Final Project

Aarti Ramani

2023-02-07

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

Introduction:

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 their 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.

Column Name	Data Type	Column Description
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)
```

How to import and clean my data

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 x 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 x 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 x 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 x 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

```
# Selecting relevant columns from flights data

carrier_performance_df <-
  merge_flights_df[c("YEAR","QUARTER","MONTH","DAY_OF_MONTH","DAY_OF_WEEK",
                     "FL_DATE","MKT_UNIQUE_CARRIER","OP_UNIQUE_CARRIER",
                     "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","DIVERTED",
                     "DISTANCE","CARRIER_DELAY","WEATHER_DELAY","NAS_DELAY",
                     "SECURITY_DELAY","LATE_AIRCRAFT_DELAY")]

# Transforming Data

#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$CANCELLATION_REASON <-
  cancellation_cd$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)]

#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)]
```

```r
#Removing null rows from the dataet
carrier_performance_df <-
  carrier_performance_df[,colSums(is.na(carrier_performance_df))<nrow(carrier_performance_df)]

#Updating blank arrival delay to 0
carrier_performance_df[is.na(carrier_performance_df$ARR_DELAY),]$ARR_DELAY = 0

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

What does the final data set look like?

```r
nrow(carrier_performance_df)
```

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

```r
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] "ORIGIN"                  "ORIGIN_CITY_NAME"
## [11] "ORIGIN_STATE_ABR"        "ORIGIN_STATE_NM"
## [13] "DEST"                    "DEST_CITY_NAME"
## [15] "DEST_STATE_ABR"          "DEST_STATE_NM"
## [17] "DEP_DELAY"               "TAXI_OUT"
## [19] "TAXI_IN"                 "ARR_DELAY"
## [21] "CANCELLED"               "CANCELLATION_CODE"
## [23] "DIVERTED"                "DISTANCE"
## [25] "CARRIER_DELAY"           "WEATHER_DELAY"
## [27] "NAS_DELAY"               "SECURITY_DELAY"
## [29] "LATE_AIRCRAFT_DELAY"     "CANCELLATION_REASON"
## [31] "MKT_UNIQUE_CARRIER_NAME" "OP_UNIQUE_CARRIER_NAME"
## [33] "ORIGIN_AIRPORT"          "DEST_AIRPORT"
## [35] "Performance"
```

What information is not self-evident?

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.

What are different ways you could look at this data ?

I would like to perform the following:

1. Percentages of flights in and out per airline.
2. Percentages of flights in 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?

Currently I am splitting dataset into 2 categories.

1. no cancellations and delays (on-time performace)
2. cancellations or 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),]
nrow(carrier_on_time_performance_df) + nrow(carrier_cancel_or_delay_df)
```

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

Splitting cancelled and delayed data.

Delayed Dataset

```
carrier_delay_df <-
  carrier_cancel_or_delay_df[carrier_cancel_or_delay_df$ARR_DELAY > 15,]
nrow(carrier_delay_df)
```

```
## [1] 1236619
```

Cancelled Dataset

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

```
## [1] 158851
```

```
nrow(carrier_delay_df)
```

```
## [1] 1236619
```

```
nrow(carrier_cancelled_df)
```

```
## [1] 158851
```

```
nrow(carrier_on_time_performance_df)
```

```
## [1] 5039717
```

```
nrow(carrier_performance_df)
```

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

Apart from the above, since the number of rows are very high at this point, I will narrow my research to flights origination from major 20 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
```

```
carrier_delay_df <-
  carrier_delay_df[carrier_delay_df$ORIGIN == "ORD"
                   | carrier_delay_df$ORIGIN == "ATL"
                   | carrier_delay_df$ORIGIN == "DFW"
```

```
                    | carrier_delay_df$ORIGIN == "DEN"
                    | carrier_delay_df$ORIGIN == "EWR"
                    | carrier_delay_df$ORIGIN == "LAX"
                    | carrier_delay_df$ORIGIN == "IAH"
                    | carrier_delay_df$ORIGIN == "PHX"
                    | carrier_delay_df$ORIGIN == "DTW"
                    | carrier_delay_df$ORIGIN == "SFO"
                    | carrier_delay_df$ORIGIN == "LAS"
                    | carrier_delay_df$ORIGIN == "DEN"
                    | carrier_delay_df$ORIGIN == "ORD"
                    | carrier_delay_df$ORIGIN == "JFK"
                    | carrier_delay_df$ORIGIN == "CLT"
                    | carrier_delay_df$ORIGIN == "LGA"
                    | carrier_delay_df$ORIGIN == "MCO"
                    | carrier_delay_df$ORIGIN == "MSP"
                    | carrier_delay_df$ORIGIN == "BOS"
                    | carrier_delay_df$ORIGIN == "PHL",]
nrow(carrier_delay_df)
```

```
## [1] 602746
```

```
carrier_cancelled_df <-
  carrier_cancelled_df[carrier_cancelled_df$ORIGIN == "ORD"
                       | carrier_cancelled_df$ORIGIN == "ATL"
                       | carrier_cancelled_df$ORIGIN == "DFW"
                       | carrier_cancelled_df$ORIGIN == "DEN"
                       | carrier_cancelled_df$ORIGIN == "EWR"
                       | carrier_cancelled_df$ORIGIN == "LAX"
                       | carrier_cancelled_df$ORIGIN == "IAH"
                       | carrier_cancelled_df$ORIGIN == "PHX"
                       | carrier_cancelled_df$ORIGIN == "DTW"
                       | carrier_cancelled_df$ORIGIN == "SFO"
                       | carrier_cancelled_df$ORIGIN == "LAS"
                       | carrier_cancelled_df$ORIGIN == "DEN"
                       | carrier_cancelled_df$ORIGIN == "ORD"
                       | carrier_cancelled_df$ORIGIN == "JFK"
                       | carrier_cancelled_df$ORIGIN == "CLT"
                       | carrier_cancelled_df$ORIGIN == "LGA"
                       | carrier_cancelled_df$ORIGIN == "MCO"
                       | carrier_cancelled_df$ORIGIN == "MSP"
                       | carrier_cancelled_df$ORIGIN == "BOS"
                       | carrier_cancelled_df$ORIGIN == "PHL",]

nrow(carrier_cancelled_df)
```

```
## [1] 75560
```

```
carrier_on_time_performance_df <-
  carrier_on_time_performance_df[carrier_on_time_performance_df$ORIGIN == "ORD"
                                 | carrier_on_time_performance_df$ORIGIN == "ATL"
                                 | carrier_on_time_performance_df$ORIGIN == "DFW"
                                 | carrier_on_time_performance_df$ORIGIN == "DEN"
```

```
                                  | carrier_on_time_performance_df$ORIGIN == "EWR"
                                  | carrier_on_time_performance_df$ORIGIN == "LAX"
                                  | carrier_on_time_performance_df$ORIGIN == "IAH"
                                  | carrier_on_time_performance_df$ORIGIN == "PHX"
                                  | carrier_on_time_performance_df$ORIGIN == "DTW"
                                  | carrier_on_time_performance_df$ORIGIN == "SFO"
                                  | carrier_on_time_performance_df$ORIGIN == "LAS"
                                  | carrier_on_time_performance_df$ORIGIN == "DEN"
                                  | carrier_on_time_performance_df$ORIGIN == "ORD"
                                  | carrier_on_time_performance_df$ORIGIN == "JFK"
                                  | carrier_on_time_performance_df$ORIGIN == "CLT"
                                  | carrier_on_time_performance_df$ORIGIN == "LGA"
                                  | carrier_on_time_performance_df$ORIGIN == "MCO"
                                  | carrier_on_time_performance_df$ORIGIN == "MSP"
                                  | carrier_on_time_performance_df$ORIGIN == "BOS"
                                  | carrier_on_time_performance_df$ORIGIN == "PHL",]

nrow(carrier_on_time_performance_df)
```

```
## [1] 2338688
```

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.

Plots & Tables

Plots that I would like to explore:

i. Scatter plot
ii. Pie chart

iii. Histogram
iv. Boxplot

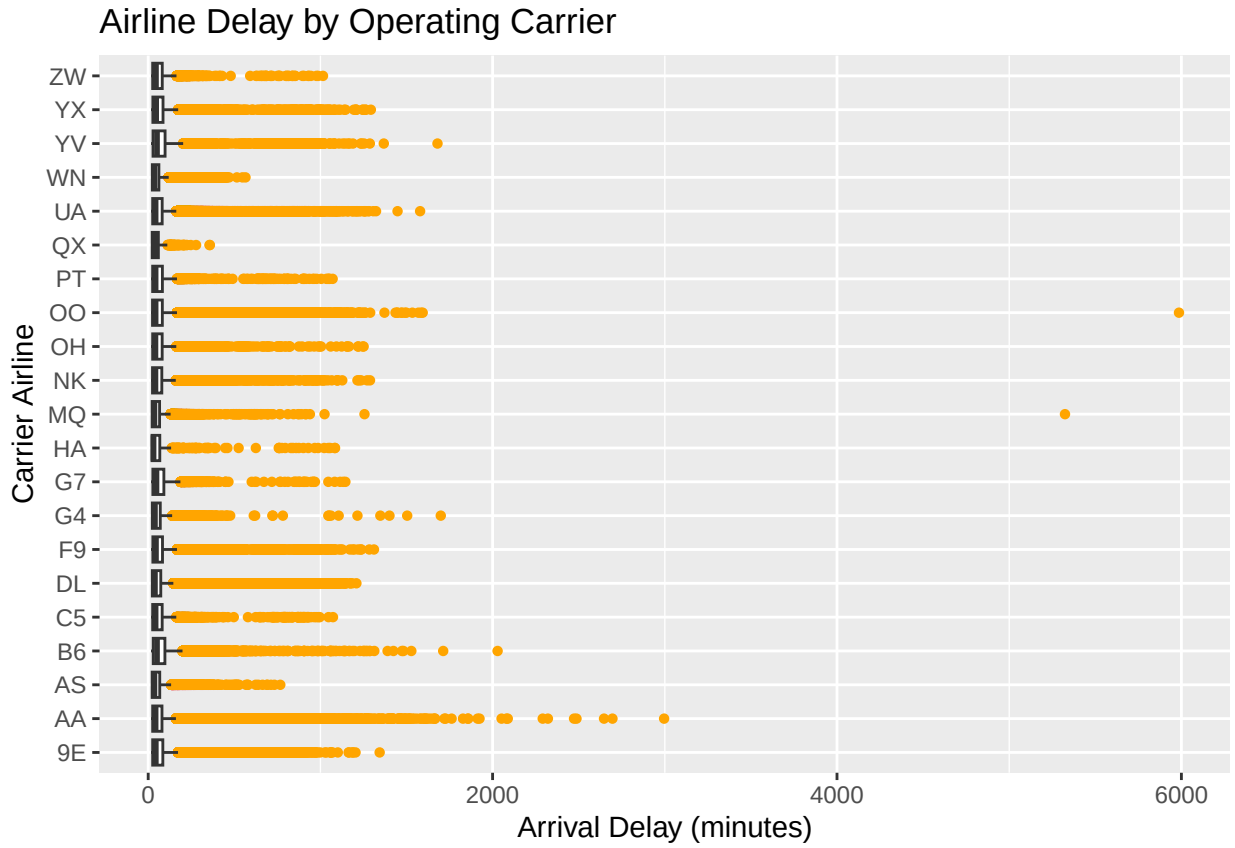
I will create tables with the following data:

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?

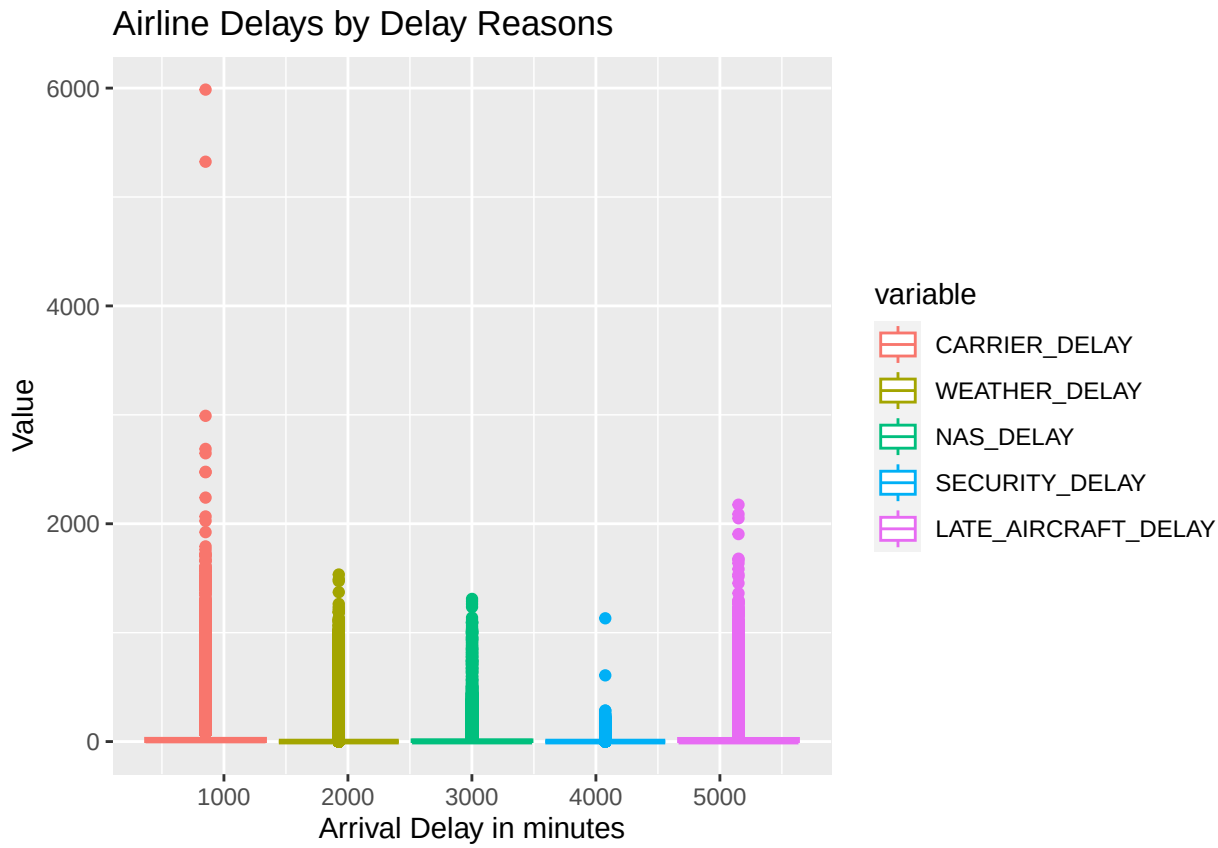
BOXPLOT

```
#Boxplot for checking Outliers in Airline Delays
ggplot(carrier_delay_df, aes(x=ARR_DELAY, y=OP_UNIQUE_CARRIER))+
  geom_boxplot(outlier.colour="orange", outlier.shape=16) +
  labs(title ="Airline Delay by Operating Carrier",
       y = "Carrier Airline",
       x = "Arrival Delay (minutes)")
```



We can see many outliers in the delay dataset which can be omitted. We can skip rows with arrival delay greater than 2000 minutes. Since 90% of the data falls below 2000 minutes. Also, we can see more outliers for the carriers American Airlines, Mesa,Skywest and Republic airlines.

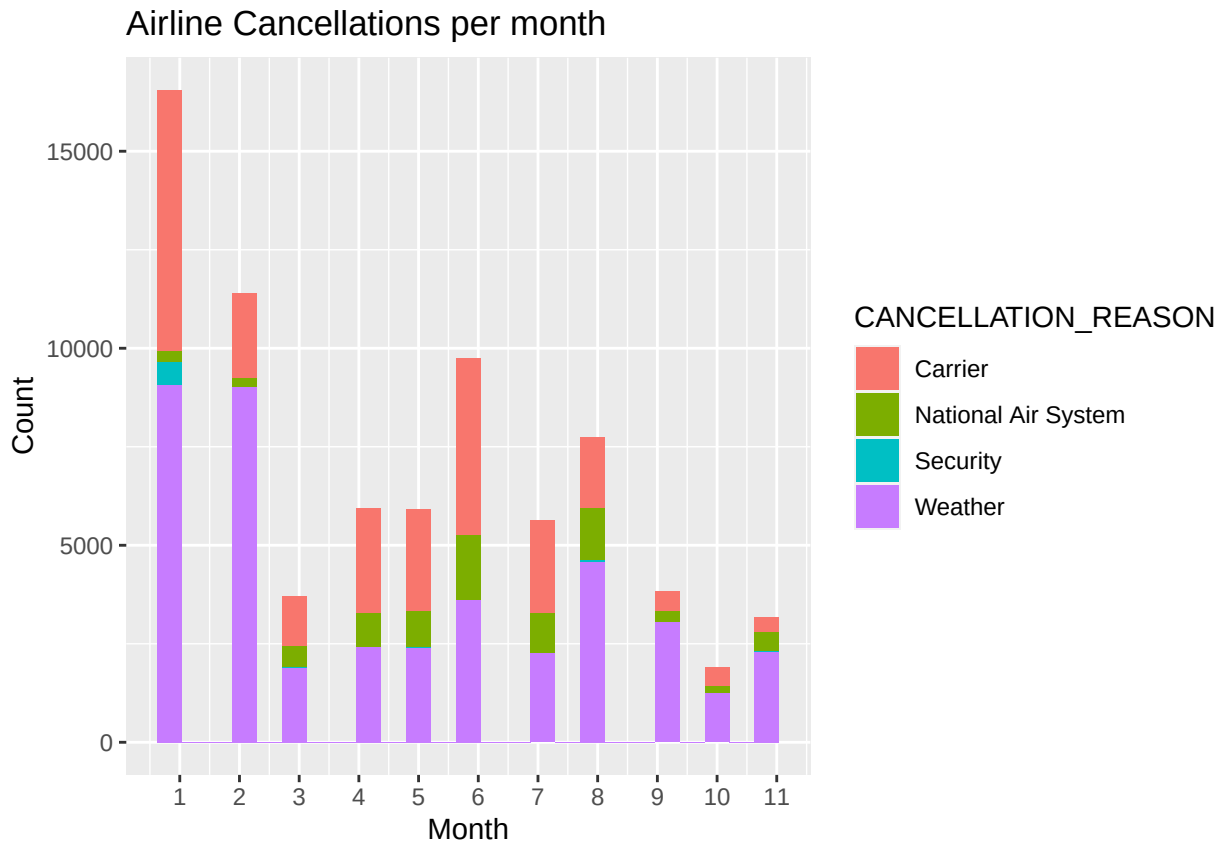
```
# Box Plot for Airline Delays by Delay Reasons

carrier_delay_df_mod <- melt(carrier_delay_df, id.vars='ARR_DELAY',
                  measure.vars=c("CARRIER_DELAY","WEATHER_DELAY","NAS_DELAY",
                                 "SECURITY_DELAY","LATE_AIRCRAFT_DELAY"))
# creating a plot
ggplot(carrier_delay_df_mod) +
  geom_boxplot(aes(x=ARR_DELAY, y=value, color=variable)) +
  xlab("Arrival Delay in minutes") + ylab("Value") + ggtitle("Airline Delays by Delay Reasons")
```



It appears there are more delays due to carrier and late aircrafts than weather, NAS or security delays. We can also see a few outliers in the delay data that can be removed for the analysis.

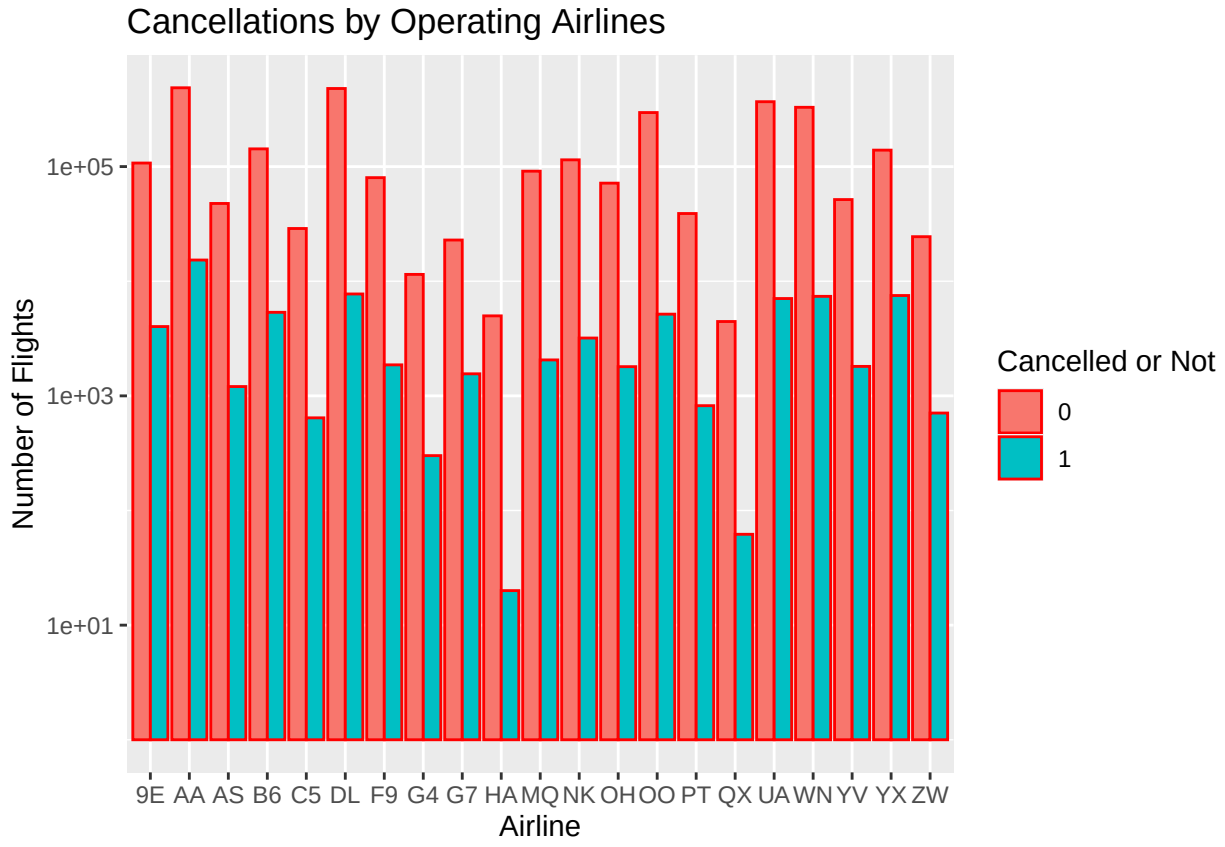
HISTOGRAM

```
# Histogram to  the Cancellations due to weather
ggplot(carrier_cancelled_df, aes(x=MONTH, fill=CANCELLATION_REASON)) +
  geom_histogram(bins=25) + xlab("Month") + ylab("Count") +
  scale_x_continuous(breaks = seq(1, 12, by = 1)) +
  ggtitle("Airline Cancellations per month")
```



We can see that weather is the major reason for cancellation, especially in the winter (January and February). Carrier cancellations are also high during this period.

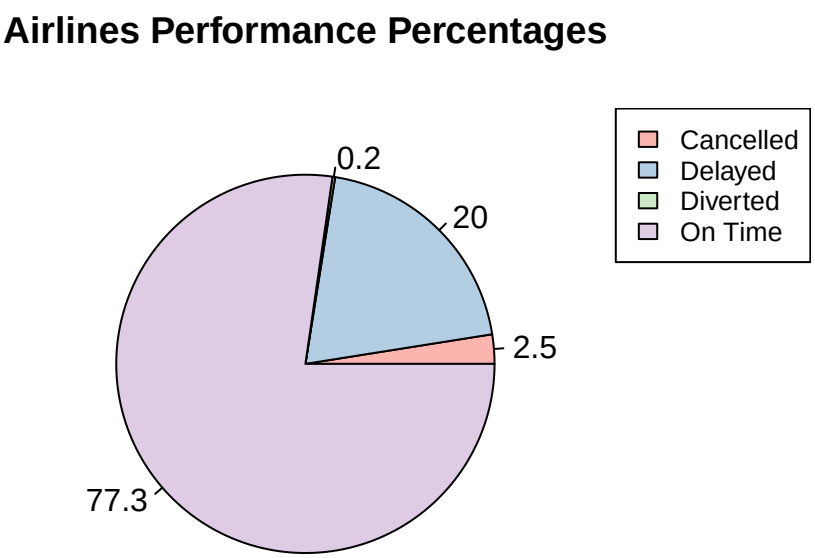
```
# Histogram for Airline Cancellations
carrier_performance_df$CANCELLED_F <-factor(carrier_performance_df$CANCELLED)

ggplot(data=carrier_performance_df,
       aes(x=OP_UNIQUE_CARRIER,fill=CANCELLED_F))+
  geom_histogram(stat = "Count",position = "dodge", col="red", bins = 15) +
  labs(title="Cancellations by Operating Airlines") +
  labs(x="Airline", y="Number of Flights") +scale_x_discrete() +
  scale_y_log10() +
  labs(fill='Cancelled or Not')
```



At this time, it appears American, Delta, United, Republic Airways etc., have more cancellations. However, we cannot conclude this to be true, since they also have more fleets in comparison to other low cost airlines. I would like to redo this plot with percentage of cancellations per airline.

PIE CHART

```
#Pie Chart for Over all performance
flight_stats <- carrier_performance_df %>% count(Performance)

#Average number of delayed flights per month since 2003
avg_flights_cancelled <- flight_stats[flight_stats$Performance == "Cancelled",]$n
avg_flights_diverted <- flight_stats[flight_stats$Performance == "Diverted",]$n
avg_flights_on_time <- flight_stats[flight_stats$Performance == "On-Time",]$n
avg_flights_delayed <- flight_stats[flight_stats$Performance == "Delayed",]$n


flights_values =
  c(avg_flights_cancelled,avg_flights_delayed,avg_flights_diverted,avg_flights_on_time)
flights_labels = c("Cancelled","Delayed","Diverted","On Time")

#Calculating the percent each will be of the total
piepercent<- round(100*flights_values/sum(flights_values), 1)

pie(flights_values, labels = piepercent, main = "Airlines Performance Percentages",
    col = brewer.pal(9, "Pastel1"))

legend("topright",flights_labels, cex = 0.8,fill = brewer.pal(9, "Pastel1"))
```



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.

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.

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.