%
  #group_by(Airports)
```

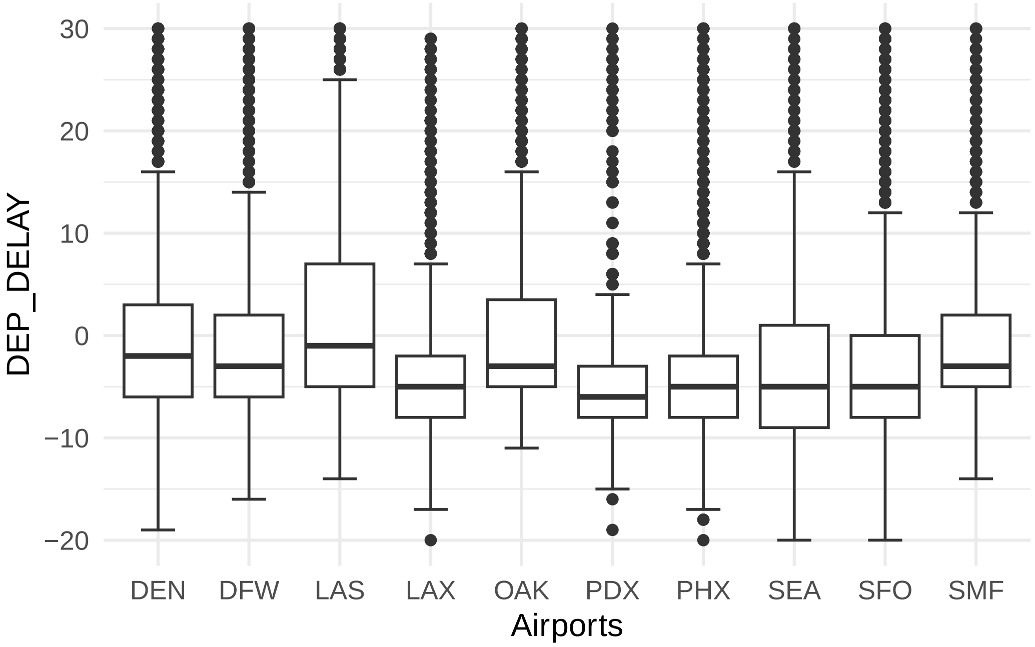
```r
sba_connections_boxplot %>%
  ggplot(aes(x = Airports,
             y = ACTUAL_ELAPSED_TIME,
             fill = Direction
             )) +
  geom_boxplot(staplewidth = 0.5, position = position_dodge()) +
  theme_minimal(base_size = 12) +
  ggtitle(
    "Distribution of Flight Durations for SBA Flights Across Airlines"
  )+
  labs(x = "Airports", y = "Flight Duration in minutes")+
  scale_fill_manual(values = c("Departure" = "red", "Destination" = "blue"),
                    labels = c("Departure" = "Leaving from SBA", "Destination" = "Arri
```

```
Warning: Removed 258 rows containing non-finite outside the scale range
(`stat_boxplot()`).
```



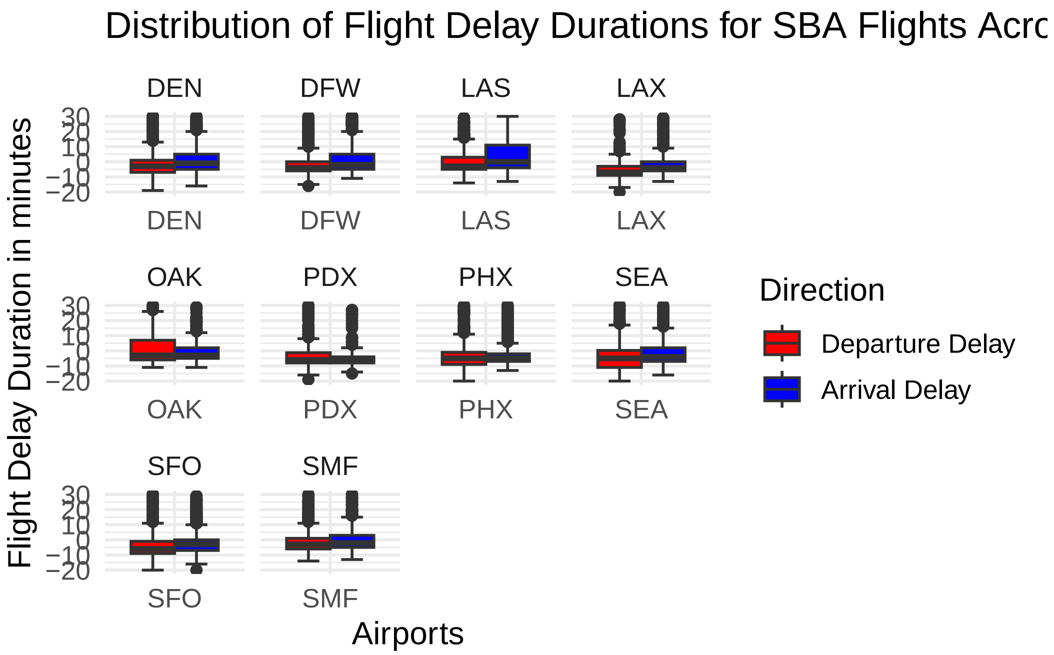
```r
# Investigate Delays
sba_connections_boxplot %>%
  ggplot(aes(x = Airports,
             y = DEP_DELAY))+
  geom_boxplot(staplewidth = 0.5, position = position_dodge()) +
  theme_minimal(base_size = 12) +
  ylim(-20,30)
```

```
Warning: Removed 1918 rows containing non-finite outside the scale range
(`stat_boxplot()`).
```



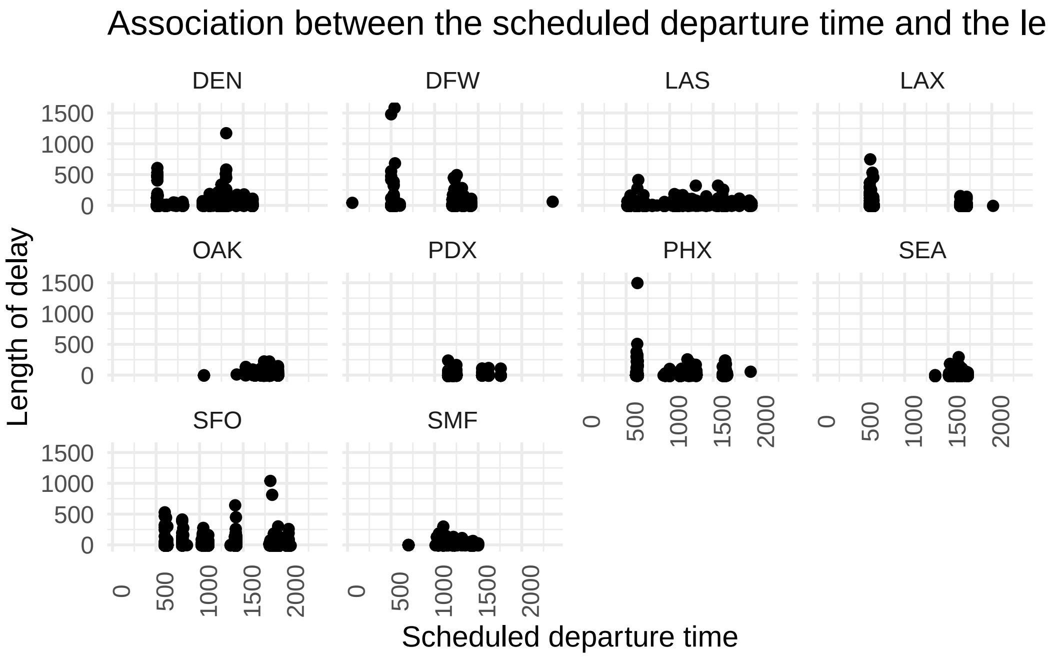
```
# Doubly-Grouped boxplot that displays delay times across airlines
sba_connections_boxplot %>%
  ggplot(aes(x = Airports,
             y = DEP_DELAY,
             fill = Direction
             )) +
  geom_boxplot(staplewidth = 0.5, position = position_dodge()) +
  theme_minimal(base_size = 12) +
   ylim(-20,30)+
  ggtitle(
    "Distribution of Flight Delay Durations for SBA Flights Across Airlines"
  )+
  labs(x = "Airports", y = "Flight Delay Duration in minutes")+
  scale_fill_manual(values = c("Departure" = "red", "Destination" = "blue"),
                    labels = c("Departure" = "Departure Delay", "Destination" = "Arriv
  facet_wrap(~Airports, scales = "free_x")
```

```
Warning: Removed 1918 rows containing non-finite outside the scale range
(`stat_boxplot()`).
```

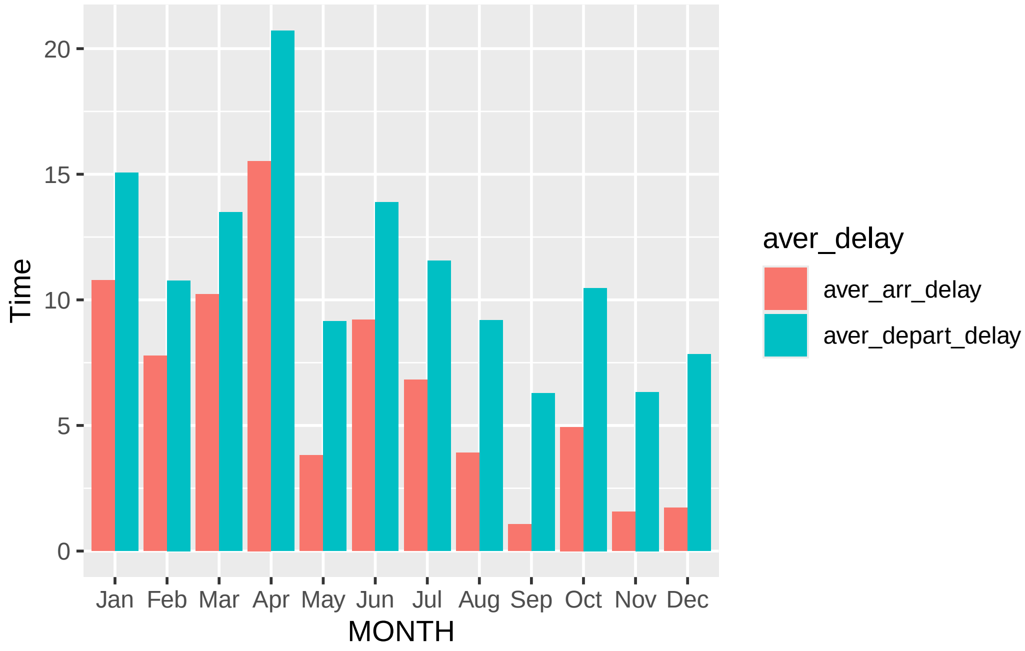


```
# Association between the scheduled departure time and the length of delay
sba_connections_boxplot <- sba_connections %>%
  mutate(Direction = ifelse(ORIGIN == "SBA", "Departure", "Destination")) %>%
  mutate(Airports = ifelse(ORIGIN == "SBA", DEST, ORIGIN) )


sba_connections_boxplot %>%
  filter(Direction == "Departure") %>%
  ggplot(aes(x = CRS_DEP_TIME,
             y = DEP_DELAY))+
  geom_point()+
  ggtitle("Association between the scheduled departure time and the length of delay fo
  labs(x = "Scheduled departure time", y = "Length of delay") +
   theme_minimal()+
  facet_wrap(~Airports)+
  theme(axis.text.x = element_text(angle = 90))
```

```
Warning: Removed 113 rows containing missing values or values outside the scale range
(`geom_point()`).
```



```
# Are there months that have higher/lower average departure delays and Arrival delays?
sba_connections_ave_delays_month <-sba_connections_boxplot %>%
  group_by(MONTH) %>%
  summarise(aver_arr_delay = mean(ARR_DELAY, na.rm = TRUE), aver_depart_delay = mean(D


sba_connections_ave_delays_month %>%
  pivot_longer(cols = c("aver_arr_delay", "aver_depart_delay"), names_to = "aver_delay
  ggplot()+
  geom_col(aes(x=MONTH, y=Time, fill=aver_delay), position = "dodge")
```



```
# fix data frame
#Include 6 columns
#Map
#Flights
#Map branch out
```