Data Science Project

Sai Krishna Kaushik Pinnelli

2/27/2022

```
library(tidyverse)
library(ggplot2)
library(dplyr)
library(tidyr)
```

Purpose

The main purpose of this project is to analyse the Dulles Flights data set in all the aspects and put forth some insights and recommendations in order to maximize the operational efficiency and minimize the flight delays.

Introduction

Airports depend on accurate flight departure and arrival estimates to maintain operations, profitability, customer satisfaction, and compliance with state and federal laws. Flight performance, including departure and arrival delays must be monitored, submitted to the Federal Aviation Agency (FAA) on a regular basis, and minimized to maintain airport operations. The FAA considered a flight to be delayed if it has an arrival delay of at least 15 minutes.

As said in the purpose of this project, the main goal is to analyze the data and draw some inferences on the reasons for the flight delays in the Dulles International Airport. The `flights_df` data frame is loaded below and consists of 33,433 flights from IAD (Dulles International) in 2016. The rows in this data frame represent a single flight with all of the associated features that are displayed in the table below.

#Dulles Flights Data

```
flights_df <- readRDS(url('https://gmubusinessanalytics.netlify.app/data/dulles_flights.rds'))
view(flights_df)
```

#Exploratory Data Analysis

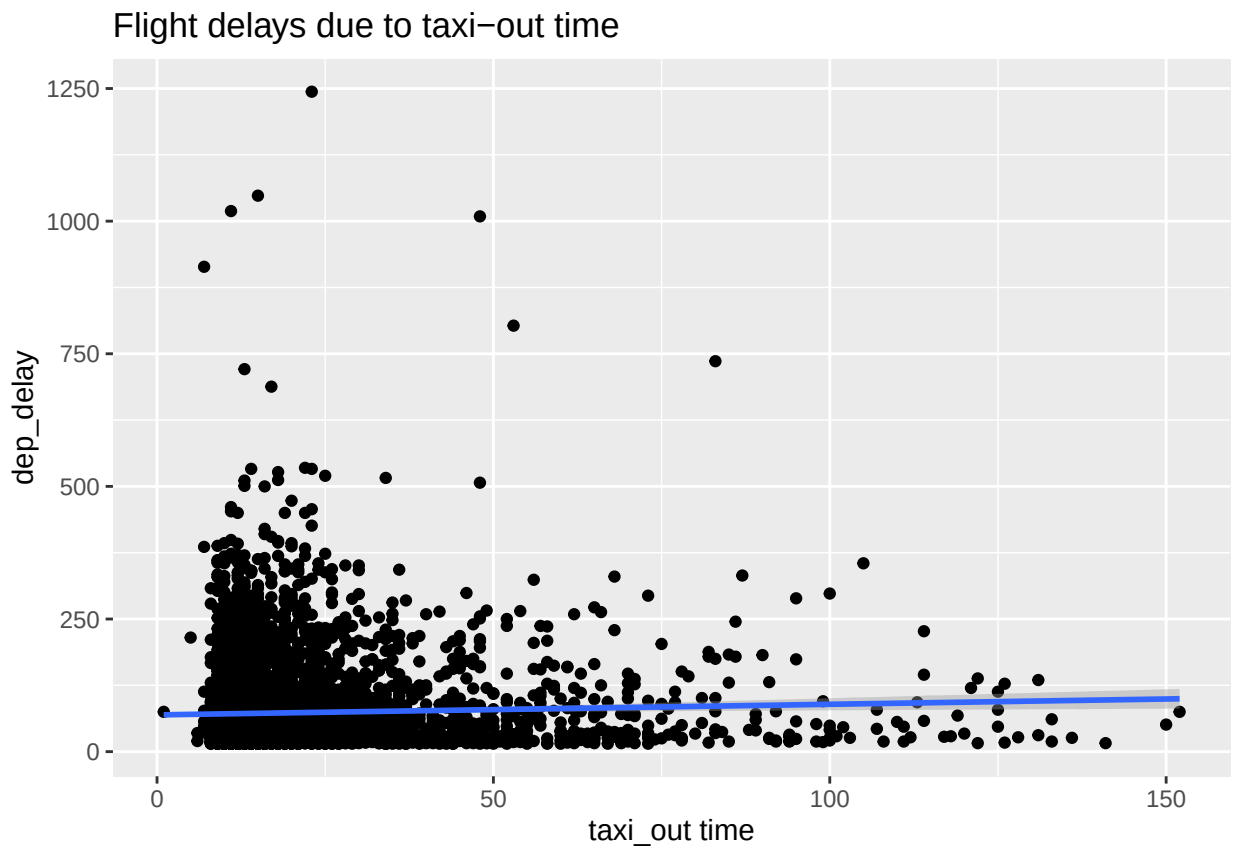
#1. Are flight delays affected by taxi-out time?

#Answer:- Yes. From the below scatterplot, we can say that there is a relationship between the departure delay and the taxi out time. This shows that there is a significant affect of taxi out time on the flight delays.

```
taxiout <- flights_df %>%
  filter(dep_delay>=15) %>%
  select(taxi_out, dep_delay, airline)

taxiout
```

```
# A tibble: 5,045 x 3
   taxi_out dep_delay airline
      <dbl>     <dbl> <fct>
 1       14        18 American
 2       11        45 American
 3       12       357 American
 4       36        47 American
 5       12        65 United
 6       10        27 United
 7       11        16 United
 8       16        27 United
 9       12        37 United
10        9        15 United
# ... with 5,035 more rows
```

```
# Scatter plot
ggplot(taxiout, aes(x=taxi_out, y= dep_delay))+
geom_point()+
  geom_smooth(method = lm)+
  labs(title = "Flight delays due to taxi-out time", x= "taxi_out time", y="dep_delay")
```



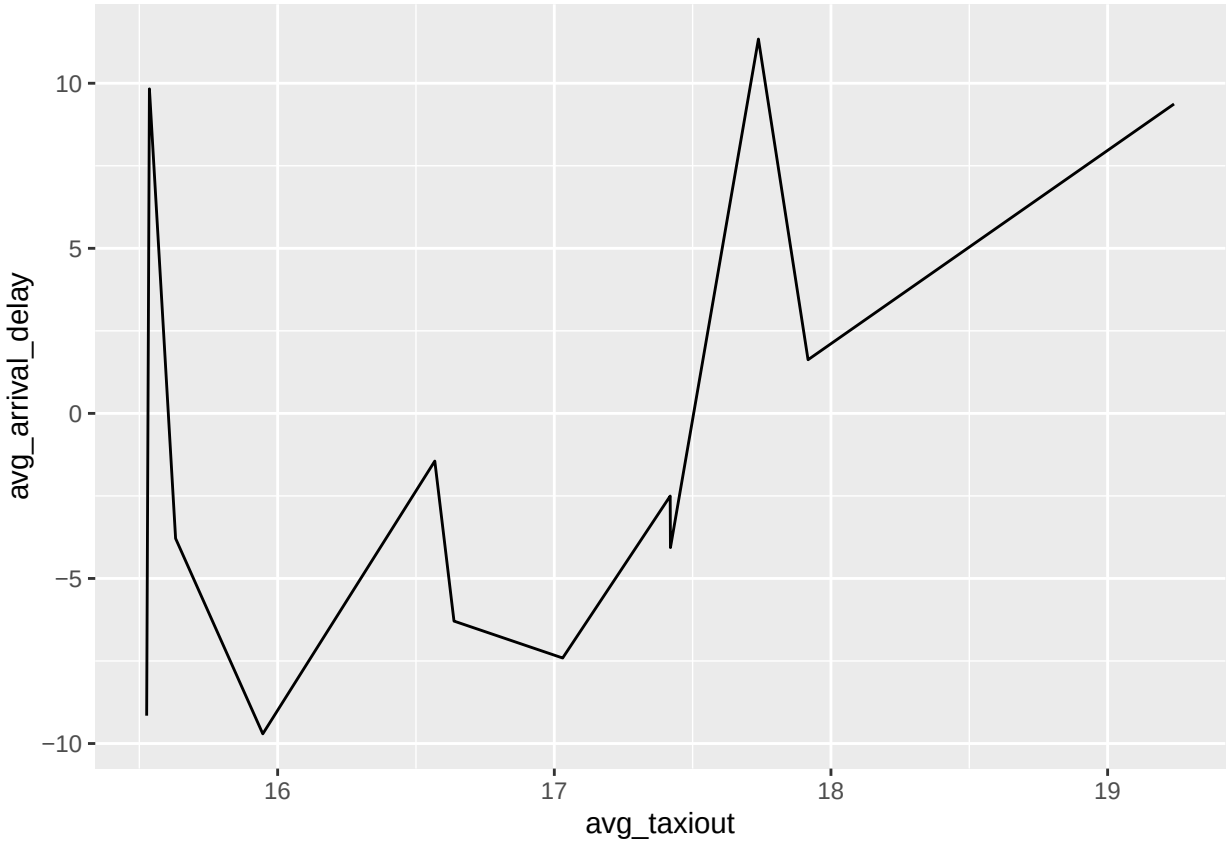
#2. How is the taxi_out time at the source affecting the arrival delay of the flight?

#Answer:- When we look at the table, we could observe that the months June, July, August and December have the positive average arrival delays. And from the line graph, we could get that when the average taxi_out time is greater than 17 and less than 18, the arrival delay is the maximum and that is in the month of July.

```
departure_delay <- flights_df %>% group_by(month) %>%
                   summarise(avg_taxiout = mean(taxi_out),
                     avg_arrival_delay = mean(arrival_delay))

departure_delay
```

```
# A tibble: 12 x 3
   month     avg_taxiout avg_arrival_delay
   <fct>           <dbl>             <dbl>
 1 January          15.5             -9.16
 2 February         17.0             -7.41
 3 March            15.6             -3.79
 4 April            16.6             -6.29
 5 May              17.4             -4.06
 6 June             19.2              9.37
 7 July             17.7             11.3
 8 August           17.9              1.63
 9 September        17.4             -2.50
10 October          16.6             -1.44
11 November         15.9             -9.71
12 December         15.5              9.83
```

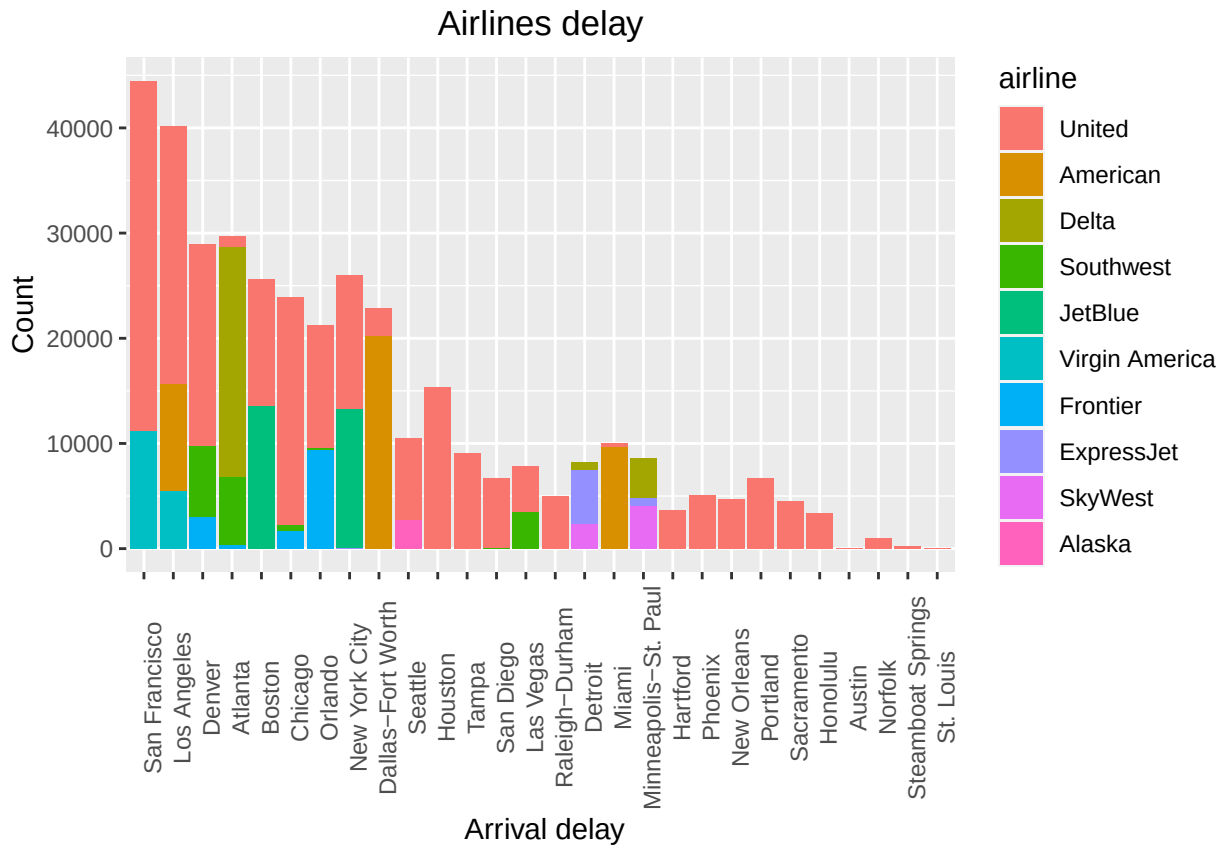
```
ggplot(departure_delay, aes(x = avg_taxiout, y = avg_arrival_delay)) +
  geom_line()
```



#3. Which airlines are prone to delay in different cities?

#Answer:- From the below chart, we can see that united airline always has flight delays than any other airline in all the cities except Detroit and Minneapolis-St.Paul. Frontier airline delays at Denver, Atlanta, Chicago, and Orlando City. Skywest airline delays at Minneapolis-St.Paul, Raleigh-Durham city. American airline delays at Los Angeles, Dallas-FortWorth, and Miami cities. No flight delays at Austin and St.Louis cities.

```
cities <- flights_df %>%
  filter(arrival_delay>=15) %>%
  select(arrival_delay,dest_airport_city,airline)

ggplot(cities, aes(x=dest_airport_city, y=arrival_delay, fill= airline)) +
  geom_bar(stat="identity")+
  labs(title = "Airlines delay", x=" Arrival delay", y="Count")+
  theme(axis.text.x = element_text(angle = 90))+
  theme(plot.title = element_text(hjust = 0.5))
```

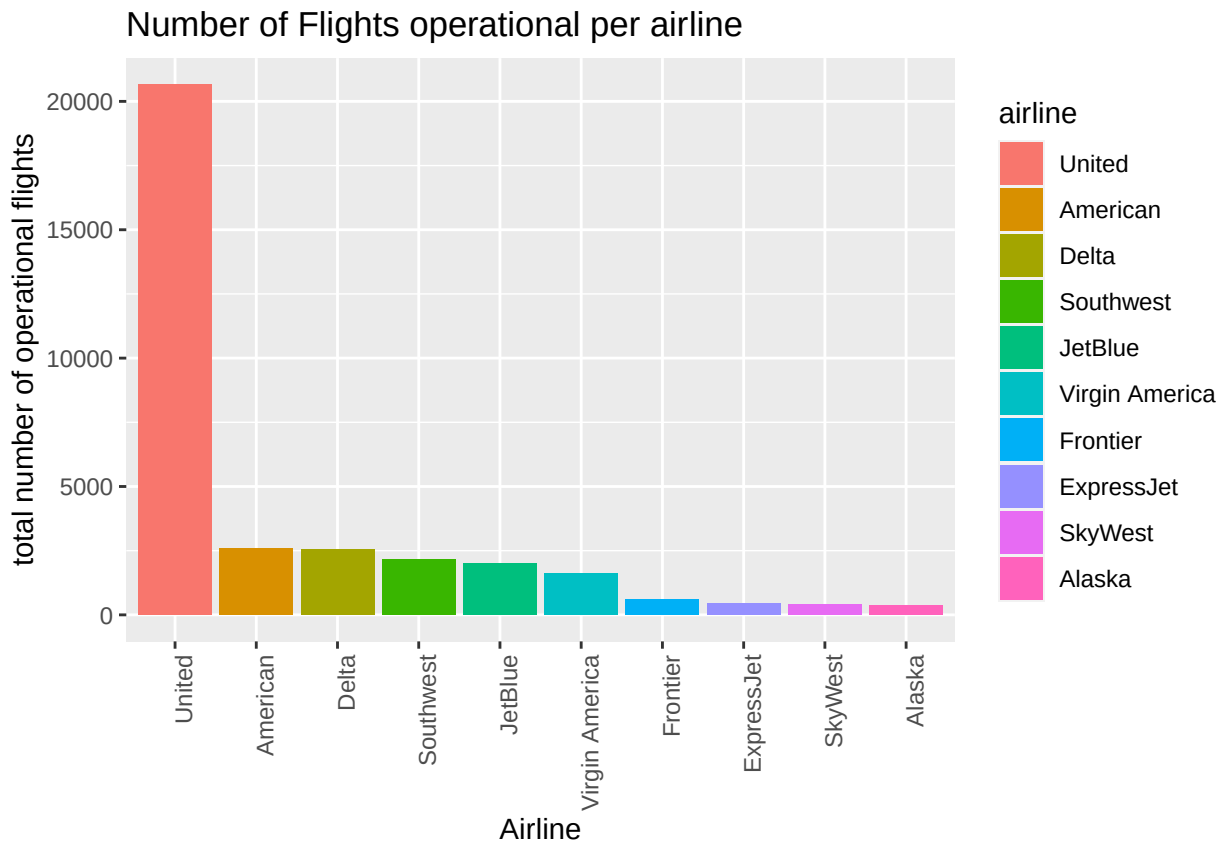


#4. Which airline has the maximum number of delayed flights?

#Answer:- United Airlines. From the below generated table total_to_delay, it says that most of the flights are run by United Airlines. When we compare the number of flights to delays, the maximum number of flight delays were by United Airlines.

```
totalflights <- flights_df %>%
  group_by(airline) %>%
  summarise(total = n())
```

```
#bar chart

ggplot(data = totalflights, mapping = aes(x = airline, y = total, fill = airline)) +
       geom_bar(stat = "identity")+
   labs(title = "Number of Flights operational per airline",
           x = "Airline",
           y = "total number of operational flights")+
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

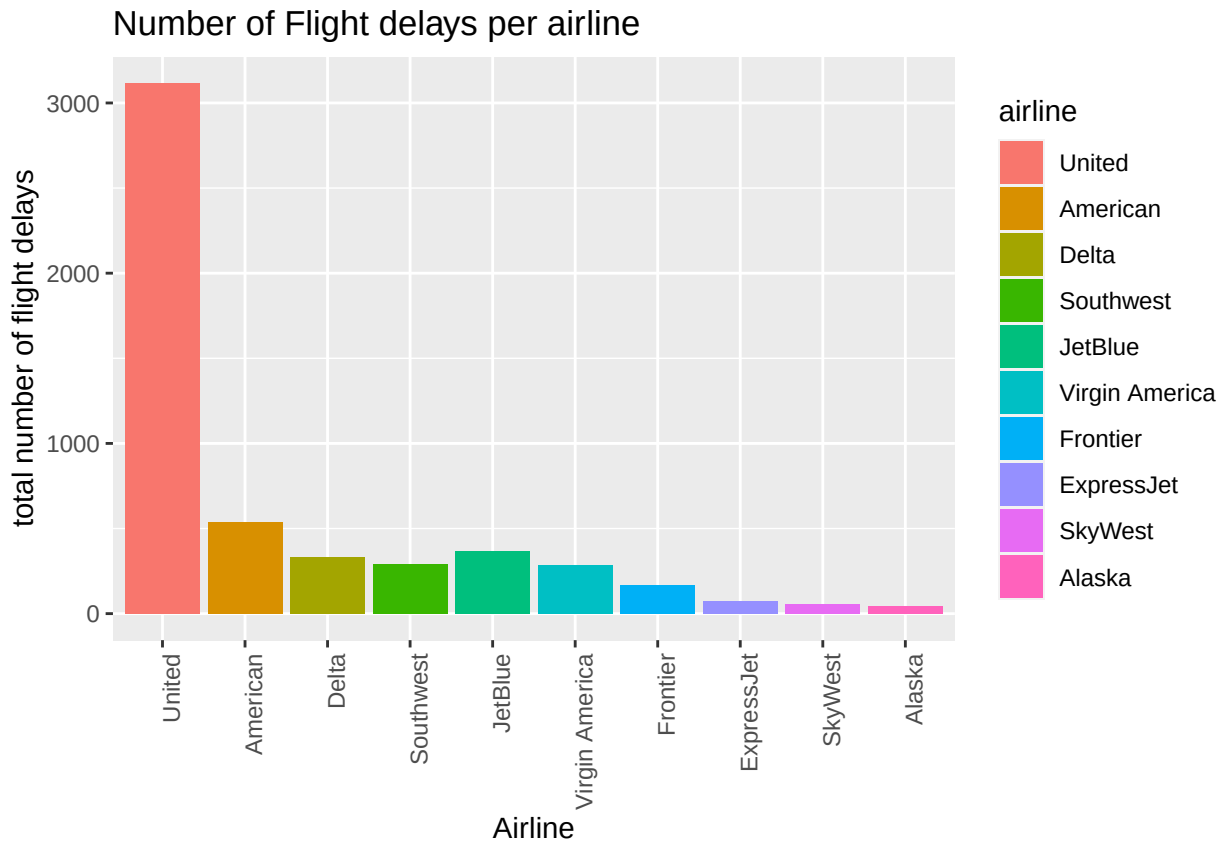



```
delays <-  flights_df %>% filter(arrival_delay >= 15) %>%
                  group_by(airline) %>%
                  summarise(delays = n())

total_to_delay <- left_join(totalflights, delays,
          by = c("airline" = "airline"))

total_to_delay
```

```
# A tibble: 10 x 3
   airline        total delays
   <fct>          <int>  <int>
 1 United         20653   3115
 2 American        2597    538
 3 Delta           2565    330
```

```
 4 Southwest       2161    288
 5 JetBlue         2013    365
 6 Virgin America  1613    285
 7 Frontier         618    167
 8 ExpressJet       453     72
 9 SkyWest          399     55
10 Alaska           361     43
```

```
#bar chart

ggplot(data = delays, mapping = aes(x = airline, y = delays, fill = airline)) +
       geom_bar(stat = "identity")+
   labs(title = "Number of Flight delays per airline",
           x = "Airline",
           y = "total number of flight delays")+
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



#5. Are the flight delays getting affected by the period of the year?

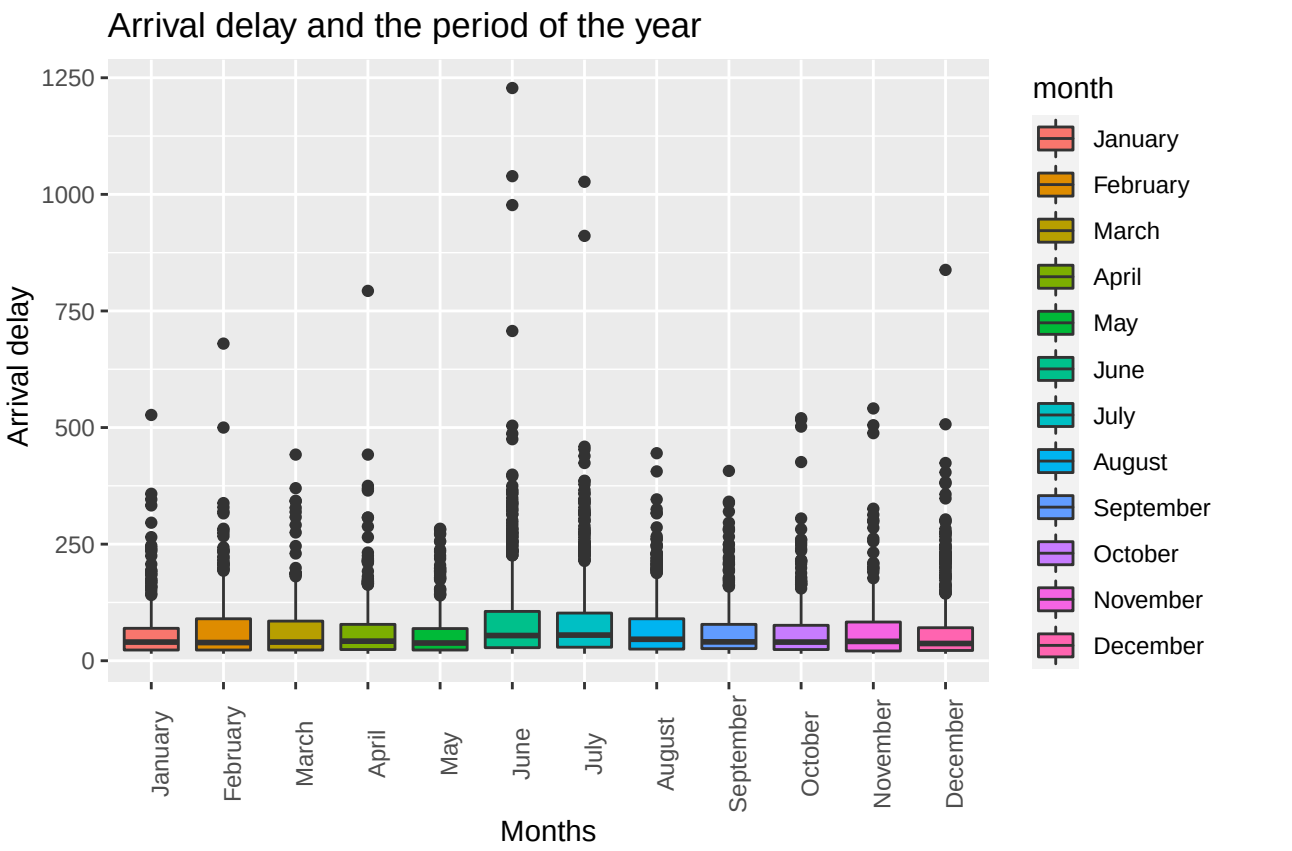
#Answer:- flight delay is affected by different periods of the year. From the analysis results above we can see that from January the flight delay tends to increase up to June. From June to December the flight delay tends to decrease. Therefore we can say that certain periods of the year affect flight delays.

```
monthdelay <- flights_df %>%
  filter(arrival_delay>=15) %>%
  select(month_numeric, month, arrival_delay)
```

```
monthdelay %>% group_by(month)%>% summarise(month_delay= mean(arrival_delay))
```

```
# A tibble: 12 x 2
   month     month_delay
   <fct>           <dbl>
 1 January          64.2
 2 February         74.0
 3 March            65.1
 4 April            67.2
 5 May              57.9
 6 June             88.4
 7 July             84.2
 8 August           68.5
 9 September        64.6
10 October          63.0
11 November         69.1
12 December         63.3
```

```
#Box Plot

ggplot(monthdelay, aes(x=month , y= arrival_delay, fill= month))+ geom_boxplot()+
  labs(title = "Arrival delay and the period of the year", x= "Months", y= " Arrival delay")+
  theme(axis.text.x = element_text(angle = 90))+
theme(plot.title = element_text(hjust = 0))
```



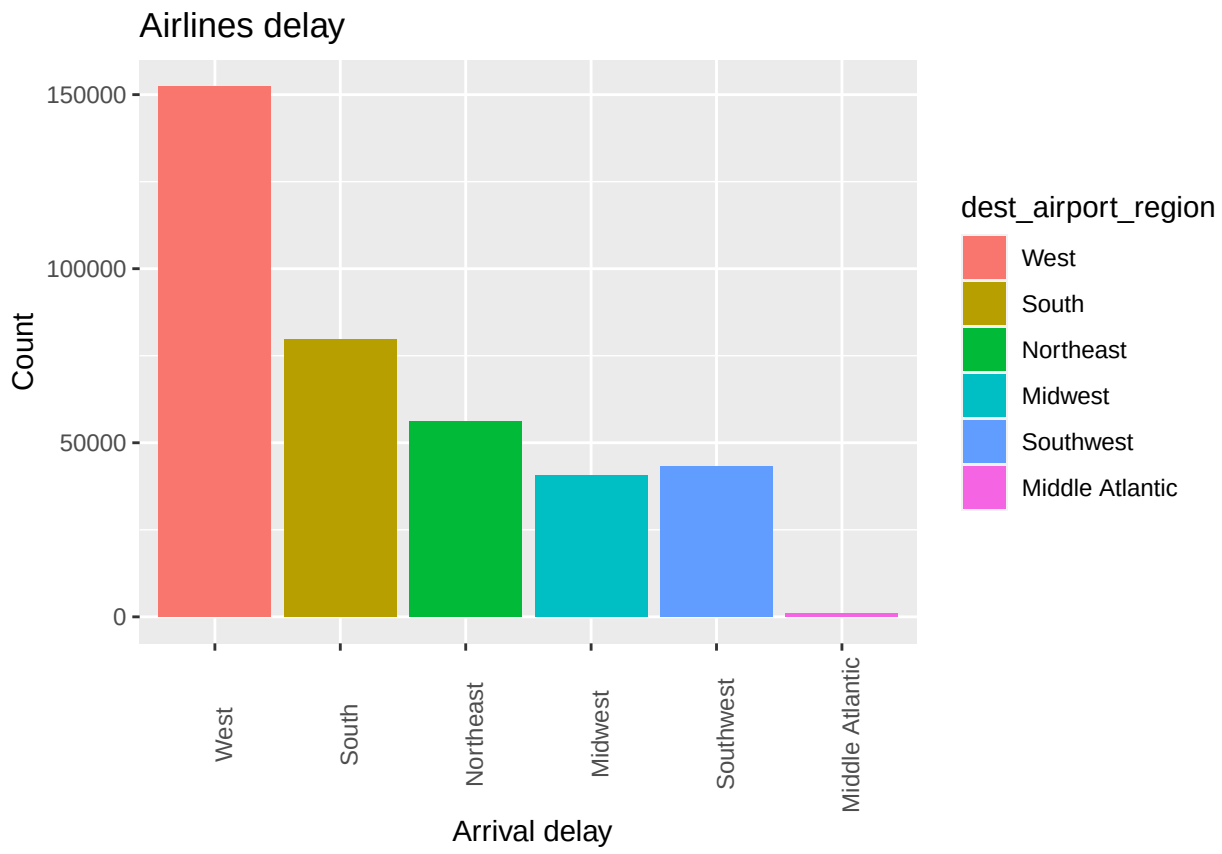
#6. In what airport regions the maximum dealys are taking place?

#Answer:- From the above graph, we could see that most of the flights travelling from the West region are prone to delays. The prime reason for this delay could be distance traveled from the west to Dulles. The next region is the South. The least is the Middle Atlantic.

```
airline <- flights_df %>%
  filter(arrival_delay>=15) %>%
  select(dest_airport_region, arrival_delay)

airline
```

```
# A tibble: 5,258 x 2
   dest_airport_region arrival_delay
   <fct>                       <dbl>
 1 West                          333
 2 West                           41
 3 Northeast                      38
 4 Northeast                      46
 5 West                           17
 6 Northeast                      23
 7 Midwest                        25
 8 Southwest                      29
 9 South                          60
10 West                           36
# ... with 5,248 more rows
```

```
ggplot(airline , aes(x= dest_airport_region, y= arrival_delay, fill=dest_airport_region)) +
  geom_bar(stat="identity")+
  labs(title = "Airlines delay", x=" Arrival delay", y="Count")+
  theme(axis.text.x = element_text(angle = 90))+
  theme(plot.title = element_text(hjust = 0))
```



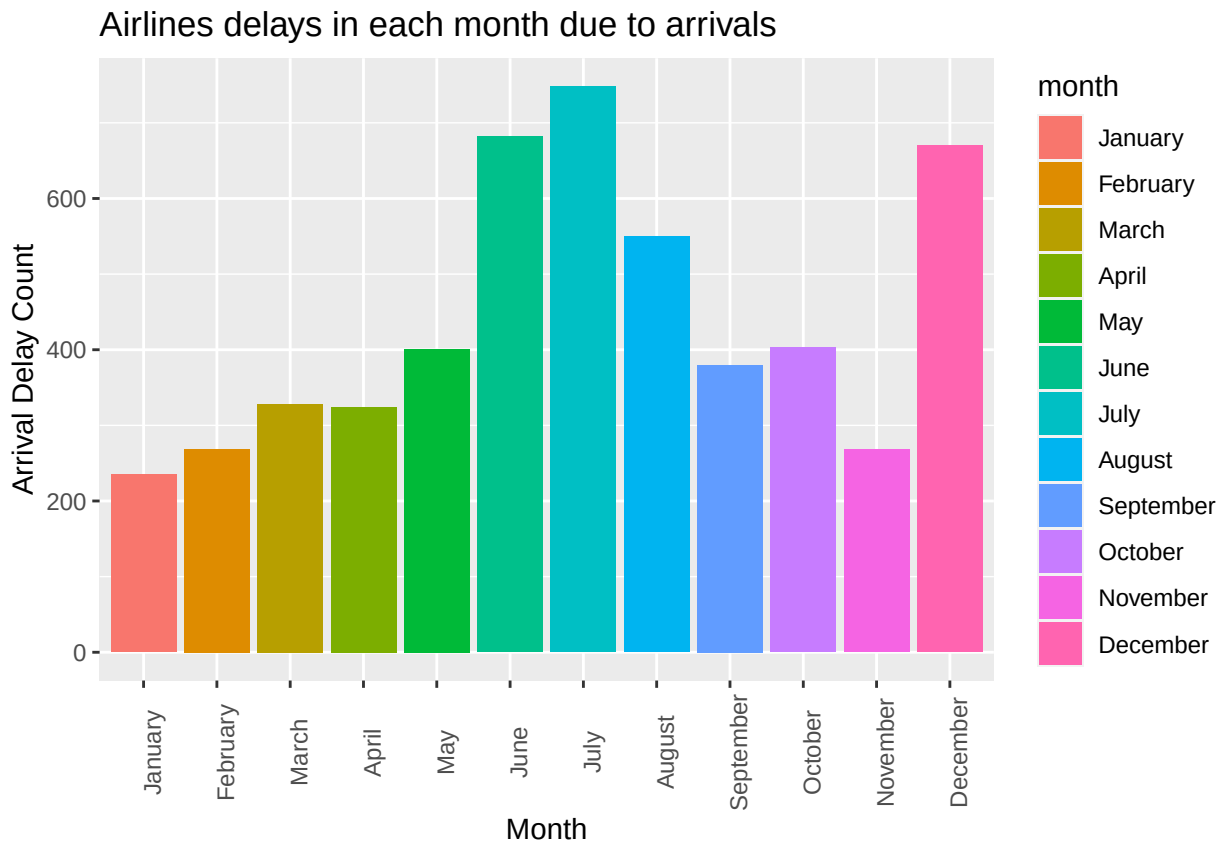
#7) Which month has the maximum number of delays with respect to arrival delay time?

#Answer:- From the below table, we can say that the maximum number of delays happened in the month of July with count 748. June has the second highest number of delayed flights. Both June and July are the hottest months of the year. This means the flight delays maybe due to the weather conditions.

```
max_month_arrival<- flights_df %>% filter(arrival_delay >= 15) %>%
  group_by(month) %>% summarise(arrival_delays = n())

max_month_arrival
```

```
# A tibble: 12 x 2
   month     arrival_delays
   <fct>              <int>
 1 January              235
 2 February             269
 3 March                328
 4 April                324
 5 May                  401
 6 June                 682
 7 July                 748
 8 August               550
 9 September            380
10 October              403
11 November             268
12 December             670
```

```
ggplot(max_month_arrival, aes(x=month, y=arrival_delays, fill = month))+
  geom_bar(stat = "identity")+
  labs(title = "Airlines delays in each month due to arrivals", x="Month", y="Arrival Delay Count")+
  theme(axis.text.x = element_text(angle = 90))+
  theme(plot.title = element_text(hjust = 0))
```



#8) Which month has the maximum number of delays with respect to departure delay time?

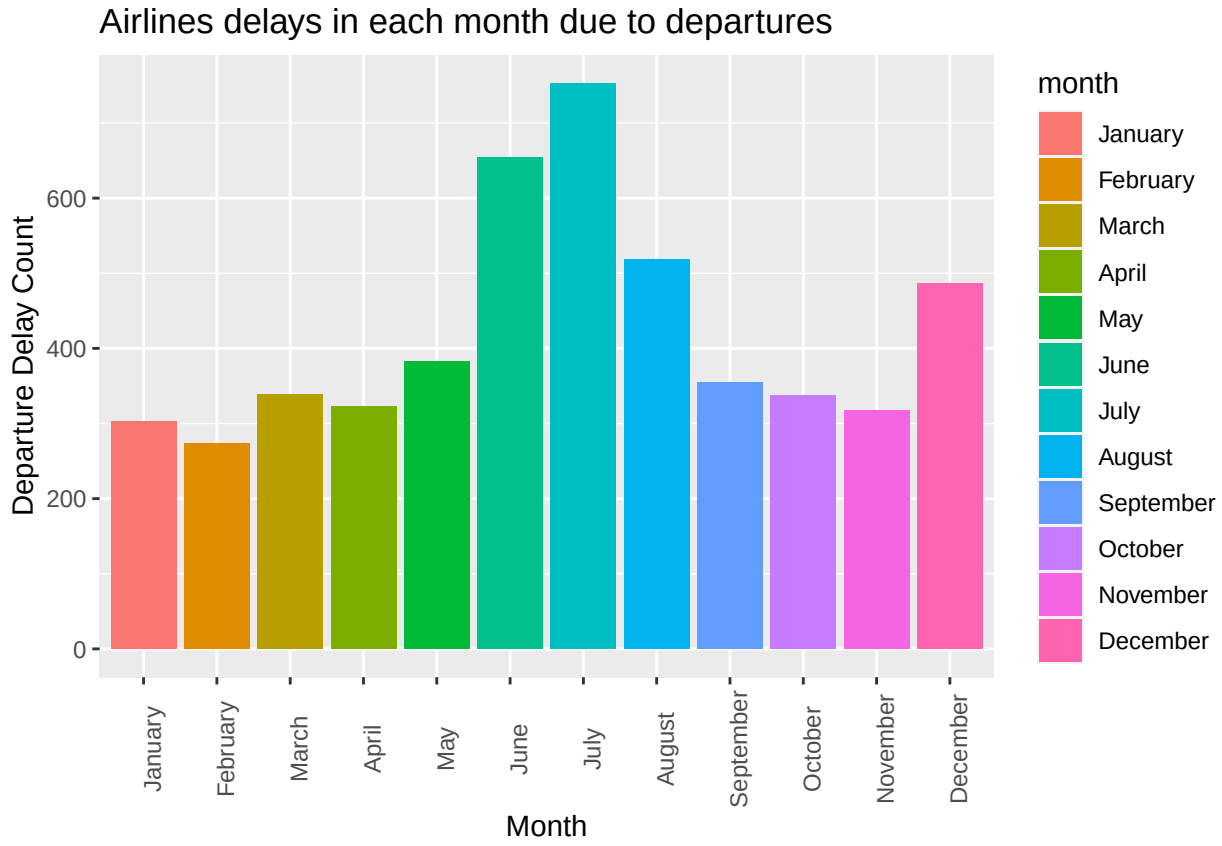
#Answer:- From the below table, we can say that the maximum number of delays happened in the month of July with count 753. June has the second highest number of delayed flights. Both June and July are the hottest months of the year. This means the flight delays maybe due to the weather conditions.

```
max_month_dep<- flights_df %>% filter(dep_delay >= 15) %>%
  group_by(month) %>% summarise(depart_delays = n())

max_month_dep
```

```
# A tibble: 12 x 2
   month     depart_delays
   <fct>             <int>
 1 January             303
 2 February            274
 3 March               339
 4 April               323
 5 May                 383
 6 June                654
```

```
 7 July                753
 8 August              518
 9 September           355
10 October             338
11 November            318
12 December            487
```

```
ggplot(max_month_dep, aes(x=month, y=depart_delays, fill = month))+
  geom_bar(stat = "identity")+
  labs(title = "Airlines delays in each month due to departures", x="Month", y="Departure Delay Count")+
  theme(axis.text.x = element_text(angle = 90))+
  theme(plot.title = element_text(hjust = 0))
```



#9. How many flights are delayed by both Departure delays and Arrival Delays?

#Answer:- Out of both departure delays and Arrival Delays, most of the flights are delayed due to departure delays. And all the flights are delayed in the months of June, July and December. So we have to take care of the Arrivals and Departures in those months specifically.

```
monthly_delays <- left_join(max_month_dep, max_month_arrival,
          by = c("month" = "month"))

monthly_delays
```

```
# A tibble: 12 x 3
   month     depart_delays arrival_delays
```

```
   <fct>             <int>          <int>
 1 January             303            235
 2 February            274            269
 3 March               339            328
 4 April               323            324
 5 May                 383            401
 6 June                654            682
 7 July                753            748
 8 August              518            550
 9 September           355            380
10 October             338            403
11 November            318            268
12 December            487            670
```

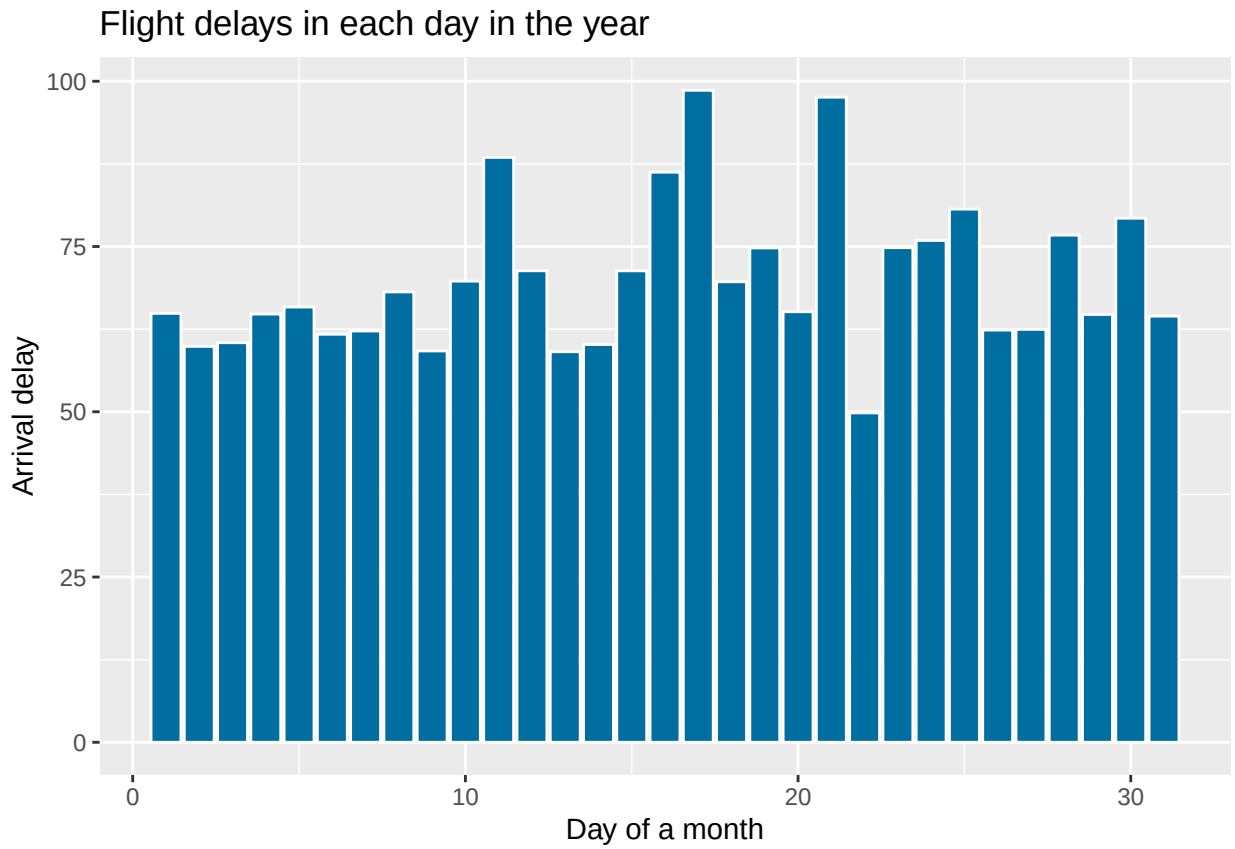
#10. Are certain times of the day of the year problematic?

#Answer:- From the below bar graph, we can observe that different days of month has different times of arrival delays. But, we we get a microscopic observation, we can notice that the 11, 17 and 21st days of the month has the highest delay in the arrival time. This cannot be the same scenario for all the months but from a generalised view. Therefore, we can conclude that the times of the day of the year could be a reason for the flight delays.

```
days <- flights_df %>%
  filter(arrival_delay>=15) %>%
  select(arrival_delay, day)

delay_days<- days%>% group_by(day)%>%
  summarise(day_mean=mean(arrival_delay))%>%
  arrange(day)

delay_days
```

```
# A tibble: 31 x 2
     day day_mean
   <dbl>    <dbl>
 1     1     64.9
 2     2     59.9
 3     3     60.4
 4     4     64.8
 5     5     65.8
 6     6     61.7
 7     7     62.2
 8     8     68.1
 9     9     59.2
10    10     69.8
# ... with 21 more rows
```

```
ggplot(delay_days, aes(x=day, y = day_mean))+
  geom_bar(stat = "identity", fill = "#006EA1", color = "white")+
  labs(title = "Flight delays in each day in the year", x= "Day of a month", y="Arrival delay")+
  theme(plot.title = element_text(hjust = 0))
```



#Summary of Results:

Executive Summary:

In every business, the main factor that makes the business a billion dollar one is the customer satisfaction. In the Airline business, customer satisfaction mainly depends on the timely operations of the flight. The timely operations in the sense perfect departure time, perfect arrival time, timely boarding and comfortable travel. All these boarding, arrival and departure are inter-dependent. if one gets delayed, rest of the two activities gets delayed. Such a same problem has been arouse at the Dulles International Airport. The executives at the Dulles International Airport have to solve this problem. They have provided us with the complete flights data in the year 2016 to analyze the problem using some visualizations and provide them some recommendations. The flights_df data frame is loaded below and consists of 33,433 flights from IAD (Dulles International) in 2016. The rows in this data frame represent a single flight with all of the associated features.

Generally the Departures and arrivals depend on the taxi in time and taxi out time. There are many reasons for the delays. The delay might be due to the air traffic issues, runway issues, technical issues in the air traffic control. To actually find out the root cause of the delay, we must first get the knowledge of the delay i.e., what are the major factors for the delays.

#Summary from all the questions:

From the data collected, Most of the flights are run by United Airlines. When we compare the number of flights to delays, the maximum number of flight delays were by United Airlines. United airline always has flight delays than any other airline in all the cities except Detroit and Minneapolis-St.Paul. Frontier airline delays at Denver, Atlanta, Chicago, and Orlando City. Skywest airline delays at Minneapolis-St.Paul, Raleigh-Durham city. American airline delays at Los Angeles, Dallas-FortWorth, and Miami cities. No flight delays at Austin and St.Louis cities.

When it comes to the taxi in and taxi out times, there is certainly a close relationship between the departure

delay and the taxi out time. This shows that there is a significant effect of taxi out time on the flight delays. In the months June, July, August and December have the positive average arrival delays. When the average taxi_out time is greater than 17 and less than 18, the arrival delay is the maximum and that is in the month of July. The other aspect of the flight delay is the period of the year. flight delay is affected by different periods of the year. From the analysis results, we can see that from January the flight delay tends to increase up to June. From June to December the flight delay tends to decrease. Therefore we can say that certain periods of the year affect flight delays. The flights travelling from the West region are prone to delays. The prime reason for this delay could be distance traveled from the west to Dulles. The next region is the South. The least is the Middle Atlantic. we can also observe that different days of month has different times of arrival delays. But, we we get a microscopic observation, we can notice that the 11, 17 and 21st days of the month has the highest delay in the arrival time. This cannot be the same scenario for all the months but from a generalized view. Therefore, we can conclude that the times of the day of the year could be a reason for the flight delays. Out of both departure delays and Arrival Delays, most of the flights are delayed due to departure delays. And all the flights are delayed in the months of June, July and December. So we have to take care of the Arrivals and Departures in those months specifically. The maximum number of delays happened in the month of July with count 748. June has the second highest number of delayed flights. Both June and July are the hottest months of the year. This means the flight delays maybe due to the weather conditions. The maximum number of delays happened in the month of July with count 753. June has the second highest number of delayed flights. Both June and July are the hottest months of the year. This means the flight delays maybe due to the weather conditions.

On the whole, from the whole analysis, the main factors for the airline delays are Taxi in time and Taxi out time. We can also infer that the delays are mostly happening while the departure of the flight which again links to the long taxi out time. Long taxi out time will lead to airline traffic on the runway. Due to the traffic, the airline gets delayed in reaching its destination. All the scheduled arrival and departure times are mostly calculated according to the capacity of the flight, type of flight, total distance and all the other factors that are needed for a timely operation.

The primary recommendation that can be made to control the delays is to reduce the taxi out time. This is completely in the hands of the Air Traffic Control. As said earlier, if the taxi out time is controlled, all the operation will be done in time and the organization will be profitable with a great customer satisfaction.