STAT 5010 Project - Flight Analysis

Darsh Shah

4/24/2022

Introduction

I love airplanes! But, before choosing a flight, I always question what is the likelihood of this flight not cancelling? What are the chances of me facing a massive delay? If i have a connecting flight, how much of a delay can i bear in my first flight as to not miss my second flight? I am always asking these questions before booking a flight and I also believe these some of the questions that any passenger would think about before choosing a flight. Fortunately, I was able to find a dataset that consists of flight arrivals and departures details of all commercial flights within the USA, from October to 1987 to April 2008. This is a large dataset: there are nearly 120 million records in total, and takes up 1.6 gigabytes of space compressed and 12 gigabytes when uncompressed. To ensure that I can properly work on the dataset, I decided to only use the data of 2007 flights. However, the methodologies used for the findings can be extended to any year's data.

Dataset URL: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/HG7NV7

Setting up Dependencies

Setting my working directory and Importing the necessary libraries

```
setwd("C:/Users/darsh/Desktop/STAT 5010/Project")
library(tidyverse)
library(tidymodels)
library(ggplot2)
library(dplyr)
library(zoo)
library(reshape2)
#install.packages("wesanderson")
#install.packages("viridis")
library(wesanderson)
library(viridis)
#machinelearning
library(mlr3)
library(mlr3learners)
library(mlr3pipelines)
library(mlr3tuning)
library(skimr)
library(plotrix)
library(caret)
library(survey)
library(pscl)
```

Loading the Data and some preprocessing

```
data = read.csv("2007.csv")
airports = read.csv("airports.csv")
carriers = read.csv("carriers.csv")
plane_data= read.csv("n_plane-data.csv")
```

```
# Rename columns within dataframe for later.
#str(plane_data) # Find col index to rename
names(plane_data)[2] <- "TailNum"
dep_airports <- airports

#str(dep_airports) # Find col index to rename
names(dep_airports)[1] <- "Origin"
names(dep_airports)[4] <- "DepState"
arr_airports <- airports

#str(arr_airports) # Find col index to rename
names(arr_airports)[1] <- "Dest"
names(arr_airports)[4] <- "ArrState"
```

Questions

I shall first try to answer some of the first questions that the Expo has asked us to think about and then I'll answer some questions as I dive deeper into the data.

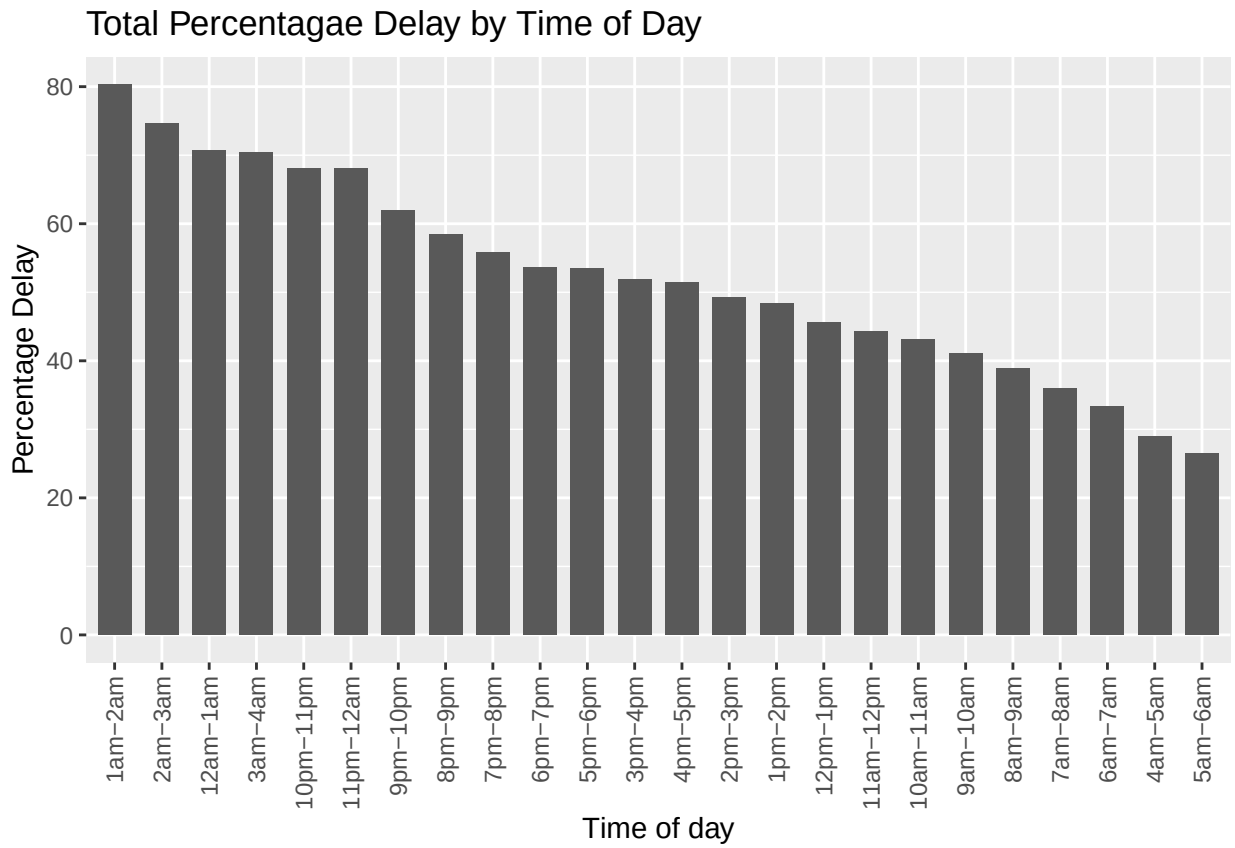
Question 1: When is the best time of the day / day of week / time of the year to fly to minimise delays?

Question 1a: When is the best of the day to fly?

```
#Creating a delay column
data  <- data %>%
  mutate(ArrStatus = case_when(ArrDelay > 0 ~ 1, T ~ 0))

#creating time groups:
data$time_group <- cut(data$DepTime,
                       breaks = c(0, 100, 200, 300, 400, 500, 600, 700, 800, 900,
                                  1000, 1100, 1200, 1300, 1400, 1500, 1600, 1700, 1800, 1900,
                                  2000, 2100, 2200, 2300, 2400),
                       labels = c("12am-1am", "1am-2am", "2am-3am", "3am-4am", "4am-5am",
                                  "5am-6am", "6am-7am", "7am-8am", "8am-9am", "9am-10am",
                                  "10am-11am", "11am-12pm", "12pm-1pm", "1pm-2pm", "2pm-3pm",
                                  "3pm-4pm", "4pm-5pm", "5pm-6pm", "6pm-7pm", "7pm-8pm",
                                  "8pm-9pm", "9pm-10pm", "10pm-11pm", "11pm-12am"))
```

```
best_time_of_day <- data %>%
  group_by(time_group) %>%
```

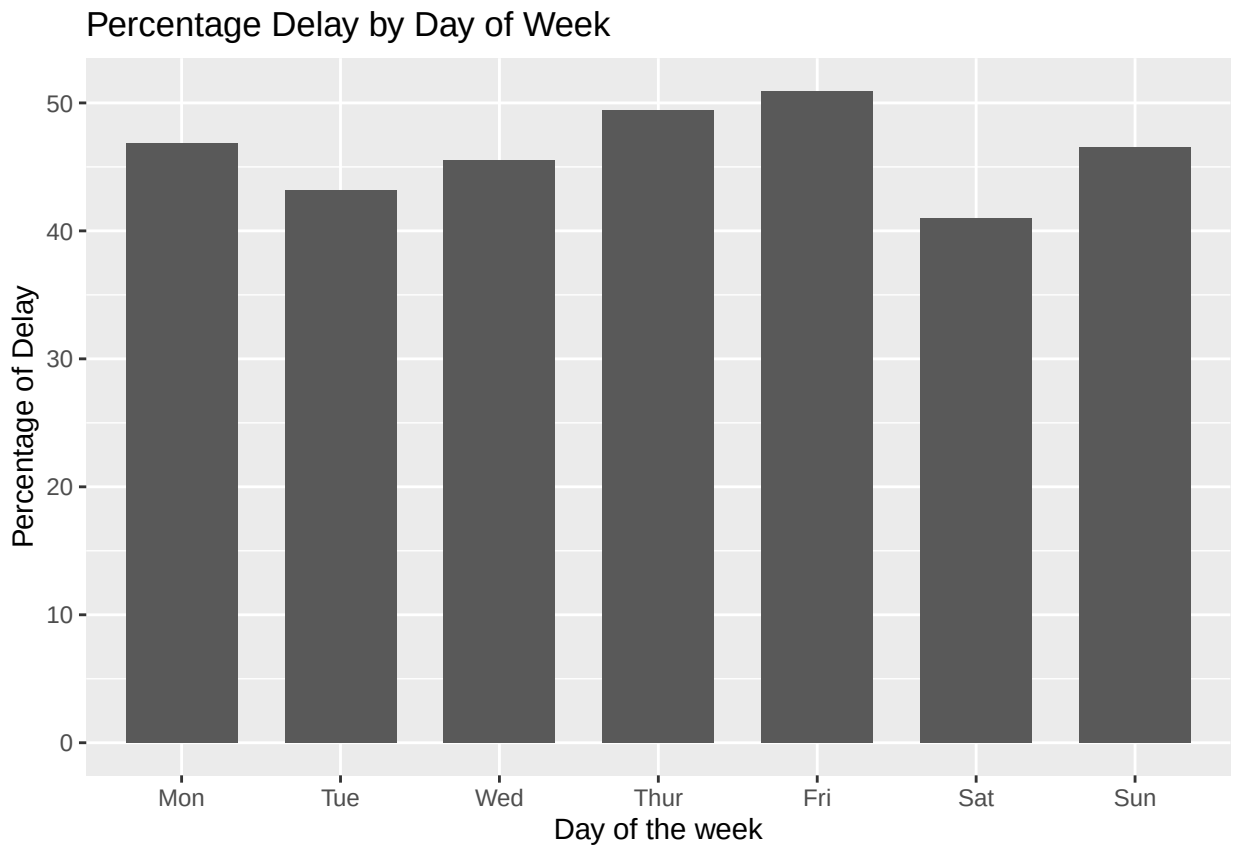
```
  summarise(total_delays =  sum(ArrStatus, na.rm = TRUE), count = n()) %>%
  mutate(percentage = round(total_delays) / (count) * 100)

#removing the NA row which is essentially when there were no flights.
best_time_of_day <- best_time_of_day[-c(25), ]

plot_timeofday <- ggplot(best_time_of_day, aes(x=reorder(time_group, -percentage),
                                               y=percentage, width=0.7)) +
  geom_bar(stat='identity', position='dodge') +
  ggtitle("Total Percentagae Delay by Time of Day") +
  xlab("Time of day") +
  ylab("Percentage Delay") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))

plot_timeofday
```



From this graph, we can see that flights between 1am to 3am have the highest delayed percentages which suggests not to take flights during that time of the day. We can also see that flights between 4am to 6am have the lowest delay percentage. That is a good time of the day to take flights to minimise delays. The number of flights also between 1am to 3am are not that much in comparison to the other times of the day. But it could be that during this the time, there are less number of ground staffs and crews which could cause delays for the airplanes.

Question 1b: What is the best day of the week to fly to minimise delays?

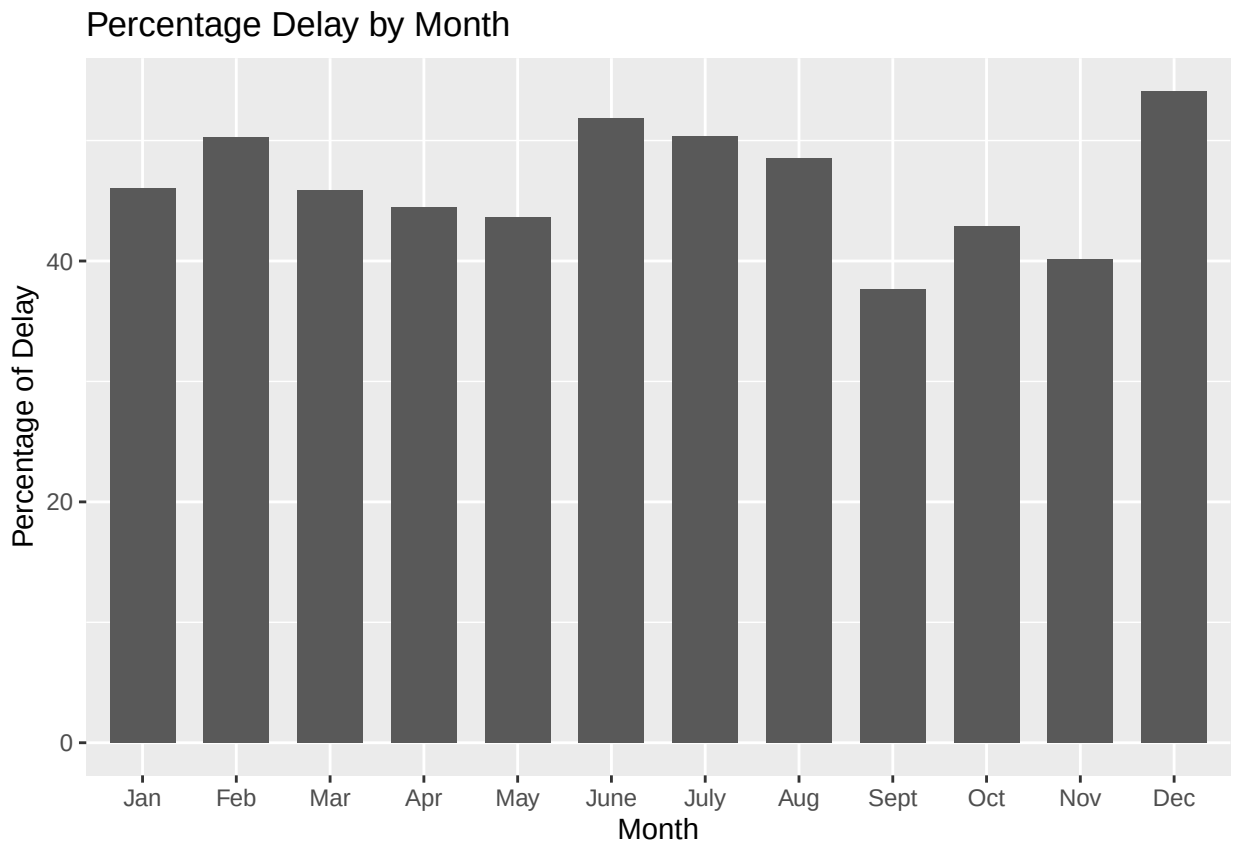
```
day_wise_delays <- data %>%
  group_by(DayOfWeek) %>%
  summarise(total_delays =  sum(ArrStatus, na.rm = TRUE), count = n()) %>%
  mutate(percentage = round(total_delays) / (count) * 100)


plot_dayofweek <- ggplot(day_wise_delays, aes(x=DayOfWeek, y=percentage,  width=0.7)) +
  geom_bar(stat='identity', position='dodge') +
  ggtitle("Percentage Delay by Day of Week") +
  scale_x_discrete(limits=c('Mon', 'Tue', 'Wed', 'Thur', 'Fri', 'Sat', 'Sun')) +
  xlab("Day of the week") +
  ylab("Percentage of Delay")

plot_dayofweek
```



From this graph we can see that Saturday looks like the best day to fly to minimse delays. However, it should also be noted that the total number of flights on Saturday are also less in comparison to other days. Less number of flights means less options to change your flights as well. So, a person has to keep that in mind as well that if for some reason, they have to change their flight, then they will be having less options in comparison to the other days of the week.

Question 1c: When is the best time of the year to fly to minimise delays?

```
month_wise_delays <- data %>%
  group_by(Month) %>%
  summarise(total_delays =  sum(ArrStatus, na.rm = TRUE), count = n()) %>%
  mutate(percentage = round(total_delays) / (count) * 100)

plot_month <- ggplot(month_wise_delays, aes(x=Month, y=percentage,  width=0.7)) +
  geom_bar(stat='identity', position='dodge') +
  ggtitle("Percentage Delay by Month") +
  scale_x_discrete(limits=c('Jan', 'Feb', 'Mar', 'Apr', 'May', 'June',
                            'July', 'Aug', 'Sept', 'Oct', 'Nov', 'Dec')) +
  xlab("Month") +
  ylab("Percentage of Delay")

plot_month
```



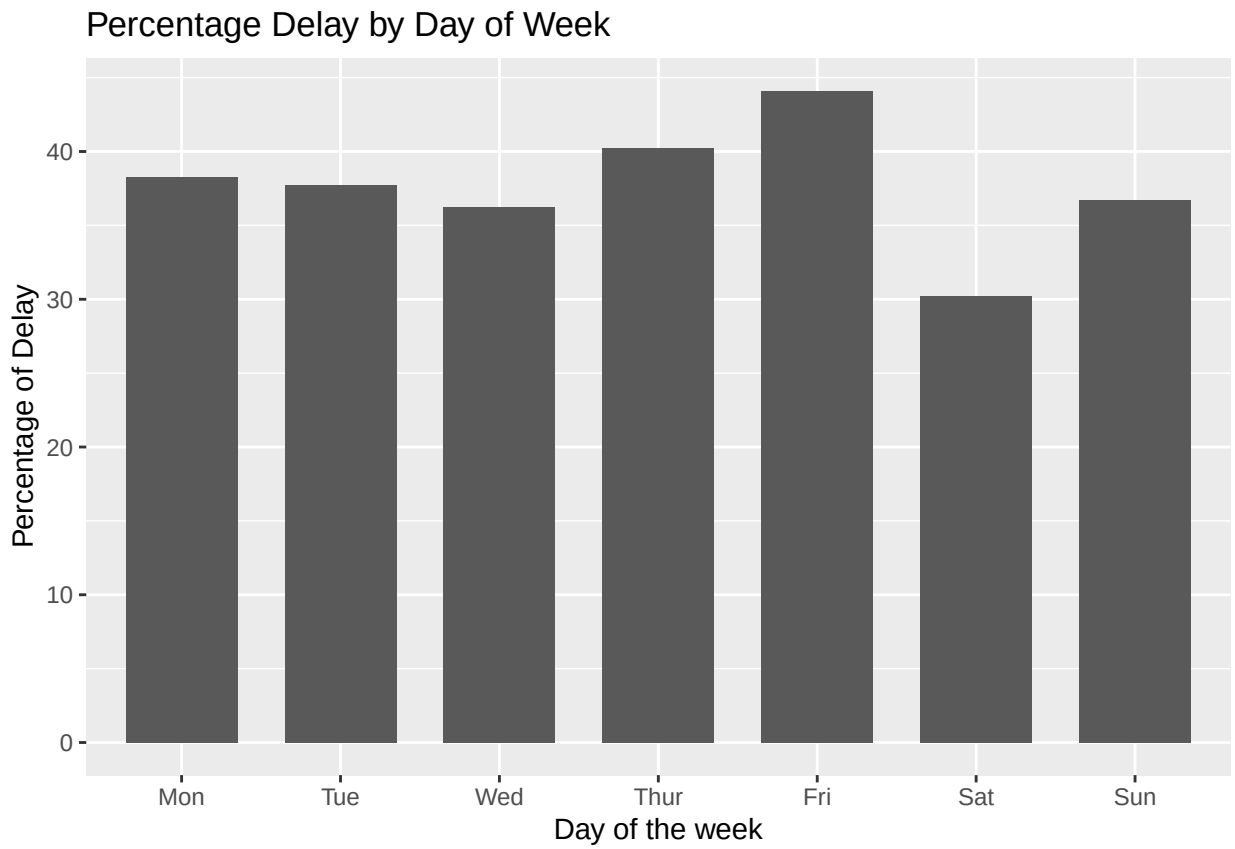
From this graph, we can see that flights during June, July, August and December have the highest delay percentages. This makes sense because June, July and August are months where a lot of students are flying across the country before the start of their fall semesters. We also see that there are more number of flights as well during those 3 months in comparison to the other months. Decemeber is the holiday period, so that also makes sense why there is such a massive percentage of delayed flights during that month.

However, the best month to fly to minimise delay is during September. I believe this might be the case because the fall semesters have already started, so hardly any students would be flying across the country. Also, there are no holidays during the month of September which eliminates any long weekends as well where

some people would go to their homes and then again come back.

So, let's see what is the best day of the week and best the time of day during the month of September to fly

```
best_day_during_september <- data %>%
  filter(Month == 9) %>%
  group_by(DayOfWeek) %>%
  summarise(total_delays =  sum(ArrStatus, na.rm = TRUE), count = n()) %>%
  mutate(percentage = round(total_delays) / (count) * 100)


plot_dayofweek_in_Sept <- ggplot(best_day_during_september, aes(x=DayOfWeek, y=percentage,  width=0.7))
  geom_bar(stat='identity', position='dodge') +
  ggtitle("Percentage Delay by Day of Week") +
  scale_x_discrete(limits=c('Mon', 'Tue', 'Wed', 'Thur', 'Fri', 'Sat', 'Sun')) +
  xlab("Day of the week") +
  ylab("Percentage of Delay")

plot_dayofweek_in_Sept
```

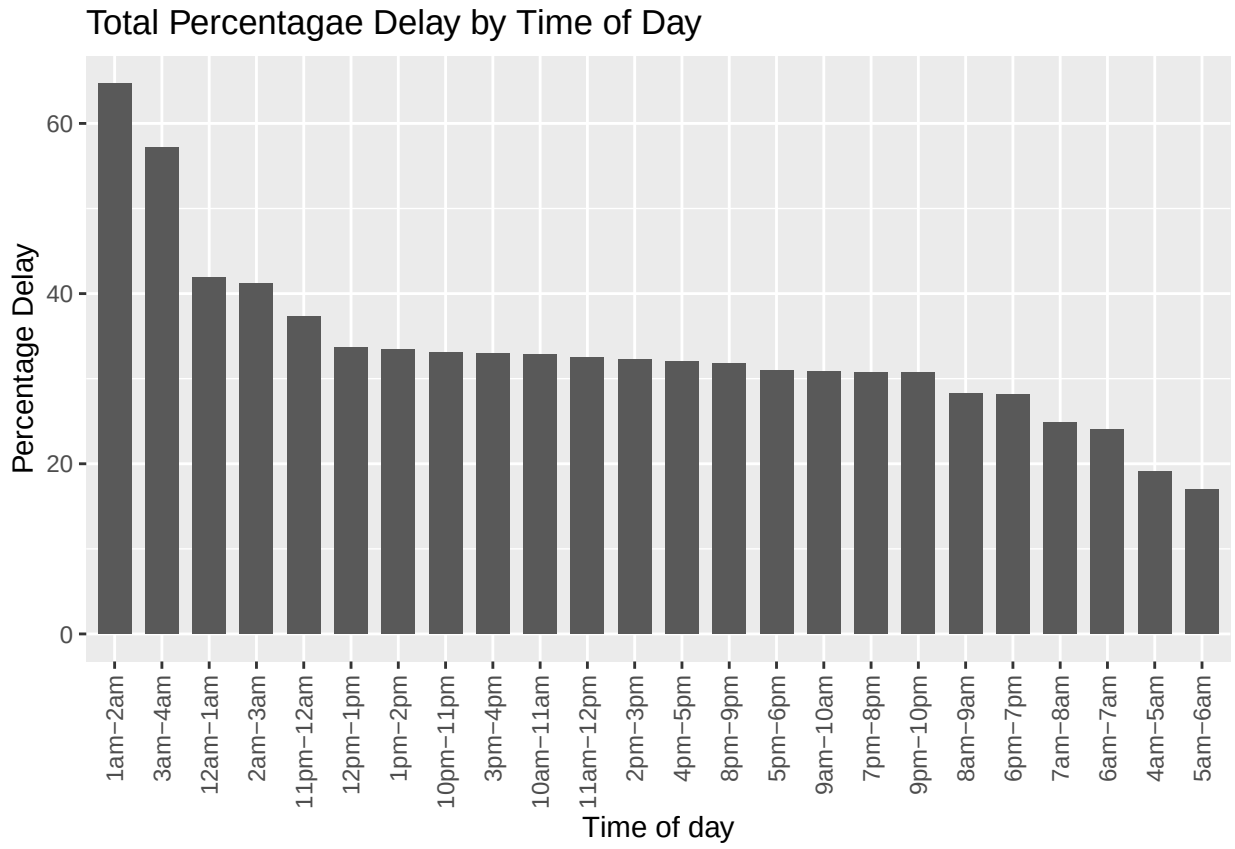


So, it also seems that Saturday is the best to fly during the Month of September!

Now, what is the best time during Saturdays in the month of September to fly?

```
best_time_during_september <- data %>%
  filter(Month == 9) %>%
  filter(DayOfWeek == 6) %>%
```

```
  group_by(time_group) %>%
  summarise(total_delays =  sum(ArrStatus, na.rm = TRUE), count = n()) %>%
  mutate(percentage = round(total_delays) / (count) * 100)


#removing the NA row which is essentially when there were no flights.
best_time_during_september <- best_time_during_september[-c(25), ]

plot_timeofday <- ggplot(best_time_during_september, aes(x=reorder(time_group, -percentage),
                                                         y=percentage, width=0.7)) +
  geom_bar(stat='identity', position='dodge') +
  ggtitle("Total Percentagae Delay by Time of Day") +
  xlab("Time of day") +
  ylab("Percentage Delay") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))

plot_timeofday
```



It looks like the best time to fly during the Month of September is between 4am to 6am on Saturdays!

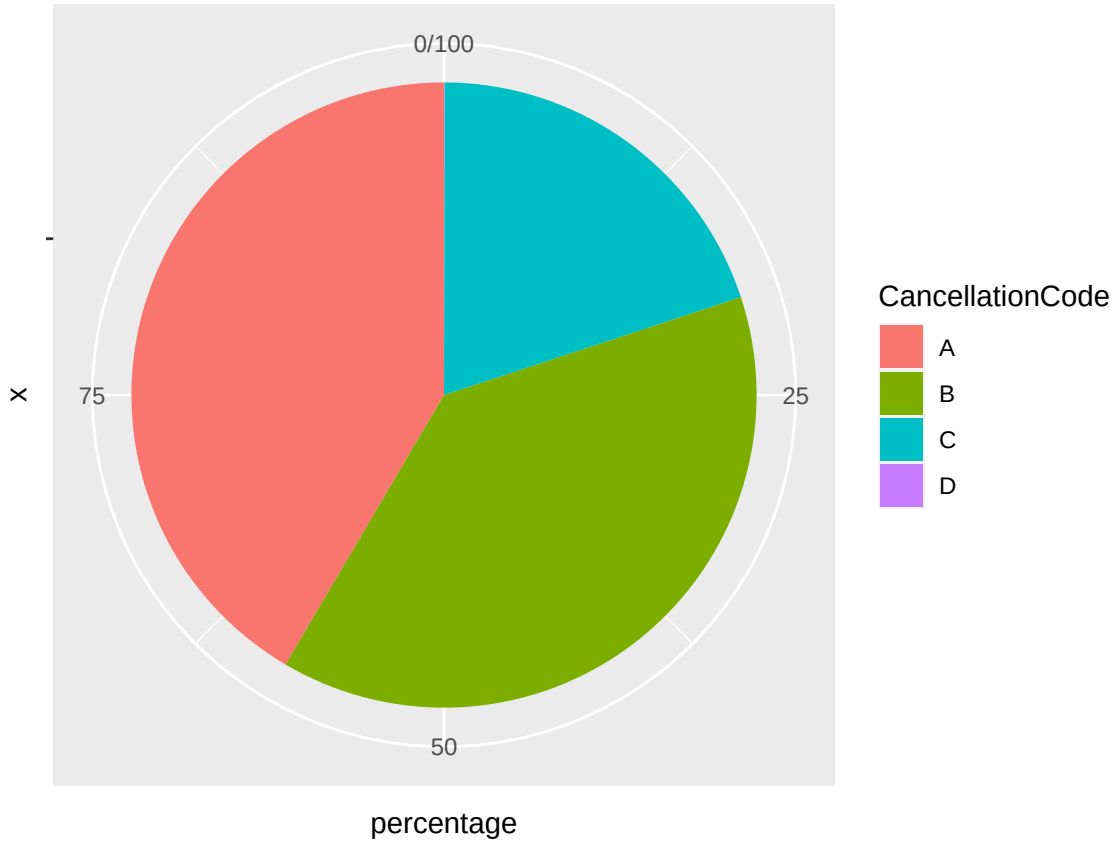
Question 2: What is the most common reason for flight Cancellations?

```
#Checking if there is a flight that has been cancelled but there is no cancellation code attached to it
rows <- which(data$Cancelled==1 & data$CancellationCode=="")
```

```
rows
```

```
## integer(0)
```

```
cancelled_flights <- data %>%
  filter(CancellationCode == "A" | CancellationCode == "B" | CancellationCode == "C" | CancellationCode
  group_by(CancellationCode) %>%
  summarise(count = n()) %>%
  mutate(percentage = round(count) / sum(count) * 100)

bp <- ggplot(cancelled_flights, aes(x="", y=percentage, fill=CancellationCode))+
  geom_bar(width = 1, stat = "identity")

pie <- bp + coord_polar("y", start=0)
pie
```

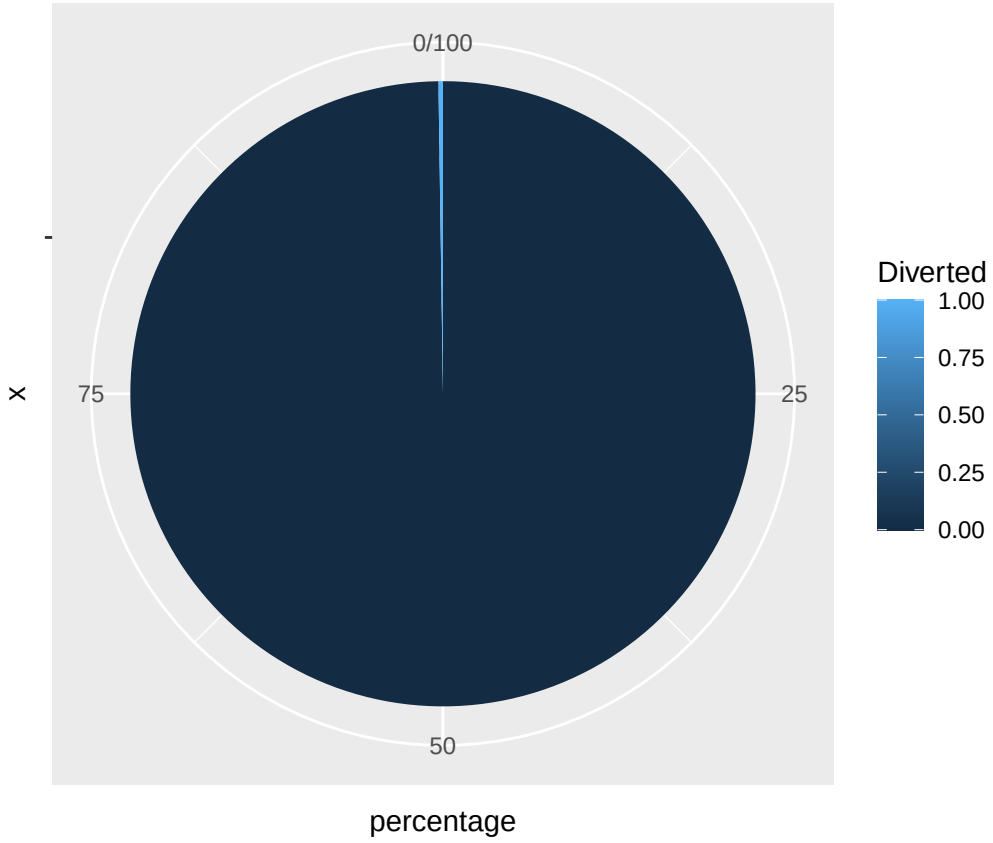


reason for cancellation (A = carrier, B = weather, C = NAS, D = security)

From this graph, we can see that the major reason for flight Cancellations is due to Carrier and Weather. Security Delays are just 0.02% which is why it is almost impossible to see that in the pie chart.

Question 3: What is the percentage of diverted flights to non-diverted flights?

```
diverted_flights <- data %>%
  select(Diverted)%>%
  group_by(Diverted) %>%
  summarise(count = n()) %>%
  mutate(percentage = round(count) / sum(count) * 100)

bp <- ggplot(diverted_flights, aes(x="", y=percentage, fill=Diverted))+
  geom_bar(width = 1, stat = "identity")

pie <- bp + coord_polar("y", start=0)
pie
```



From this graph, we can see that about 99% of the flights in 2007 were not diverted which is great news because out 7.4 Million flights in 2007, only 17,000 didnt reach their original destination and were diverted to some other airports or had to return back to their originating airports.

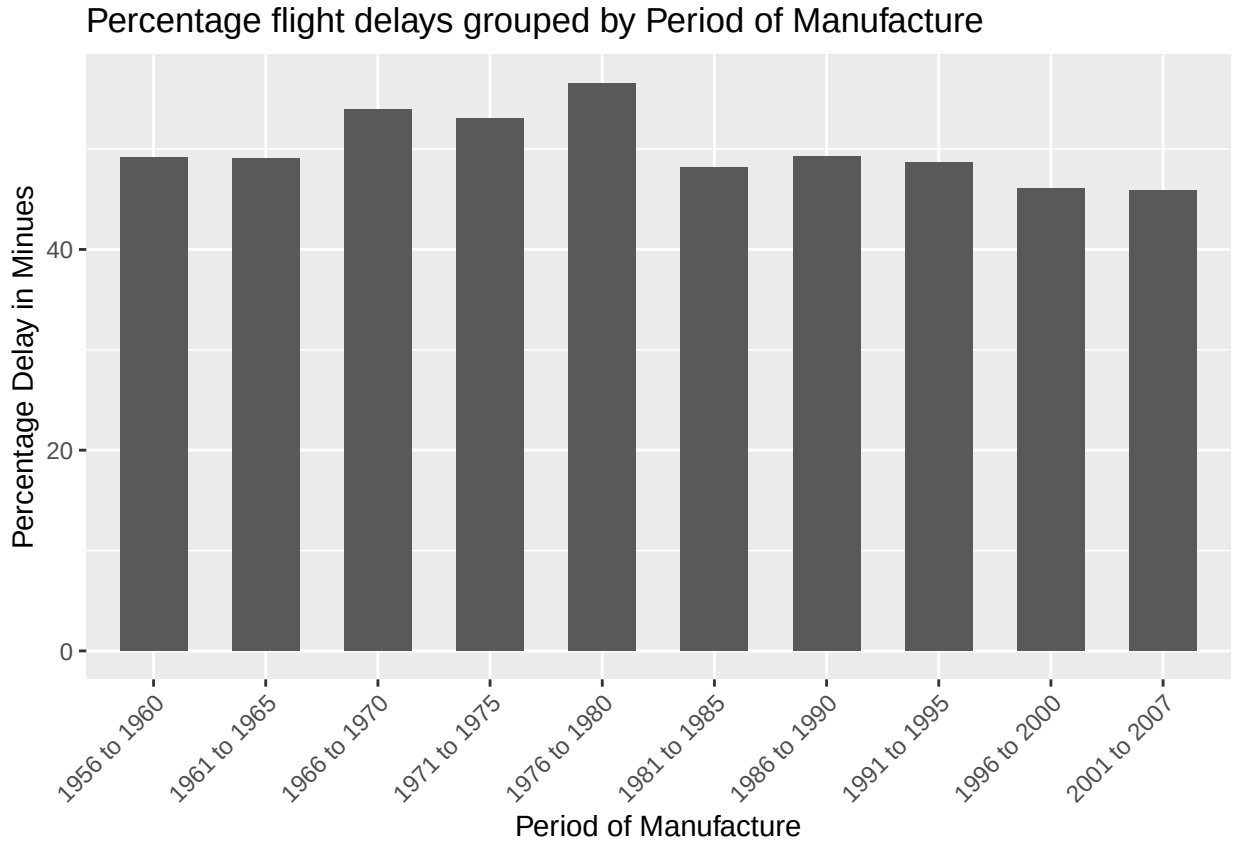
Question 4: Do Older planes suffer more delays?

```
flight_planes_data <- merge(data, plane_data, by="TailNum")
```

```
flight_planes_data$PlanePeriod <- cut(flight_planes_data$PlaneYear,
                                      breaks = c(1955, 1960, 1965, 1970, 1975,
                                                 1980,1985, 1990, 1995, 2000,
```

```
                                                2007),
                                    labels=c("1956 to 1960", "1961 to 1965",
                                             "1966 to 1970", "1971 to 1975",
                                              "1976 to 1980","1981 to 1985",
                                              "1986 to 1990", "1991 to 1995",
                                              "1996 to 2000", "2001 to 2007"))
```

```
planes_delay <- flight_planes_data %>%
  group_by(PlanePeriod) %>%
  summarise(total_delays =  sum(ArrStatus, na.rm = TRUE), count = n()) %>%
  mutate(percentage = round(total_delays) / (count) * 100)

planes_delay_plot <- ggplot(planes_delay, aes(y=percentage, x=PlanePeriod,
                                              width=0.6)) +
  geom_bar(stat='identity', position='dodge') +
  ggtitle("Percentage flight delays grouped by Period of Manufacture") +
  ylab("Percentage Delay in Minues") +
  xlab("Period of Manufacture") +
  theme(axis.text.x = element_text(angle = 45, hjust=1))

planes_delay_plot
```

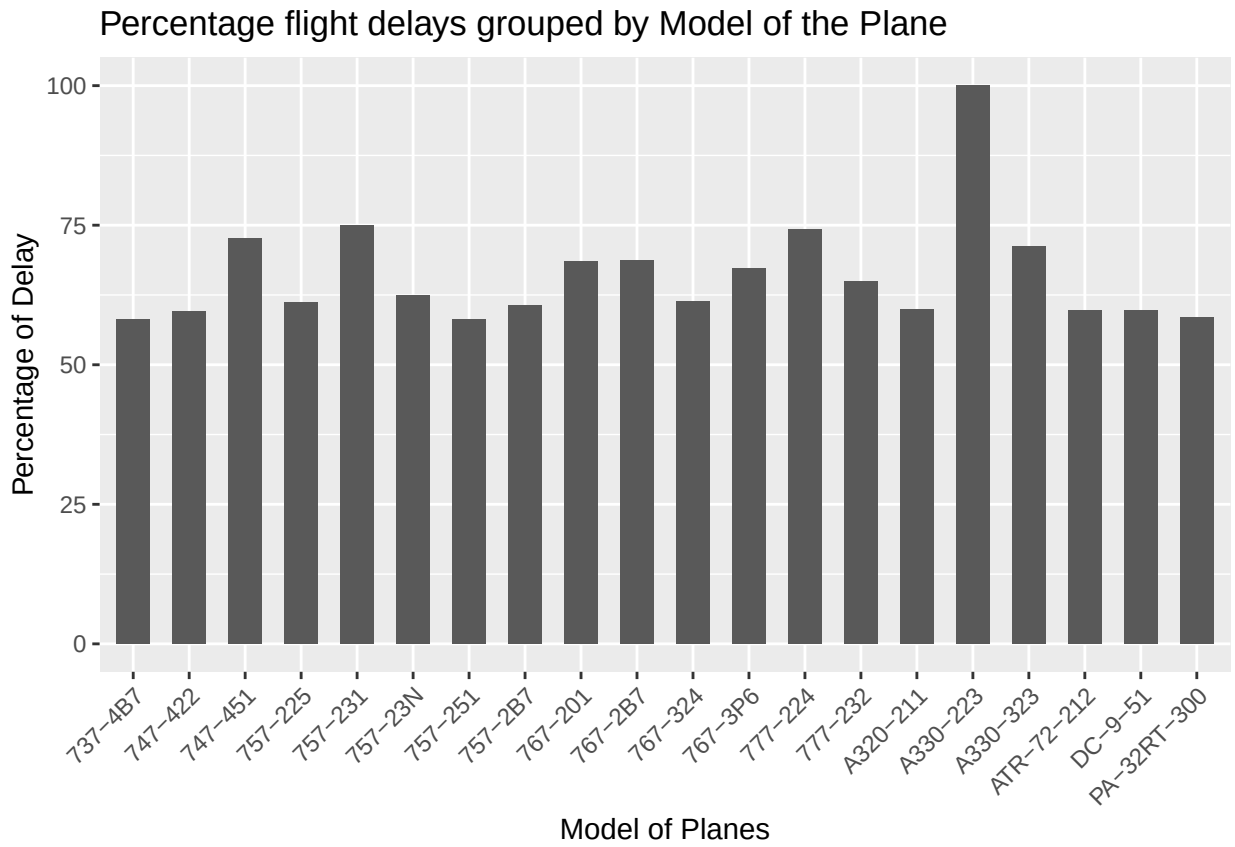


From this graph, we can see that planes manufactured between 1966 to 1970 and 1976 to 1980 suffer the most delays. This does answer the question, that yes, older planes are prone to suffer more delays than planes manufactured between 2001 to 2007. The reason for that could that older planes are less fuel efficient, they

are prone to more fault due to older parts. Plus they also require more maintenance to keep them running meaning there could be many flights which are delayed due to some fault in the plane at the ground and it could have taken a long time to either get it fixed or get a spare part.

Let's also take a look at which model of the plane suffers the most delays. This is an interesting question because when booking a flight, a passenger gets to see what model of plane would they be flying on. Based on these results, they could guage what percentage of flights for that particular model of plane have been delayed.

```
plane_model_delays <- flight_planes_data %>%
  group_by(model) %>%
  summarise(total_delays =  sum(ArrStatus, na.rm = TRUE), count = n()) %>%
  mutate(percentage = round(total_delays) / (count) * 100)


#There are 156 Models of planes which will not look good while plotting.
#Let's see the top 20 Models of Planes prone to delays.

plane_model_delays_plot <- top_n(plane_model_delays, n=20, percentage) %>%
  ggplot(., aes(y=percentage, x=model, width=0.6)) +
  geom_bar(stat='identity', position='dodge') +
  ggtitle("Percentage flight delays grouped by Model of the Plane") +
  ylab("Percentage of Delay") +
  xlab("Model of Planes") +
  theme(axis.text.x = element_text(angle = 45, hjust=1))

plane_model_delays_plot
```



The two most delayed aircraft models are Airbus A330-300 series and Boeing 777-200 series.

I couldn't find the exact date of Manufacture for the A330-323, but based on the model of the plane, it is an Airbus A330-300 series variant. Now, Airbus introduced the first A330-100 series variant in 1994 which suggests that the 300 series variant definitely came after 1994. Similarly, Boeing started the 777 series of planes after 1994. However, from the plane manufacture period graphs, we can see that planes manufactured after 1994 have a lower delayed percentage rate than planes manufactured before 1994.

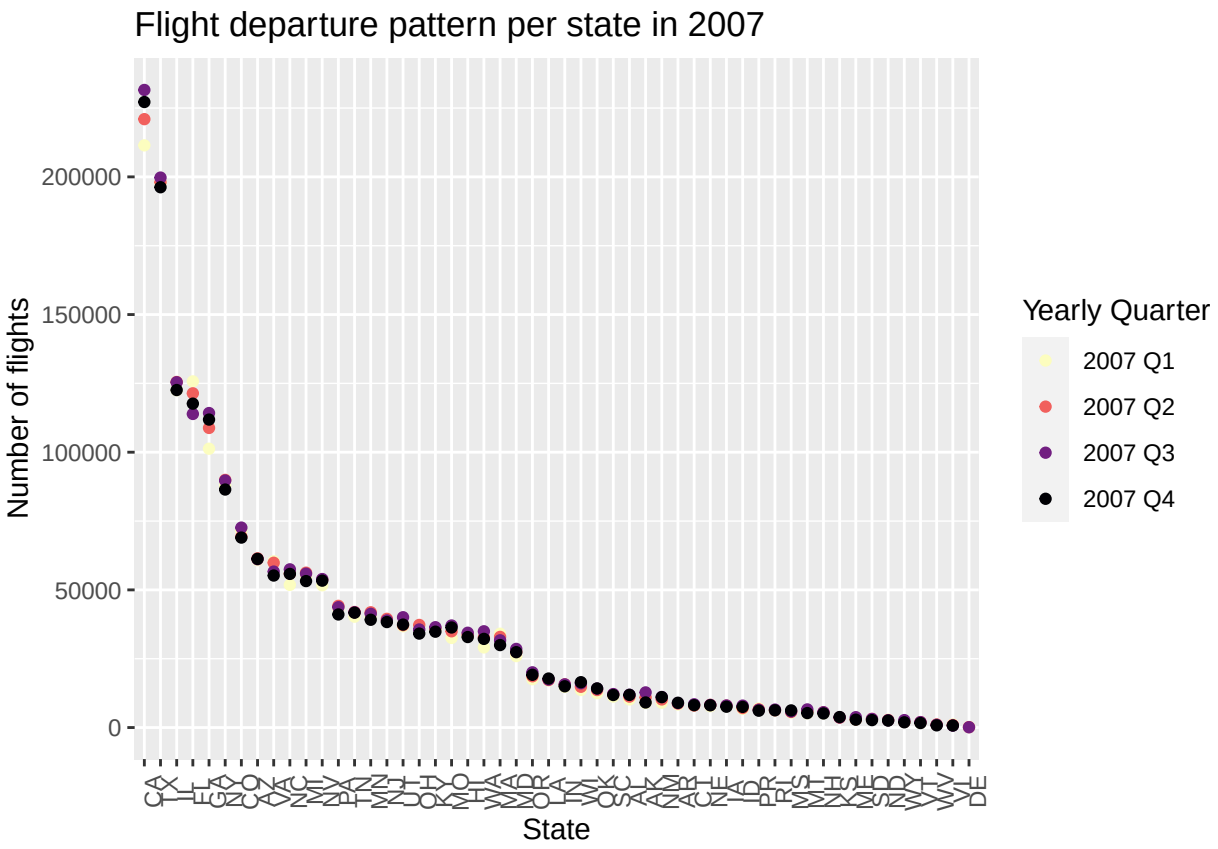
So, this puts the question why are these bigger wide body jets more delayed? The answer to this could be because Wide Body aircraft require more preparation time at the gate, they require more fuel, and can carry many passengers and bags. So, there could be delays while cleaning the aircraft or loading the food in the aircraft. There could be delays because of some fuellign issues. There could also be delays because of passengers. Since, these planes carry close to 300 passengers in one flight, it is highly likely that one or more passenger is delayed at the airport due to security reasons or other and is taking time to board the plane which essentially delays the plane and the flight. It could also be that one passenger could not make it to the flight. In such a case, their bags must be taken out as well which causes a massive problem since the ground staff have to then check each and every container and de-board the bags of that one passenger. They then have to load the containers again which would also cause massive delays for the plane. In smaller planes, these factors cause less delays because it is easier to carry out these things with less number of passengers for every flight.

Question 5: How does the number of people flying between different locations change over time?

```
#For Departing Flights
flights_Depart <- merge(data, dep_airports, by="Origin")
```

```
flights_Depart <- flights_Depart %>%
  unite(Date, Year, Month, DayofMonth, sep = "-") %>%
  mutate(Date = paste(Date))

Passengers_leaving <- flights_Depart %>%
  select(contains(c('Date','DepState'))) %>%
  mutate(Date = as.Date(Date, format =  "%Y-%m-%d")) %>%
  mutate(yearQuarter = as.yearqtr(Date, format)) %>%
  group_by(yearQuarter, DepState) %>%
  summarise(StateCount = n()) %>%
  na.omit('DepState')
```

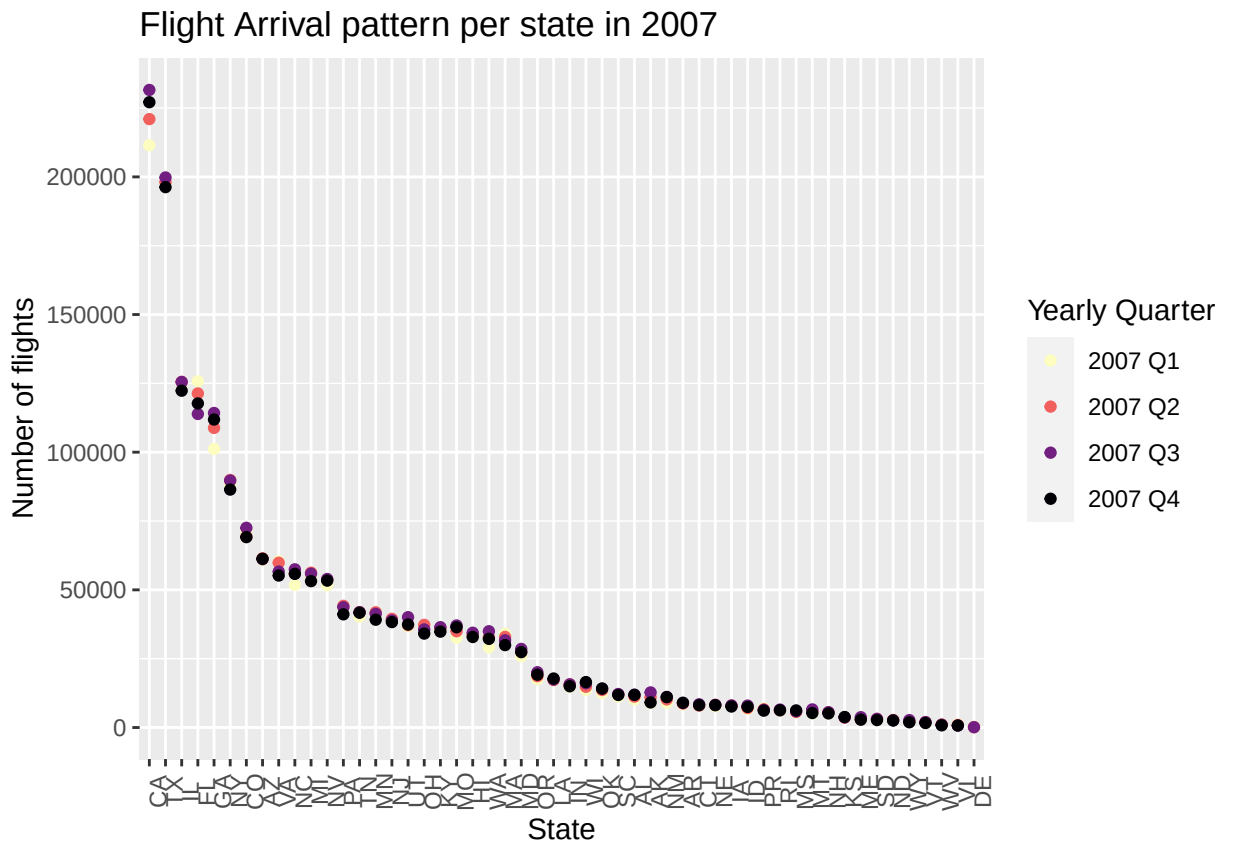
```
plot_dep <- ggplot(Passengers_leaving, aes(x=reorder(DepState, -StateCount), y=StateCount))
plot_dep + geom_point(aes(colour = factor(yearQuarter))) +
  scale_color_viridis(option='magma', direction=-1, discrete=TRUE, name="Yearly Quarter") +
  ggtitle("Flight departure pattern per state in 2007") +
  xlab("State") +
  ylab("Number of flights") +
  theme(axis.text.x = element_text(angle = 90))
```



We can see that California (CA), Texas (TX), Illinois (IL), Floria (FL),and Georgia (GA) have the highest number of outgoing flights. But Overall across the US, the number of outgoing flights have been decreasing in 2007.

```
#For Arriving Flights
flights_arriving <- merge(data, arr_airports, by="Dest")
flights_arriving <- flights_arriving %>%
  unite(Date, Year, Month, DayofMonth, sep = "-") %>%
  mutate(Date = paste(Date))
```

```
Passengers_arriving <- flights_arriving %>%
  select(contains(c('Date','ArrState'))) %>%
  mutate(Date = as.Date(Date, format =  "%Y-%m-%d")) %>%
  mutate(yearQuarter = as.yearqtr(Date, format)) %>%
  group_by(yearQuarter, ArrState) %>%
  summarise(StateCount = n()) %>%
  na.omit('ArrState')


plot_dep <- ggplot(Passengers_arriving, aes(x=reorder(ArrState, -StateCount), y=StateCount))
plot_dep + geom_point(aes(colour = factor(yearQuarter))) +
  scale_color_viridis(option='magma', direction=-1, discrete=TRUE, name="Yearly Quarter") +
  ggtitle("Flight Arrival pattern per state in 2007") +
  xlab("State") +
  ylab("Number of flights") +
  theme(axis.text.x = element_text(angle = 90))
```



We can see that California (CA), Texas (TX), Illinois (IL), Floria (FL),and Georgia (GA) have the highest number of icnoming flights as well. But Overall across the US, the number of incomings domestic flights have been decreasing in 2007.

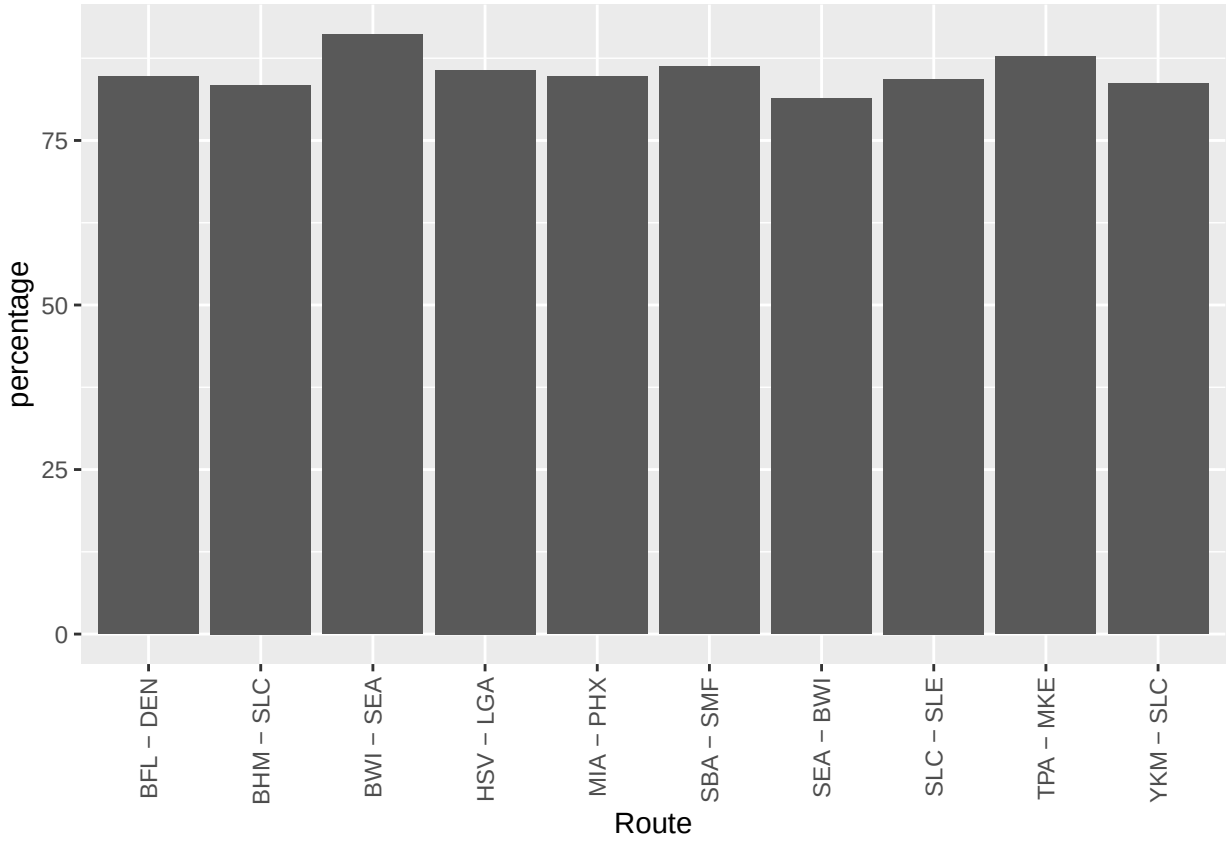
Question 6: Which Route faces the most delays?

```
data$Route <- paste(data$Origin, "-", data$Dest)

route_delays <- data %>%
  group_by(Route) %>%
  summarise(total_delays = sum(ArrStatus, na.rm = TRUE), count = n()) %>%
  mutate(percentage = round(total_delays) / (count) * 100)



route_delays <- route_delays[order(route_delays$percentage, decreasing = TRUE), ]
```

There are 104 routes with 100% delay percentage but the total number of flights on these routs are like 1 or 2 or 3 and those flights are only delayed. This isn't a good way to judge those routes since they are least travelled routes. So i decided to remove those rows where the count of flights is less than 50. Because having less than 50 flights on a particular route is like 1 or no flight in a week as well which is not a good metric to judge a route. So we judge the delayedness of a route with flight count more than 50.

```
route_delays <- subset(route_delays, count >= 50)
```

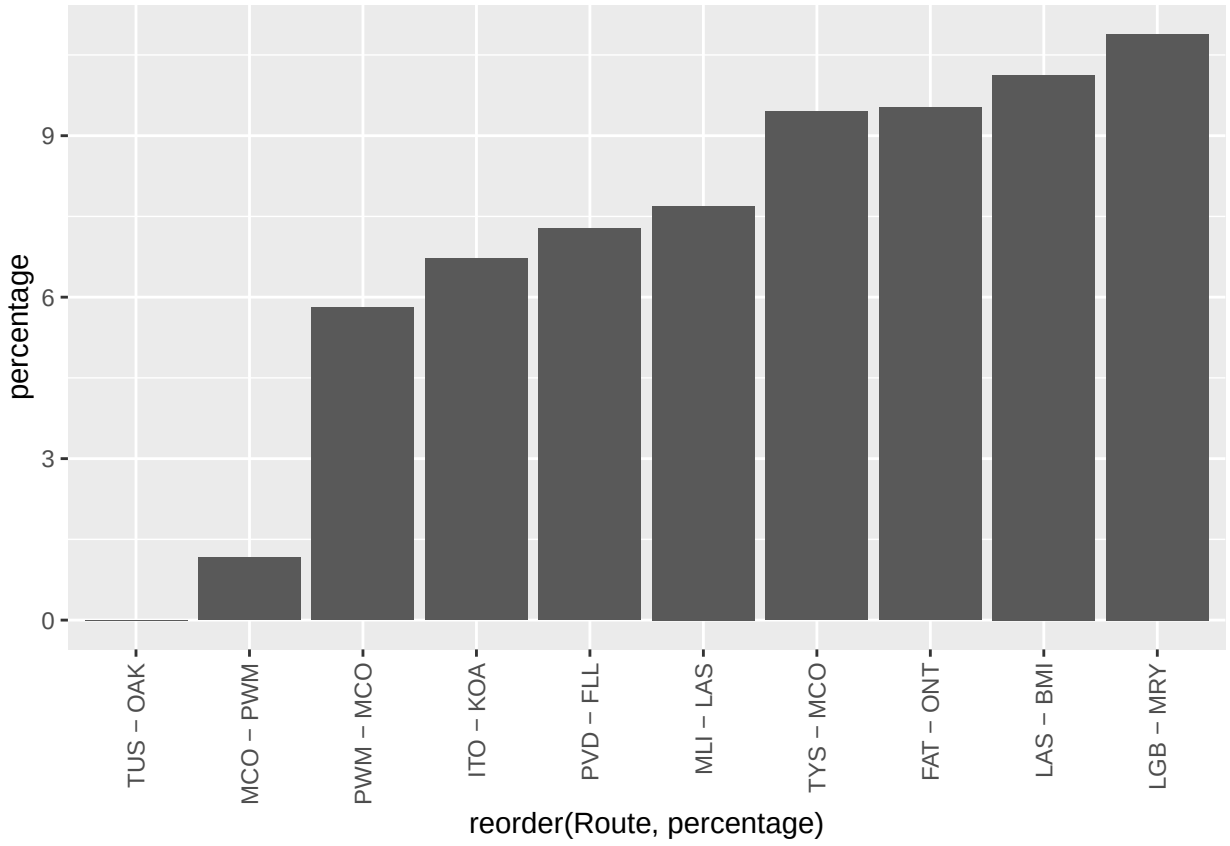
```
route_delays_percentage_plot <- top_n(route_delays, n=10, percentage) %>%
  ggplot(., aes(x=Route, y=percentage)) + geom_bar(stat='identity') +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))

route_delays_percentage_plot
```



Based on this graph, we can see that Baltimore to Seattle is the route with highest percentage of delayed flights. The second most delayed route is Tampa, Florida to Milwaukee, Wisconsin.

Let's also check the best routes with the least delay percentage

```
best_route_percentage_plot <- top_n(route_delays, n=-10, percentage) %>%
  ggplot(., aes(x=reorder(Route, percentage), y=percentage)) + geom_bar(stat='identity') +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))

best_route_percentage_plot
```



The two best flights are Tuscon to Oakland with 0 flights delayed out of 58 flights on that route in 2007. Coming second is the route from Orlando to Portland with 1 flight delayed out of 86 flights on that route in 2007.

Let's take a look at which airline does the worst in the Baltimore to Seattle Route.

```
colnames(carriers)[1] <- "UniqueCarrier"

flight_and_carrier_data <- merge(data, carriers, by="UniqueCarrier")
```

```
bwi_to_sea <- flight_and_carrier_data %>%
  group_by(Route, Description) %>%
  filter(Route == "BWI - SEA") %>%
  summarise(total_delays = sum(ArrStatus, na.rm = TRUE), count = n()) %>%
  mutate(percentage = round(total_delays)/(count) * 100)

bwi_to_sea
```

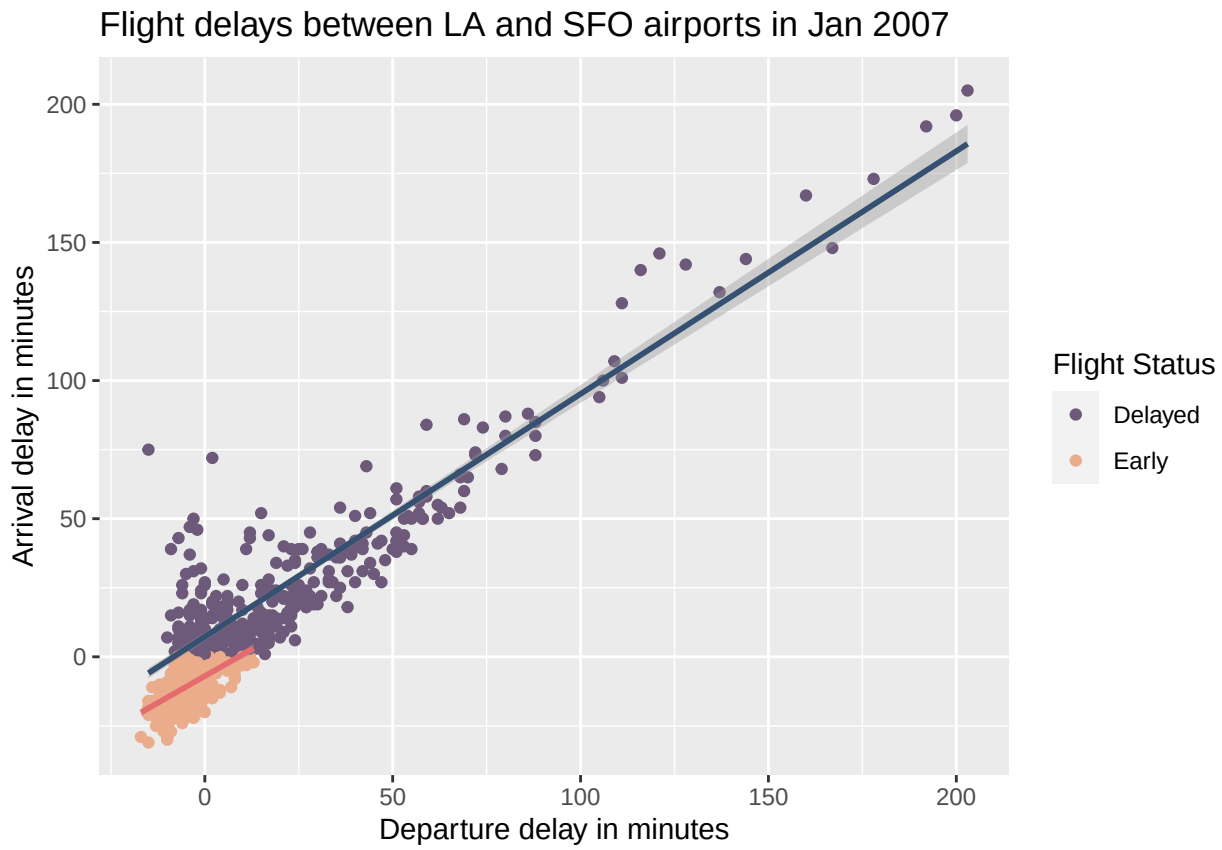
```
## # A tibble: 1 x 5
## # Groups:   Route [1]
##   Route     Description                total_delays count percentage
##   <chr>     <chr>                             <dbl> <int>      <dbl>
## 1 BWI - SEA AirTran Airways Corporation          92   101       91.1
```

Looks like, there is only one airline that flies this route and it is AirTran Airways.

Question 7: Can you detect Cascading Failures as delays in one airport create delays in another?

For this question, I wanted to see how does connection delays affect the two of busiest airports in USA which are Los Angeles Airport and San Francisco Airport.

```
la_to_sfo_flights <- data %>%
  filter((Year == 2007), (Month == 1), (Origin == 'LAX'), (Dest == 'SFO')) %>%
  select(contains(c('DepDelay','ArrDelay'))) %>%
  mutate(arrStatus = case_when(ArrDelay > 0 ~ "Delayed", T ~ "Early"))


la_to_sfo_cascading_plot <- ggplot(la_to_sfo_flights, aes(x=DepDelay, y=ArrDelay))
la_to_sfo_cascading_plot + geom_point(aes(colour = factor(arrStatus)), size=1.5 ) +
  scale_color_manual(values=c("#6d597a", "#eaac8b"), name="Flight Status") +
  geom_smooth(method='lm', formula=y~x, colour='#355070',
              data=filter(la_to_sfo_flights, arrStatus == 'Delayed')) +
  geom_smooth(method='lm', formula=y~x, colour='#e56b6f',
              data=filter(la_to_sfo_flights, arrStatus == 'Early')) +
  ggtitle("Flight delays between LA and SFO airports in Jan 2007") +
  xlab("Departure delay in minutes") +
  ylab("Arrival delay in minutes")
```



From the graph, we can see that there is a direct correlation between Departure Delay and Arrival Delays. Hence, we can say that if an incoming flight is late at SFO, then the outgoing flight from SFO will also be late and this would cause a chain reaction causing further delays as well.

Modelling

Let's Construct a model that could predict delays based on certain Predictors.

The flights data that has been used for this model is retrieved from the Source that I have referenced at the end of the report.

Unfortunately, Rstudio was not able to knit the outputs of the log models. I therefore had to set the evals of the below chunks to false, so that it does not try to display the output of the log models.

```
flights <- read.csv("flights.csv")
```

```
#Since we are only working with 2007's data, we can ignore the year column.
flights_ml_data <- left_join(flights,plane_data,by='TailNum') %>%
  drop_na() %>%
  filter(ArrDelay < 600)  %>%  # prevent outlier data from getting chosen
  slice_sample(n= 100000) %>% # randomized sample size of 100,000

  #selecting predictors
  select(c(Month,DayofMonth,DayOfWeek,DepTime,DepDelay,
           Origin,Dest,Distance,ArrDelay,PlaneYear)) %>%
  mutate(ArrStatus = case_when(ArrDelay > 0 ~ "Delayed", T ~ "Early")) %>%
  # create our target variable
  as.data.frame()
```

```
#Factoring the Predictors
flights_ml_data$ArrStatus <- factor(flights_ml_data$ArrStatus)

# convert to factor and set abbreviation of months
flights_ml_data$Month <- factor(flights_ml_data$Month,1:12,month.abb)

# convert to factor and set labels
flights_ml_data$DayOfWeek <- factor(flights_ml_data$DayOfWeek,
                           labels= c("Mon","Tues","Wed","Thurs","Fri","Sat",
                                     "Sun"))


flights_ml_data$DayofMonth <- factor(flights_ml_data$DayofMonth)

flights_ml_data$Origin <- factor(flights_ml_data$Origin)
flights_ml_data$Dest <- factor(flights_ml_data$Dest)
```

```
#Split the data into 70% Training and 30% Testing
set.seed(1)
dt = sort(sample(nrow(flights_ml_data), nrow(flights_ml_data)*.7))
train<-flights_ml_data[dt,]
test<-flights_ml_data[-dt,]
```

```
#Creating a Logistic Regression Model
log_model <- glm(ArrStatus ~ Month + DayofMonth + DayOfWeek + DepTime +
                 Origin + Dest + Distance + PlaneYear, data=train,
                 family="binomial")
summary(log_model)
```

I thought of making another log model with only the Origin and Destination as the predictors. The thinking behind this was many times people would just look at the origin and destination and make a call if the flight would arrive delayed or on Early. Also, it would be interesting as well to see how both the log models compare against each other.

```
orig_dest_log_model <- glm(ArrStatus ~ Origin + Dest, data = train, family="binomial")
summary(orig_dest_log_model)
```

Doing an ANOVA test to compare the full log_model and the orig_dest_log_model.

```
anova(log_model, orig_dest_log_model, test="Chisq")
```

Natrually, the full log_model is much more significant than the lesser model. This is understandable because predicting if a flight is going to be arriving delayed or not by just the origin and destination is not enough. There are countless other factors which would go into consideration in predicting.

Calculating the McFadden's R_Squared Value.

```
pR2(log_model)
```

The measure ranges from 0 to just under 1, with values closer to zero indicating that the model has no predictive power.

For our model, the value has come to somewhere around 0.05 which i would i say is somewhat more close to zero than being far from it which would indicate that our model has less predictive power.

Let's look at the Important Predictors of our log_model

```
Imp_variables <- as.data.frame(varImp(log_model))

Imp_variables <- Imp_variables %>%
  arrange(desc(Overall))

head(Imp_variables)
```

The Log Model has considered DepTime as the predictor with the highest importance followed by the Month of September. We also saw that Month of September had the lowest Delay percetanges of flight. The model also considered the year of the plane as significant in the prediction. This also makes sense because an older plane could suffer more delays than a newer plane. However, we also saw that two models of planes that were manufactured in the later years had the highest delay percentages as well.

Conclusions

- The best time to fly in 2007 was between 4am to 6am on Saturdays of September.
- 99% of the flights in 2007 were not diverted, reaching to their pre-planned destination
- Carrier and Weather were the predominant reasons of flight Cancellations.
- Planes Manufactured between 1966 to 1970 and 1976 to 1980 suffer the most delays.
- However, few models of planes manufactured after 1994 also suffered the most delays but those reasons could be not because of the plane but because of external factors like Passengers, Baggage or Ground Staffs.
- California, Texas, Illinois and Georgia experience the most of number of incoming and outgoing flights.
- Overall, the number of flights across various states have decreased in 2007.

- There are many routes with 100% delays but the number of flights on those routes are very less.
- Baltimore to Seattle is the route with the highest percentage of delayed flights.
- Tuscon to Oakland is the route with 0% of delayed flights.
- In the case of LAX to SFO, there was a direct correlation between an incoming flight getting delayed causing the departing flight from SFO to also get delayed. The findings can be extended to other more popular routes across USA.
- The logistic Regression Model didn't do a good job in predicting based on the predictors it was given. However, with more suitable data, there is room for improvement of the model.

Future Scope

- Combine Weather data of 2007 along with this data to understand what forms of Weather delays affected different routes.
- Build a better model to predict if a flight is going to be delayed or not.
- Get passenger data for different flights to understand the passenger traffic in various routes. Currently, we are going by the assumption that each flight is full but that is not always the case.
- Get a better understanding why planes manufactured after 1994 also have a high percentage of delays.

References

[1] https://github.com/Royleejy/Flight-Analysis-UOL-2022