Final Project

Aarti Ramani

2023-02-28

Project Topic: Airlines On-Time Performance, Delays, and Cancellations

Introduction:

Airline cancellations or delays are one of the major causes for passenger inconvenience. With the publicly available dataset (huge datasets with around 16 million flights flown annually), using datascience I am hoping to gain meaningful insights into the best performing airlines and understanding the causes for delays and cancellations across different airline carriers.

For the final project I would like to analyze airline data to identify different factors and their effects on a carrier's performance. As a performance measure, we would be exploring on-time arrivals, number of cancellations by carrier and also explore different reasons for a carrier delay. Data Science can help identify the major causes of delay and cancellations per carrier. Based on the outcome, carriers can take necessary actions to focus on the problem areas.

Problem statement addressed:

This study would benefit airlines by comparing airline performances and predicting possibilities of delay based on aircraft/origin/destination and apply corrective measures to reduce cancellations and delays and to improve on-time performance.

Research Questions

Following are the topics I would like to focus on as part of this project.

1. Are small carriers reliable in terms of lesser cancellations and delays?
2. Are the delays seasonal? If yes, which regions are most affected?
3. Does the time of day have any significance on delays?
4. Which carrier has the best on-time performance.
5. Which carrier has the least on-time performance.
6. Identifying the most common cancellation reason for all carriers.
7. Which carrier has the most number of cancellations.
8. Which carrier has the most number of delays.
9. What is the percentage of delays by reason.

Approach:

I will be performing the following steps:

1. Data analysis - Gathering and understanding different datasets.
2. Data Cleaning and Transforming

3. Merge transformed/cleansed datasets
4. Data visualization/plotting

Addressing the problem

Based on the outcomes from data analysis and visualization, I would like to identify the following:

- Which carriers are more likely to cause delays or cancellations.
- Which carriers are more reliable in terms of on-time performance.

Datasets

Below data submitted by major carriers to department of transportation (DOT).

- Flights.csv
- UniqueCarriers.csv
- Airports.csv

Data was collected by DOT's Bureau of Transportation Statistics for the year 2022. The purpose of this data is to analyze airline on-time performance reported by carriers. The datasets has around 40 fields in total of which I will be considering between 15 to 25 columns for analysis.

Datasets and Relationships:

TABLE: **Flights.csv**

Column Name	Data Type	Column Description
Year	Integer	Year of extracted flight data
Quarter	Integer	Quarter
Month	Integer	Month of extracted flight data
DayofMonth	Integer	Day of month
DayOfWeek	Integer	Day of Week
FlightDate	Date	Flight Date
Marketing_Airline_Network	Character	Marketing Carrier Airline Code
Flight_Number_Marketing_Airline	Integer	Marketing Carrier Flight Number
Operating_Airline	Character	Operating Carrier Airline Code
Tail_Number	Integer	Operating Carrier Tail Number
Flight_Number_Operating_Airline	Integer	Operating Carrier Flight Number
Origin	Character	Origin Airport Code(Airports.csv)
OriginCityName	Character	Origin Airport City Name
OriginState	Character	Origin Airport State Code

OriginStateName	Character	Origin Airport State Name
OriginWac	Integer	Origin Airport Worlde Area Code
Dest	Character	Destination Airport Code(Airports.csv)
DestCityName	Character	Destination Airport City Name
DestState	Character	Destination Airport State Code
DestStateName	Character	Destination Airport State Name
DestWac	Integer	Destination Airport Worlde Area Code
CRSDepTime	Integer	CRS Departure Time (local time: hhmm)
DepTime	Integer	Actual Departure Time(local time: hhmm)
DepDelay	Integer	Difference in minutes between scheduled and actual departure time. Early departures show negative numbers.
DepDelayMinutes	Integer	Difference in minutes between scheduled and actual departure time. Early departures set to 0
DepDel15	Integer	Departure Delay Indicator, 15 Minutes or More (1=Yes)
TaxiOut	Integer	Taxi Out Time, in Minutes
WheelsOff	Integer	Wheels Off Time (local time: hhmm)
WheelsOn	Integer	Wheels On Time (local time: hhmm)
TaxiIn	Integer	Taxi In Time, in Minutes
CRSArrTime	Integer	CRS Arrival Time (local time: hhmm)
ArrTime	Integer	Actual Arrival Time (local time: hhmm)
ArrDelay	Integer	Difference in minutes between scheduled and actual arrival time. Early arrivals show negative numbers.
ArrDelayMinutes	Integer	Difference in minutes between scheduled and actual arrival time. Early arrivals set to 0.
ArrDel15	Integer	Arrival Delay Indicator, 15 Minutes or More (1=Yes)
Cancelled	Integer	Cancelled Flight Indicator (1=Yes)
CancellationCode	Integer	Specifies The Reason For Cancellation
Diverted	Integer	Diverted Flight Indicator (1=Yes)
CRSElapsedTime	Integer	CRS Elapsed Time of Flight, in Minutes
ActualElapsedTime	Integer	Elapsed Time of Flight, in Minutes
AirTime	Integer	Flight Time, in Minutes
Flights	Integer	Number of Flights
Distance	Integer	Distance between airports (miles)

DistanceGroup	Integer	Distance Intervals, every 250 Miles, for Flight Segment
CarrierDelay	Integer	Carrier Delay, in Minutes
WeatherDelay	Integer	Weather Delay, in Minutes
NASDelay	Integer	National Air System Delay, in Minutes
SecurityDelay	Integer	Security Delay, in Minutes
LateAircraftDelay	Integer	Late Aircraft Delay, in Minutes

TABLE: **UniqueCarriers.csv**

Column Name	Data Type	Column Description
Code	Character	Unique Airline Carrier Code
Description	Character	Airline Carrier Code Description

TABLE: **Airports.csv**

Column Name	Data Type	Column Description
Code	Character	Airport Code (IATA)
Description	Character	Airport Code Description

Data Considerations:

The following rows will be dropped from the dataset:

- Rows that do not qualify for delay or cancellation
- Rows with missing values for carrier, origin, destination, date and time of departure and arrival will be dropped.

Packages

Following packages are required for the project:

i. dplyr
ii. ggplot2
iii. readr
iv. tidyr

Data importing and cleaning

Packages

```
library(readr)
library(dplyr)
library(ggplot2)
library(RColorBrewer)
library(reshape2)
library(pastecs)
library(psych)
library(plotly)
library(corrplot)
library(webshot)
#library(shiny)
```

Data importing

```
## Set the working directory to the root of your DSC 520 directory
setwd("C:/Masters/GitHub/Winter2022/Ramani-DSC520/assignments/FinalProject/")


#Merge flight data from Jan through Nov 2022 into a single dataframe
list_of_files <- list.files(path="Data/DOT_Flight_Data/",
                            recursive = TRUE,
                            pattern = "\\.csv$",
                            full.names = TRUE)

merge_flights_df <- readr::read_csv(list_of_files, id = "fl_date")
nrow(merge_flights_df)
```

```
## [1] 6435187
```

```
head(merge_flights_df,5)
```

```
## # A tibble: 5 × 40
##   fl_date    YEAR QUARTER MONTH DAY_O…¹ DAY_O…² FL_DATE MKT_U…³ OP_UN…⁴ ORIGI…⁵
##   <chr>     <dbl>   <dbl> <dbl>   <dbl>   <dbl> <chr>   <chr>   <chr>     <dbl>
## 1 Data/DOT_…  2022       2     4       1       5 4/1/20… AA      AA        10140
## 2 Data/DOT_…  2022       2     4       1       5 4/1/20… AA      AA        10140
## 3 Data/DOT_…  2022       2     4       1       5 4/1/20… AA      AA        10140
## 4 Data/DOT_…  2022       2     4       1       5 4/1/20… AA      AA        10140
## 5 Data/DOT_…  2022       2     4       1       5 4/1/20… AA      AA        10140
## # … with 30 more variables: ORIGIN <chr>, ORIGIN_CITY_NAME <chr>,
## #   ORIGIN_STATE_ABR <chr>, ORIGIN_STATE_NM <chr>, ORIGIN_WAC <dbl>,
## #   DEST_AIRPORT_ID <dbl>, DEST <chr>, DEST_CITY_NAME <chr>,
## #   DEST_STATE_ABR <chr>, DEST_STATE_NM <chr>, DEST_WAC <dbl>, DEP_DELAY <dbl>,
## #   DEP_DELAY_NEW <dbl>, TAXI_OUT <dbl>, TAXI_IN <dbl>, ARR_TIME <chr>,
## #   ARR_DELAY <dbl>, ARR_DELAY_NEW <dbl>, CANCELLED <dbl>,
## #   CANCELLATION_CODE <chr>, DIVERTED <dbl>, ACTUAL_ELAPSED_TIME <dbl>, …
```

```
cancellation_cd <- read_csv(file="Data/DOT/L_CANCELLATION.csv")
nrow(cancellation_cd)
```

```
## [1] 4
```

```
head(cancellation_cd,2)
```

```
## # A tibble: 2 × 2
##   Code  Description
##   <chr> <chr>
## 1 A     Carrier
## 2 B     Weather
```

```
unique_carrier <- read_csv(file="Data/DOT/L_UNIQUE_CARRIERS.csv")
nrow(unique_carrier)
```

```
## [1] 1714
```

```
head(unique_carrier,2)
```

```
## # A tibble: 2 × 2
##   Code  Description
##   <chr> <chr>
## 1 02Q   Titan Airways
## 2 04Q   Tradewind Aviation
```

```
airport_cd <- read_csv(file="Data/DOT/L_AIRPORT.csv")
```

```
nrow(airport_cd)
```

```
## [1] 6666
```

```
head(airport_cd,2)
```

```
## # A tibble: 2 × 2
##   Code  Description
##   <chr> <chr>
## 1 01A   Afognak Lake, AK: Afognak Lake Airport
## 2 03A   Granite Mountain, AK: Bear Creek Mining Strip
```

Data Transformation and Cleaning

```
#Removing null rows from the dataset
merge_flights_df <- merge_flights_df[,colSums(is.na(merge_flights_df))<nrow(merge_flights_df)]

#Cancellation reason in the flight dataset is represented as A, B, C and D.
#Looking up the cancellation code against the cancellation dataset and adding
#cancellation description to the flight dataframe.
carrier_performance_df <- merge_flights_df

carrier_performance_df$CANCELLATION_REASON <-
  cancellation_cd$Description[match(carrier_performance_df$CANCELLATION_CODE,
                                    cancellation_cd$Code)]

#Carrier codes in flight dataset are represented as 2 character airline carrier codes.
#Looking up the carrier code against the unique carrier dataset and updating the
#code by carrier name in the flight dataframe for both operating and marketing carriers.

carrier_performance_df$MKT_UNIQUE_CARRIER_NAME <-
  unique_carrier$Description[match(carrier_performance_df$MKT_UNIQUE_CARRIER,
                                   unique_carrier$Code)]
carrier_performance_df$OP_UNIQUE_CARRIER_NAME <-
  unique_carrier$Description[match(carrier_performance_df$OP_UNIQUE_CARRIER,
                                   unique_carrier$Code)]

#Updating blank arrival delay to 0
carrier_performance_df[is.na(merge_flights_df$DISTANCE),]$DISTANCE = 0
carrier_performance_df[is.na(merge_flights_df$ARR_DELAY),]$ARR_DELAY = 0
carrier_performance_df[is.na(merge_flights_df$CARRIER_DELAY),]$CARRIER_DELAY = 0
carrier_performance_df[is.na(merge_flights_df$WEATHER_DELAY),]$WEATHER_DELAY = 0
carrier_performance_df[is.na(merge_flights_df$NAS_DELAY),]$NAS_DELAY = 0
carrier_performance_df[is.na(merge_flights_df$SECURITY_DELAY),]$SECURITY_DELAY = 0
carrier_performance_df[is.na(merge_flights_df$LATE_AIRCRAFT_DELAY),]$LATE_AIRCRAFT_DELAY = 0

# Transforming Data

# Update day_of_week from a number to Day
carrier_performance_df <- carrier_performance_df %>% mutate(DAY_OF_WEEK = case_when(
  DAY_OF_WEEK==1~"Monday",
```

```r
  DAY_OF_WEEK==2~"Tuesday",
  DAY_OF_WEEK==3~"Wednesday",
  DAY_OF_WEEK==4~"Thursday",
  DAY_OF_WEEK==5~"Friday",
  DAY_OF_WEEK==6~"Saturday",
  DAY_OF_WEEK==7~"Sunday"))

# Add a new column with the performance status
carrier_performance_df <- carrier_performance_df %>% mutate(
  STATUS = case_when(
    CANCELLED==1~"Cancelled",
    DIVERTED==1~"Diverted",
    ARR_DELAY<=15~"On-Time",
    ARR_DELAY>15~"Delayed"))


# Add a new column with the Delay Flag
carrier_performance_df <- carrier_performance_df %>% mutate(
  DELAYED = case_when(
    ARR_DELAY>15~TRUE,
    ARR_DELAY<=15~FALSE))


# Add a new column with the Delay Reason
carrier_performance_df <- carrier_performance_df %>% mutate(
  DELAY_REASON = case_when(
    ((DELAYED == TRUE) & (CARRIER_DELAY!=0))~"Carrier",
    ((DELAYED == TRUE) & (LATE_AIRCRAFT_DELAY!=0))~"LateAircraft",
    ((DELAYED == TRUE) & (WEATHER_DELAY!=0))~"Weather",
    ((DELAYED == TRUE) & (NAS_DELAY!=0))~"Nas",
    ((DELAYED == TRUE) & (SECURITY_DELAY!=0))~"Security"))

#Since the number of rows are very high (over 6 million),
#we'll narrow the research to flights between 20 major airports.

#Filtering ORIGIN airports
carrier_performance_df <-
  carrier_performance_df[carrier_performance_df$ORIGIN == "ORD"
                                | carrier_performance_df$ORIGIN == "ATL"
                                | carrier_performance_df$ORIGIN == "DFW"
                                | carrier_performance_df$ORIGIN == "DEN"
                                | carrier_performance_df$ORIGIN == "EWR"
                                | carrier_performance_df$ORIGIN == "LAX"
                                | carrier_performance_df$ORIGIN == "IAH"
                                | carrier_performance_df$ORIGIN == "PHX"
                                | carrier_performance_df$ORIGIN == "DTW"
                                | carrier_performance_df$ORIGIN == "SFO"
                                | carrier_performance_df$ORIGIN == "LAS"
                                | carrier_performance_df$ORIGIN == "DEN"
                                | carrier_performance_df$ORIGIN == "ORD"
                                | carrier_performance_df$ORIGIN == "JFK"
                                | carrier_performance_df$ORIGIN == "CLT"
                                | carrier_performance_df$ORIGIN == "LGA"
                                | carrier_performance_df$ORIGIN == "MCO"
```

```
                                | carrier_performance_df$ORIGIN == "MSP"
                                | carrier_performance_df$ORIGIN == "BOS"
                                | carrier_performance_df$ORIGIN == "PHL",]

nrow(carrier_performance_df)
```

```
## [1] 3016994
```

```
#Filtering DESTINATION airports
carrier_performance_df <-
  carrier_performance_df[carrier_performance_df$DEST == "ORD"
                                | carrier_performance_df$DEST == "ATL"
                                | carrier_performance_df$DEST == "DFW"
                                | carrier_performance_df$DEST == "DEN"
                                | carrier_performance_df$DEST == "EWR"
                                | carrier_performance_df$DEST == "LAX"
                                | carrier_performance_df$DEST == "IAH"
                                | carrier_performance_df$DEST == "PHX"
                                | carrier_performance_df$DEST == "DTW"
                                | carrier_performance_df$DEST == "SFO"
                                | carrier_performance_df$DEST == "LAS"
                                | carrier_performance_df$DEST == "DEN"
                                | carrier_performance_df$DEST == "ORD"
                                | carrier_performance_df$DEST == "JFK"
                                | carrier_performance_df$DEST == "CLT"
                                | carrier_performance_df$DEST == "LGA"
                                | carrier_performance_df$DEST == "MCO"
                                | carrier_performance_df$DEST == "MSP"
                                | carrier_performance_df$DEST == "BOS"
                                | carrier_performance_df$DEST == "PHL",]

nrow(carrier_performance_df)
```

```
## [1] 1073457
```

```
#Airport codes in flight dataset are represented as 3 character airport codes.
#Looking up the airport codes against the airport dataset and updating the
#airport code by name in the flight dataframe for origin and destination columns.

#carrier_performance_df$ORIGIN_AIRPORT <-
#  airport_cd$Description[match(carrier_performance_df$ORIGIN, airport_cd$Code)]
#carrier_performance_df$DEST_AIRPORT <-
#  airport_cd$Description[match(carrier_performance_df$DEST, airport_cd$Code)]

# Selecting relevant columns from flights data
carrier_performance_df <-
  carrier_performance_df[c("YEAR","QUARTER","MONTH","DAY_OF_MONTH","DAY_OF_WEEK",
                      "FL_DATE","MKT_UNIQUE_CARRIER","OP_UNIQUE_CARRIER",
                      "OP_UNIQUE_CARRIER_NAME","MKT_UNIQUE_CARRIER_NAME",
                      "ORIGIN","ORIGIN_CITY_NAME","ORIGIN_STATE_ABR",
                      "ORIGIN_STATE_NM","DEST","DEST_CITY_NAME","DEST_STATE_ABR",
```

```
                         "DEST_STATE_NM","DEP_DELAY","TAXI_OUT","TAXI_IN","ARR_DELAY",
                         "CANCELLED","CANCELLATION_CODE","CANCELLATION_REASON",
                         "DIVERTED","DISTANCE","CARRIER_DELAY","WEATHER_DELAY",
                         "NAS_DELAY","SECURITY_DELAY","LATE_AIRCRAFT_DELAY",
                         "DELAYED" ,"DELAY_REASON","STATUS")]
```

What does the final data set look like?

```
head(carrier_performance_df,5)
```

```
## # A tibble: 5 × 35
##    YEAR QUARTER MONTH DAY_OF_M…¹ DAY_O…² FL_DATE MKT_U…³ OP_UN…⁴ OP_UN…⁵ MKT_U…⁶
##   <dbl>   <dbl> <dbl>      <dbl> <chr>   <chr>   <chr>   <chr>   <chr>   <chr>
## 1  2022       2     4          1 Friday  4/1/20… AA      AA      Americ… Americ…
## 2  2022       2     4          1 Friday  4/1/20… AA      AA      Americ… Americ…
## 3  2022       2     4          1 Friday  4/1/20… AA      AA      Americ… Americ…
## 4  2022       2     4          1 Friday  4/1/20… AA      AA      Americ… Americ…
## 5  2022       2     4          1 Friday  4/1/20… AA      AA      Americ… Americ…
## # … with 25 more variables: ORIGIN <chr>, ORIGIN_CITY_NAME <chr>,
## #   ORIGIN_STATE_ABR <chr>, ORIGIN_STATE_NM <chr>, DEST <chr>,
## #   DEST_CITY_NAME <chr>, DEST_STATE_ABR <chr>, DEST_STATE_NM <chr>,
## #   DEP_DELAY <dbl>, TAXI_OUT <dbl>, TAXI_IN <dbl>, ARR_DELAY <dbl>,
## #   CANCELLED <dbl>, CANCELLATION_CODE <chr>, CANCELLATION_REASON <chr>,
## #   DIVERTED <dbl>, DISTANCE <dbl>, CARRIER_DELAY <dbl>, WEATHER_DELAY <dbl>,
## #   NAS_DELAY <dbl>, SECURITY_DELAY <dbl>, LATE_AIRCRAFT_DELAY <dbl>, …
```

```
names(carrier_performance_df)
```

```
##  [1] "YEAR"                      "QUARTER"
##  [3] "MONTH"                     "DAY_OF_MONTH"
##  [5] "DAY_OF_WEEK"               "FL_DATE"
##  [7] "MKT_UNIQUE_CARRIER"        "OP_UNIQUE_CARRIER"
##  [9] "OP_UNIQUE_CARRIER_NAME"    "MKT_UNIQUE_CARRIER_NAME"
## [11] "ORIGIN"                    "ORIGIN_CITY_NAME"
## [13] "ORIGIN_STATE_ABR"          "ORIGIN_STATE_NM"
## [15] "DEST"                      "DEST_CITY_NAME"
## [17] "DEST_STATE_ABR"            "DEST_STATE_NM"
## [19] "DEP_DELAY"                 "TAXI_OUT"
## [21] "TAXI_IN"                   "ARR_DELAY"
## [23] "CANCELLED"                 "CANCELLATION_CODE"
## [25] "CANCELLATION_REASON"       "DIVERTED"
## [27] "DISTANCE"                  "CARRIER_DELAY"
## [29] "WEATHER_DELAY"             "NAS_DELAY"
## [31] "SECURITY_DELAY"            "LATE_AIRCRAFT_DELAY"
## [33] "DELAYED"                   "DELAY_REASON"
## [35] "STATUS"
```

What information is not self-evident?

Initial thoughts: I would like to see if there are weather delays or cancellations specific to a time of year. If yes, I would like to see if it can be isolated to a particular airport or carrier.Also, I am hoping to evaluate the reason reported. Was it reported as a weather delay or a NAS delay. This would probably give an option to see which carrier has reported the most number of NAS delays during bad weather.

Current thoughts: There is not sufficient data for weather to relate to delay/cancellation reason. It would be good to have weather information in the dataset to build a relation and analyze further.

What are different ways you could look at this data ?

I would like to perform the following:

1. Percentages of flights scheduled and flown per airline.
2. Percentages of flights scheduled vs delayed per airline.
3. Identify the correlations between variables and perform further analysis based on the outcomes.

Do you plan to slice and dice the data?

For the purposes of this analysis, I am considering flights with arrival time less than 15 minutes as on-time.

I am splitting dataset into 2 categories.

1. no cancellations and delays (on-time performace)
2. cancellations, delays

```
carrier_on_time_performance_df <-
  carrier_performance_df[(is.na(carrier_performance_df$CANCELLATION_CODE)&
                            carrier_performance_df$ARR_DELAY <= 15),]

carrier_cancel_or_delay_df <-
  carrier_performance_df[!(is.na(carrier_performance_df$CANCELLATION_CODE)&
                            carrier_performance_df$ARR_DELAY <= 15),]

#Further splitting dataframes for delays and cancellations.

#Delays Dataset

carrier_delay_df <-
  carrier_cancel_or_delay_df[carrier_cancel_or_delay_df$ARR_DELAY > 15,]

#Cancel Dataset

carrier_cancelled_df <-
  carrier_cancel_or_delay_df[!is.na(carrier_cancel_or_delay_df$CANCELLATION_CODE),]
```

```
## [1] "No. of rows in complte DF :  1073457"
```

```
## [1] "No. of rows in delay DF :  215522"
```

```
## [1] "No. of rows in cancelled DF :  27655"
```

```
## [1] "No. of rows in on-time performance DF :  830280"
```

How could you summarize your data to answer key questions?

Calculating the correlation and covariance are great ways to summarize my data to answer key questions. Results from the summary function would also help. In addition, finding the maximum, minimum, mean, and median values for delays will provide some more information.

STATISTICAL ANALYSIS

```
summary(carrier_performance_df)
```

```
##       YEAR         QUARTER          MONTH         DAY_OF_MONTH
##  Min.   :2022   Min.   :1.000   Min.   : 1.000   Min.   : 1.00
##  1st Qu.:2022   1st Qu.:1.000   1st Qu.: 3.000   1st Qu.: 8.00
##  Median :2022   Median :2.000   Median : 6.000   Median :16.00
##  Mean   :2022   Mean   :2.399   Mean   : 6.106   Mean   :15.72
##  3rd Qu.:2022   3rd Qu.:3.000   3rd Qu.: 9.000   3rd Qu.:23.00
##  Max.   :2022   Max.   :4.000   Max.   :11.000   Max.   :31.00
##
##  DAY_OF_WEEK          FL_DATE          MKT_UNIQUE_CARRIER OP_UNIQUE_CARRIER
##  Length:1073457     Length:1073457     Length:1073457     Length:1073457
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  OP_UNIQUE_CARRIER_NAME MKT_UNIQUE_CARRIER_NAME    ORIGIN
##  Length:1073457         Length:1073457          Length:1073457
##  Class :character       Class :character        Class :character
##  Mode  :character       Mode  :character        Mode  :character
##
##
##
##
##  ORIGIN_CITY_NAME   ORIGIN_STATE_ABR   ORIGIN_STATE_NM        DEST
##  Length:1073457     Length:1073457     Length:1073457     Length:1073457
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  DEST_CITY_NAME     DEST_STATE_ABR     DEST_STATE_NM        DEP_DELAY
##  Length:1073457     Length:1073457     Length:1073457     Min.   : -78.00
##  Class :character   Class :character   Class :character   1st Qu.:  -5.00
##  Mode  :character   Mode  :character   Mode  :character   Median :  -1.00
##                                                           Mean   :  13.61
##                                                           3rd Qu.:  11.00
##                                                           Max.   :2991.00
```

```
##                                                              NA's   :26923
##     TAXI_OUT         TAXI_IN          ARR_DELAY          CANCELLED
##  Min.   :  1.00   Min.   :  1.000   Min.   : -87.000   Min.   :0.00000
##  1st Qu.: 13.00   1st Qu.:  6.000   1st Qu.: -16.000   1st Qu.:0.00000
##  Median : 16.00   Median :  8.000   Median :  -6.000   Median :0.00000
##  Mean   : 18.64   Mean   :  9.487   Mean   :   6.281   Mean   :0.02576
##  3rd Qu.: 21.00   3rd Qu.: 11.000   3rd Qu.:   9.000   3rd Qu.:0.00000
##  Max.   :197.00   Max.   :253.000   Max.   :2996.000   Max.   :1.00000
##  NA's   :27602    NA's   :27827
##  CANCELLATION_CODE  CANCELLATION_REASON    DIVERTED           DISTANCE
##  Length:1073457     Length:1073457      Min.   :0.000000   Min.   :  80
##  Class :character   Class :character    1st Qu.:0.000000   1st Qu.: 602
##  Mode  :character   Mode  :character    Median :0.000000   Median : 907
##                                         Mean   :0.002243   Mean   :1067
##                                         3rd Qu.:0.000000   3rd Qu.:1440
##                                         Max.   :1.000000   Max.   :2704
##
##  CARRIER_DELAY      WEATHER_DELAY         NAS_DELAY        SECURITY_DELAY
##  Min.   :   0.000   Min.   :   0.0000   Min.   :   0.000   Min.   :  0.00000
##  1st Qu.:   0.000   1st Qu.:   0.0000   1st Qu.:   0.000   1st Qu.:  0.00000
##  Median :   0.000   Median :   0.0000   Median :   0.000   Median :  0.00000
##  Mean   :   5.623   Mean   :   0.6368   Mean   :   3.053   Mean   :  0.02397
##  3rd Qu.:   0.000   3rd Qu.:   0.0000   3rd Qu.:   0.000   3rd Qu.:  0.00000
##  Max.   :2991.000   Max.   :1491.0000   Max.   :1310.000   Max.   :255.00000
##
##  LATE_AIRCRAFT_DELAY  DELAYED        DELAY_REASON          STATUS
##  Min.   :   0.000    Mode :logical   Length:1073457     Length:1073457
##  1st Qu.:   0.000    FALSE:857935    Class :character   Class :character
##  Median :   0.000    TRUE :215522    Mode  :character   Mode  :character
##  Mean   :   4.951
##  3rd Qu.:   0.000
##  Max.   :2175.000
##
```

```
## [1] "                VARIANCE                       "
```

```
## [1] "Distance            :  428311.109704511"
```

```
## [1] "Arrival Delay       :  2970.88989969246"
```

```
## [1] "Carrier Delay       :  1249.63070092016"
```

```
## [1] "Weather Delay       :  128.724722963208"
```

```
## [1] "NAS Delay           :  262.724559984357"
```

```
## [1] "Security Dela       :  1.3815269689925"
```

```
## [1] "Late Aircraft Delay :  730.252958009103"
```

```
## [1] "           STANDARD DEVIATION            "
```

```
## [1] "Distance            :  654.454818688434"
```

```
## [1] "Arrival Delay       :  54.5058703232272"
```

```
## [1] "Carrier Delay       :  35.3501159958516"
```

```
## [1] "Weather Delay       :  11.3456918239131"
```

```
## [1] "NAS Delay           :  16.2087803361128"
```

```
## [1] "Security Dela       :  1.17538375392571"
```

```
## [1] "Late Aircraft Delay :  27.0231929647313"
```

The average arrival delay is only around 6 minutes. We can see that the median value is -5 minutes, suggesting the majority of flights actually arrive earlier than their expected time of arrival.

```
## SUMMARY
describe(head(carrier_performance_df,5000))
```

```
##                            vars    n    mean    sd median trimmed   mad  min
## YEAR                          1 5000 2022.00  0.00   2022 2022.00  0.00 2022
## QUARTER                       2 5000    2.00  0.00      2    2.00  0.00    2
## MONTH                         3 5000    4.00  0.00      4    4.00  0.00    4
## DAY_OF_MONTH                  4 5000    1.33  0.47      1    1.29  0.00    1
## DAY_OF_WEEK*                  5 5000    1.33  0.47      1    1.29  0.00    1
## FL_DATE*                      6 5000    1.33  0.47      1    1.29  0.00    1
## MKT_UNIQUE_CARRIER*           7 5000    3.53  2.30      4    3.36  4.45    1
## OP_UNIQUE_CARRIER*            8 5000    6.40  4.89      5    5.82  4.45    1
## OP_UNIQUE_CARRIER_NAME*       9 5000    6.83  5.86      3    6.18  1.48    1
## MKT_UNIQUE_CARRIER_NAME*     10 5000    3.79  2.19      3    3.52  1.48    1
## ORIGIN*                      11 5000    9.17  5.27      9    9.12  7.41    1
## ORIGIN_CITY_NAME*            12 5000    8.56  4.89      9    8.49  5.93    1
## ORIGIN_STATE_ABR*            13 5000    7.95  4.61      7    7.88  5.93    1
## ORIGIN_STATE_NM*             14 5000    7.93  4.57      7    7.85  5.93    1
## DEST*                        15 5000    9.19  5.27      9    9.15  7.41    1
## DEST_CITY_NAME*              16 5000    8.58  4.89      9    8.51  5.93    1
## DEST_STATE_ABR*              17 5000    7.94  4.61      7    7.86  5.93    1
## DEST_STATE_NM*               18 5000    7.91  4.58      7    7.83  5.93    1
## DEP_DELAY                    19 4654   24.37 66.41      1   10.31  8.90  -20
## TAXI_OUT                     20 4644   18.42 11.38     16   16.40  4.45    5
## TAXI_IN                      21 4643    9.46  8.00      7    7.99  2.97    1
## ARR_DELAY                    22 5000   15.38 65.79     -2    2.86 19.27  -49
```

```
## CANCELLED                  23 5000    0.07   0.26      0    0.00   0.00    0
## CANCELLATION_CODE*         24  357    2.03   0.79      2    2.03   1.48    1
## CANCELLATION_REASON*       25  357    2.09   0.82      2    2.11   1.48    1
## DIVERTED                   26 5000    0.00   0.02      0    0.00   0.00    0
## DISTANCE                   27 5000 1078.22 658.84    925 1012.68 610.83   80
## CARRIER_DELAY              28 5000    8.90  43.68      0    1.03   0.00    0
## WEATHER_DELAY              29 5000    0.79  13.87      0    0.00   0.00    0
## NAS_DELAY                  30 5000    3.93  21.54      0    0.02   0.00    0
## SECURITY_DELAY             31 5000    0.02   0.72      0    0.00   0.00    0
## LATE_AIRCRAFT_DELAY        32 5000    8.24  32.17      0    0.36   0.00    0
## DELAYED                    33 5000     NaN     NA     NA     NaN     NA  Inf
## DELAY_REASON*              34 1310    1.54   0.92      1    1.35   0.00    1
## STATUS*                    35 5000    3.26   1.07      4    3.42   0.00    1
##                          max range  skew kurtosis   se
## YEAR                    2022     0   NaN      NaN 0.00
## QUARTER                    2     0   NaN      NaN 0.00
## MONTH                      4     0   NaN      NaN 0.00
## DAY_OF_MONTH               2     1  0.70    -1.51 0.01
## DAY_OF_WEEK*               2     1  0.70    -1.51 0.01
## FL_DATE*                   2     1  0.70    -1.51 0.01
## MKT_UNIQUE_CARRIER*        8     7  0.40    -1.06 0.03
## OP_UNIQUE_CARRIER*        17    16  0.89    -0.69 0.07
## OP_UNIQUE_CARRIER_NAME*   17    16  0.76    -1.16 0.08
## MKT_UNIQUE_CARRIER_NAME*   8     7  0.97    -0.55 0.03
## ORIGIN*                   18    17  0.02    -1.25 0.07
## ORIGIN_CITY_NAME*         17    16  0.06    -1.22 0.07
## ORIGIN_STATE_ABR*         15    14  0.09    -1.35 0.07
## ORIGIN_STATE_NM*          15    14  0.09    -1.34 0.06
## DEST*                     18    17  0.01    -1.25 0.07
## DEST_CITY_NAME*           17    16  0.05    -1.22 0.07
## DEST_STATE_ABR*           15    14  0.09    -1.35 0.07
## DEST_STATE_NM*            15    14  0.09    -1.34 0.06
## DEP_DELAY               1421  1441  7.35    96.35 0.97
## TAXI_OUT                 157   152  4.81    35.59 0.17
## TAXI_IN                  109   108  4.72    33.42 0.12
## ARR_DELAY               1398  1447  7.04    91.69 0.93
## CANCELLED                  1     1  3.33     9.08 0.00
## CANCELLATION_CODE*         3     2 -0.05    -1.38 0.04
## CANCELLATION_REASON*       3     2 -0.17    -1.50 0.04
## DIVERTED                   1     1 49.96  2494.00 0.00
## DISTANCE                2704  2624  0.79    -0.27 9.32
## CARRIER_DELAY           1398  1398 15.34   351.53 0.62
## WEATHER_DELAY            671   671 31.22  1248.35 0.20
## NAS_DELAY                799   799 16.30   451.71 0.30
## SECURITY_DELAY            30    30 33.44  1181.33 0.01
## LATE_AIRCRAFT_DELAY      546   546  6.27    54.04 0.45
## DELAYED                 -Inf  -Inf    NA       NA   NA
## DELAY_REASON*              5     4  1.93     3.70 0.03
## STATUS*                    4     3 -0.88    -0.92 0.02
```

```
stat.desc(head(carrier_performance_df$ARR_DELAY,5000), basic = TRUE, norm = TRUE)
```

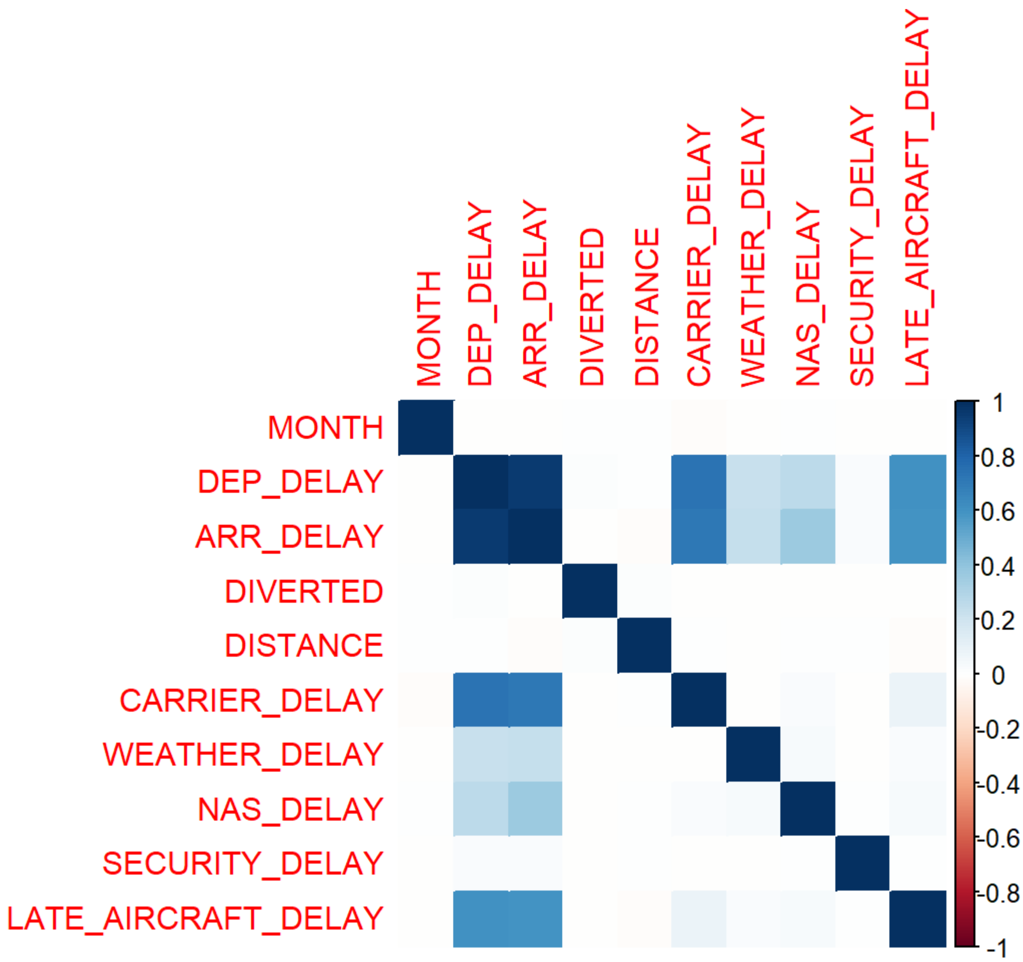
```
##       nbr.val      nbr.null         nbr.na            min            max
```

```
##   5.000000e+03  4.250000e+02  0.000000e+00 -4.900000e+01  1.398000e+03
##          range           sum        median          mean       SE.mean
##   1.447000e+03  7.692200e+04 -2.000000e+00  1.538440e+01  9.304249e-01
##   CI.mean.0.95           var       std.dev      coef.var      skewness
##   1.824041e+00  4.328452e+03  6.579097e+01  4.276473e+00  7.043388e+00
##       skew.2SE      kurtosis      kurt.2SE    normtest.W    normtest.p
##   1.016930e+02  9.169439e+01  6.620779e+02  5.154294e-01  1.793159e-79
```

Skew and Kurtosis are both non-zero and positive for the top 5000 rows. A positive kurtosis represents a pointy and heavy-tailed distribution and a positive skew represents a right skew.

CORRELATION

```
delay_cormatrix <- cor(carrier_performance_df$DEP_DELAY,
                       carrier_performance_df$ARR_DELAY,
                       use = "complete.obs")

corr_df <-
    carrier_performance_df[,c("MONTH","DEP_DELAY","ARR_DELAY","DIVERTED",
                              "DISTANCE","CARRIER_DELAY","WEATHER_DELAY",
                              "NAS_DELAY","SECURITY_DELAY","LATE_AIRCRAFT_DELAY")]


cormatrix <- cor(corr_df,use = "complete.obs")
corrplot(cormatrix, method="color")
```



Plots & Tables

Plots that I would like to explore:

i. Scatter plot
ii. Pie chart
iii. Histogram
iv. Boxplot

I will create tables with the following data: A summary table of on-time performance, delays, and cancellations per carrier.

What types of plots and tables will help you to illustrate the findings to your questions?

TABLES

```
flight_totals_df
```

```
## # A tibble: 19 × 4
## # Groups:   OP_UNIQUE_CARRIER, OP_UNIQUE_CARRIER_NAME [19]
##    OP_UNIQUE_CARRIER OP_UNIQUE_CARRIER_NAME                  TOTAL PERCENTAGE
##    <chr>             <chr>                                   <int>      <dbl>
##  1 9E                Endeavor Air Inc.                       12575       1.17
##  2 AA                American Airlines Inc.                 256452      23.9 
##  3 AS                Alaska Airlines Inc.                    12626       1.18
##  4 B6                JetBlue Airways                         76435       7.12
##  5 DL                Delta Air Lines Inc.                   228512      21.3 
##  6 F9                Frontier Airlines Inc.                  38985       3.63
##  7 G4                Allegiant Air                               5       0   
##  8 G7                GoJet Airlines LLC d/b/a United Express  1823       0.17
##  9 MQ                Envoy Air                                5077       0.47
## 10 NK                Spirit Air Lines                        59970       5.59
## 11 OH                PSA Airlines Inc.                        4743       0.44
## 12 OO                SkyWest Airlines Inc.                   38685       3.6 
## 13 PT                Piedmont Airlines                         154       0.01
## 14 QX                Horizon Air                               911       0.08
## 15 UA                United Air Lines Inc.                  208725      19.4 
## 16 WN                Southwest Airlines Co.                  75171       7   
## 17 YV                Mesa Airlines Inc.                       9686       0.9 
## 18 YX                Republic Airline                        42877       3.99
## 19 ZW                Air Wisconsin Airlines Corp                45       0   
```

```
flight_stats
```

```
## # A tibble: 104 × 5
## # Groups:   OP_UNIQUE_CARRIER, OP_UNIQUE_CARRIER_NAME, DELAY_REASON [104]
##    OP_UNIQUE_CARRIER OP_UNIQUE_CARRIER_NAME DELAY_REASON COUNT PERCENTAGE
##    <chr>             <chr>                  <chr>        <int>      <dbl>
##  1 9E                Endeavor Air Inc.      Carrier        829       6.59
##  2 9E                Endeavor Air Inc.      LateAircraft   556       4.42
```

```
## 3 9E Endeavor Air Inc. Nas 559 4.45
## 4 9E Endeavor Air Inc. Security 1 0.01
## 5 9E Endeavor Air Inc. Weather 77 0.61
## 6 9E Endeavor Air Inc. <NA> 10553 83.9
## 7 AA American Airlines Inc. Carrier 30736 12.0
## 8 AA American Airlines Inc. LateAircraft 10606 4.14
## 9 AA American Airlines Inc. Nas 7621 2.97
## 10 AA American Airlines Inc. Security 70 0.03
## # … with 94 more rows
```

```
flight_cancel
```

```
## # A tibble: 72 × 5
## # Groups: OP_UNIQUE_CARRIER, OP_UNIQUE_CARRIER_NAME, CANCELLATION_REASON [72]
## OP_UNIQUE_CARRIER OP_UNIQUE_CARRIER_NAME CANCELLATION_REASON COUNT PERCENT…¹
## <chr> <chr> <chr> <int> <dbl>
## 1 9E Endeavor Air Inc. Carrier 118 0.94
## 2 9E Endeavor Air Inc. National Air System 303 2.41
## 3 9E Endeavor Air Inc. Weather 210 1.67
## 4 9E Endeavor Air Inc. <NA> 11944 95.0
## 5 AA American Airlines Inc. Carrier 2585 1.01
## 6 AA American Airlines Inc. National Air System 442 0.17
## 7 AA American Airlines Inc. Weather 4813 1.88
## 8 AA American Airlines Inc. <NA> 248612 96.9
## 9 AS Alaska Airlines Inc. Carrier 357 2.83
## 10 AS Alaska Airlines Inc. National Air System 3 0.02
## # … with 62 more rows, and abbreviated variable name ¹PERCENTAGE
```

```
flight_status
```

```
## # A tibble: 72 × 5
## # Groups: OP_UNIQUE_CARRIER, OP_UNIQUE_CARRIER_NAME, STATUS [72]
## OP_UNIQUE_CARRIER OP_UNIQUE_CARRIER_NAME STATUS COUNT PERCENTAGE
## <chr> <chr> <chr> <int> <dbl>
## 1 9E Endeavor Air Inc. Cancelled 631 5.02
## 2 9E Endeavor Air Inc. Delayed 2022 16.1
## 3 9E Endeavor Air Inc. Diverted 29 0.23
## 4 9E Endeavor Air Inc. On-Time 9893 78.7
## 5 AA American Airlines Inc. Cancelled 7840 3.06
## 6 AA American Airlines Inc. Delayed 50919 19.9
## 7 AA American Airlines Inc. Diverted 648 0.25
## 8 AA American Airlines Inc. On-Time 197045 76.8
## 9 AS Alaska Airlines Inc. Cancelled 377 2.99
## 10 AS Alaska Airlines Inc. Delayed 2830 22.4
## # … with 62 more rows
```

```
status_percentage
```

```
## # A tibble: 4 × 3
## # Groups: STATUS [4]
```

```
##   STATUS     COUNT PERCENTAGE
##   <chr>      <int>      <dbl>
## 1 Cancelled  27655       2.58
## 2 Delayed   215522      20.1
## 3 Diverted    2408       0.22
## 4 On-Time   827872      77.1
```

```
flight_origin_totals_df
```

```
## # A tibble: 18 × 3
## # Groups:   ORIGIN [18]
##    ORIGIN TOTAL PERCENTAGE
##    <chr>  <int>      <dbl>
##  1 ATL    76828       7.16
##  2 BOS    64461       6
##  3 CLT    52401       4.88
##  4 DEN    72364       6.74
##  5 DFW    67146       6.26
##  6 DTW    45822       4.27
##  7 EWR    53907       5.02
##  8 IAH    52286       4.87
##  9 JFK    49075       4.57
## 10 LAS    62311       5.8
## 11 LAX    83480       7.78
## 12 LGA    56552       5.27
## 13 MCO    61543       5.73
## 14 MSP    41347       3.85
## 15 ORD    87089       8.11
## 16 PHL    38535       3.59
## 17 PHX    52665       4.91
## 18 SFO    55645       5.18
```

```
cancelled_status
```

```
## # A tibble: 71 × 5
## # Groups:   ORIGIN, CANCELLATION_REASON, STATUS [71]
##    ORIGIN CANCELLATION_REASON STATUS    COUNT PERCENTAGE
##    <chr>  <chr>               <chr>     <int>      <dbl>
##  1 ATL    Carrier             Cancelled   643       0.84
##  2 ATL    National Air System Cancelled   159       0.21
##  3 ATL    Security            Cancelled     8       0.01
##  4 ATL    Weather             Cancelled   655       0.85
##  5 BOS    Carrier             Cancelled   700       1.09
##  6 BOS    National Air System Cancelled   317       0.49
##  7 BOS    Security            Cancelled    16       0.02
##  8 BOS    Weather             Cancelled  1141       1.77
##  9 CLT    Carrier             Cancelled   474       0.9
## 10 CLT    National Air System Cancelled   167       0.32
## # … with 61 more rows
```

```
delayed_status
```

```
## # A tibble: 72 × 4
## # Groups:   ORIGIN, STATUS [72]
##    ORIGIN STATUS    COUNT PERCENTAGE
##    <chr>  <chr>     <int>      <dbl>
##  1 ATL    Cancelled  1465       1.91
##  2 ATL    Delayed   14683      19.1
##  3 ATL    Diverted    182       0.24
##  4 ATL    On-Time   60498      78.7
##  5 BOS    Cancelled  2174       3.37
##  6 BOS    Delayed   12271      19.0
##  7 BOS    Diverted    136       0.21
##  8 BOS    On-Time   49880      77.4
##  9 CLT    Cancelled  1617       3.09
## 10 CLT    Delayed   10452      20.0
## # … with 62 more rows
```

```
delayed_reason_status
```

```
## # A tibble: 144 × 5
## # Groups:   ORIGIN, DELAY_REASON, STATUS [144]
##    ORIGIN DELAY_REASON STATUS    COUNT PERCENTAGE
##    <chr>  <chr>        <chr>     <int>      <dbl>
##  1 ATL    Carrier      Delayed    9777      12.7
##  2 ATL    LateAircraft Delayed    1828       2.38
##  3 ATL    Nas          Delayed    2643       3.44
##  4 ATL    Security     Delayed      19       0.02
##  5 ATL    Weather      Delayed     416       0.54
##  6 ATL    <NA>         Cancelled  1465       1.91
##  7 ATL    <NA>         Diverted    182       0.24
##  8 ATL    <NA>         On-Time   60498      78.7
##  9 BOS    Carrier      Delayed    7384      11.4
## 10 BOS    LateAircraft Delayed    1793       2.78
## # … with 134 more rows
```

Plots

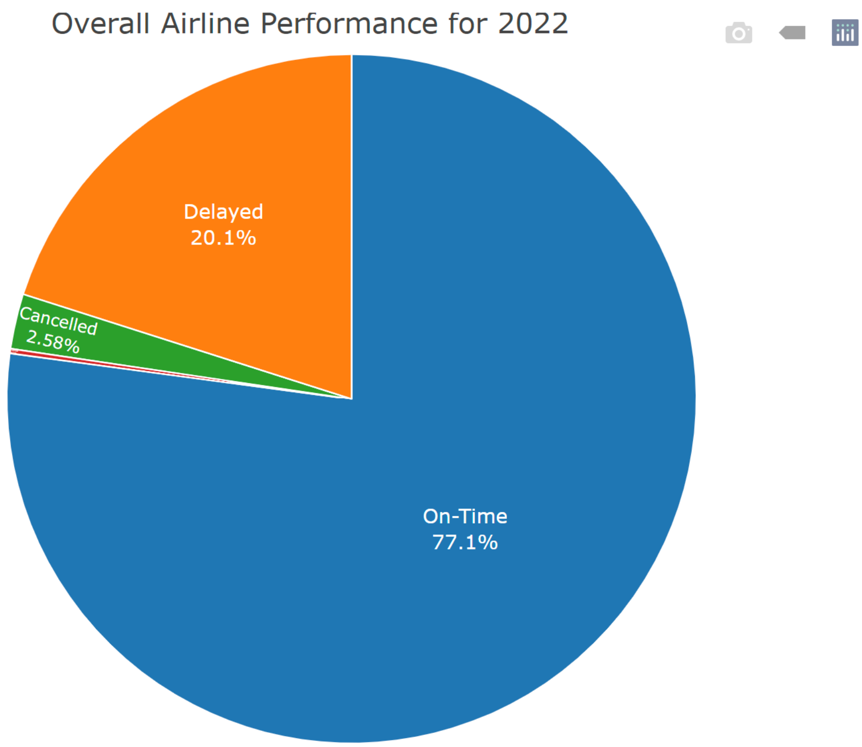
Pie Plot

Airline Performance

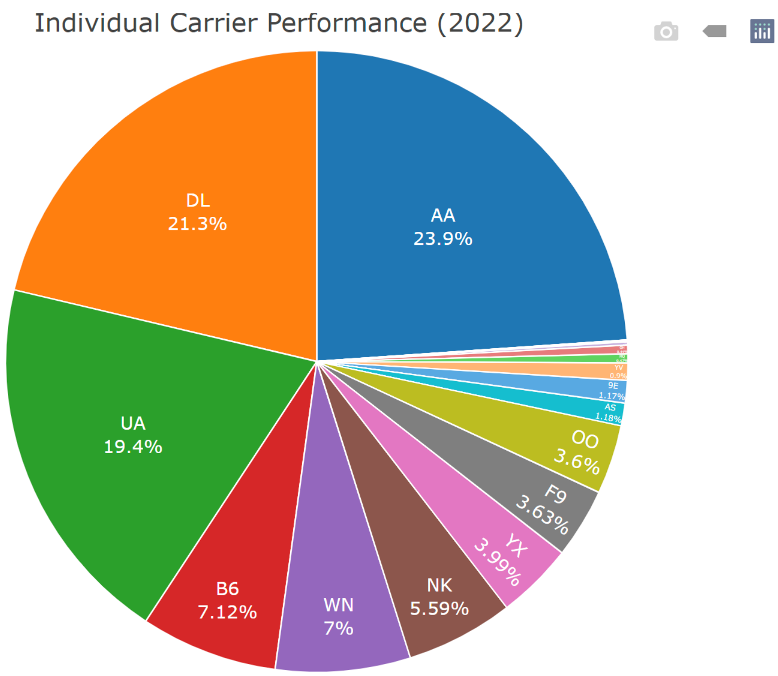
As observed from the statistical analysis, the average arrival delay is only around 6 minutes. To capture this further, I've created a bar chart with the percentage of airline performance in 2022. Only 20.1% Delayed are delayed while 77.1% are on-time.

```
fig1 <- plot_ly(status_percentage, labels = ~STATUS, values = ~PERCENTAGE, type = 'pie',
        textposition = 'inside',
        textinfo = 'STATUS + PERCENTAGE',
        text = ~paste(STATUS),
```

```
        insidetextfont = list(color = '#FFFFFF'),
        marker = list(colors = colors,line = list(color = '#FFFFFF', width = 1)),
        showlegend = FALSE)
fig1 <- fig1 %>% layout(title = 'Overall Airline Performance for 2022',
         xaxis = list(showgrid = FALSE, zeroline = FALSE, showticklabels = FALSE),
         yaxis = list(showgrid = FALSE, zeroline = FALSE, showticklabels = FALSE))
fig1
```




```
#tmpFile <- tempfile(fileext = ".png")
#export(fig1, file = #tmpFile)

fig2 <- plot_ly(flight_totals_df, labels = ~OP_UNIQUE_CARRIER, values = ~PERCENTAGE, type = 'p
ie',
        textposition = 'inside',
        textinfo = 'OP_UNIQUE_CARRIER + PERCENTAGE',
        insidetextfont = list(color = '#FFFFFF'),
        hoverinfo = 'text',
        text = ~paste(OP_UNIQUE_CARRIER),
        marker = list(colors = colors,line = list(color = '#FFFFFF', width = 1)),
        showlegend = FALSE)
fig2 <- fig2 %>% layout(title = 'Individual Carrier Performance (2022)',
         xaxis = list(showgrid = FALSE, zeroline = FALSE, showticklabels = FALSE),
         yaxis = list(showgrid = FALSE, zeroline = FALSE, showticklabels = FALSE))

fig2
```

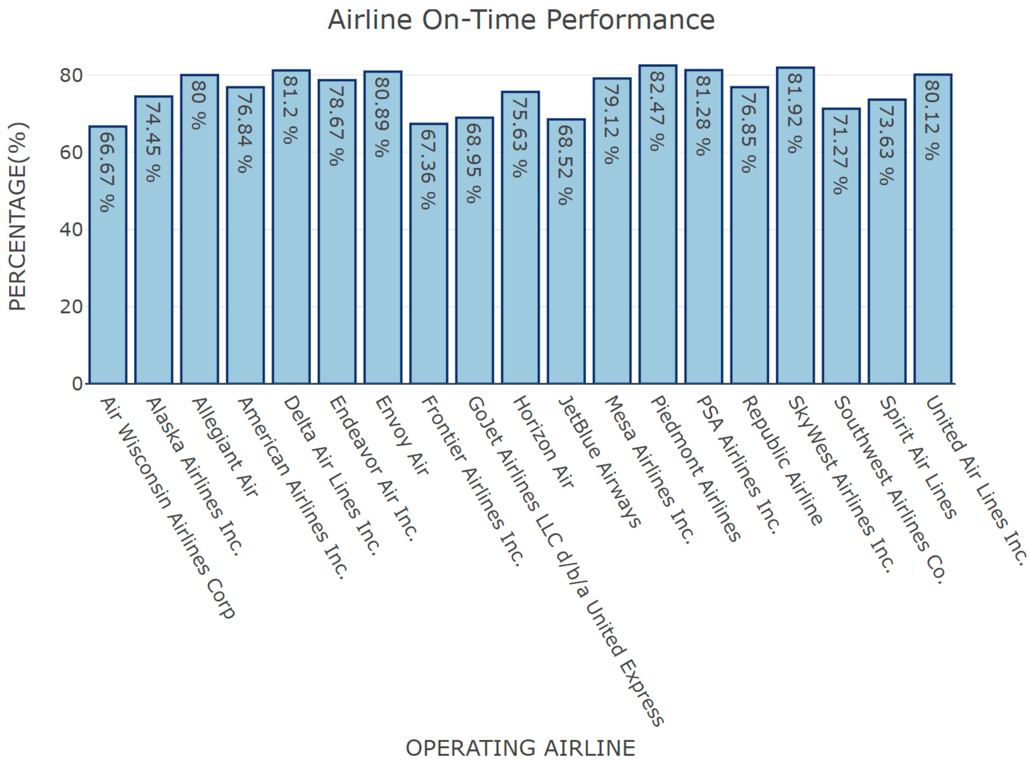



```
#tmpFile <- tempfile(fileext = ".png")
#export(fig2, file = #tmpFile)
```

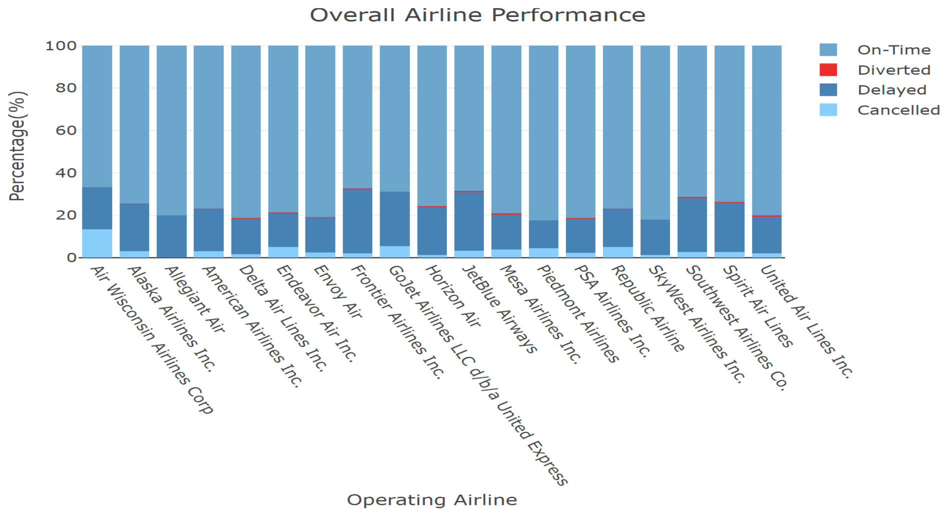
Bar Plot

Flight Stats by Operating Carrier

```
airline_on_time_performance <- flight_status[flight_status$STATUS=='On-Time',]

airline_on_time_performance <- airline_on_time_performance[order(airline_on_time_performance$P
ERCENTAGE,decreasing=TRUE),]



fig1 <- plot_ly(airline_on_time_performance, x = ~OP_UNIQUE_CARRIER_NAME,
                y = ~PERCENTAGE, type = 'bar',text = ~paste(PERCENTAGE,'%'),
                textposition = 'auto',
                marker = list(color = 'rgb(158,202,225)',
                              line = list(color = 'rgb(8,48,107)', width = 1.5)))
fig1 <- fig1 %>% layout(title = "Airline On-Time Performance",
         xaxis = list(title = "OPERATING AIRLINE",tickangle=60),
         yaxis = list(title = "PERCENTAGE(%)"))
fig1
```




```
#tmpFile <- tempfile(fileext = ".png")
#export(fig1, file = #tmpFile)


fig2 <- plot_ly(flight_status, x = ~OP_UNIQUE_CARRIER_NAME, y = ~PERCENTAGE,
               type = 'bar',text = ~paste(STATUS),  textposition = 'auto',
               name = ~STATUS, color = ~STATUS, colors = c('lightskyblue','steelblue','Firebri
ck2','skyblue3'))
fig2 <- fig2 %>% layout(yaxis = list(title='Percentage(%)'),title = 'Overall Airline Performan
ce',
                      xaxis=list(title='Operating Airline',tickangle=60), barmode = 'stack')
fig2
```

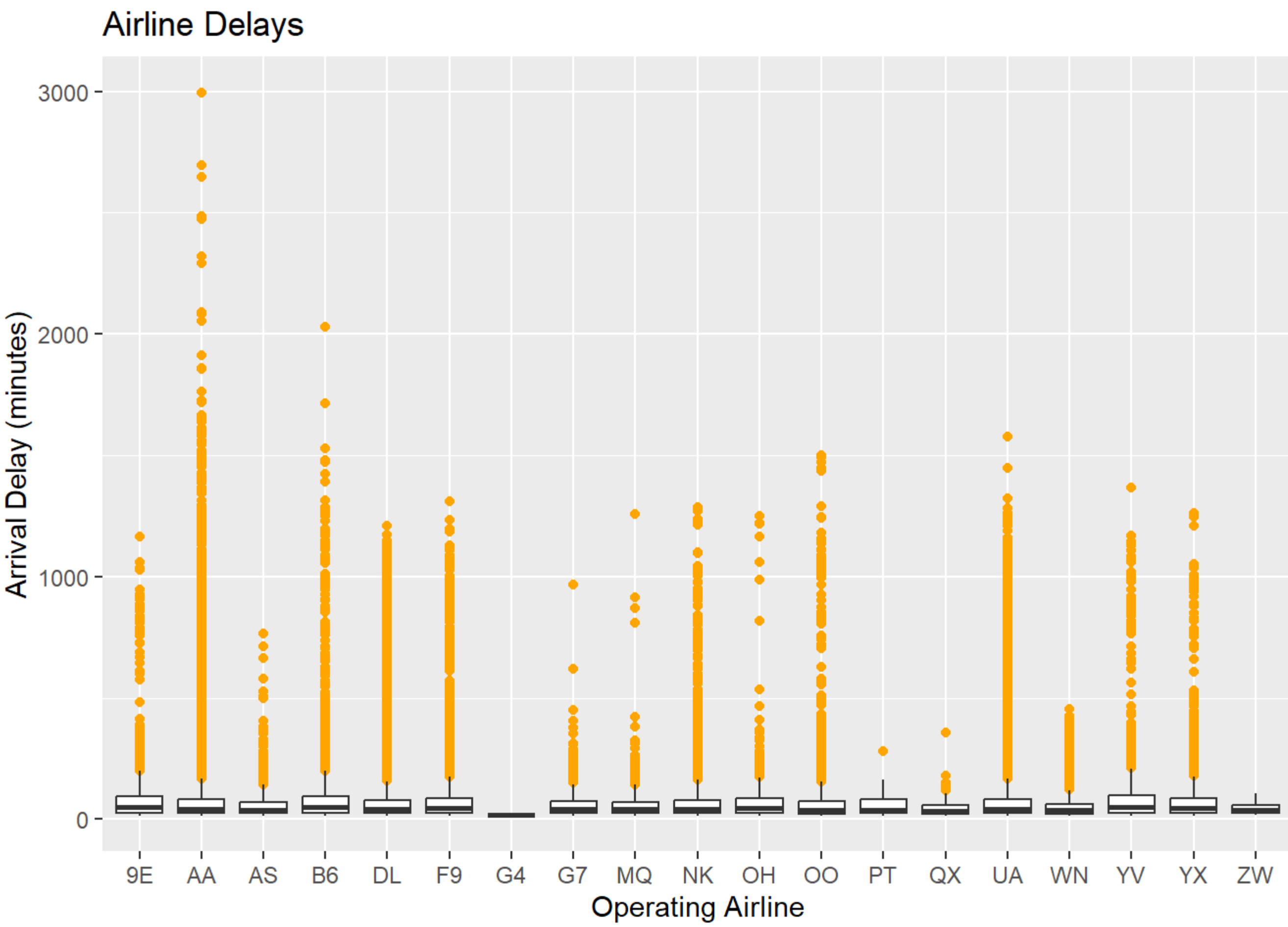



```
#tmpFile <- tempfile(fileext = ".png")
#export(fig2, file = #tmpFile)
```

Delays

Box Plot - Overall Delays per carrier

```
ggplot(carrier_delay_df, aes(x=ARR_DELAY, y=OP_UNIQUE_CARRIER))+
  geom_boxplot(outlier.colour="orange", outlier.shape=16) +
  labs(title ="Airline Delays",
       y = "Operating Airline",
       x = "Arrival Delay (minutes)") + coord_flip()
```



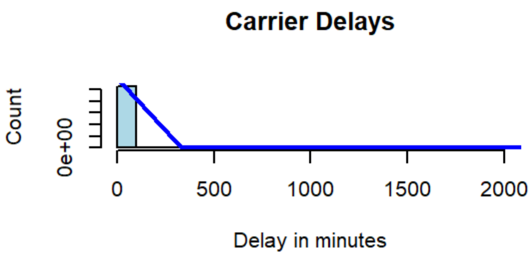
There are more delays due to carrier and late aircrafts than weather, NAS or security delays.

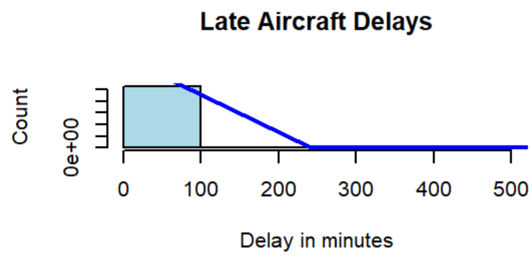
Histogram

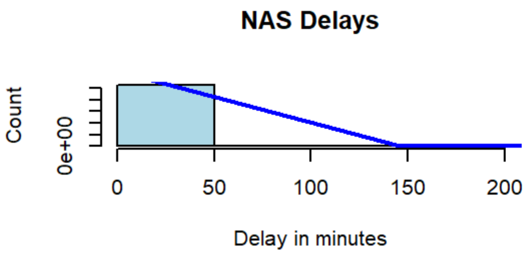
Histogram for Delay Reasons

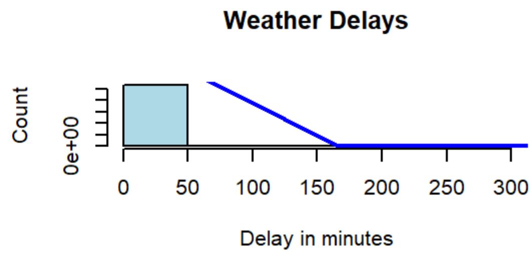
```
carrier_performance_df[is.na(carrier_performance_df$CARRIER_DELAY),]$CARRIER_DELAY <- 0
Carrier_Delay <- carrier_performance_df$CARRIER_DELAY

carrier_performance_df[is.na(carrier_performance_df$LATE_AIRCRAFT_DELAY),]$LATE_AIRCRAFT_DELAY
<-0
LateAircraft_Delay <- carrier_performance_df$LATE_AIRCRAFT_DELAY

carrier_performance_df[is.na(carrier_performance_df$NAS_DELAY),]$NAS_DELAY<-0
NAS_Delay <- carrier_performance_df$NAS_DELAY

carrier_performance_df[is.na(carrier_performance_df$WEATHER_DELAY),]$WEATHER_DELAY<-0
Weather_Delay <- carrier_performance_df$WEATHER_DELAY

carrier_performance_df[is.na(carrier_performance_df$SECURITY_DELAY),]$SECURITY_DELAY<-0
Security_Delay <- carrier_performance_df$SECURITY_DELAY

par(mar=c(5,5,5,5))
par(mfrow = c(3, 2))
h <- hist(Carrier_Delay, main = "Carrier Delays",xlab ="Delay in minutes", ylab="Count",col="L
ight Blue",xlim = c(0,2000))
```

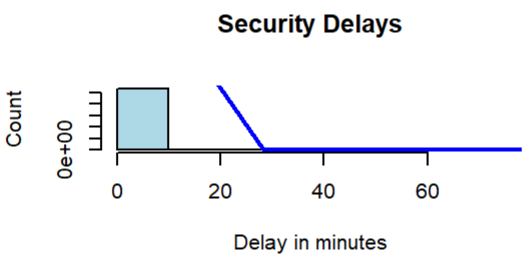
```
xfit<-seq(min(Carrier_Delay),max(Carrier_Delay),length=10)
yfit<-dnorm(xfit,mean=mean(Carrier_Delay),sd=sd(Carrier_Delay))
yfit <- yfit*diff(h$mids[1:2])*length(Carrier_Delay)
lines(xfit, yfit, col="blue", lwd=2)

h <- hist(LateAircraft_Delay, main = "Late Aircraft Delays", xlab = "Delay in minutes",ylab="C
ount",col="Light Blue",xlim = c(0,500))
xfit<-seq(min(LateAircraft_Delay),max(LateAircraft_Delay),length=10)
yfit<-dnorm(xfit,mean=mean(LateAircraft_Delay),sd=sd(LateAircraft_Delay))
yfit <- yfit*diff(h$mids[1:2])*length(LateAircraft_Delay)
lines(xfit, yfit, col="blue", lwd=2)


h <- hist(NAS_Delay, main = "NAS Delays",xlab = "Delay in minutes",ylab="Count",col="Light Blu
e", xlim = c(0,200))
xfit<-seq(min(NAS_Delay),max(NAS_Delay),length=10)
yfit<-dnorm(xfit,mean=mean(NAS_Delay),sd=sd(NAS_Delay))
yfit <- yfit*diff(h$mids[1:2])*length(NAS_Delay)
lines(xfit, yfit, col="blue", lwd=2)

h <- hist(Weather_Delay, main = "Weather Delays",xlab = "Delay in minutes", ylab="Count",col="
Light Blue",xlim = c(0,300)) #, ylim=c(0, 100000),
xfit<-seq(min(Weather_Delay),max(Weather_Delay),length=10)
yfit<-dnorm(xfit,mean=mean(Weather_Delay),sd=sd(Weather_Delay))
yfit <- yfit*diff(h$mids[1:2])*length(Weather_Delay)
lines(xfit, yfit, col="blue", lwd=2)

h <- hist(Security_Delay, main = "Security Delays", xlab = "Delay in minutes", ylab="Count",co
l="Light Blue",xlim = c(0,75))
xfit<-seq(min(Security_Delay),max(Security_Delay),length=10)
yfit<-dnorm(xfit,mean=mean(Security_Delay),sd=sd(Security_Delay))
yfit <- yfit*diff(h$mids[1:2])*length(Security_Delay)
lines(xfit, yfit, col="blue", lwd=2)
```





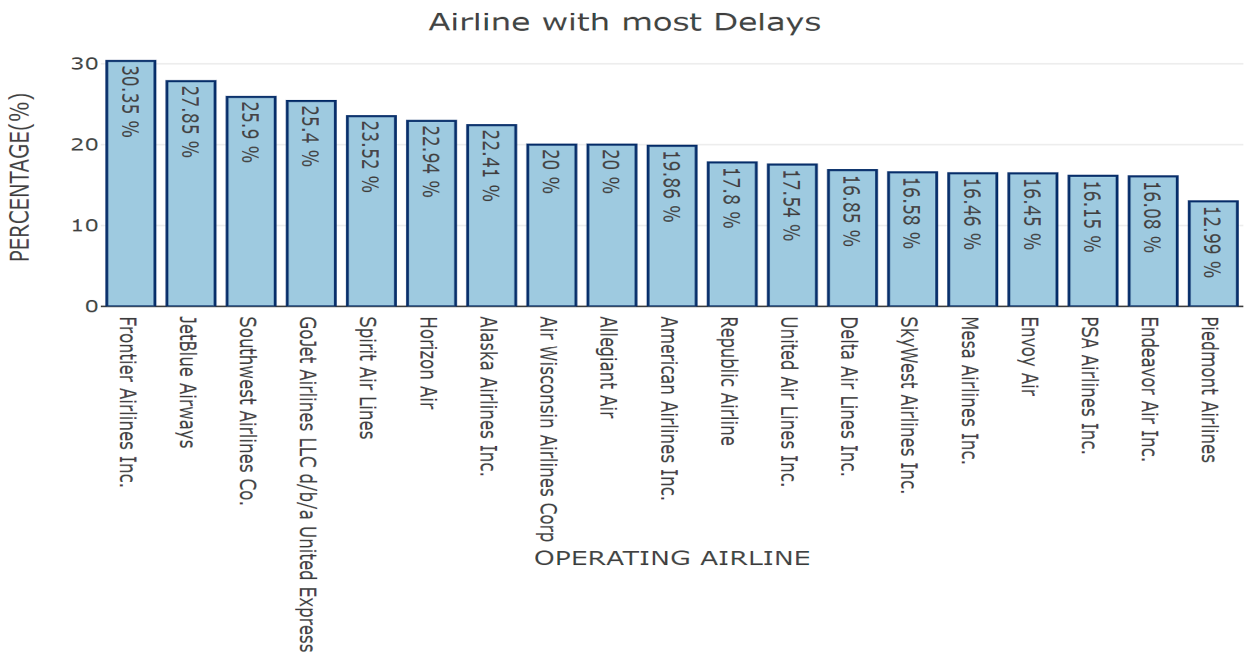






Carrier with most delays by performance percentage

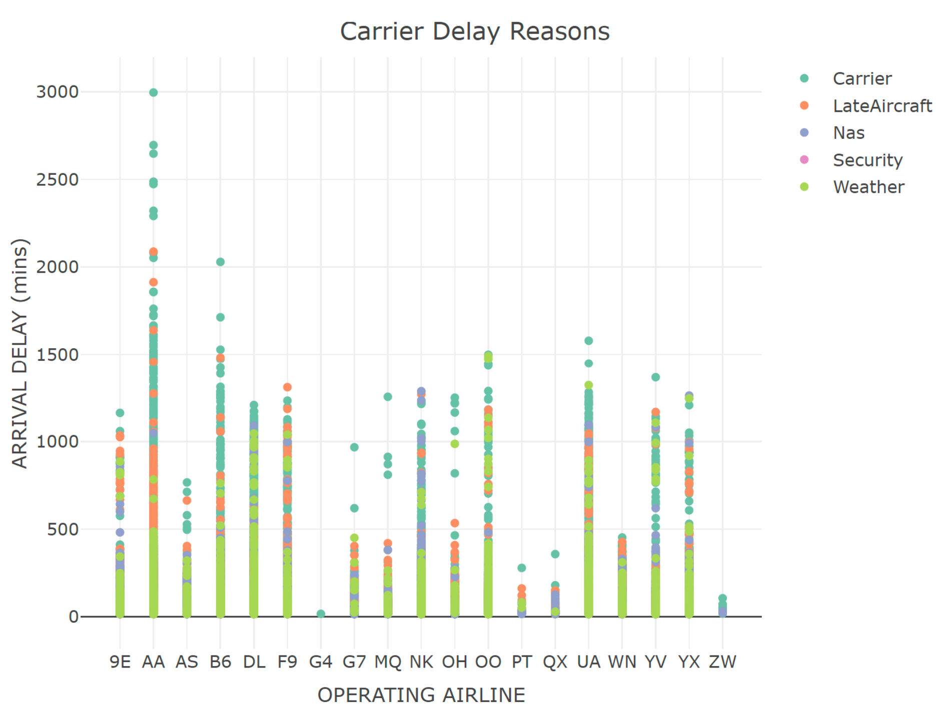
```
delayed_performance <- flight_status[flight_status$STATUS == 'Delayed',]
delayed_performance <- delayed_performance[order(delayed_performance$PERCENTAGE,decreasing=TRU
E),]


fig <- plot_ly(delayed_performance, x = ~OP_UNIQUE_CARRIER_NAME, y = ~PERCENTAGE,
               type = 'bar',text = ~paste(PERCENTAGE,'%'), textposition = 'auto',
               marker = list(color = 'rgb(158,202,225)',
                            line = list(color = 'rgb(8,48,107)', width = 1.5)))
fig <- fig %>% layout(title = "Airline with most Delays",
         xaxis = list(title = "OPERATING AIRLINE"),
         yaxis = list(title = "PERCENTAGE(%)"))
fig <- fig %>% layout(xaxis = list(categoryorder = "total descending"))

fig
```




```
#tmpFile <- tempfile(fileext = ".png")
#export(fig, file = #tmpFile)
```

Carriers vs Delay Reasons

```
fig <- plot_ly(carrier_delay_df, x = ~OP_UNIQUE_CARRIER,y = ~ARR_DELAY,
               color=~DELAY_REASON, type = 'scatter')
fig <- fig %>% layout(title = "Carrier Delay Reasons",
                      xaxis = list(title = "OPERATING AIRLINE"),
                      yaxis = list(title = "ARRIVAL DELAY (mins)"))

fig
```

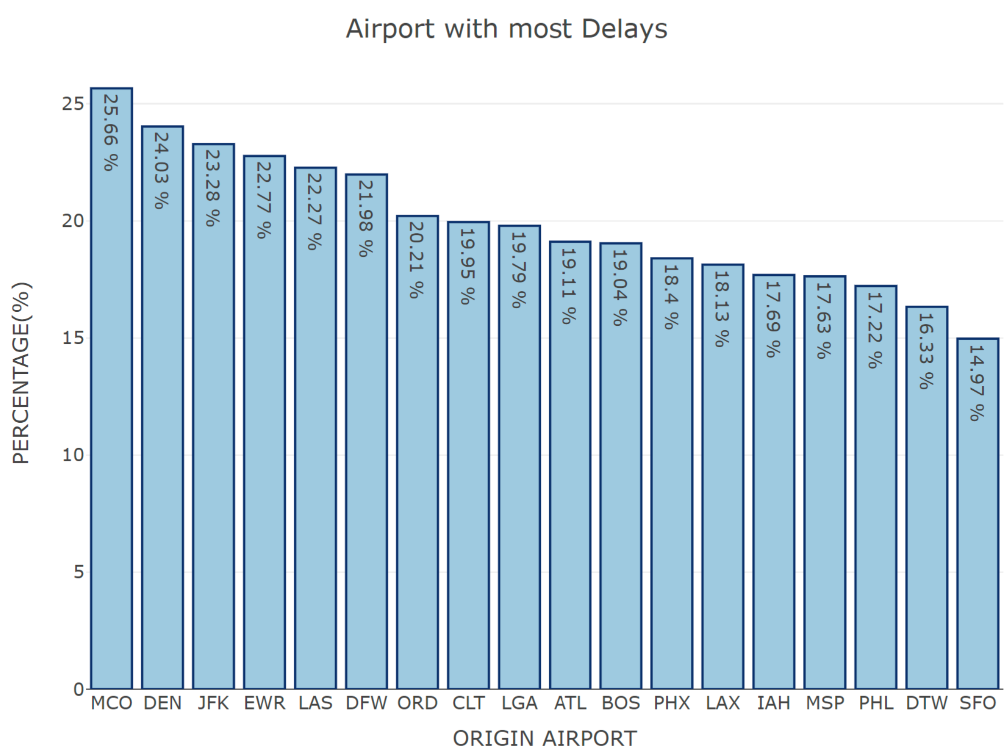


```
#tmpFile <- tempfile(fileext = ".png")
#export(fig, file = #tmpFile)
```

Frontier Airlines has the most number of delays, followed by JetBlue Airways. Piedmont and Endeavor air have the least delays.

Origin Airport vs Arrival Delays

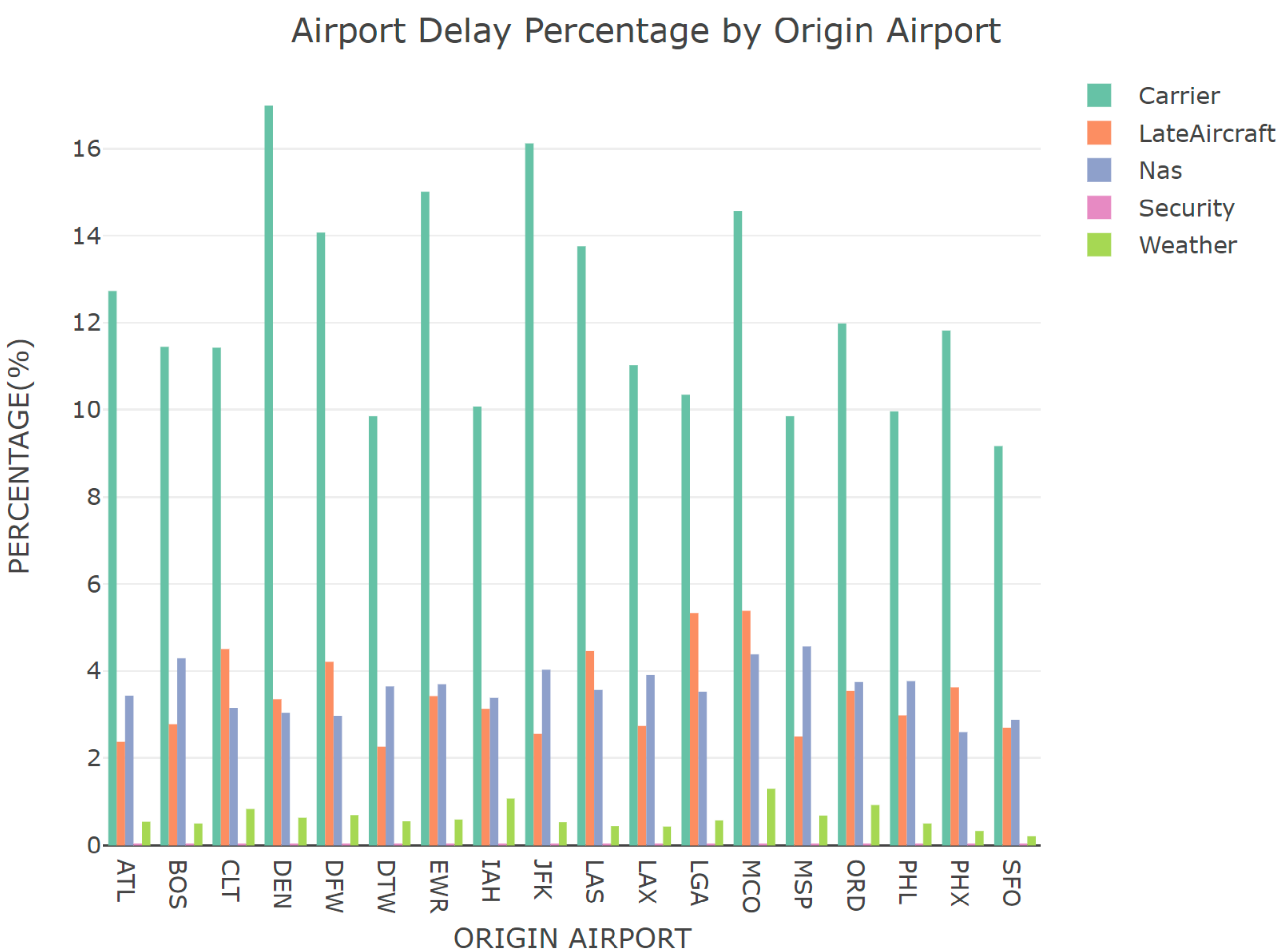
```
#tmpFile <- tempfile(fileext = ".png")
#export(fig, file = tmpFile)
df <- delayed_status[delayed_status$STATUS=='Delayed',]

fig <- plot_ly(df, x = ~ORIGIN, y = ~PERCENTAGE,
               type = 'bar',text = ~paste(PERCENTAGE,'%'), textposition = 'auto',
               marker = list(color = 'rgb(158,202,225)',
                            line = list(color = 'rgb(8,48,107)', width = 1.5)))
fig <- fig %>% layout(title = "Airport with most Delays",
         xaxis = list(title = "ORIGIN AIRPORT"),
         yaxis = list(title = "PERCENTAGE(%)"))
fig <- fig %>% layout(xaxis = list(categoryorder = "total descending"))

fig
```



Orlando has the most number of delays for the given origin destination pairs available in the dataset. San Francisco (SFO) although one of the busiest airports has fewer delays than other airports.

Origin Airport vs Delay Reasons

```
df <- delayed_reason_status[delayed_reason_status$STATUS=="Delayed",]

fig <- plot_ly(df, x = ~ORIGIN,y = ~PERCENTAGE,color=~DELAY_REASON,type = 'bar')
fig <- fig %>% layout(title = "Airport Delay Percentage by Origin Airport",
                      xaxis = list(title = "ORIGIN AIRPORT"),
                      yaxis = list(title = "PERCENTAGE(%)"))
fig
```

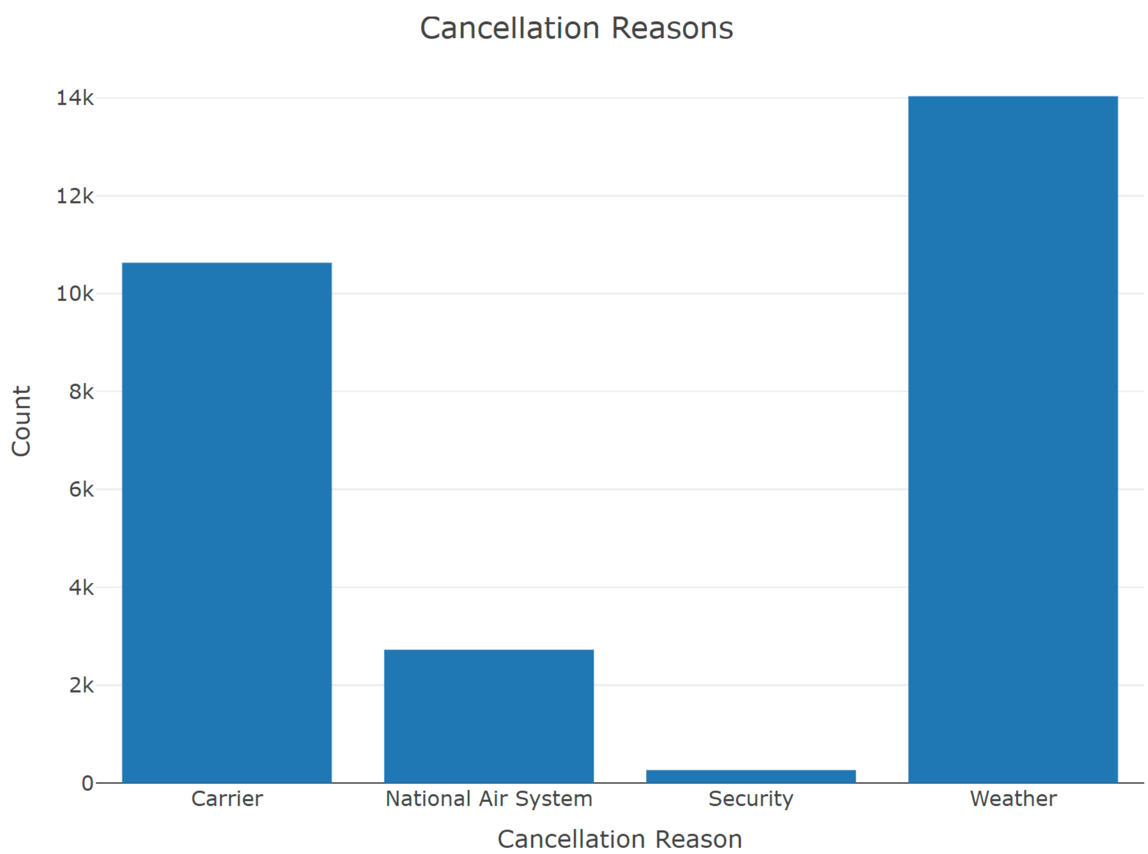


Cancellations

Histogram - Overall cancellations

```
fig <- plot_ly(carrier_cancelled_df, x = ~CANCELLATION_REASON,type = 'histogram')
```

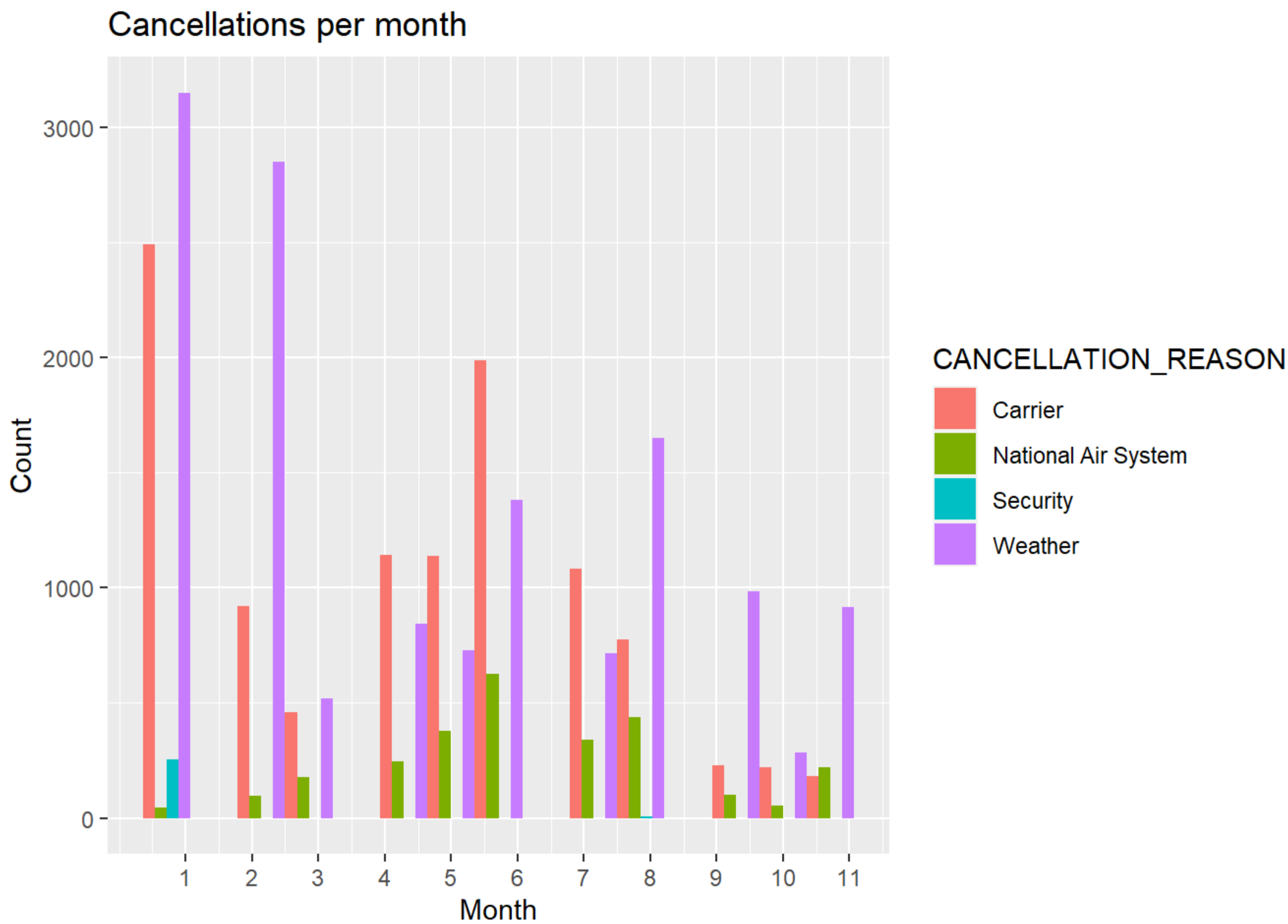
```
fig <- fig %>% layout(title = "Cancellation Reasons",
                      xaxis = list(title = "Cancellation Reason"),
                      yaxis = list(title = "Count"))
fig
```



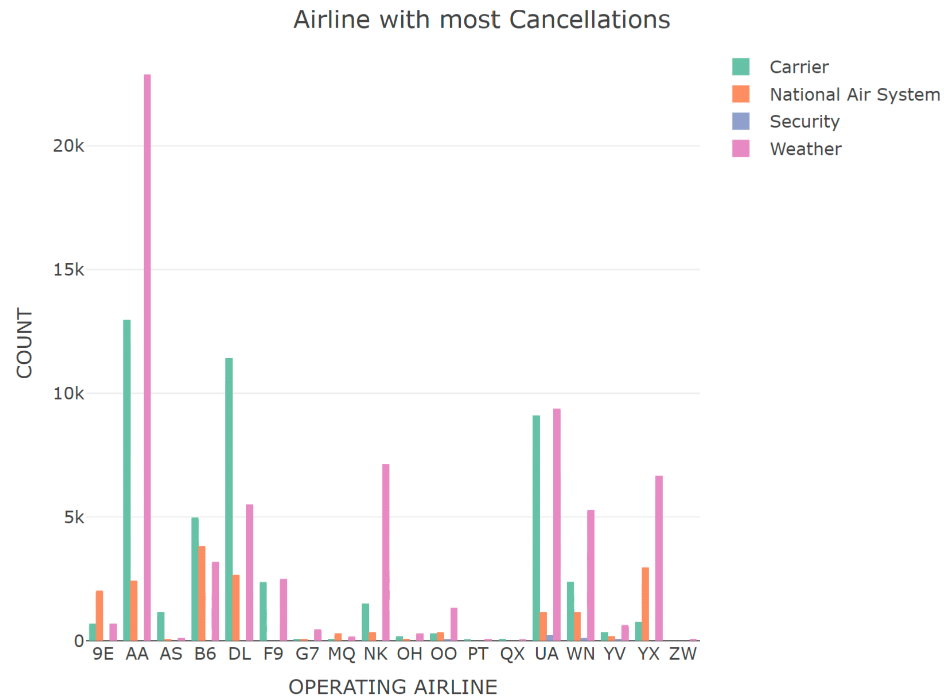

```
#tmpFile <- tempfile(fileext = ".png")
#export(fig, file = #tmpFile)
```

Carriers vs Cancellation Reasons

```
ggplot(carrier_cancelled_df, aes(x=MONTH, fill=CANCELLATION_REASON)) +
  geom_histogram(bins=15,position = "dodge") + xlab("Month") + ylab("Count") +
  scale_x_continuous(breaks = seq(1, 12, by = 1)) +
  ggtitle("Cancellations per month")
```



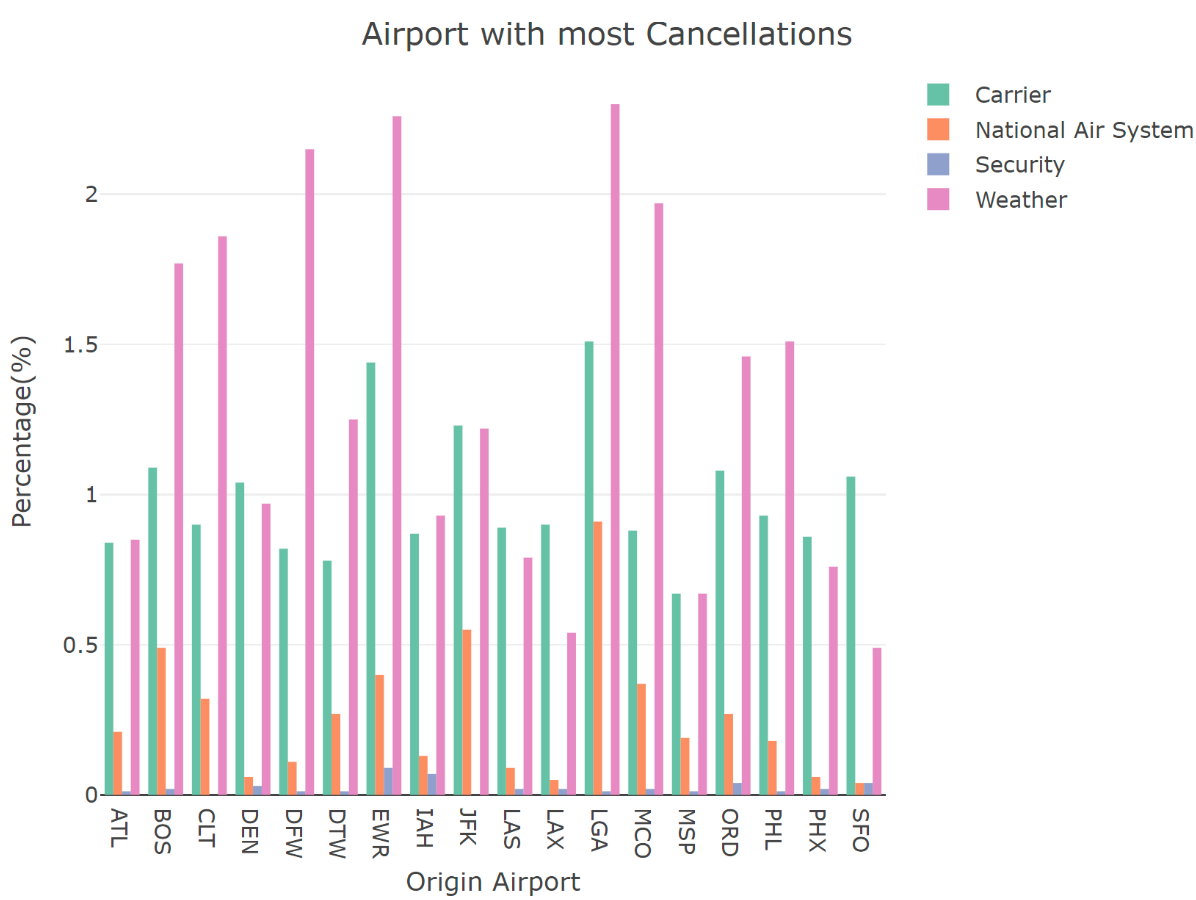

```
#ggplot(carrier_cancelled_df, aes(OP_UNIQUE_CARRIER,MONTH,fill=CANCELLATION_REASON)) +
#   geom_bar(stat="identity", position = "dodge") +
#   ggtitle("Cancellations by Carrier") + xlab("Operating Airline") + ylab("Month")



fig <- plot_ly(carrier_cancelled_df, x = ~OP_UNIQUE_CARRIER,
               y = ~MONTH,color=~CANCELLATION_REASON,type = 'bar')
fig <- fig %>% layout(title = "Airline with most Cancellations",
                      xaxis = list(title = "OPERATING AIRLINE"),
                      yaxis = list(title = "COUNT"))
fig
```



Weather is a major reason for cancellation in winters (January and February). Carrier cancellations are also high during this period.

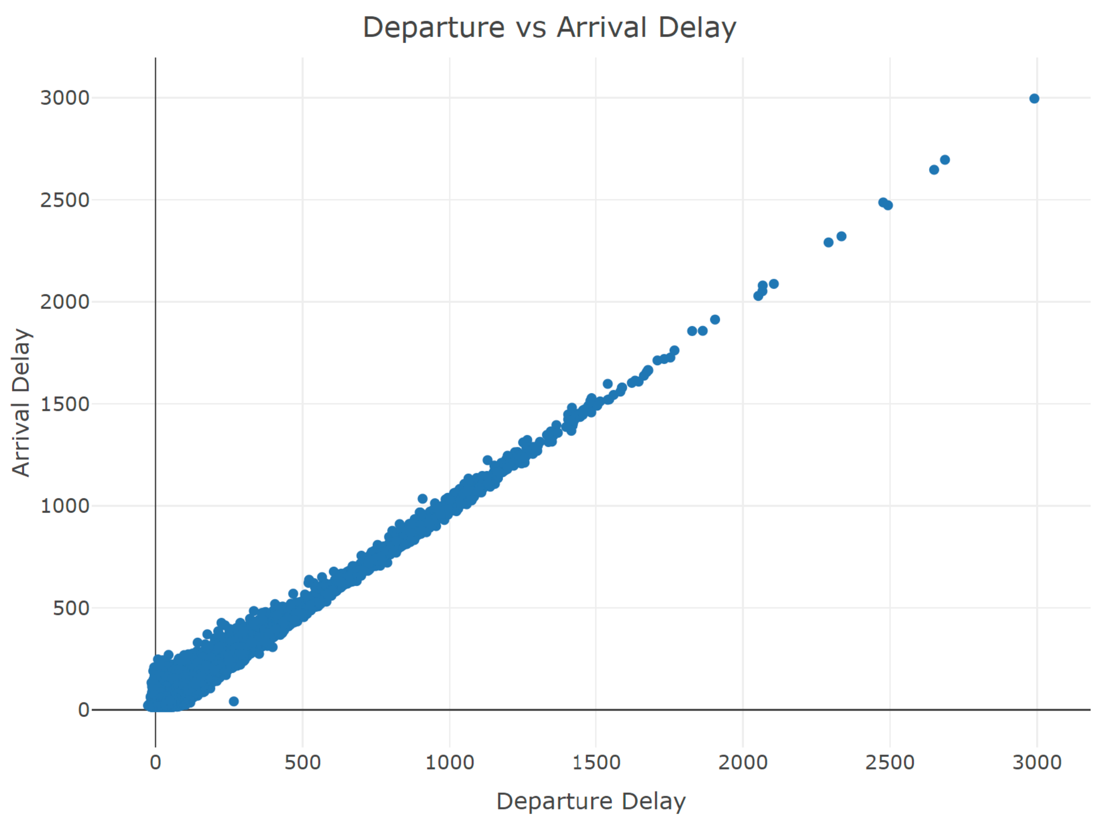
Airport vs Cancellation Reasons

```
fig <- plot_ly(cancelled_status, x = ~ORIGIN,
               y = ~PERCENTAGE,color=~CANCELLATION_REASON,type = 'bar')
fig <- fig %>% layout(title = "Airport with most Cancellations",
                      xaxis = list(title = "Origin Airport"),
                      yaxis = list(title = "Percentage(%)"))
fig
```



Scatter Plot

```
fig <- plot_ly(carrier_delay_df, x = ~DEP_DELAY,
               y = ~ARR_DELAY,type = 'scatter')
fig <- fig %>% layout(title = "Departure vs Arrival Delay",
                      xaxis = list(title = "Departure Delay"),
                      yaxis = list(title = "Arrival Delay"))
fig
```




```
#tmpFile <- tempfile(fileext = ".png")
#export(fig, file = #tmpFile)
```

What do you not know how to do right now that you need to learn to answer your questions?

I would like to learn more on the machine learning concepts to use in my final project.

Do you plan on incorporating any machine learning techniques to answer your research questions? Explain.

Not at this time, but would like to consider incorporating them based on week 11 and 12 learnings.

Questions

It is unclear if I would be able to recommend the right area of focus for better performance, to the airlines. Delays are high For example: If the majority of delays are due to NAS - National Air System Delay, it could mean there was an issue in one or more areas such as mechanical, crew, airport operations etc. I would need to identify another dataset that logs the maintenance or operational issues by carrier. This information could be hard to get as it is carrier specific and probably not allowed to be made public.

Outcomes

1. Are small carriers reliable in terms of lesser cancellations and delays? Frontier has the maximum number of delays whereas Piedmont has the least delays. It is unclear if small carriers are more reliable.

2. Which carrier has the best on-time performance. American Airlines Inc, Delta Airlines and United Airlines have the best performance.

3. Which carrier has the least on-time performance. Allegiant Air,Air Wisconsin Airlines Corp , Piedmont Airlines , Horizon Air ,GoJet Airlines LLC have the least on-time performance

4. Identifying the most common cancellation reason for all carriers. Based on thre 1 million rows of data, weather cancellations are the most common.

5. Which carrier has the most number of cancellations. Air Wisconsin has the most cancellations.

6. Which carrier has the most number of delays. Frontier Airlines has the most delays.

7. What is the percentage of delays by reason.

```
head(flight_stats,20)
```

```
## # A tibble: 20 × 5
## # Groups:   OP_UNIQUE_CARRIER, OP_UNIQUE_CARRIER_NAME, DELAY_REASON [20]
##    OP_UNIQUE_CARRIER OP_UNIQUE_CARRIER_NAME DELAY_REASON  COUNT PERCENTAGE
##    <chr>             <chr>                  <chr>         <int>      <dbl>
##  1 9E                Endeavor Air Inc.      Carrier         829       6.59
##  2 9E                Endeavor Air Inc.      LateAircraft    556       4.42
##  3 9E                Endeavor Air Inc.      Nas             559       4.45
##  4 9E                Endeavor Air Inc.      Security          1       0.01
##  5 9E                Endeavor Air Inc.      Weather          77       0.61
##  6 9E                Endeavor Air Inc.      <NA>          10553      83.9
##  7 AA                American Airlines Inc. Carrier       30736      12.0
##  8 AA                American Airlines Inc. LateAircraft  10606       4.14
##  9 AA                American Airlines Inc. Nas            7621       2.97
## 10 AA                American Airlines Inc. Security         70       0.03
## 11 AA                American Airlines Inc. Weather        1886       0.74
```

```
## 12 AA              American Airlines Inc. <NA>         205533      80.1
## 13 AS              Alaska Airlines Inc.   Carrier         865       6.85
## 14 AS              Alaska Airlines Inc.   LateAircraft    783       6.2
## 15 AS              Alaska Airlines Inc.   Nas            1126       8.92
## 16 AS              Alaska Airlines Inc.   Security          5       0.04
## 17 AS              Alaska Airlines Inc.   Weather          51       0.4
## 18 AS              Alaska Airlines Inc.   <NA>           9796      77.6
## 19 B6              JetBlue Airways        Carrier       15932      20.8
## 20 B6              JetBlue Airways        LateAircraft   1877       2.46
```

Limitations

The dataset used for this analysis has around 6 million rows. For purposes of analysis, I stripped data to 1 million rows. The outcomes mentioned could change with more data. Restricting analysis to major airports could be omitting many performance aspects of airlines. It would be nice to run the analysis with years of data to average the findings. The huge size of dataset made the process extremely slow with several application crashes. Moreover, another inherent challenge of the dataset was that there were limited variables that could be used. Many columns were inapplicable to the analysis (i.e. TAXI_OUT, TAXI_IN, AIR_TIME etc.) and so the analysis was done on limited variables. Additional information such as weather, NAS issue etc., could open more areas for analysis.

Conclusion

It was very exciting for me to analyze this dataset. I found myself surprised at several instances. I assumed most cancellations would be because of weather but on adding more parameters in the process of data cleaning, I noticed that most cancellations are actually due to Carrier and not weather. I wasn''t able to show this in the analysis due to data size restrictions. This was a great experience in gaining better understanding on how to work with datasets and understanding the significance of each step. As next steps, I would like to calculate the delay percentage of flights at each interval of arrival delay such as (0-15, <15, >15 - <30, >30) to validate the average delay time.

Citations

(*Airline on-Time Statistics and Delay Causes*, n.d.)

Airline on-Time Statistics and Delay Causes. n.d. https://www.transtats.bts.gov/OT_Delay/OT_DelayCause1.asp?20=E.