Exploring NYC flights data in R

Shree Priya

Setup:

Do whatever setup you do here, such as loading libraries

```
# Load standard libraries
library("tidyverse")
library("nycflights13")
data(flights)
library(plotly)
```

Problem 1: Exploring the NYC Flights Data

(a) Importing and Inspecting Data:

```
# Load standard libraries
library("tidyverse")
library("nycflights13")
library(dplyr)
mydata = filter(flights, year == 2013)
#Removing all the null values in the arrival and departure delay
data1 = mydata %>% filter(is.na(arr_delay) == FALSE) %>% filter(is.na(dep_delay) == FALSE)

#Checking the number of total number of flights by each carrier
cat(paste("The total number of flights by each carrier are:"))
```

```
## The total number of flights by each carrier are:
```

```
table(data1$carrier)
```

```
##
##    9E    AA    AS    B6    DL    EV    F9    FL    HA    MQ    OO    UA
## 17294 31947   709 54049 47658 51108   681  3175   342 25037    29 57782
##    US    VX    WN    YV
## 19831  5116 12044   544
```

```
#Checking the total number of carriers
cat(paste("The total number of carriers are: \n"))
```

```
## The total number of carriers are:
```

```
dim(table(data1$carrier))
```

```
## [1] 16
```

```
cat(paste("The total number of rows and columns are: \n"))
```

```
## The total number of rows and columns are:
```

```
dim(data1)
```

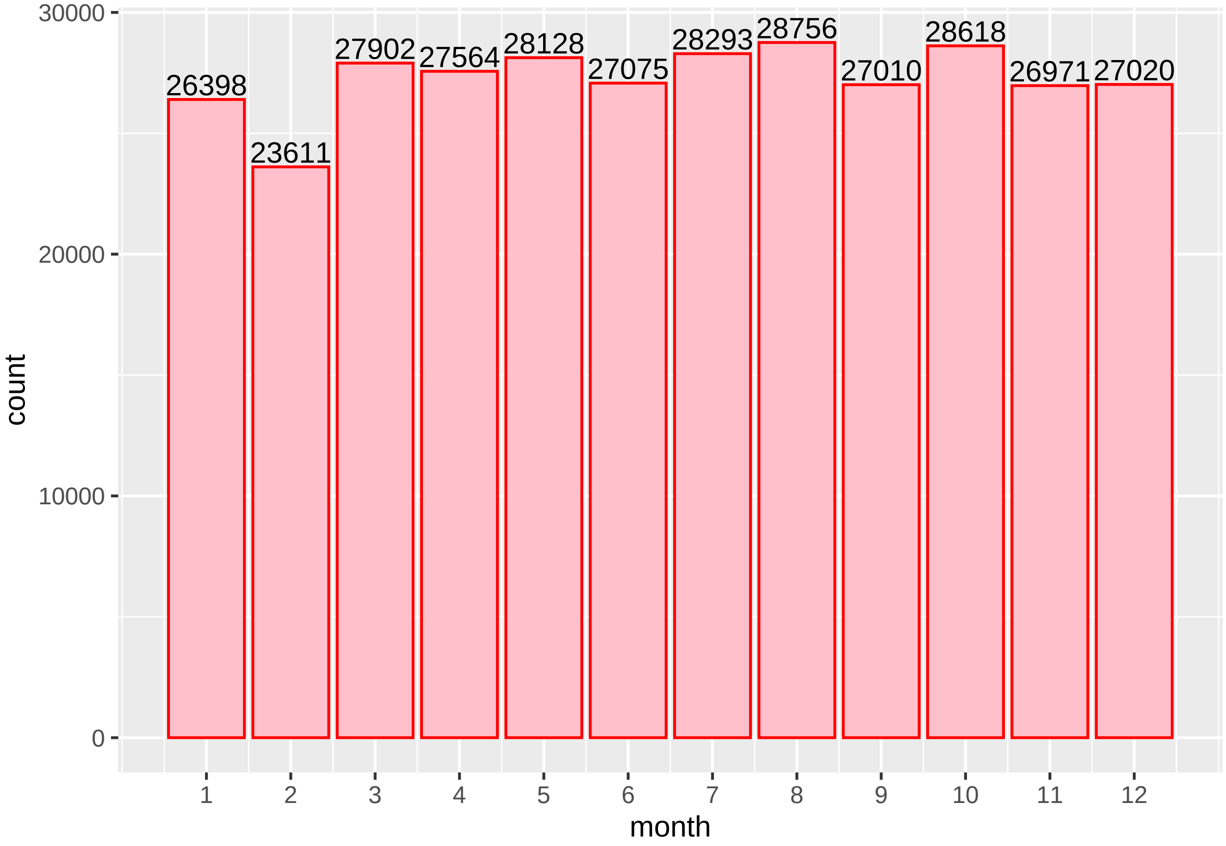
```
## [1] 327346     19
```

```
cat(paste("The total number of flights each month are"))
```

```
## The total number of flights each month are
```

```
ggplot(data1, aes(month)) + geom_histogram(stat = "count", color="Red", fill="Pink") + scale_x_continuou
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



This data set consists of all on-time data for all flights that departed NYC(i.e JFK, LGA, EWR) in 2013. For the date of departure there is –> year, month, day For the actual departure and arrival times there is –> dep_time, arr_time For the schedule dep and arr time there is –> sched_dep_time, sched_arr_time for the departure and arrival delays there is –> dep_delay, arr_delay There is carrier for the carrier of the flight There is the flight number, tail number, origin, dest, air time, distance of the flight.

(b) Formulating Questions:

1. Exploring arrival delay

   a. Which are the top 3 months with the highest arrival delay?
   b. Why do you think these were the months with most delay?
   c. Which airline carrier has the most number of delays in those months?
   d. Choose a month and check if the airline had the most number of arrival delays that month

2. Exploring departure delay

   a. Which are the top 3 months with the highest departure delay?
   b. Why do you think these were the months with most delay?
   c. Which airline carrier has the most number of delays in those months?
   d. Choose a month and check if the airline had the most number of departure delays that month

3. Was there any correlation between 1 and 2

4. Which airports have the highest departure delay?

5. Did more number of flights that departed from 6am to 6pm get delayed?

(c) Exploring Data:

Question 1:Exploring arrival delay

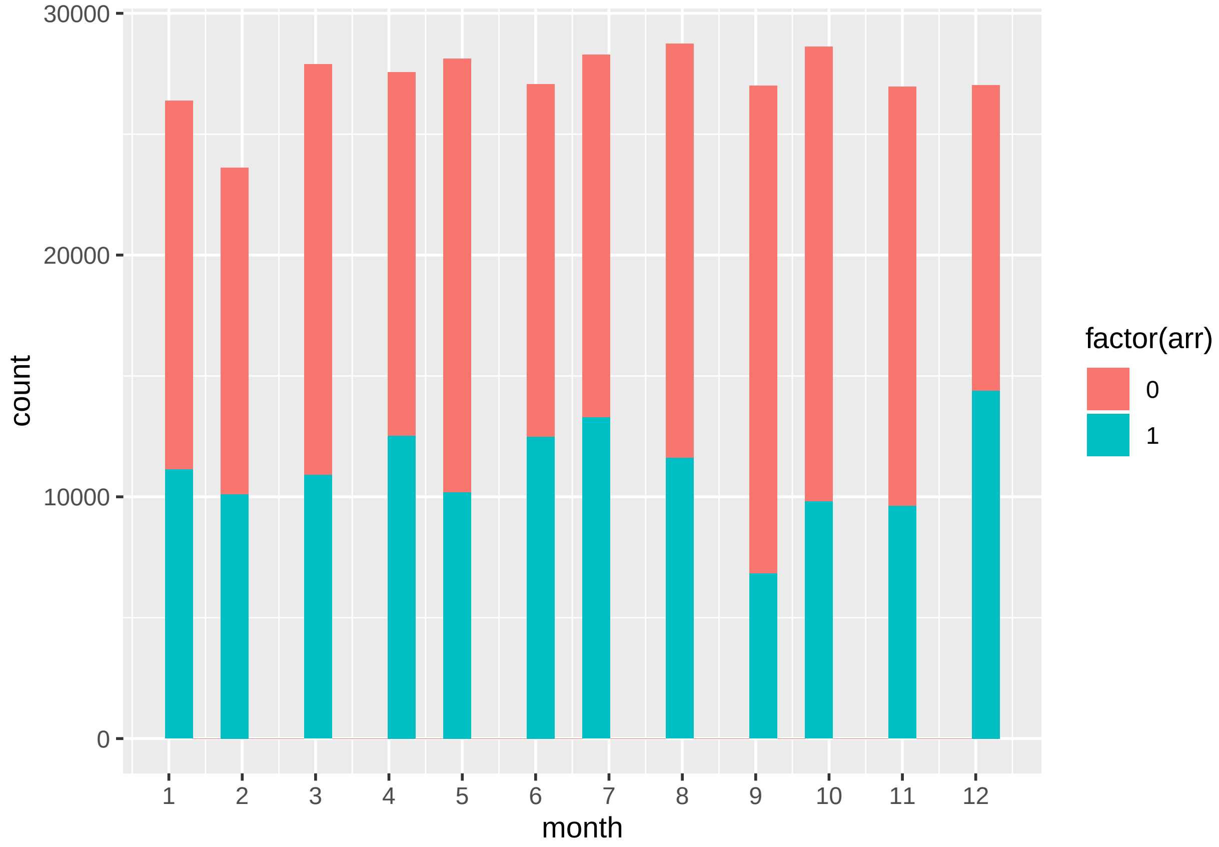
Part a :Which are the top 3 months with the highest arrival delay?

```r
#data1 no missing values
head(data1)
```

```
## # A tibble: 6 x 19
##    year month   day dep_time sched_dep_time dep_delay arr_time
##   <int> <int> <int>    <int>          <int>     <dbl>    <int>
## 1  2013     1     1      517            515         2      830
## 2  2013     1     1      533            529         4      850
## 3  2013     1     1      542            540         2      923
## 4  2013     1     1      544            545        -1     1004
## 5  2013     1     1      554            600        -6      812
## 6  2013     1     1      554            558        -4      740
## # ... with 12 more variables: sched_arr_time <int>, arr_delay <dbl>,
## #   carrier <chr>, flight <int>, tailnum <chr>, origin <chr>, dest <chr>,
## #   air_time <dbl>, distance <dbl>, hour <dbl>, minute <dbl>,
## #   time_hour <dttm>
```

```r
#Mutating the data add a column arr which is 1 if the arrival delay is greater than 0 and 0 if the arri
arrival_flights <- data1 %>% group_by(month) %>% mutate( arr = ifelse(arr_delay >0 , 1, 0))

#Plotting the histogram with a factor of whether it was delayed or not.
ggplot(arrival_flights, aes(month, fill = factor(arr)), labels = TRUE) +
   geom_histogram() + scale_x_continuous(breaks = c(1:12))
```



According to this plot we can see that the maximum delays happen in month 6, 7 and 12.

Part b: Why do you think these were the months with most delay?

June, July and December tend to be the holidays in the US. Most of the flights are overbooked and there are too many flights. This usually causes delaying of flights.

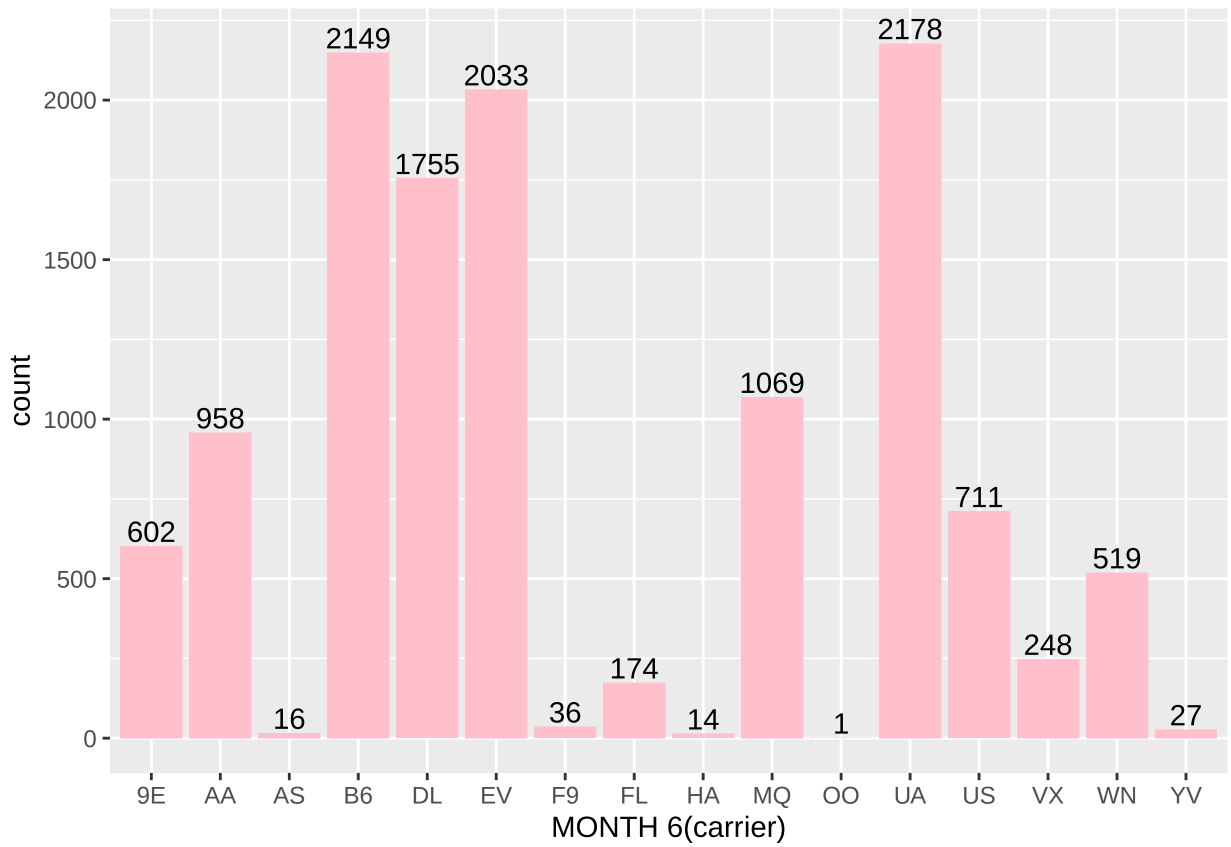
Part c: Which airline carrier has the most number of delays in those months?

```
#Filtering data of month 6 with positive arrival delay
df1 = data1 %>% filter(month == 6 & arr_delay >0)
dim(df1)
```

```
## [1] 12490    19
```

```
#Plotting the data
ggplot(df1, aes(carrier, fill = arr_delay)) + geom_histogram(stat="count", fill="pink") + xlab("MONTH 6
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

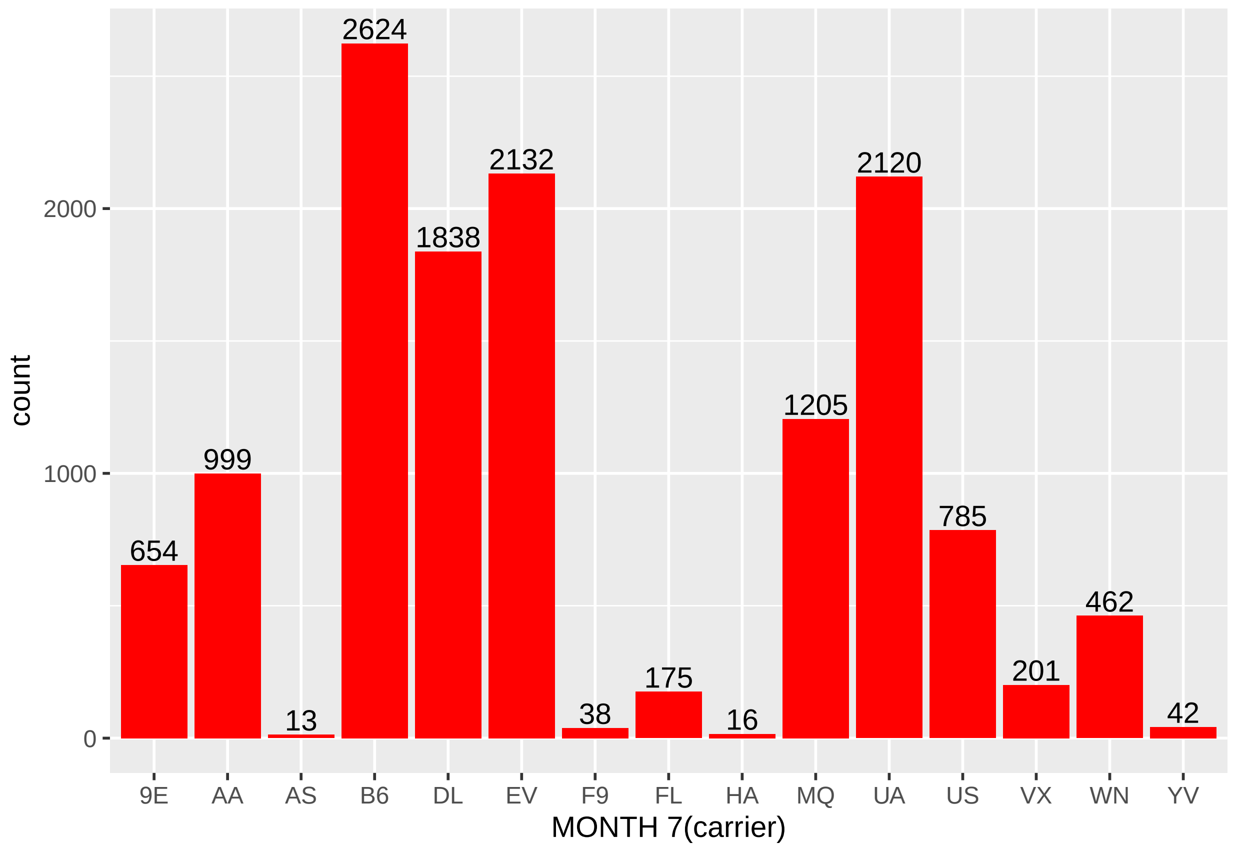


```
#Filtering data of month 7 with positive arrival delay
df2 = data1 %>% filter(month == 7 & arr_delay >0)
dim(df2)
```

```
## [1] 13304    19
```

```
#Plotting the data
ggplot(df2, aes(carrier, fill = arr_delay)) + geom_histogram(stat="count", fill="red") + xlab("MONTH 7(
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

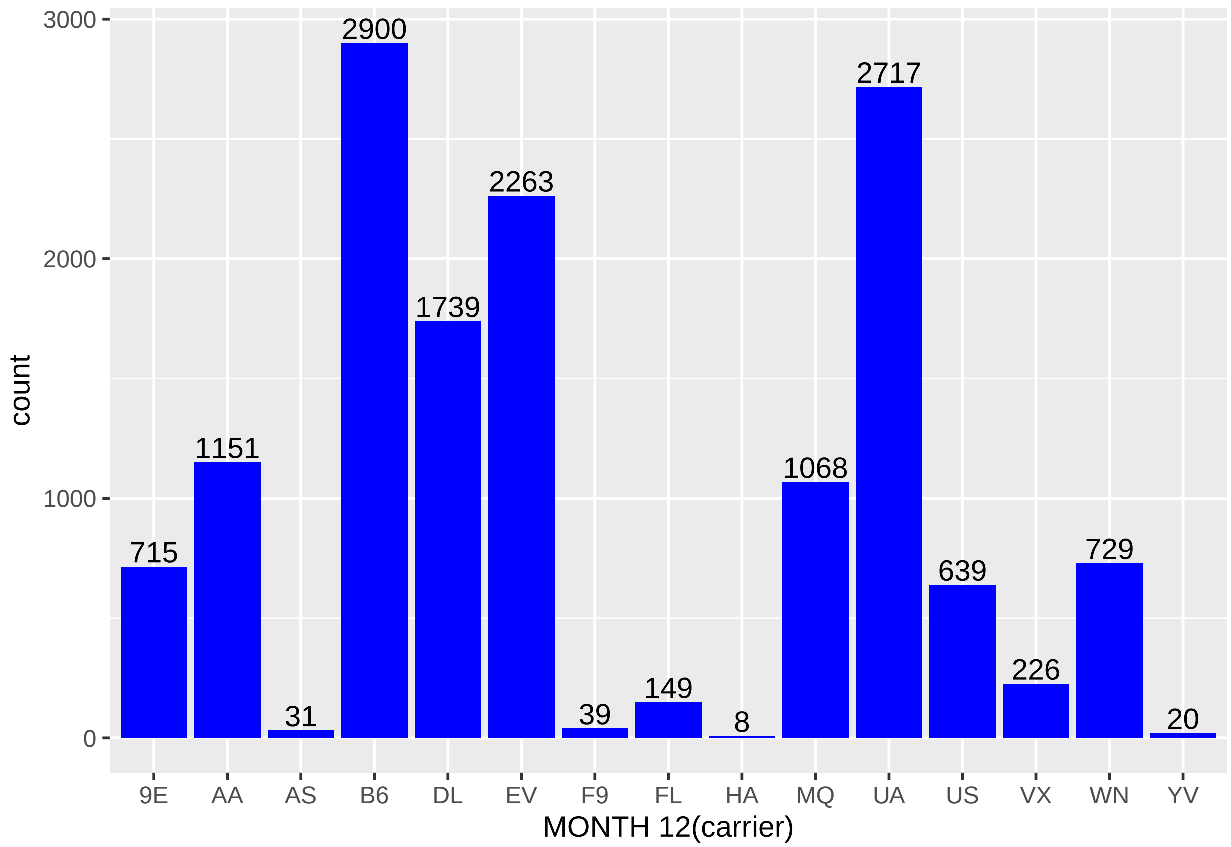


```
#Filtering data of month 12 with positive arrival delay
df3 = data1 %>% filter(month == 12 & arr_delay >0)
dim(df3)
```

```
## [1] 14394    19
```

```
#Plotting the data
ggplot(df3, aes(carrier, fill = arr_delay)) + geom_histogram(stat="count", fill="blue") + xlab("MONTH 1
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



We can clearly see that most of the delays happen in carrier B6 and UA

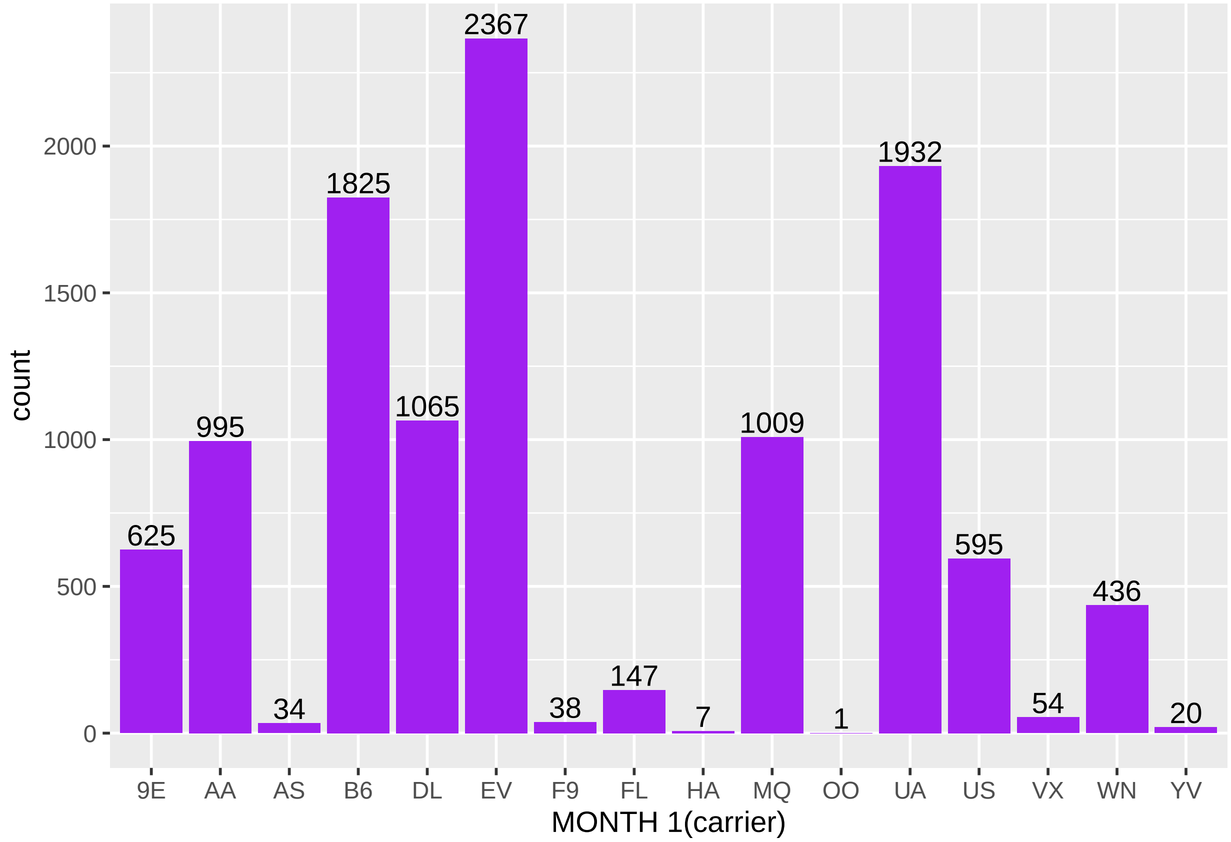
Part d: Choose a month and check if the airline had the most number of arrival delays that month.

```
#Choosing month 1 for checking the arrival delays that month
df4 = data1 %>% filter(month == 1 & arr_delay >0)
dim(df4)
```

```
## [1] 11150    19
```

```
#Plotting the data
ggplot(df4, aes(carrier, fill = arr_delay)) + geom_histogram(stat="count", fill="purple") + xlab("MONTH
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



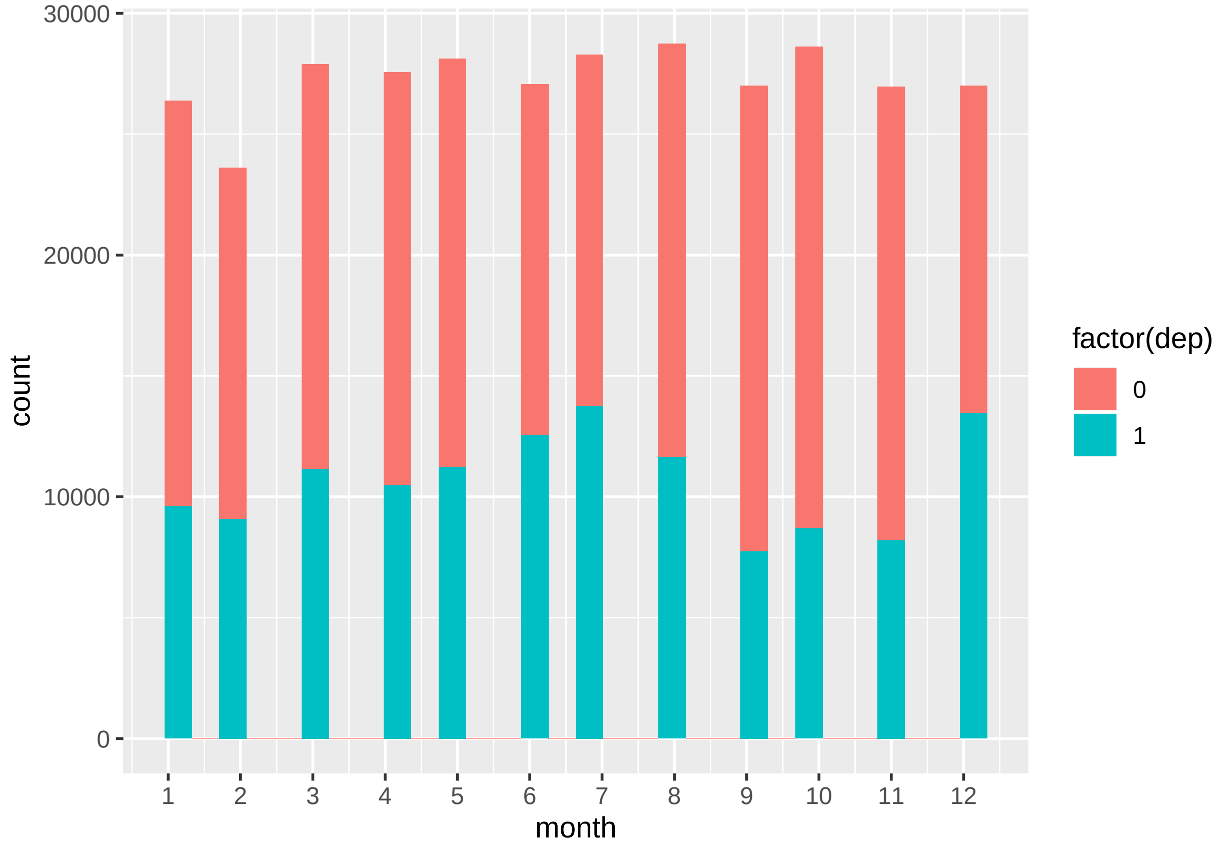
We can see that neither UA nor B6 has the highest arrival delays in month 1.

Question 2: Exploring departure delay

a. Which are the top 3 months with the highest departure delay?

```
#Mutating the data add a column dep which is 1 if the departure delay is greater than 0 and 0 if the dep
dep_flights <- data1 %>% group_by(month) %>% mutate( dep = ifelse(dep_delay >0 , 1, 0))

#Plotting the data
ggplot(dep_flights, aes(month, fill = factor(dep)), labels = TRUE) +
   geom_histogram() + scale_x_continuous(breaks = c(1:12))
```



We can see that months 6, 7 and 12 again have the most number of departure delays.

Part b: Why do you think these were the months with most delay?

June, July and December are holiday months, therefore, like the arrival delay even the departure delay is the most in these months.

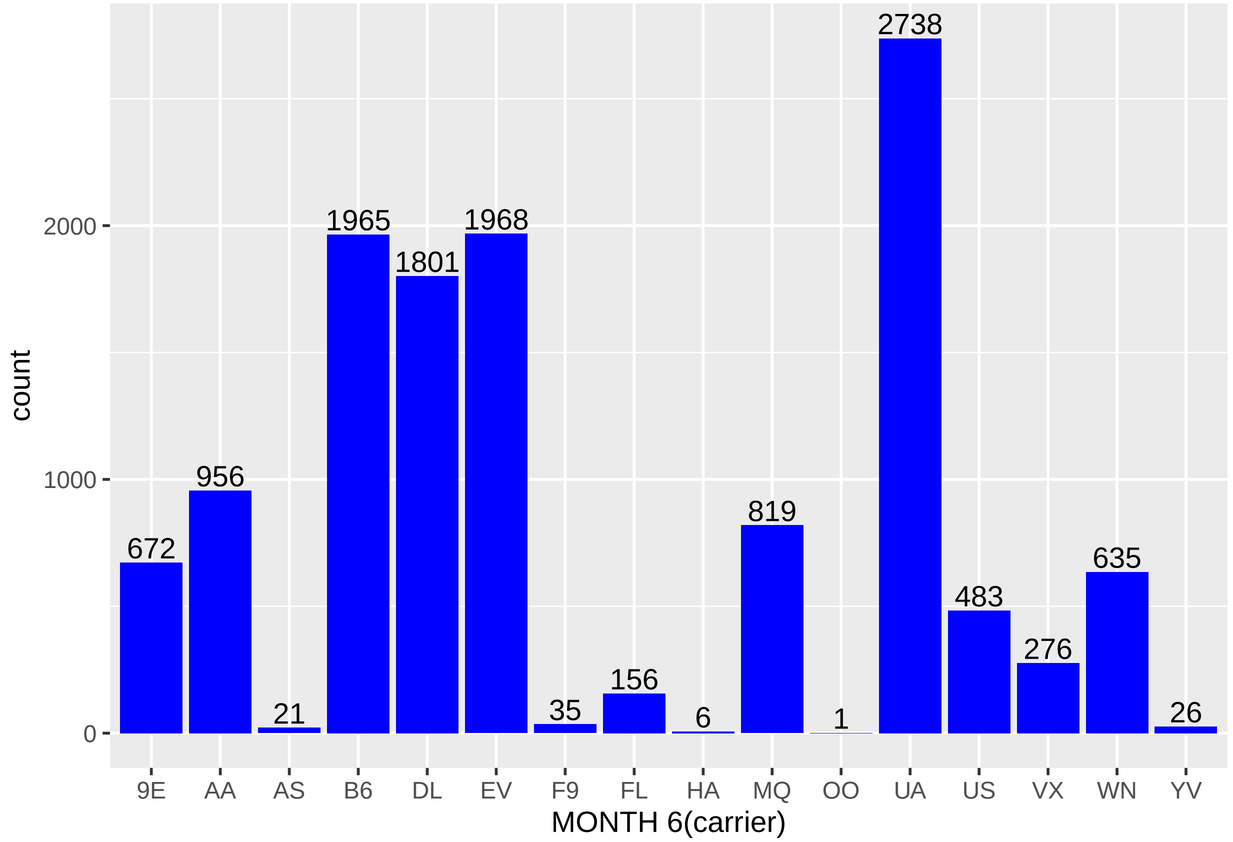
Part c: Which airline carrier has the most number of delays in those months?

```
#Filtering data of month 6 with positive departure delay
df1 = data1 %>% filter(month == 6 & dep_delay >0)
dim(df1)
```

```
## [1] 12558    19
```

```
#Plotting the data
ggplot(df1, aes(carrier, fill = dep_delay)) + geom_histogram(stat="count", fill="blue") + xlab("MONTH 6
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

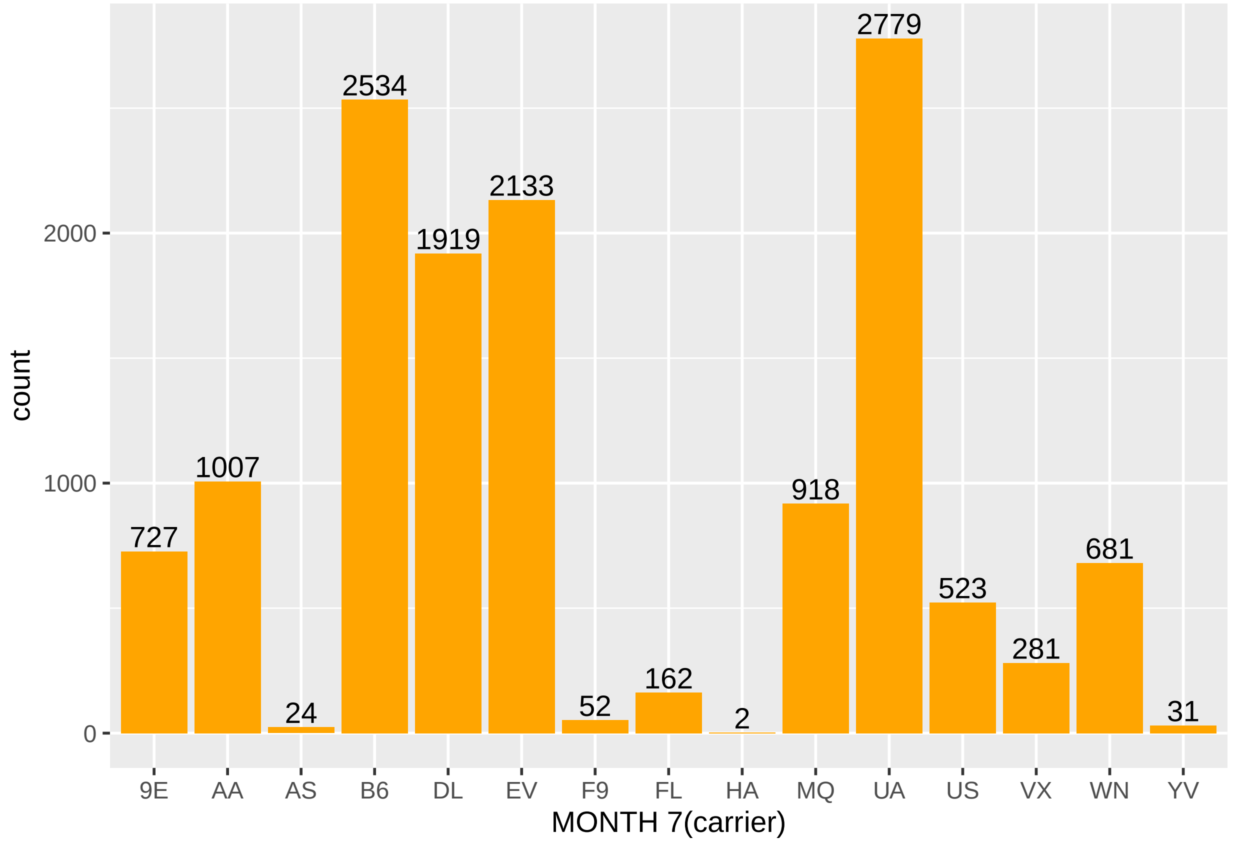


```
#Filtering data of month 7 with positive arrival delay
df2 = data1 %>% filter(month == 7 & dep_delay >0)
dim(df2)
```

```
## [1] 13773    19
```

```
#Plotting the data
ggplot(df2, aes(carrier, fill = dep_delay)) + geom_histogram(stat="count", fill="orange") + xlab("MONTH
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

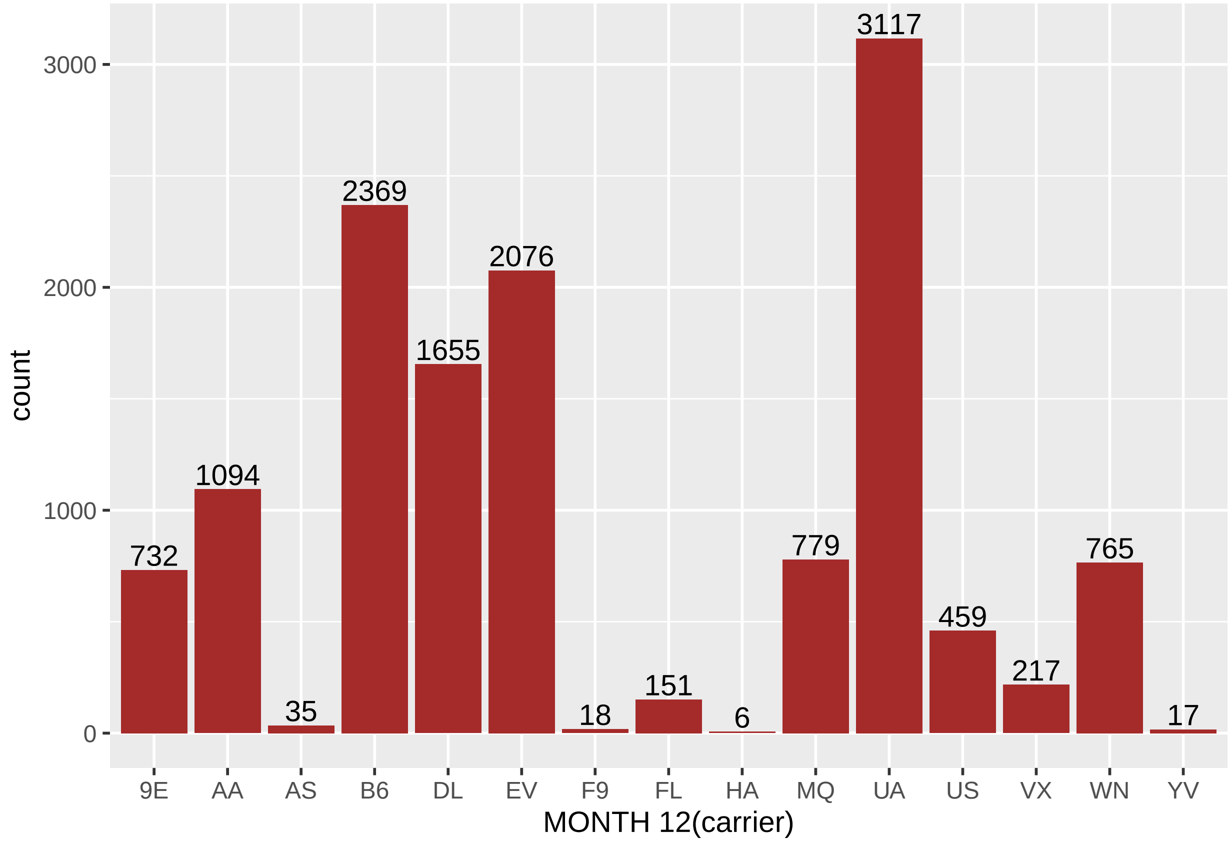


```
#Filtering data of month 12 with positive arrival delay
df3 = data1 %>% filter(month == 12 & dep_delay >0)
dim(df3)
```

```
## [1] 13490    19
```

```
#Plotting the data
ggplot(df3, aes(carrier, fill = dep_delay)) + geom_histogram(stat="count", fill="brown") + xlab("MONTH 
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



UA flights have the highest departure delays.

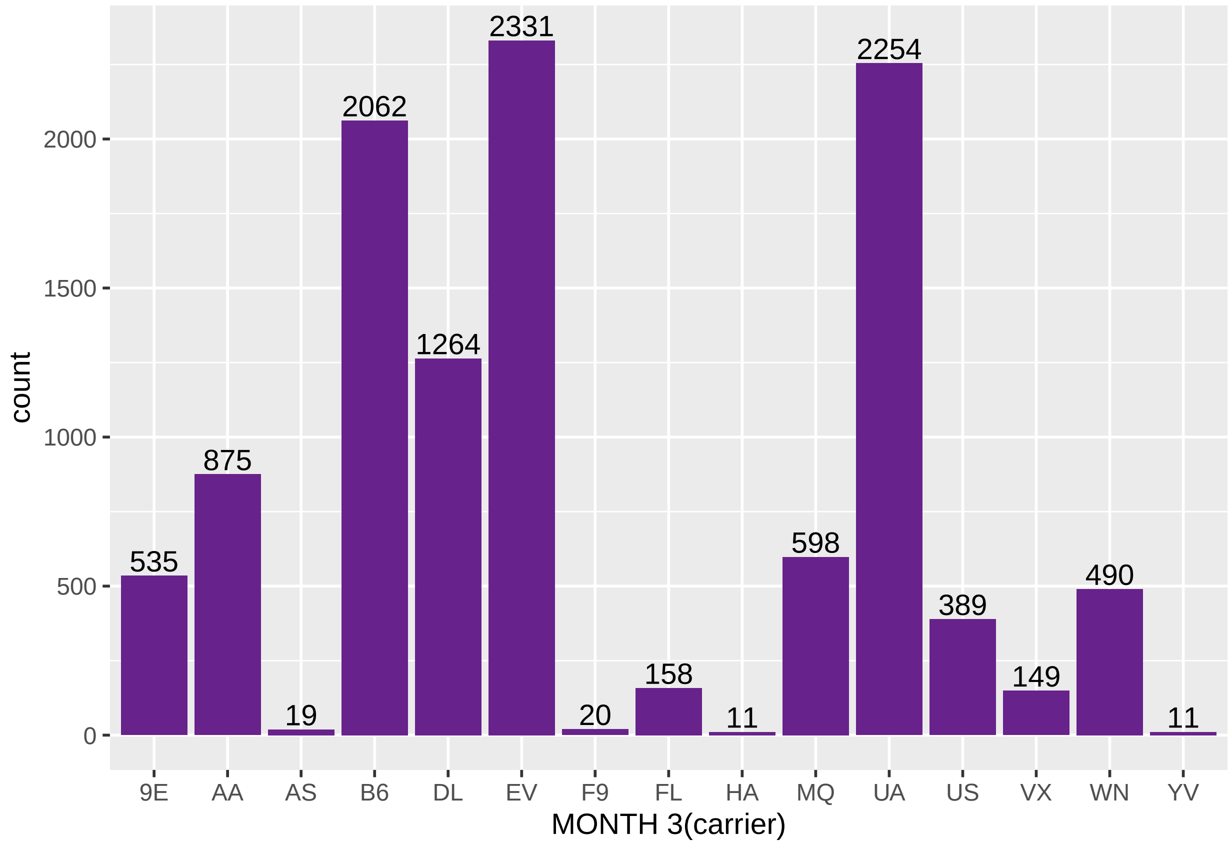
Part d: Choose a month and check if the airline had the most number of departure delays that month

```
#Choosing month 3 to check the departure delays
df4 = data1 %>% filter(month == 3 & dep_delay >0)
dim(df4)
```

```
## [1] 11166    19
```

```
#Plotting the data
ggplot(df4, aes(carrier, fill = dep_delay)) + geom_histogram(stat="count", fill="darkorchid4") + xlab("
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



No, EV had the highest delay in other months as well.

Question 3: Was there any correlation between 1 and 2

Since months 6, 7 and 12 are holiday months, the arrival and departure delays are the highest in these months. Other than that, there is no correlation between the other two parts.

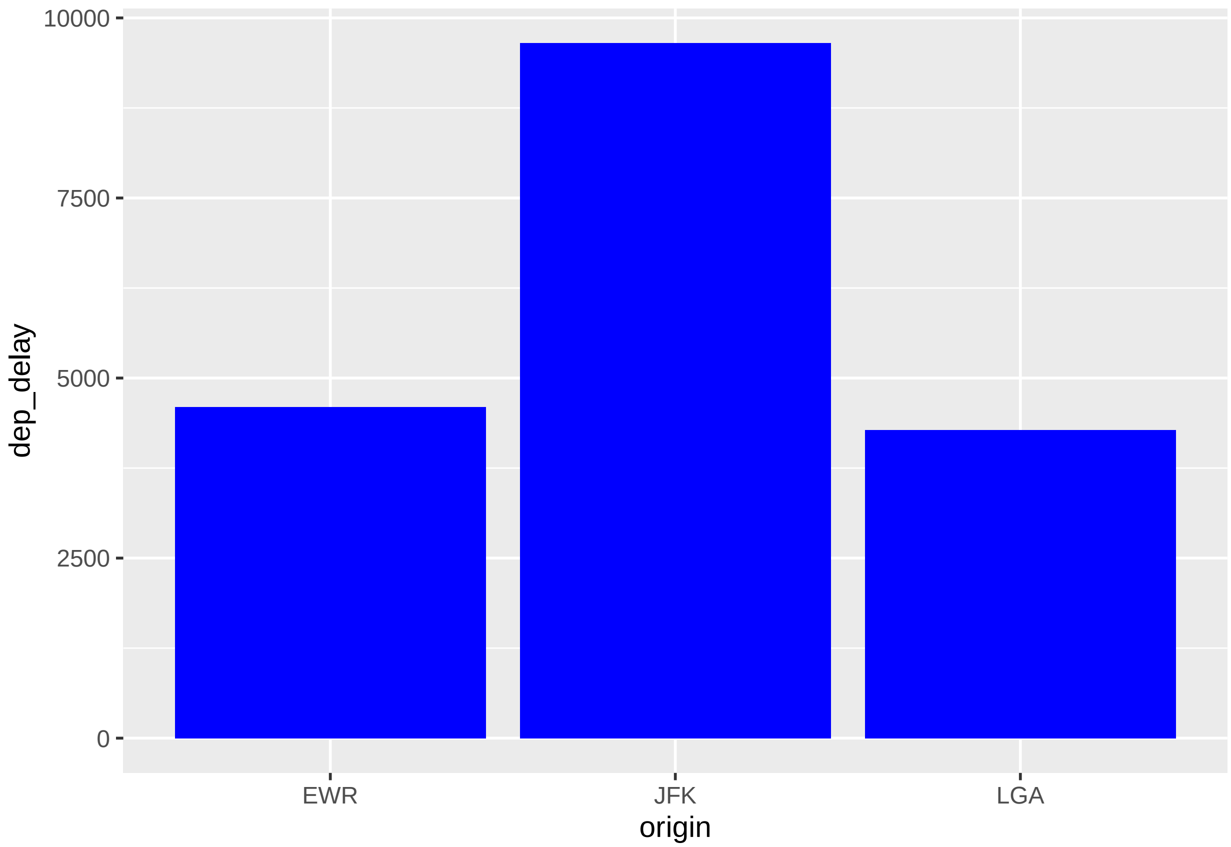
Question 4: Which airports have the highest departure delay?

```r
#Filtering the data for positive departure delay
df6 = data1 %>% filter(dep_delay >0)

#Getting the top 20 values for the plot
df6 = head(arrange(df6, desc(dep_delay)),20)

#Plotting the data
ggplot(df6,aes(x=origin, y=dep_delay)) + geom_histogram(stat="identity", fill="blue")
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



From the above graph we can see that JFK, EWR and LGA have the highest departure delay. All these airport are in NYC.

Question 5: Did more number of flights that departed from 6am to 6pm get delayed?

```
#Filtering out all the flights that had a delay in departure
data1 = data1 %>% filter(dep_delay>0)

#Grouping by month and factor whether it departed between 6am to 6pm.
dep_times <- data1 %>% group_by(month) %>% mutate( dep = ifelse(dep_time>600& dep_time<1800 , 1, 0))

#Plotting the data
p <- ggplot(dep_times, aes(month, fill = factor(dep)), labels = TRUE) +
   geom_histogram() + scale_x_continuous(breaks = c(1:12))

#Using plotly to make it an interactive plot!
ggplotly(p)
```

Therefore, from the above data we can see that most number of flights that departed between 6am and 6pm got delayed.

(d) Challenge Your Results:

The challenges with this data set are :

1. It is not well structured
2. The data is not sufficient to reach thorough conclusions.
3. Few columns in the data set are not useful.
4. There is no data if we want to explore the delay due to stop-overs.