STAT3355 Group Project- Group 1

Ved Nigam, Hiba Khan, Monique Picott

2022-11-02

Cleaning and Subsetting Data

```
# Reading in data
nineteenJune_data <- read.csv("airTravel2019June.csv", header = TRUE)
twentyJune_data <- read.csv("airTravel2020June.csv", header = TRUE)

nineteenJuly_data <- read.csv("airTravel2019July.csv", header = TRUE)
twentyJuly_data <- read.csv("airTravel2020July.csv", header = TRUE)

nineteenAug_data <- read.csv("airTravel2019Aug.csv", header = TRUE)
twentyAug_data <- read.csv("airTravel2020Aug.csv", header = TRUE)

# Condensing data by year
nineteen_flights <-
  rbind(nineteenJune_data, nineteenJuly_data, nineteenAug_data)
twenty_flights <-
  rbind(twentyJune_data, twentyJuly_data, twentyAug_data)

# All data considered
total_flights <- rbind(nineteen_flights, twenty_flights)
total_flight_count <- nrow(total_flights)

# Count of all flights considered in the year
nineteen_flights_count <- nrow(nineteen_flights)
twenty_flights_count <- nrow(twenty_flights)

#total number of flights in summer of 2019 and 2020:
print(nineteen_flights_count + twenty_flights_count)
```

```
## [1] 2073540
```

```
# Used to find the airline codes in the dataset
unique(nineteen_flights$OP_UNIQUE_CARRIER)
```

```
##  [1] "DL" "UA" "OH" "YX" "AS" "YV" "F9" "OO" "B6" "NK" "G4" "HA" "EV" "WN" "AA"
## [16] "MQ" "9E"
```

```r
# Subsetting data based on airline and year
# Summer of 2019
Delta_nineteen_data <-
  subset(nineteen_flights, nineteen_flights$OP_UNIQUE_CARRIER == "DL")
Delta_flights_nineteen <- nrow(Delta_nineteen_data)

United_nineteen_data <-
  subset(nineteen_flights, nineteen_flights$OP_UNIQUE_CARRIER == "UA")
United_flights_nineteen <- nrow(United_nineteen_data)

American_nineteen_data <-
  subset(nineteen_flights, nineteen_flights$OP_UNIQUE_CARRIER == "AA")
American_flights_nineteen <- nrow(American_nineteen_data)

Southwest_nineteen_data <-
  subset(nineteen_flights, nineteen_flights$OP_UNIQUE_CARRIER == "WN")
Southwest_flights_nineteen <- nrow(Southwest_nineteen_data)

Spirit_nineteen_data <-
  subset(nineteen_flights, nineteen_flights$OP_UNIQUE_CARRIER == "NK")
Spirit_flights_nineteen <- nrow(Spirit_nineteen_data)

# Summer of 2020
Delta_twenty_data <-
  subset(twenty_flights, twenty_flights$OP_UNIQUE_CARRIER == "DL")
Delta_flights_twenty <- nrow(Delta_twenty_data)

United_twenty_data <-
  subset(twenty_flights, twenty_flights$OP_UNIQUE_CARRIER == "UA")
United_flights_twenty <- nrow(United_twenty_data)

American_twenty_data <-
  subset(twenty_flights, twenty_flights$OP_UNIQUE_CARRIER == "AA")
American_flights_twenty <- nrow(American_twenty_data)

Southwest_twenty_data <-
  subset(twenty_flights, twenty_flights$OP_UNIQUE_CARRIER == "WN")
Southwest_flights_twenty <- nrow(Southwest_twenty_data)

Spirit_twenty_data <-
  subset(twenty_flights, twenty_flights$OP_UNIQUE_CARRIER == "NK")
Spirit_flights_twenty <- nrow(Spirit_twenty_data)

# Data we are considering
flights_considered_nineteen_data <-
  rbind(Delta_nineteen_data, United_nineteen_data, American_nineteen_data,
        Southwest_nineteen_data, Spirit_nineteen_data)

flights_considered_nineteen <-
  c(Delta_flights_nineteen, United_flights_nineteen, American_flights_nineteen,
    Southwest_flights_nineteen, Spirit_flights_nineteen)

flights_considered_twenty_data <-
```

```
  rbind(Delta_twenty_data, United_twenty_data, American_twenty_data,
        Southwest_twenty_data, Spirit_nineteen_data)

flights_considered_twenty <-
  c(Delta_flights_twenty, United_flights_twenty, American_flights_twenty,
    Southwest_flights_twenty, Spirit_flights_twenty)

# Formatting the date
flights_considered_nineteen_data$Date <-
  as.Date(with(flights_considered_nineteen_data,
               paste(2019, MONTH, DAY_OF_MONTH,sep="-")),"%Y-%m-%d")

flights_considered_twenty_data$Date <-
  as.Date(with(flights_considered_twenty_data,
               paste(2020, MONTH, DAY_OF_MONTH, sep = "-")), "%Y-%m-%d")

# airlines we are considering
airlines <- c("Delta", "United", "American", "Southwest", "Spirit")

# Data cleaning for cancellations between summer of 2019 and summer of 2020
nineteen_cancellation_data <-
  subset(flights_considered_nineteen_data,
         flights_considered_nineteen_data$CANCELLED == 1)
nineteen_cancellations <- nrow(nineteen_cancellation_data)

twenty_cancellation_data <-
  subset(flights_considered_twenty_data,
         flights_considered_twenty_data$CANCELLED == 1)
twenty_cancellations <- nrow(twenty_cancellation_data)

Delta_cancelled_19_data <-
  subset(Delta_nineteen_data, Delta_nineteen_data$CANCELLED == 1)
United_cancelled_19_data <-
  subset(United_nineteen_data, United_nineteen_data$CANCELLED == 1)
American_cancelled_19_data <-
  subset(American_nineteen_data, American_nineteen_data$CANCELLED == 1)
Southwest_cancelled_19_data <-
  subset(Southwest_nineteen_data, Southwest_nineteen_data$CANCELLED == 1)
Spirit_cancelled_19_data <-
  subset(Spirit_nineteen_data, Spirit_nineteen_data$CANCELLED == 1)

Delta_cancelled_20_data <-
  subset(Delta_twenty_data, Delta_twenty_data$CANCELLED == 1)
United_cancelled_20_data <-
  subset(United_twenty_data, United_twenty_data$CANCELLED == 1)
American_cancelled_20_data <-
  subset(American_twenty_data, American_twenty_data$CANCELLED == 1)
Southwest_cancelled_20_data <-
  subset(Southwest_twenty_data, Southwest_twenty_data$CANCELLED == 1)
Spirit_cancelled_20_data <-
  subset(Spirit_twenty_data, Spirit_twenty_data$CANCELLED == 1)

cancelled_19 <- c(nrow(Delta_cancelled_19_data),
```

```
                  nrow(United_cancelled_19_data),
                  nrow(American_cancelled_19_data),
                  nrow(Southwest_cancelled_19_data),
                  nrow(Spirit_cancelled_19_data))

cancelled_20 <- c(nrow(Delta_cancelled_20_data),
                  nrow(United_cancelled_20_data),
                  nrow(American_cancelled_20_data),
                  nrow(Southwest_cancelled_20_data),
                  nrow(Spirit_cancelled_20_data))

all_cancelled <- cbind(cancelled_19, cancelled_20)

Delta_19_cancellation_ratio <- (nrow(Delta_cancelled_19_data) /
                                  nrow(Delta_nineteen_data))
Delta_19_cancellation_percent <- (Delta_19_cancellation_ratio * 100)

United_19_cancellation_ratio <- (nrow(United_cancelled_19_data) /
                                   nrow(United_nineteen_data))
United_19_cancellation_percent <- (United_19_cancellation_ratio * 100)

American_19_cancellation_ratio <- (nrow(American_cancelled_19_data) /
                                     nrow(American_nineteen_data))
American_19_cancellation_percent <- (American_19_cancellation_ratio * 100)

Southwest_19_cancellation_ratio <- (nrow(Southwest_cancelled_19_data) /
                                      nrow(Southwest_nineteen_data))
Southwest_19_cancellation_percent <- (Southwest_19_cancellation_ratio * 100)

Spirit_19_cancellation_ratio <- (nrow(Spirit_cancelled_19_data) /
                                   nrow(Spirit_nineteen_data))
Spirit_19_cancellation_percent <- (Spirit_19_cancellation_ratio * 100)


Delta_20_cancellation_ratio <- (nrow(Delta_cancelled_20_data) /
                                  nrow(Delta_twenty_data))
Delta_20_cancellation_percent <- (Delta_20_cancellation_ratio * 100)

United_20_cancellation_ratio <- (nrow(United_cancelled_20_data) /
                                   nrow(United_twenty_data))
United_20_cancellation_percent <- (United_20_cancellation_ratio * 100)

American_20_cancellation_ratio <- (nrow(American_cancelled_20_data) /
                                     nrow(American_twenty_data))
American_20_cancellation_percent <- (American_20_cancellation_ratio * 100)

Southwest_20_cancellation_ratio <- (nrow(Southwest_cancelled_20_data) /
                                      nrow(Southwest_twenty_data))
Southwest_20_cancellation_percent <- (Southwest_20_cancellation_ratio * 100)

Spirit_20_cancellation_ratio <- (nrow(Spirit_cancelled_20_data) /
                                   nrow(Spirit_twenty_data))
Spirit_20_cancellation_percent <- (Spirit_20_cancellation_ratio * 100)
```

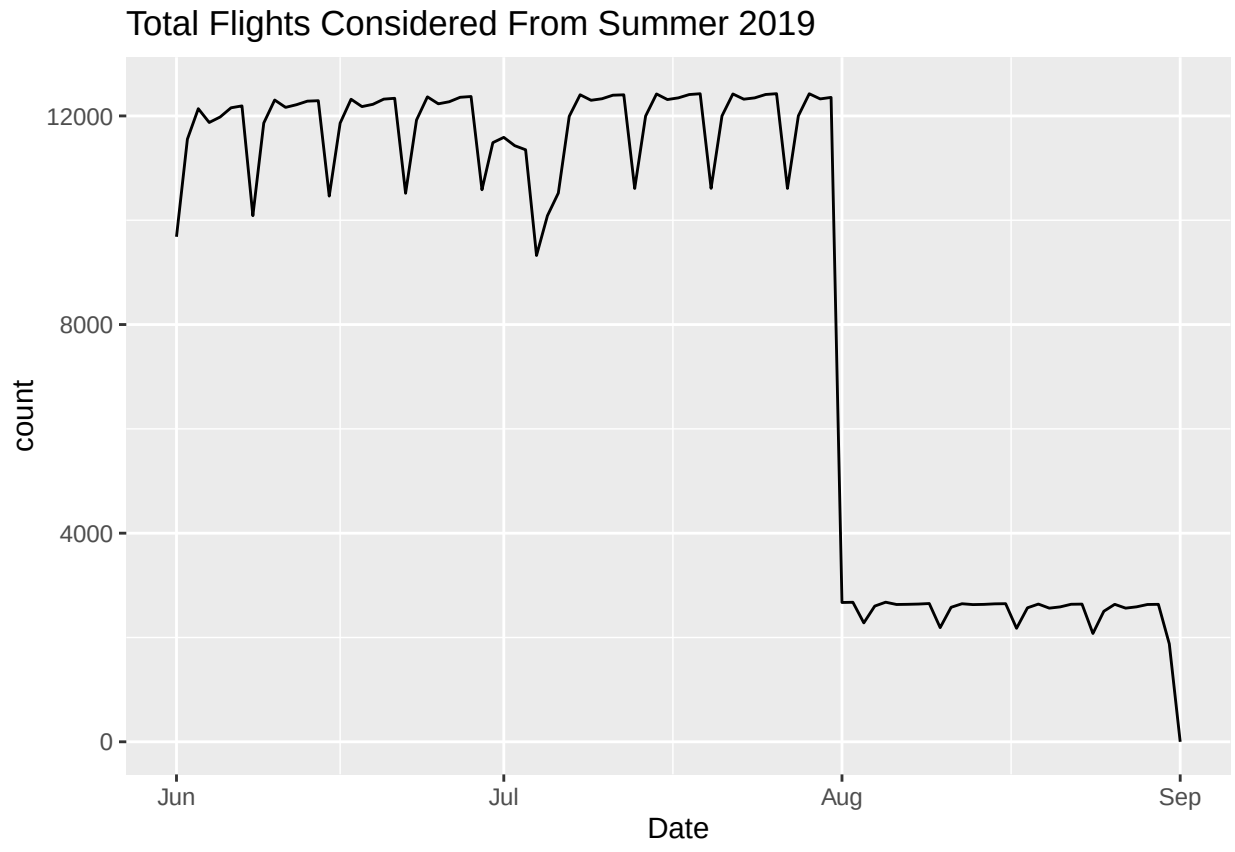
```r
# CODE A - CANCELLED DUE TO CARRIER
cancellation_reason_A_19_data <-
  subset(nineteen_cancellation_data,
         nineteen_cancellation_data$CANCELLATION_CODE == "A" )
cancellation_reason_A_19 <- nrow(cancellation_reason_A_19_data)

# CODE B - CANCELLED DUE TO WEATHER
cancellation_reason_B_19_data <-
  subset(nineteen_cancellation_data,
         nineteen_cancellation_data$CANCELLATION_CODE == "B" )
cancellation_reason_B_19 <- nrow(cancellation_reason_B_19_data)

# CODE C - CANCELLED DUE TO NATIONAL AIR SYSTEM
cancellation_reason_C_19_data <-
subset(nineteen_cancellation_data,
      nineteen_cancellation_data$CANCELLATION_CODE == "C" )
cancellation_reason_C_19 <- nrow(cancellation_reason_C_19_data)

# CODE D - CANCELLED DUE TO SECUIRTY ISSUES
cancellation_reason_D_19_data <-
  subset(nineteen_cancellation_data,
         nineteen_cancellation_data$CANCELLATION_CODE == "D" )
cancellation_reason_D_19 <- nrow(cancellation_reason_D_19_data)


# CREATING SUBSETS BASED ON REASON FOR CANCELLED (CANCELLATION CODE) 2020
# unique(Delta_cancelled_20_data$CANCELLATION_CODE)

# CODE A - CANCELLED DUE TO CARRIER
cancellation_reason_A_20_data <-
  subset(twenty_cancellation_data,
         twenty_cancellation_data$CANCELLATION_CODE == "A" )
cancellation_reason_A_20 <- nrow(cancellation_reason_A_20_data)

# CODE B - CANCELLED DUE TO WEATHER
cancellation_reason_B_20_data <-
  subset(twenty_cancellation_data,
         twenty_cancellation_data$CANCELLATION_CODE == "B" )
cancellation_reason_B_20 <- nrow(cancellation_reason_B_20_data)

# CODE C - CANCELLED DUE TO NATIONAL AIR SYSTEM
cancellation_reason_C_20_data <-
  subset(twenty_cancellation_data,
         twenty_cancellation_data$CANCELLATION_CODE == "C" )
cancellation_reason_C_20 <- nrow(cancellation_reason_C_20_data)

# CODE D - CANCELLED DUE TO SECUIRTY ISSUES
cancellation_reason_D_20_data <-
  subset(twenty_cancellation_data,
         twenty_cancellation_data$CANCELLATION_CODE == "D" )
cancellation_reason_D_20 <- nrow(cancellation_reason_D_20_data)
```

```r
# DATA TO BE CONSIDERED FOR CANCELLATIONS
cancellation_considered_nineteen_data <-
  rbind(cancellation_reason_A_19_data,
        cancellation_reason_B_19_data,
        cancellation_reason_C_19_data,
        cancellation_reason_D_19_data)

cancellation_considered_nineteen_data$Date <-
  as.Date(with(cancellation_considered_nineteen_data,
               paste(2019, MONTH, DAY_OF_MONTH,sep="-")), "%Y-%m-%d")

cancellation_considered_nineteen <-
  c(cancellation_reason_A_19,
    cancellation_reason_B_19,
    cancellation_reason_C_19,
    cancellation_reason_D_19)

cancellation_considered_twenty_data <-
  rbind(cancellation_reason_A_20_data,
        cancellation_reason_B_20_data,
        cancellation_reason_C_20_data,
        cancellation_reason_D_20_data)
cancellation_considered_twenty_data$Date <-
  as.Date(with(cancellation_considered_twenty_data,
               paste(2020, MONTH, DAY_OF_MONTH,sep="-")), "%Y-%m-%d")

cancellation_considered_twenty <-
  c(cancellation_reason_A_20,
    cancellation_reason_B_20,
    cancellation_reason_C_20,
    cancellation_reason_D_20)

# REASONS FOR CANCELLATION NAMED
cancellation_reasons <-
  c("Carrier Issues",
    "Weather Risk",
    "National Air System Issue",
    "Security Breach")

popular_city_dep_19 <-
  subset(flights_considered_nineteen_data,
         flights_considered_nineteen_data$ORIGIN_CITY_NAME ==
           "Chicago, IL" |
           flights_considered_nineteen_data$ORIGIN_CITY_NAME ==
           "Dallas/Fort Worth, TX" |
           flights_considered_nineteen_data$ORIGIN_CITY_NAME ==
           "Los Angeles, CA" |
           flights_considered_nineteen_data$ORIGIN_CITY_NAME ==
           "New York, NY")

popular_city_dep_20 <-
  subset(flights_considered_twenty_data,
         flights_considered_twenty_data$ORIGIN_CITY_NAME ==
```

```
  "Chicago, IL" |
          flights_considered_twenty_data$ORIGIN_CITY_NAME ==
  "Dallas/Fort Worth, TX" |
          flights_considered_twenty_data$ORIGIN_CITY_NAME ==
  "Los Angeles, CA" |
          flights_considered_twenty_data$ORIGIN_CITY_NAME ==
  "New York, NY")
```

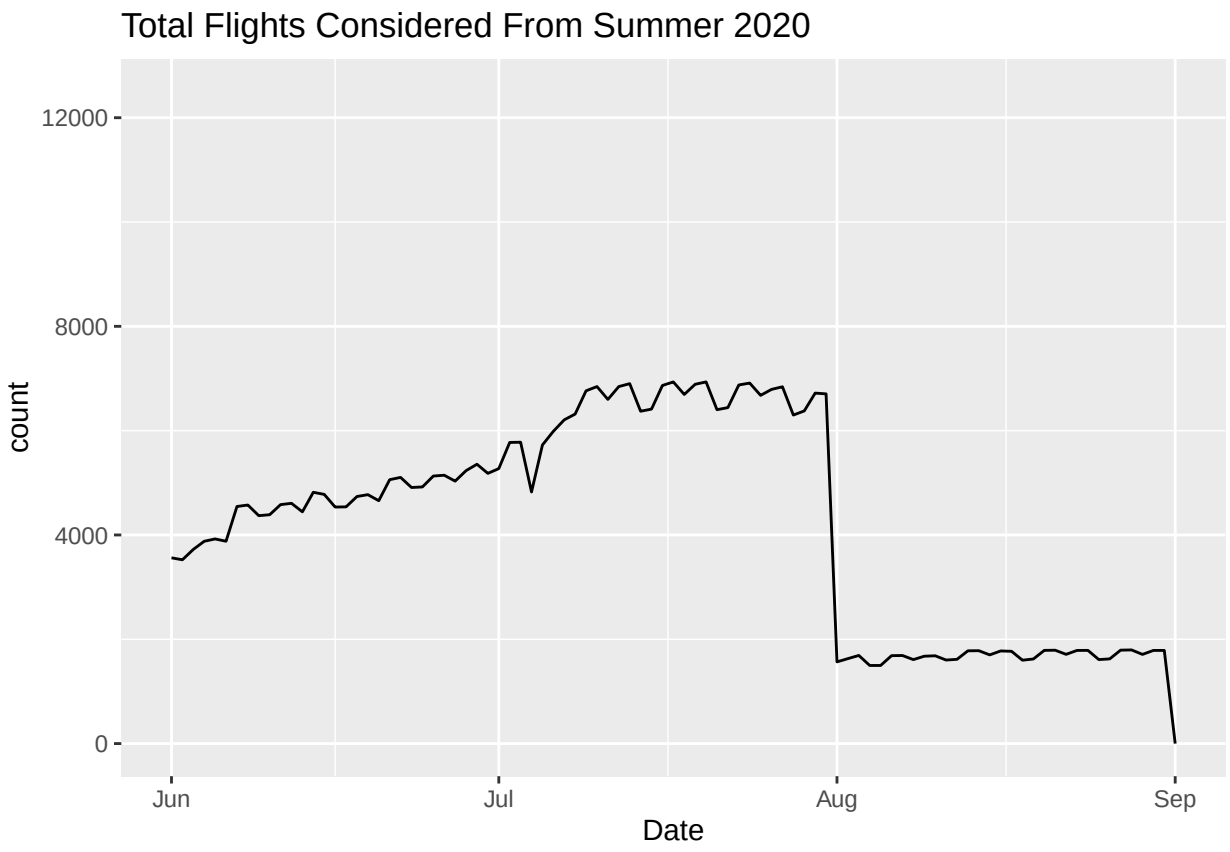
Code with all the plots

```
library(ggplot2)

# ggplot line graph of total flights considered 2019
nineteen_line_gg <- ggplot(flights_considered_nineteen_data) +
  geom_line(aes(x = Date, y = ..count..), stat = "bin", binwidth = 1) +
  scale_x_date(limits = as.Date(c("2019-06-01", "2019-09-01"))) +
  ylim(0, 12500) +
  ggtitle("Total Flights Considered From Summer 2019")
(nineteen_line_gg)
```

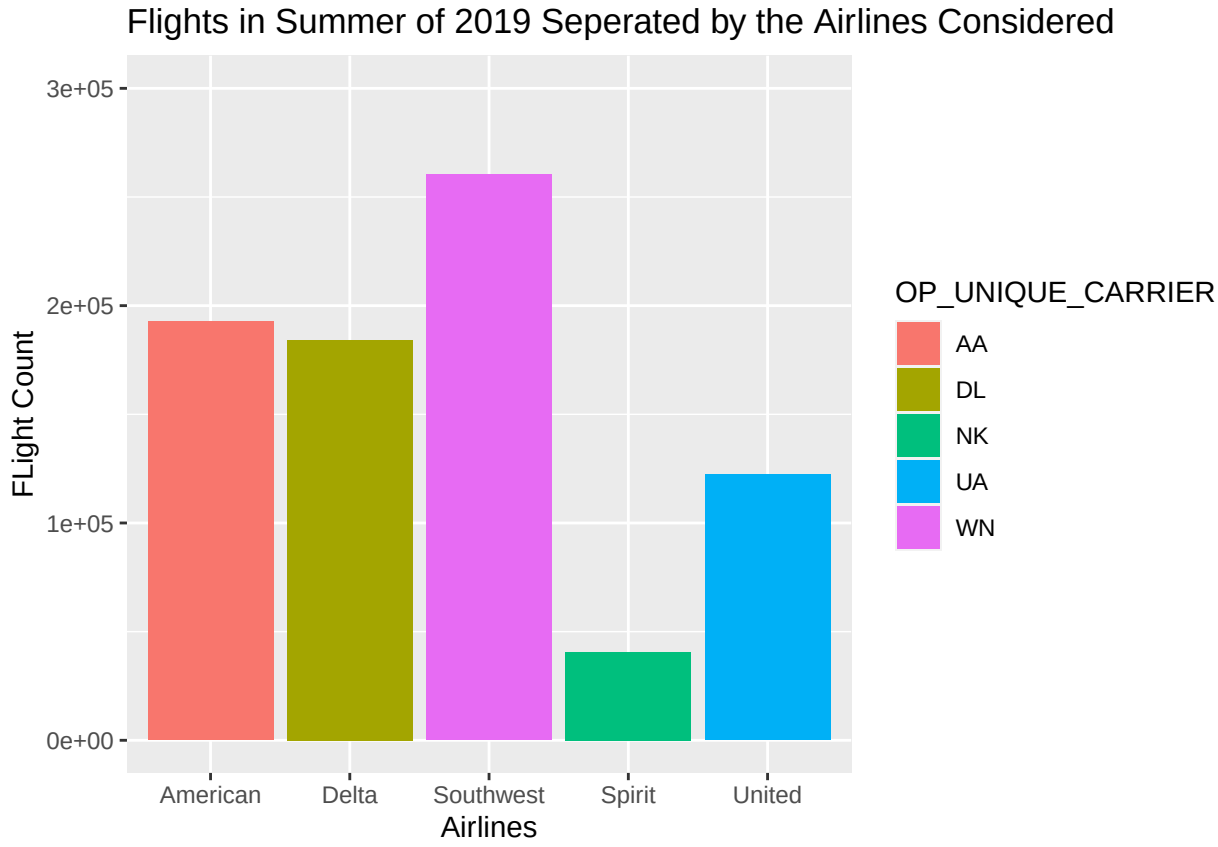



```
# ggplot line graph of total flights considered 2020
twenty_line_gg <- ggplot(flights_considered_twenty_data) +
```

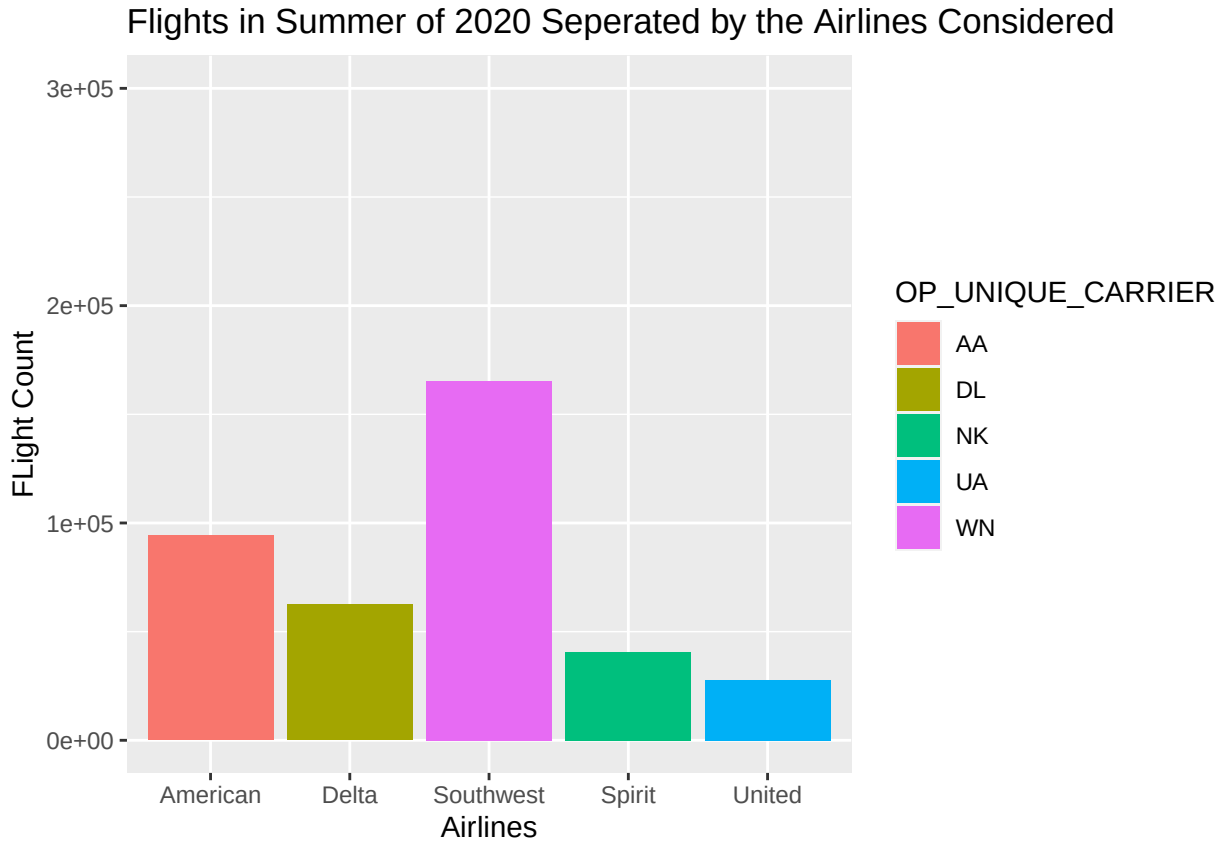
```
  geom_line(aes(x = Date, y = ..count..), stat = "bin", binwidth = 1) +
  scale_x_date(limits = as.Date(c("2020-06-01", "2020-09-01"))) +
  ylim(0,12500) +
  ggtitle("Total Flights Considered From Summer 2020")
(twenty_line_gg)
```



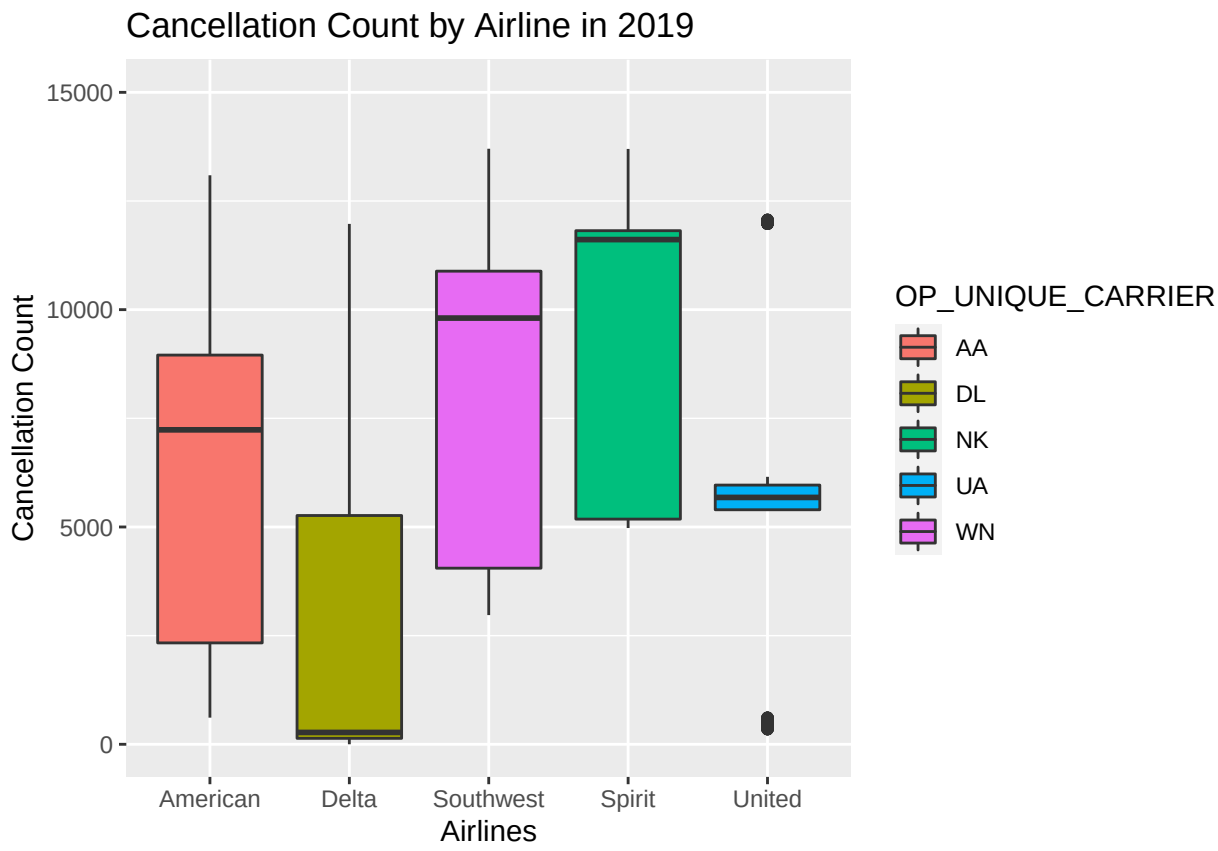

```
# ggplot bar graph of total flights in 2019 for the airlines we are considering
nineteen_bar_gg <- ggplot(flights_considered_nineteen_data) +
  geom_bar(aes(x = factor(OP_UNIQUE_CARRIER,
                          levels = c('AA', 'DL', 'WN', 'NK', 'UA')),
               y = ..count..,
               fill = OP_UNIQUE_CARRIER)) +
  scale_x_discrete(labels = c("AA" = "American",
                              "DL" = "Delta",
                              "WN" = "Southwest",
                              "NK" = "Spirit",
                              "UA" = "United")) +
  ylim(0, 300000) +
  xlab("Airlines") +
  ylab("FLight Count") +
  ggtitle("Flights in Summer of 2019 Seperated by the Airlines Considered")
(nineteen_bar_gg)
```



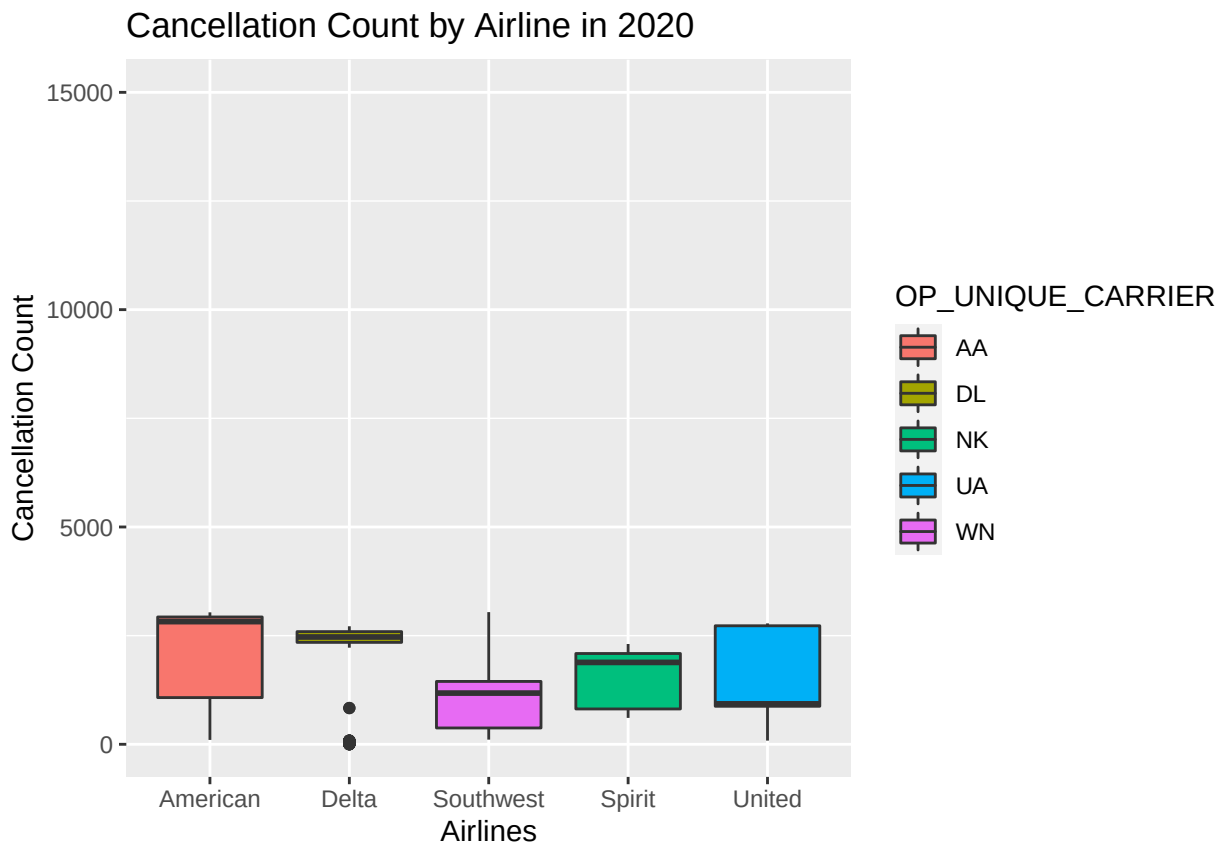

```
# ggplot bar graph of total flights in 2020 for the airlines we are considering
twenty_bar_gg <- ggplot(flights_considered_twenty_data) +
  geom_bar(aes(x = factor(OP_UNIQUE_CARRIER,
                          levels = c('AA', 'DL', 'WN', 'NK', 'UA')),
               y = ..count..,
               fill = OP_UNIQUE_CARRIER)) +
  scale_x_discrete(labels = c("AA" = "American",
                              "DL" = "Delta",
                              "WN" = "Southwest",
                              "NK" = "Spirit",
                              "UA" = "United")) +
  ylim(0, 300000) +
  xlab("Airlines") +
  ylab("FLight Count") +
  ggtitle("Flights in Summer of 2020 Seperated by the Airlines Considered") +
  labs(color = "Airlines")
twenty_bar_gg
```



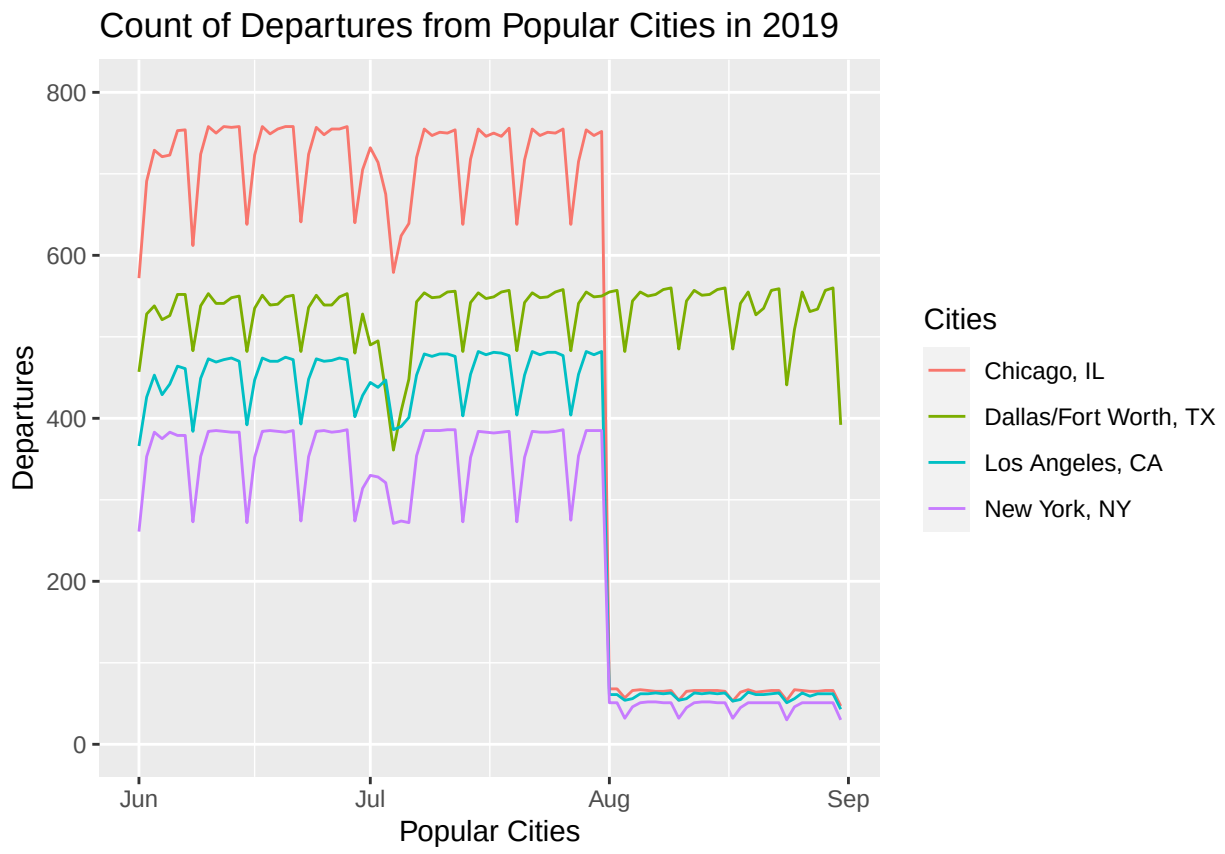

```
# ggplot boxplot of flight cancellations by airline
cancellation_19_box <- ggplot(data = cancellation_considered_nineteen_data) +
  geom_boxplot(aes(x = factor(OP_UNIQUE_CARRIER,
                              levels = c('AA', 'DL', 'WN', 'NK', 'UA')),
                   y = which(CANCELLED == 1),
                   fill = OP_UNIQUE_CARRIER)) +
  scale_x_discrete(labels = c("AA" = "American",
                              "DL" = "Delta",
                              "WN" = "Southwest",
                              "NK" = "Spirit",
                              "UA" = "United")) +
  ylim(0, 15000) +
  xlab("Airlines") +
  ylab("Cancellation Count") +
  ggtitle("Cancellation Count by Airline in 2019") +
  labs(color = "Airlines")
cancellation_19_box
```



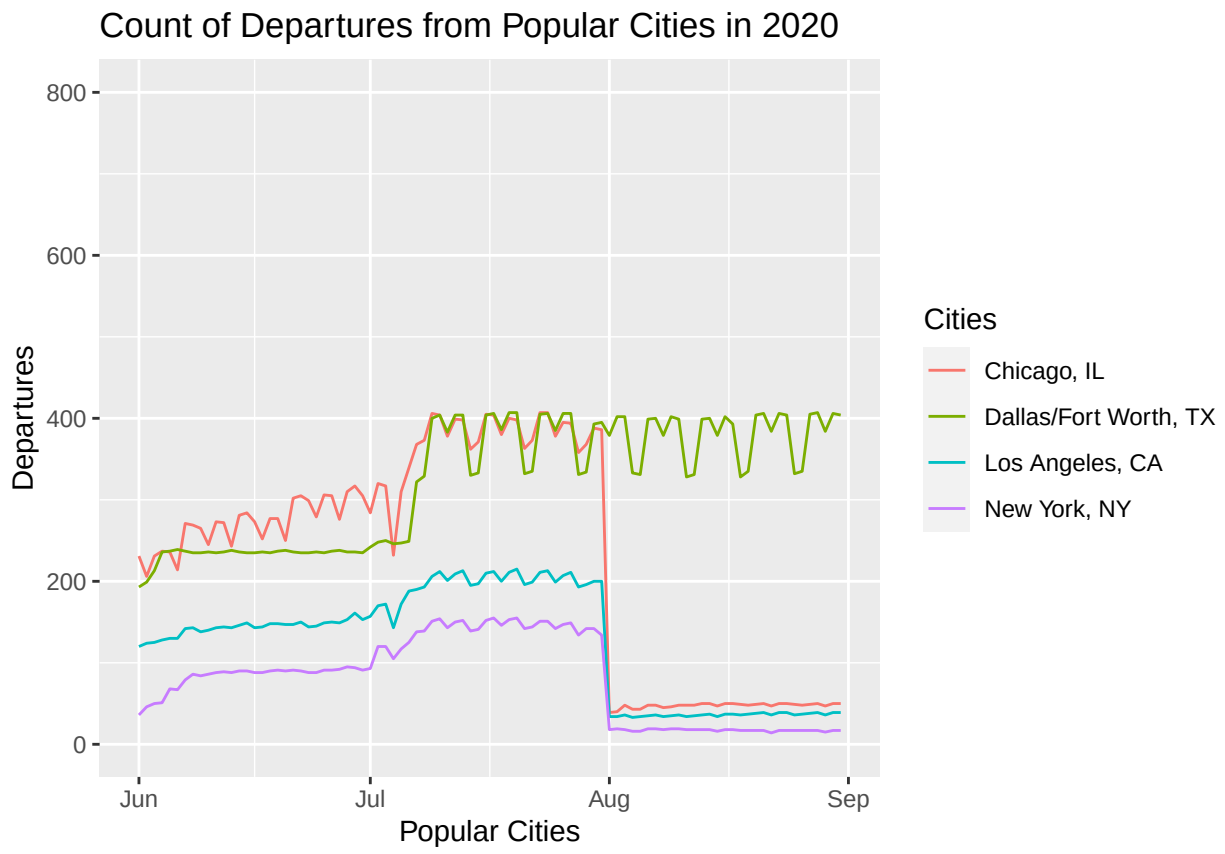

```
cancellation_20_box <- ggplot(data = cancellation_considered_twenty_data) +
  geom_boxplot(aes(x = factor(OP_UNIQUE_CARRIER,
                              levels = c('AA', 'DL', 'WN', 'NK', 'UA')),
                   y = which(CANCELLED == 1),
                   fill = OP_UNIQUE_CARRIER)) +
  scale_x_discrete(labels = c("AA" = "American",
                              "DL" = "Delta",
                              "WN" = "Southwest",
                              "NK" = "Spirit",
                              "UA" = "United")) +
  ylim(0, 15000) +
  xlab("Airlines") +
  ylab("Cancellation Count") +
  ggtitle("Cancellation Count by Airline in 2020")
cancellation_20_box
```



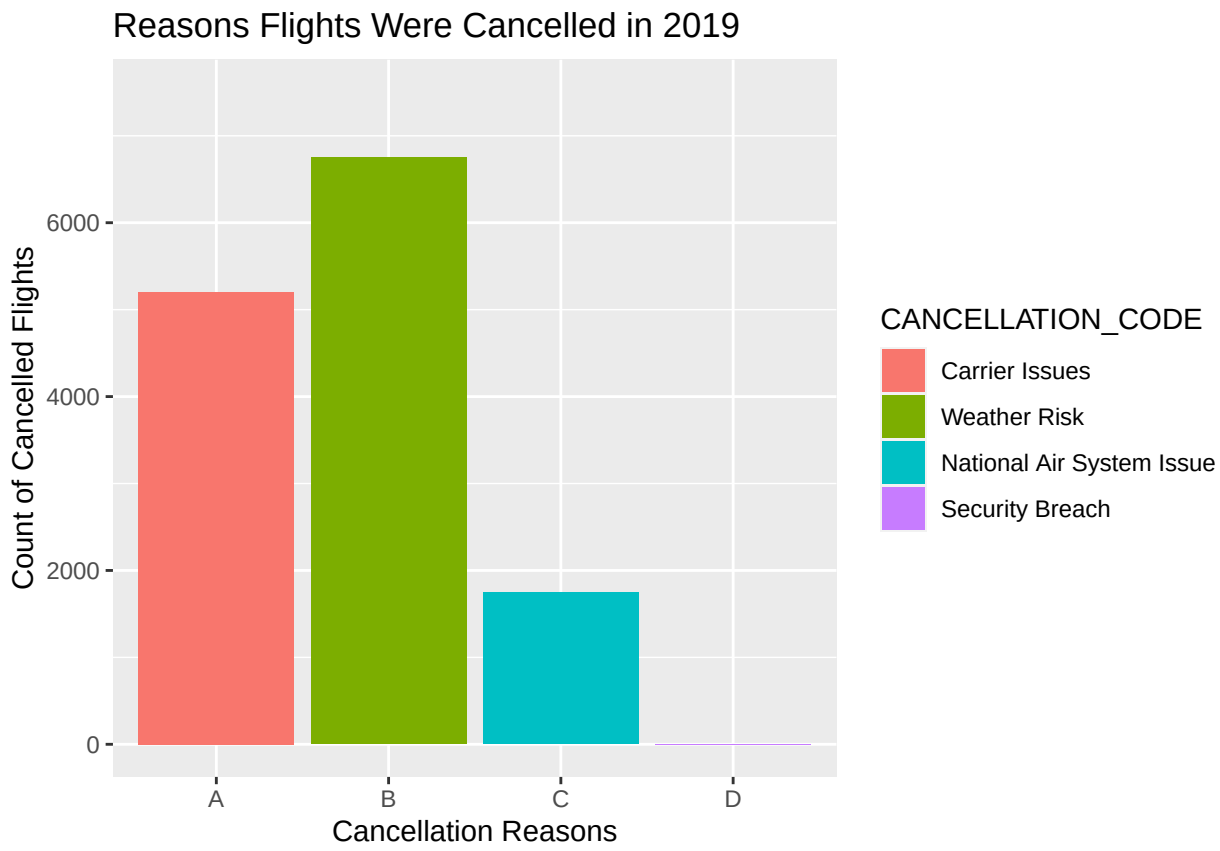

```
# ggplot line plot of departures from popular cities
departures_19_line <- ggplot(popular_city_dep_19) +
  geom_line(aes(x = Date,
                y = ..count..,
                color = as.factor(ORIGIN_CITY_NAME)),
            stat = "bin",
            binwidth = 1) +
  xlab("Popular Cities") +
  ylab("Departures") +
  ylim(0, 800) +
  ggtitle("Count of Departures from Popular Cities in 2019") +
  labs(color = "Cities")
departures_19_line
```



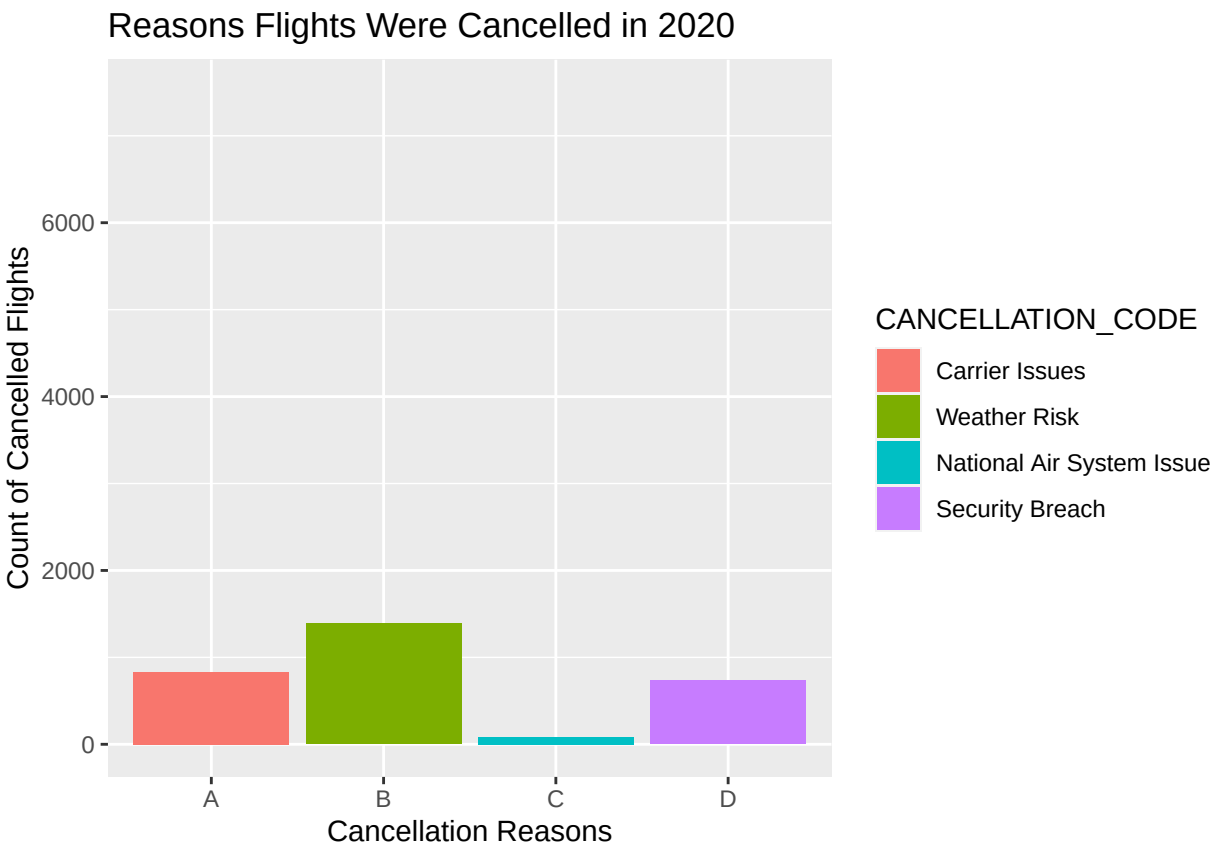

```
departures_20_line <- ggplot(popular_city_dep_20) +
  geom_line(aes(x = Date,
                y = ..count..,
                color = as.factor(ORIGIN_CITY_NAME)),
            stat = "bin",
            binwidth = 1) +
  xlab("Popular Cities") +
  ylab("Departures") +
  ylim(0, 800) +
  ggtitle("Count of Departures from Popular Cities in 2020") +
  labs(color = "Cities")
departures_20_line
```




```
# ggplot line graph of the reasons for cancellations
cancellation19_bar_gg <- ggplot(data = cancellation_considered_nineteen_data) +
  geom_bar(aes(x = factor(CANCELLATION_CODE,
                          levels = c('A', 'B', 'C', 'D')),
               y = ..count..,
              fill = CANCELLATION_CODE)) +
  xlab("Cancellation Reasons") +
  ylab("Count of Cancelled Flights") +
  labs(color = "Cancellation Code") +
  scale_fill_discrete(labels = c("Carrier Issues",
                                 "Weather Risk",
                                 "National Air System Issue",
                                 "Security Breach")) +
  ggtitle("Reasons Flights Were Cancelled in 2019") +
  ylim(0, 7500)
cancellation19_bar_gg
```




```
cancellation20_bar_gg <- ggplot(data = cancellation_considered_twenty_data) +
  geom_bar(aes(x = factor(CANCELLATION_CODE,
                          levels = c('A', 'B', 'C', 'D')),
               y = ..count..,
              fill = CANCELLATION_CODE)) +
  xlab("Cancellation Reasons") +
  ylab("Count of Cancelled Flights") +
  labs(color = "Cancellation Code") +
  scale_fill_discrete(labels = c("Carrier Issues",
                                 "Weather Risk",
                                 "National Air System Issue",
                                 "Security Breach")) +
  ggtitle("Reasons Flights Were Cancelled in 2020") +
  ylim(0, 7500)
cancellation20_bar_gg
```


Reasons Flights Were Cancelled in 2020
Count of Cancelled Flights
6000
4000
2000
0
A
B
C
D
Cancellation Reasons
CANCELLATION_CODE
Carrier Issues
Weather Risk
National Air System Issue
Security Breach