IMT 573: Problem Set 1 - Exploring Data

Miloni Desai

Due: Tuesday, October 08, 2019

Collaborators:

Instructions:

Before beginning this assignment, please ensure you have access to R and RStudio; this can be on your own personal computer or on the IMT 573 R Studio Server.

1. Download the `problemset1.Rmd` file from Canvas or save a copy to your local directory on RStudio Server. Open `problemset1.Rmd` in RStudio and supply your solutions to the assignment by editing `problemset1.Rmd`.
2. Replace the "Insert Your Name Here" text in the `author:` field with your own full name. Any collaborators must be listed on the top of your assignment.
3. Be sure to include well-documented (e.g. commented) code chucks, figures, and clearly written text chunk explanations as necessary. Any figures should be clearly labeled and appropriately referenced within the text. Be sure that each visualization adds value to your written explanation; avoid redundancy – you do no need four different visualizations of the same pattern.
4. Collaboration on problem sets is fun and useful, and we encourage it, but each student must turn in an individual write-up in their own words as well as code/work that is their own. Regardless of whether you work with others, what you turn in must be your own work; this includes code and interpretation of results. The names of all collaborators must be listed on each assignment. Do not copy-and-paste from other students' responses or code.
5. All materials and resources that you use (with the exception of lecture slides) must be appropriately referenced within your assignment.
6. Remember partial credit will be awarded for each question for which a serious attempt at finding an answer has been shown. Students are *strongly* encouraged to attempt each question and to document their reasoning process even if they cannot find the correct answer. If you would like to include R code to show this process, but it does not run withouth errors you can do so with the `eval=FALSE` option as follows:

```
a + b # these object dont' exist
# if you run this on its own it with give an error
```

7. When you have completed the assignment and have **checked** that your code both runs in the Console and knits correctly when you click `Knit PDF`, rename the knitted PDF file to `Yps1_ourLastName_YourFirstName.pdf`, and submit the PDF file on Canvas.

Setup:

In this problem set you will need, at minimum, the following R packages.

```
# Load standard libraries
library(tidyverse)
library(nycflights13)
```

Problem 1: Exploring the NYC Flights Data

In this problem set we will use the data on all flights that departed NYC (i.e. JFK, LGA or EWR) in 2013. You can find this data in the `nycflights13` R package.

```
# Load the nycflights13 library which includes data on all
# flights departing NYC
data(flights)
# Note the data itself is called flights, we will make it into a local df
# for readability
flights <- tbl_df(flights)
# Look at the help file for information about the data
# ?flights
flights
```

```
## # A tibble: 336,776 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
##  7  2013     1     1      555            600        -5      913
##  8  2013     1     1      557            600        -3      709
##  9  2013     1     1      557            600        -3      838
## 10  2013     1     1      558            600        -2      753
## # ... with 336,766 more rows, and 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>
```

```
# summary(flights)
summary(flights)
```

```
##       year          month             day           dep_time
##  Min.   :2013   Min.   : 1.000   Min.   : 1.00   Min.   :   1
##  1st Qu.:2013   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.: 907
##  Median :2013   Median : 7.000   Median :16.00   Median :1401
##  Mean   :2013   Mean   : 6.549   Mean   :15.71   Mean   :1349
##  3rd Qu.:2013   3rd Qu.:10.000   3rd Qu.:23.00   3rd Qu.:1744
##  Max.   :2013   Max.   :12.000   Max.   :31.00   Max.   :2400
##                                                  NA's   :8255
##  sched_dep_time   dep_delay          arr_time    sched_arr_time
##  Min.   : 106   Min.   : -43.00   Min.   :   1   Min.   :   1
##  1st Qu.: 906   1st Qu.:  -5.00   1st Qu.:1104   1st Qu.:1124
##  Median :1359   Median :  -2.00   Median :1535   Median :1556
##  Mean   :1344   Mean   :  12.64   Mean   :1502   Mean   :1536
##  3rd Qu.:1729   3rd Qu.:  11.00   3rd Qu.:1940   3rd Qu.:1945
##  Max.   :2359   Max.   :1301.00   Max.   :2400   Max.   :2359
##                 NA's   :8255      NA's   :8713
##    arr_delay          carrier              flight       tailnum
##  Min.   : -86.000   Length:336776      Min.   :   1   Length:336776
##  1st Qu.: -17.000   Class :character   1st Qu.: 553   Class :character
##  Median :  -5.000   Mode  :character   Median :1496   Mode  :character
```

```
##  Mean   :   6.895                     Mean   :1972
##  3rd Qu.:  14.000                     3rd Qu.:3465
##  Max.   :1272.000                     Max.   :8500
##  NA's   :9430
##     origin              dest              air_time        distance
##  Length:336776      Length:336776      Min.   : 20.0   Min.   :  17
##  Class :character   Class :character   1st Qu.: 82.0   1st Qu.: 502
##  Mode  :character   Mode  :character   Median :129.0   Median : 872
##                                        Mean   :150.7   Mean   :1040
##                                        3rd Qu.:192.0   3rd Qu.:1389
##                                        Max.   :695.0   Max.   :4983
##                                        NA's   :9430
##       hour           minute        time_hour
##  Min.   : 1.00   Min.   : 0.00   Min.   :2013-01-01 05:00:00
##  1st Qu.: 9.00   1st Qu.: 8.00   1st Qu.:2013-04-04 13:00:00
##  Median :13.00   Median :29.00   Median :2013-07-03 10:00:00
##  Mean   :13.18   Mean   :26.23   Mean   :2013-07-03 05:22:54
##  3rd Qu.:17.00   3rd Qu.:44.00   3rd Qu.:2013-10-01 07:00:00
##  Max.   :23.00   Max.   :59.00   Max.   :2013-12-31 23:00:00
##
```

(a) Importing and Inspecting Data

Load the data and describe in a short paragraph how the data was collected and what each variable represents. Perform a basic inspection of the data and discuss what you find.

```
#Load the dataset
flights <- tbl_df(flights)
#view it as a table sepeartely
view(flights)
#summary of the data to understand parameters
summary(flights)
```

```
##       year          month             day           dep_time
##  Min.   :2013   Min.   : 1.000   Min.   : 1.00   Min.   :   1
##  1st Qu.:2013   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.: 907
##  Median :2013   Median : 7.000   Median :16.00   Median :1401
##  Mean   :2013   Mean   : 6.549   Mean   :15.71   Mean   :1349
##  3rd Qu.:2013   3rd Qu.:10.000   3rd Qu.:23.00   3rd Qu.:1744
##  Max.   :2013   Max.   :12.000   Max.   :31.00   Max.   :2400
##                                                  NA's   :8255
##  sched_dep_time   dep_delay          arr_time    sched_arr_time
##  Min.   : 106   Min.   : -43.00   Min.   :   1   Min.   :   1
##  1st Qu.: 906   1st Qu.:  -5.00   1st Qu.:1104   1st Qu.:1124
##  Median :1359   Median :  -2.00   Median :1535   Median :1556
##  Mean   :1344   Mean   :  12.64   Mean   :1502   Mean   :1536
##  3rd Qu.:1729   3rd Qu.:  11.00   3rd Qu.:1940   3rd Qu.:1945
##  Max.   :2359   Max.   :1301.00   Max.   :2400   Max.   :2359
##                 NA's   :8255      NA's   :8713
##    arr_delay          carrier              flight       tailnum
##  Min.   : -86.000   Length:336776      Min.   :   1   Length:336776
##  1st Qu.: -17.000   Class :character   1st Qu.: 553   Class :character
##  Median :  -5.000   Mode  :character   Median :1496   Mode  :character
##  Mean   :   6.895                      Mean   :1972
##  3rd Qu.:  14.000                      3rd Qu.:3465
```

```
##  Max.   :1272.000                     Max.   :8500
##  NA's   :9430
##     origin              dest              air_time        distance
##  Length:336776      Length:336776      Min.   : 20.0   Min.   :  17
##  Class :character   Class :character   1st Qu.: 82.0   1st Qu.: 502
##  Mode  :character   Mode  :character   Median :129.0   Median : 872
##                                        Mean   :150.7   Mean   :1040
##                                        3rd Qu.:192.0   3rd Qu.:1389
##                                        Max.   :695.0   Max.   :4983
##                                        NA's   :9430
##       hour            minute        time_hour
##  Min.   : 1.00   Min.   : 0.00   Min.   :2013-01-01 05:00:00
##  1st Qu.: 9.00   1st Qu.: 8.00   1st Qu.:2013-04-04 13:00:00
##  Median :13.00   Median :29.00   Median :2013-07-03 10:00:00
##  Mean   :13.18   Mean   :26.23   Mean   :2013-07-03 05:22:54
##  3rd Qu.:17.00   3rd Qu.:44.00   3rd Qu.:2013-10-01 07:00:00
##  Max.   :23.00   Max.   :59.00   Max.   :2013-12-31 23:00:00
##
```

```
head(flights)
```

```
## # 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>
```

```
tail(flights)
```

```
## # 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     9    30       NA           1842        NA       NA
## 2  2013     9    30       NA           1455        NA       NA
## 3  2013     9    30       NA           2200        NA       NA
## 4  2013     9    30       NA           1210        NA       NA
## 5  2013     9    30       NA           1159        NA       NA
## 6  2013     9    30       NA            840        NA       NA
## # ... 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>
```

```
#Gives us the names of the columns and their datatypes
names(flights)
```

```
##  [1] "year"           "month"          "day"            "dep_time"
##  [5] "sched_dep_time" "dep_delay"      "arr_time"       "sched_arr_time"
```

```
##  [9] "arr_delay"      "carrier"        "flight"         "tailnum"
## [13] "origin"         "dest"           "air_time"       "distance"
## [17] "hour"           "minute"         "time_hour"
```

```
str(flights)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    336776 obs. of  19 variables:
##  $ year          : int  2013 2013 2013 2013 2013 2013 2013 2013 2013 2013 ...
##  $ month         : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ day           : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ dep_time      : int  517 533 542 544 554 554 555 557 557 558 ...
##  $ sched_dep_time: int  515 529 540 545 600 558 600 600 600 600 ...
##  $ dep_delay     : num  2 4 2 -1 -6 -4 -5 -3 -3 -2 ...
##  $ arr_time      : int  830 850 923 1004 812 740 913 709 838 753 ...
##  $ sched_arr_time: int  819 830 850 1022 837 728 854 723 846 745 ...
##  $ arr_delay     : num  11 20 33 -18 -25 12 19 -14 -8 8 ...
##  $ carrier       : chr  "UA" "UA" "AA" "B6" ...
##  $ flight        : int  1545 1714 1141 725 461 1696 507 5708 79 301 ...
##  $ tailnum       : chr  "N14228" "N24211" "N619AA" "N804JB" ...
##  $ origin        : chr  "EWR" "LGA" "JFK" "JFK" ...
##  $ dest          : chr  "IAH" "IAH" "MIA" "BQN" ...
##  $ air_time      : num  227 227 160 183 116 150 158 53 140 138 ...
##  $ distance      : num  1400 1416 1089 1576 762 ...
##  $ hour          : num  5 5 5 5 6 5 6 6 6 6 ...
##  $ minute        : num  15 29 40 45 0 58 0 0 0 0 ...
##  $ time_hour     : POSIXct, format: "2013-01-01 05:00:00" "2013-01-01 05:00:00" ...
```

```
#Find out how many unique destinations flights from the airports fly out to
view(distinct(flights['dest']))
#Help function
?flights
```

###The data set contains the flight details of flights from 3 New York airports to a 105 destinations.The data contains 16 descriptors, few of which have been futher broken down to aid analysis. The data has been collected over a period of 12 months and it contains the following attributes collected by the Bureau of Transportation Statistics ###Year,month and date as the date of departure,departure and arrival times in the local time zone,departure and arrival delays in minutes where negatives represent early flights, carrier, tail number, flight number, which of the three New York airports it originated from, destination airport, amount of time spent in air, distance flown in miles and the departure time broken in hours, minutes and seconds. ###From the basic inspection of the data, we can observe what data we have about the flights,we can say that the median departure time is that it tends to departe 2 minutes earlier which usually leads to a median arrival time of 5 minutes earlier than the arrival time. The avg flight time is approximately 2 hours and covers a distance of 800 to 1000 miles in that time.

(b) Formulating Questions

Consider the NYC flights data. Formulate two motivating questions you want to explore using this data. Describe why these questions are interesting and how you might go about answering them. ### The first question that I believe this data can answer is of which airport sees the maximum delays, here I am referring to the departure airport. The same could also be done for a particular arrival airport. ###I think this question is particularly interesting as it is possible that the biggest of the three may not have the most delays. This would also be a tool to identify where and what is causing the delays.When the same is found for a particular airport, we could futher try to see if the flight took of earlier and still did not arrive on time, also different airlines could be identified from the carriers to find if the delay is due to airport traffic or inefficiency of the airline.To go about answering this question,once we load the dataset into a dataframe we will arrange

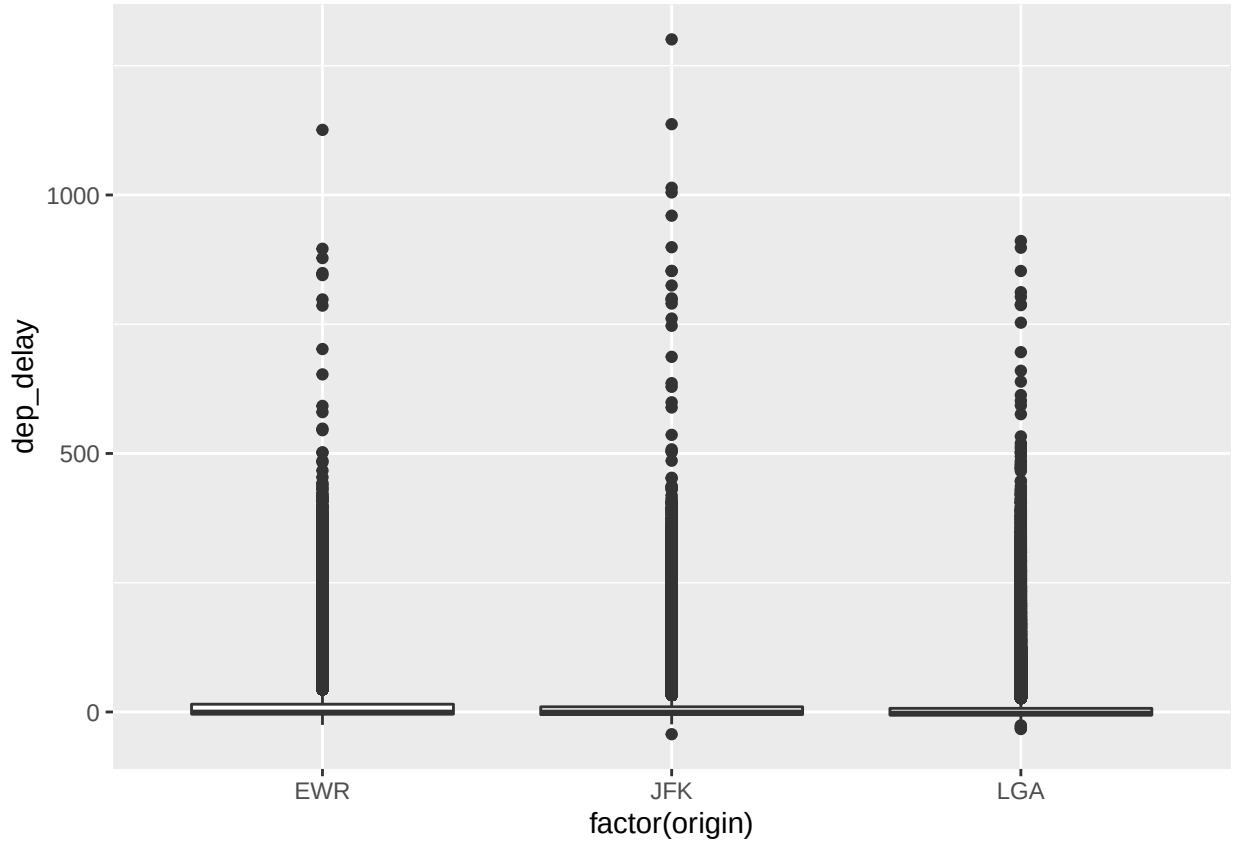
or sort the data by the departure time delays in descending order.Post that we will group by origin and visualise this in a bar graph. ###The second question that I find interesting if which month sees the most delays.Using this information we can find out what seasons see the maximum delay and can predict if its due to weather conditions and airports can prepare for that.I find this particularly interesting because the data will be able to give us peak delays and monthwise and will be interesting to find two months having the highest delay that dont map to the same season.For this analysis, we simply group by month and find the mean of the departure times and visulize the same.

(c) Exploring Data

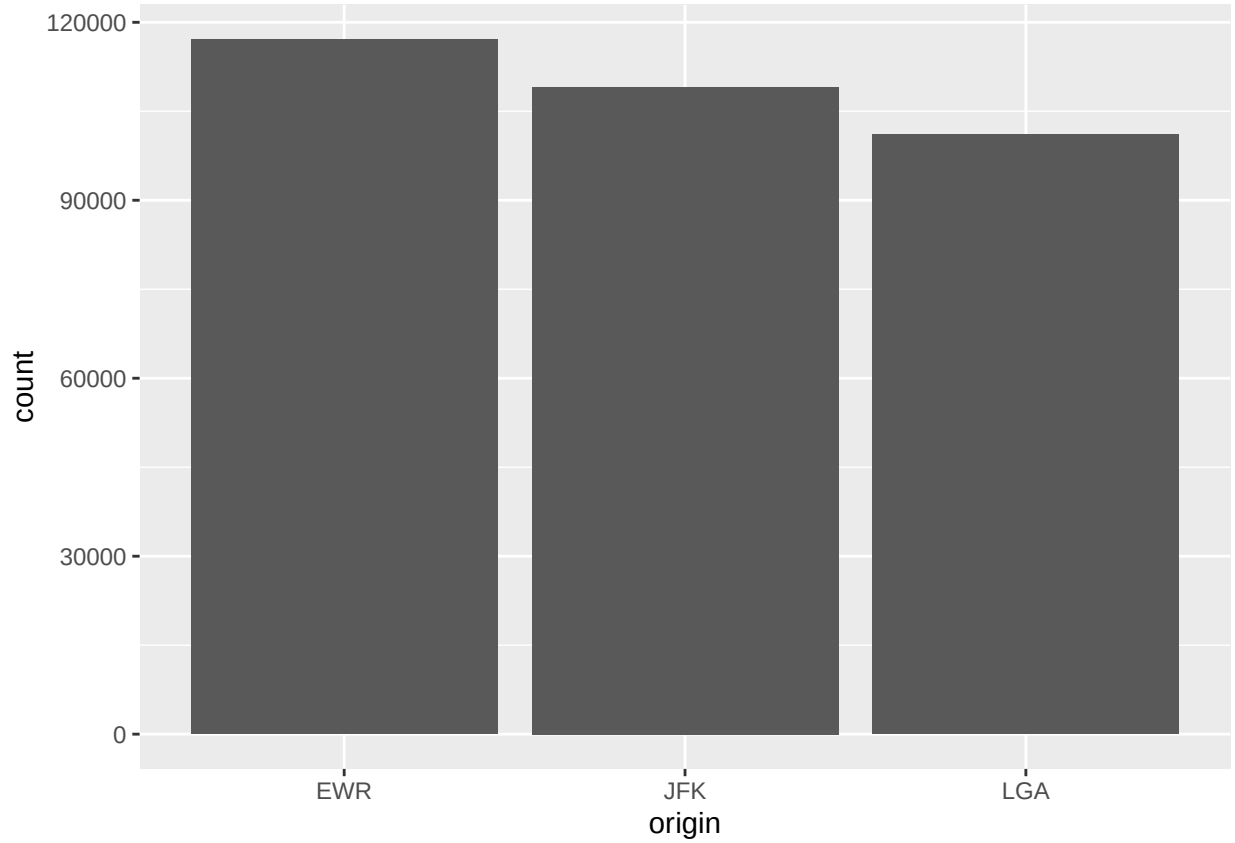
For each of the questions you proposed in Problem 1b, perform an exploratory data analysis designed to address the question. At a minimum, you should produce two visualizations (graphics or tables) related to each question. Be sure to describe what the visuals show and how they speak to your question of interest.

#Question 1

```
#Remove Nan values
flights <- na.omit(flights)
ggplot(flights, aes(x = factor(origin), y = dep_delay)) +
  geom_boxplot()
```




```
#From this box plot we can see that JFK seems to have experienced the most delays

ggplot(data = flights, aes(x = origin, fill = dep_delay)) +
  geom_bar()
```




```
#From this bar chart we can see the EWR seems to have the most delays

flights%>%
  arrange(desc(dep_delay))%>%
  filter(origin == "EWR")%>%
  summary(flights)
```

```
##       year          month             day           dep_time
##  Min.   :2013   Min.   : 1.000   Min.   : 1.00   Min.   :   1
##  1st Qu.:2013   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.: 858
##  Median :2013   Median : 7.000   Median :16.00   Median :1341
##  Mean   :2013   Mean   : 6.511   Mean   :15.73   Mean   :1336
##  3rd Qu.:2013   3rd Qu.: 9.000   3rd Qu.:23.00   3rd Qu.:1732
##  Max.   :2013   Max.   :12.000   Max.   :31.00   Max.   :2400
##  sched_dep_time   dep_delay          arr_time    sched_arr_time
##  Min.   : 500   Min.   : -25.00   Min.   :   1   Min.   :   1
##  1st Qu.: 859   1st Qu.:  -4.00   1st Qu.:1102   1st Qu.:1124
##  Median :1329   Median :  -1.00   Median :1522   Median :1539
##  Mean   :1318   Mean   :  15.01   Mean   :1492   Mean   :1524
##  3rd Qu.:1720   3rd Qu.:  15.00   3rd Qu.:1928   3rd Qu.:1930
##  Max.   :2345   Max.   :1126.00   Max.   :2400   Max.   :2359
##    arr_delay          carrier              flight       tailnum
##  Min.   : -86.000   Length:117127      Min.   :   1   Length:117127
##  1st Qu.: -16.000   Class :character   1st Qu.: 798   Class :character
##  Median :  -4.000   Mode  :character   Median :1624   Mode  :character
##  Mean   :   9.107                      Mean   :2341
```

```
##  3rd Qu.:  16.000                     3rd Qu.:4204
##  Max.   :1109.000                     Max.   :6181
##     origin             dest              air_time        distance
##  Length:117127      Length:117127      Min.   : 20.0   Min.   :  80
##  Class :character   Class :character   1st Qu.: 89.0   1st Qu.: 533
##  Mode  :character   Mode  :character   Median :130.0   Median : 872
##                                        Mean   :153.3   Mean   :1065
##                                        3rd Qu.:195.0   3rd Qu.:1400
##                                        Max.   :695.0   Max.   :4963
##       hour            minute       time_hour
##  Min.   : 5.00   Min.   : 0.0   Min.   :2013-01-01 05:00:00
##  1st Qu.: 8.00   1st Qu.:11.0   1st Qu.:2013-04-03 21:00:00
##  Median :13.00   Median :29.0   Median :2013-07-01 08:00:00
##  Mean   :12.91   Mean   :27.2   Mean   :2013-07-02 01:50:00
##  3rd Qu.:17.00   3rd Qu.:41.0   3rd Qu.:2013-09-30 08:00:00
##  Max.   :23.00   Max.   :59.0   Max.   :2013-12-31 23:00:00
```

```
#Average Delay in departure for EWR is 15.01


flights%>%
  arrange(desc(dep_delay))%>%
  filter(origin == "JFK")%>%
  summary(flights)
```

```
##       year          month             day           dep_time
##  Min.   :2013   Min.   : 1.000   Min.   : 1.00   Min.   :   1
##  1st Qu.:2013   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.: 913
##  Median :2013   Median : 7.000   Median :16.00   Median :1500
##  Mean   :2013   Mean   : 6.514   Mean   :15.76   Mean   :1398
##  3rd Qu.:2013   3rd Qu.: 9.000   3rd Qu.:23.00   3rd Qu.:1825
##  Max.   :2013   Max.   :12.000   Max.   :31.00   Max.   :2400
##  sched_dep_time   dep_delay          arr_time    sched_arr_time
##  Min.   : 540   Min.   : -43.00   Min.   :   1   Min.   :   1
##  1st Qu.: 915   1st Qu.:  -5.00   1st Qu.:1059   1st Qu.:1118
##  Median :1459   Median :  -1.00   Median :1625   Median :1645
##  Mean   :1399   Mean   :  12.02   Mean   :1520   Mean   :1562
##  3rd Qu.:1815   3rd Qu.:  10.00   3rd Qu.:2015   3rd Qu.:2027
##  Max.   :2359   Max.   :1301.00   Max.   :2400   Max.   :2359
##    arr_delay          carrier              flight         tailnum
##  Min.   : -79.000   Length:109079      Min.   :   1.0   Length:109079
##  1st Qu.: -18.000   Class :character   1st Qu.: 225.5   Class :character
##  Median :  -6.000   Mode  :character   Median : 775.0   Mode  :character
##  Mean   :   5.551                      Mean   :1341.8
##  3rd Qu.:  13.000                      3rd Qu.:2041.0
##  Max.   :1272.000                      Max.   :5765.0
##     origin             dest              air_time        distance
##  Length:109079      Length:109079      Min.   : 21.0   Min.   :  94
##  Class :character   Class :character   1st Qu.: 73.0   1st Qu.: 427
##  Mode  :character   Mode  :character   Median :149.0   Median :1069
##                                        Mean   :178.3   Mean   :1275
##                                        3rd Qu.:303.0   3rd Qu.:2248
##                                        Max.   :691.0   Max.   :4983
##       hour            minute        time_hour
##  Min.   : 5.00   Min.   : 0.00   Min.   :2013-01-01 05:00:00
```

```
##  1st Qu.: 9.00   1st Qu.:10.00   1st Qu.:2013-04-03 12:00:00
##  Median :14.00   Median :30.00   Median :2013-07-02 19:00:00
##  Mean   :13.71   Mean   :27.49   Mean   :2013-07-02 05:39:50
##  3rd Qu.:18.00   3rd Qu.:45.00   3rd Qu.:2013-09-29 02:00:00
##  Max.   :23.00   Max.   :59.00   Max.   :2013-12-31 23:00:00
```

```
#Average Delay in departure for JFK is 12.02

flights%>%
  arrange(desc(dep_delay))%>%
  filter(origin == "LGA")%>%
  summary(flights)
```

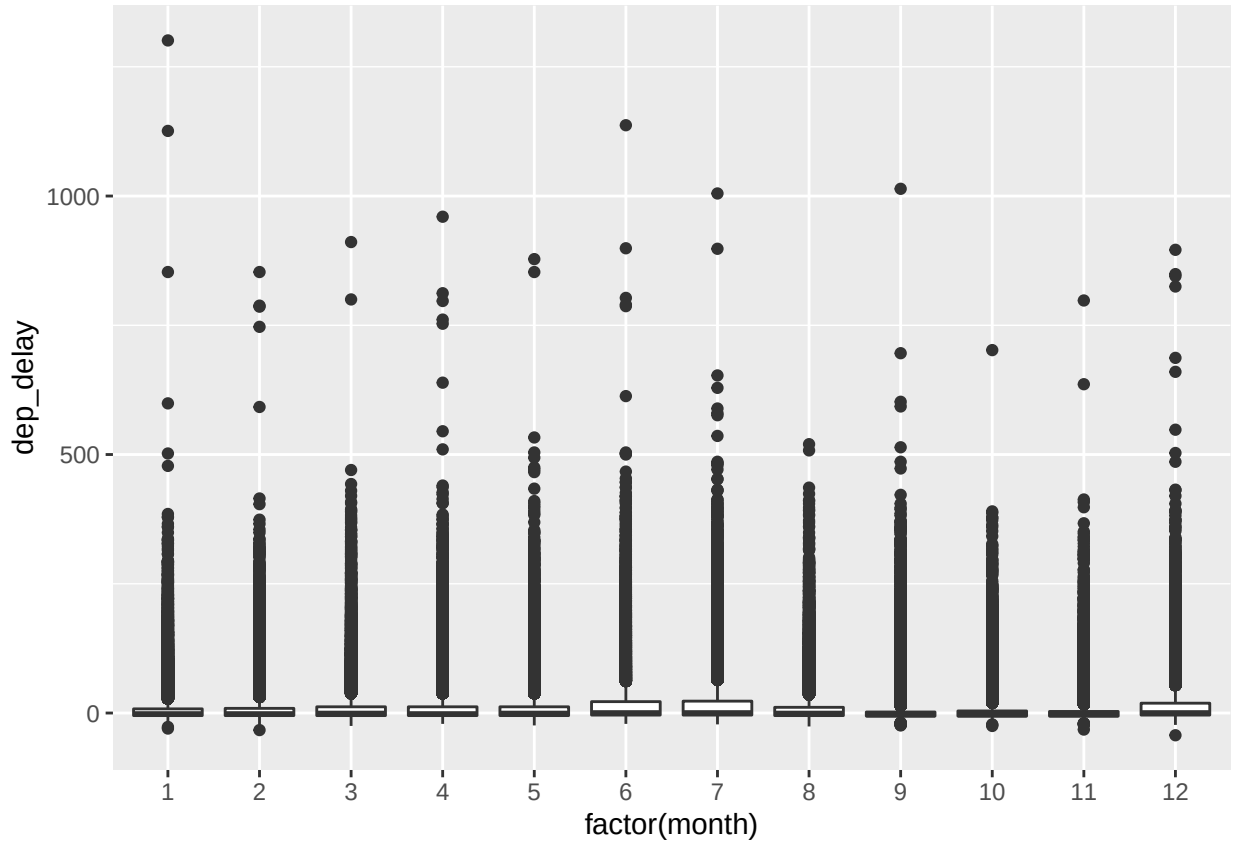
```
##       year          month             day           dep_time
##  Min.   :2013   Min.   : 1.000   Min.   : 1.00   Min.   :   1
##  1st Qu.:2013   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.: 913
##  Median :2013   Median : 7.000   Median :16.00   Median :1315
##  Mean   :2013   Mean   : 6.683   Mean   :15.73   Mean   :1310
##  3rd Qu.:2013   3rd Qu.:10.000   3rd Qu.:23.00   3rd Qu.:1713
##  Max.   :2013   Max.   :12.000   Max.   :31.00   Max.   :2400
##  sched_dep_time   dep_delay         arr_time    sched_arr_time
##  Min.   : 529   Min.   :-33.00   Min.   :   1   Min.   :  14
##  1st Qu.: 910   1st Qu.: -6.00   1st Qu.:1112   1st Qu.:1123
##  Median :1305   Median : -3.00   Median :1509   Median :1515
##  Mean   :1303   Mean   : 10.29   Mean   :1494   Mean   :1512
##  3rd Qu.:1700   3rd Qu.:  7.00   3rd Qu.:1913   3rd Qu.:1911
##  Max.   :2225   Max.   :911.00   Max.   :2400   Max.   :2349
##    arr_delay         carrier              flight       tailnum
##  Min.   :-68.000   Length:101140      Min.   :   1   Length:101140
##  1st Qu.:-17.000   Class :character   1st Qu.: 793   Class :character
##  Median : -5.000   Mode  :character   Median :1875   Mode  :character
##  Mean   :  5.784                      Mean   :2131
##  3rd Qu.: 12.000                      3rd Qu.:3384
##  Max.   :915.000                      Max.   :8500
##     origin              dest              air_time        distance
##  Length:101140      Length:101140      Min.   : 21.0   Min.   :  96.0
##  Class :character   Class :character   1st Qu.: 81.0   1st Qu.: 502.0
##  Mode  :character   Mode  :character   Median :115.0   Median : 762.0
##                                        Mean   :117.8   Mean   : 784.8
##                                        3rd Qu.:148.0   3rd Qu.:1035.0
##                                        Max.   :331.0   Max.   :1620.0
##       hour          minute        time_hour
##  Min.   : 5.0   Min.   : 0.00   Min.   :2013-01-01 05:00:00
##  1st Qu.: 9.0   1st Qu.: 0.00   1st Qu.:2013-04-08 11:00:00
##  Median :13.0   Median :25.00   Median :2013-07-10 07:00:00
##  Mean   :12.8   Mean   :23.76   Mean   :2013-07-07 07:31:44
##  3rd Qu.:17.0   3rd Qu.:40.00   3rd Qu.:2013-10-06 12:00:00
##  Max.   :22.0   Max.   :59.00   Max.   :2013-12-31 21:00:00
```

```
#Average Delay in departure for LGA is 10.29
#Thus looking at these statistics we can see that EWR does have the higher departure delay times.
#The first plot contains outliers and is not the best visualization to
#distinctly give us which among the three has the most departure delays.

#Question 2
```

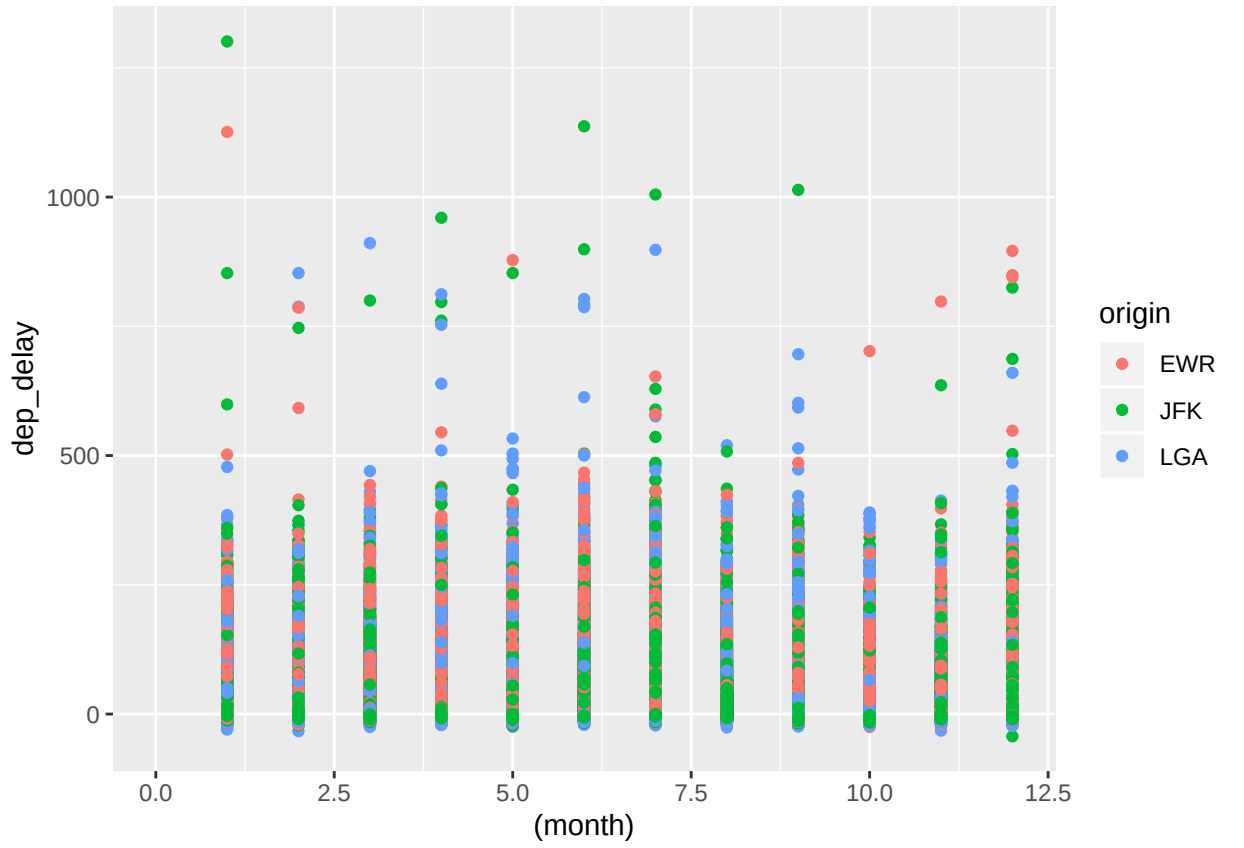
```r
flights %>%
  group_by(month) %>%
  summarise(mean_dd = mean(dep_delay)) %>%
  arrange(desc(mean_dd))
```

```
## # A tibble: 12 x 2
##    month mean_dd
##    <int>   <dbl>
##  1     7   21.5
##  2     6   20.7
##  3    12   16.5
##  4     4   13.8
##  5     3   13.2
##  6     5   12.9
##  7     8   12.6
##  8     2   10.8
##  9     1    9.99
## 10     9    6.63
## 11    10    6.23
## 12    11    5.42
```

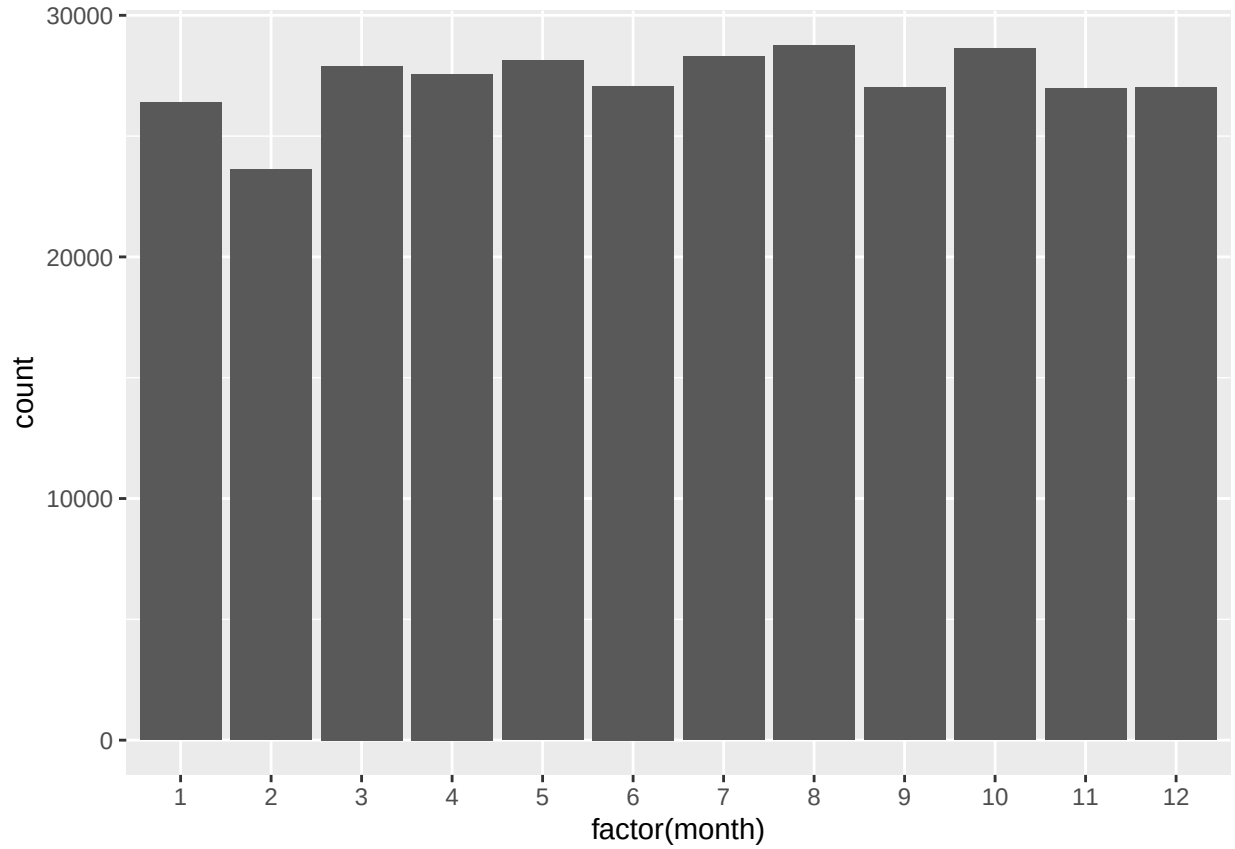
```r
#As we can see July seems to have the most delays in departures, followed by June and December

ggplot(flights, aes(x = factor(month), y = dep_delay)) +
  geom_boxplot()
```



```
ggplot(flights, aes(x = (month), y = dep_delay, color = origin)) +
        xlim(0, 12) +
        geom_point()
```




```
ggplot(data = flights, aes(x = factor(month), fill = dep_delay)) +
  geom_bar()
```




```
# Thus through these visualizations we can see the months that have the max delays ie July, June and De
flights%>%
  filter(month == 7)%>%
  summary(flights)
```

```
##       year          month        day           dep_time    sched_dep_time
##  Min.   :2013   Min.   :7   Min.   : 1.00   Min.   :   1   Min.   : 500
##  1st Qu.:2013   1st Qu.:7   1st Qu.: 9.00   1st Qu.: 904   1st Qu.: 901
##  Median :2013   Median :7   Median :16.00   Median :1357   Median :1354
##  Mean   :2013   Mean   :7   Mean   :16.23   Mean   :1352   Mean   :1339
##  3rd Qu.:2013   3rd Qu.:7   3rd Qu.:24.00   3rd Qu.:1753   3rd Qu.:1726
##  Max.   :2013   Max.   :7   Max.   :31.00   Max.   :2400   Max.   :2359
##    dep_delay          arr_time    sched_arr_time   arr_delay
##  Min.   : -22.00   Min.   :   1   Min.   :   1   Min.   :-66.00
##  1st Qu.:  -4.00   1st Qu.:1034   1st Qu.:1108   1st Qu.:-16.00
##  Median :   0.00   Median :1502   Median :1540   Median : -2.00
##  Mean   :  21.52   Mean   :1455   Mean   :1512   Mean   : 16.71
##  3rd Qu.:  23.00   3rd Qu.:1927   3rd Qu.:1932   3rd Qu.: 27.00
##  Max.   :1005.00   Max.   :2400   Max.   :2359   Max.   :989.00
##    carrier              flight       tailnum             origin
##  Length:28293       Min.   :   1   Length:28293       Length:28293
##  Class :character   1st Qu.: 580   Class :character   Class :character
##  Mode  :character   Median :1473   Mode  :character   Mode  :character
##                     Mean   :1921
##                     3rd Qu.:3353
##                     Max.   :6177
```

```
##      dest              air_time        distance         hour
##  Length:28293       Min.   : 23.0   Min.   :  94   Min.   : 5.00
##  Class :character   1st Qu.: 79.0   1st Qu.: 529   1st Qu.: 9.00
##  Mode  :character   Median :124.0   Median : 937   Median :13.00
##                     Mean   :146.7   Mean   :1070   Mean   :13.12
##                     3rd Qu.:190.0   3rd Qu.:1411   3rd Qu.:17.00
##                     Max.   :629.0   Max.   :4983   Max.   :23.00
##      minute        time_hour
##  Min.   : 0.00   Min.   :2013-07-01 05:00:00
##  1st Qu.:10.00   1st Qu.:2013-07-09 07:00:00
##  Median :29.00   Median :2013-07-16 18:00:00
##  Mean   :27.08   Mean   :2013-07-16 18:38:12
##  3rd Qu.:45.00   3rd Qu.:2013-07-24 14:00:00
##  Max.   :59.00   Max.   :2013-07-31 23:00:00
```

```
  #Average departure delays in July are 21.52
flights%>%
  filter(month == 6)%>%
  summary(flights)
```

```
##       year          month        day           dep_time    sched_dep_time
##  Min.   :2013   Min.   :6   Min.   : 1.00   Min.   :   1   Min.   : 500
##  1st Qu.:2013   1st Qu.:6   1st Qu.: 8.00   1st Qu.: 902   1st Qu.: 900
##  Median :2013   Median :6   Median :16.00   Median :1357   Median :1355
##  Mean   :2013   Mean   :6   Mean   :15.47   Mean   :1350   Mean   :1336
##  3rd Qu.:2013   3rd Qu.:6   3rd Qu.:23.00   3rd Qu.:1753   3rd Qu.:1727
##  Max.   :2013   Max.   :6   Max.   :30.00   Max.   :2400   Max.   :2359
##    dep_delay          arr_time    sched_arr_time   arr_delay
##  Min.   : -21.00   Min.   :   1   Min.   :   1   Min.   : -64.00
##  1st Qu.:  -4.00   1st Qu.:1041   1st Qu.:1113   1st Qu.: -15.00
##  Median :   0.00   Median :1507   Median :1541   Median :  -2.00
##  Mean   :  20.73   Mean   :1468   Mean   :1518   Mean   :  16.48
##  3rd Qu.:  22.00   3rd Qu.:1936   3rd Qu.:1937   3rd Qu.:  26.00
##  Max.   :1137.00   Max.   :2400   Max.   :2359   Max.   :1127.00
##    carrier              flight       tailnum             origin
##  Length:27075       Min.   :   1   Length:27075       Length:27075
##  Class :character   1st Qu.: 541   Class :character   Class :character
##  Mode  :character   Median :1403   Mode  :character   Mode  :character
##                     Mean   :1883
##                     3rd Qu.:3365
##                     Max.   :6177
##      dest              air_time        distance         hour
##  Length:27075       Min.   : 21.0   Min.   :  94   Min.   : 5.00
##  Class :character   1st Qu.: 81.0   1st Qu.: 529   1st Qu.: 9.00
##  Mode  :character   Median :127.0   Median : 937   Median :13.00
##                     Mean   :150.3   Mean   :1074   Mean   :13.09
##                     3rd Qu.:192.0   3rd Qu.:1416   3rd Qu.:17.00
##                     Max.   :650.0   Max.   :4983   Max.   :23.00
##      minute        time_hour
##  Min.   : 0.00   Min.   :2013-06-01 05:00:00
##  1st Qu.:10.00   1st Qu.:2013-06-08 15:00:00
##  Median :29.00   Median :2013-06-16 06:00:00
##  Mean   :27.18   Mean   :2013-06-16 00:22:01
##  3rd Qu.:45.00   3rd Qu.:2013-06-23 10:00:00
##  Max.   :59.00   Max.   :2013-06-30 23:00:00
```

```r
#Average departure delays in July are 20.73


flights%>%
  filter(month == 12)%>%
  summary(flights)
```

```
##       year          month          day           dep_time    sched_dep_time
##  Min.   :2013   Min.   :12   Min.   : 1.00   Min.   :   1   Min.   : 500
##  1st Qu.:2013   1st Qu.:12   1st Qu.: 8.00   1st Qu.: 913   1st Qu.: 905
##  Median :2013   Median :12   Median :16.00   Median :1406   Median :1355
##  Mean   :2013   Mean   :12   Mean   :15.87   Mean   :1357   Mean   :1342
##  3rd Qu.:2013   3rd Qu.:12   3rd Qu.:23.00   3rd Qu.:1750   3rd Qu.:1729
##  Max.   :2013   Max.   :12   Max.   :31.00   Max.   :2400   Max.   :2359
##    dep_delay         arr_time    sched_arr_time   arr_delay
##  Min.   :-43.00   Min.   :   1   Min.   :   3   Min.   :-68.00
##  1st Qu.: -4.00   1st Qu.:1114   1st Qu.:1130   1st Qu.:-11.00
##  Median :  0.00   Median :1544   Median :1605   Median :  2.00
##  Mean   : 16.48   Mean   :1505   Mean   :1541   Mean   : 14.87
##  3rd Qu.: 19.00   3rd Qu.:1948   3rd Qu.:1951   3rd Qu.: 25.00
##  Max.   :896.00   Max.   :2400   Max.   :2359   Max.   :878.00
##    carrier              flight       tailnum             origin
##  Length:27020       Min.   :   1   Length:27020       Length:27020
##  Class :character   1st Qu.: 507   Class :character   Class :character
##  Mode  :character   Median :1456   Mode  :character   Mode  :character
##                     Mean   :1886
##                     3rd Qu.:2932
##                     Max.   :6181
##      dest              air_time        distance         hour
##  Length:27020       Min.   : 21.0   Min.   :  94   Min.   : 5.00
##  Class :character   1st Qu.: 93.0   1st Qu.: 541   1st Qu.: 9.00
##  Mode  :character   Median :142.0   Median : 944   Median :13.00
##                     Mean   :162.6   Mean   :1076   Mean   :13.16
##                     3rd Qu.:202.0   3rd Qu.:1400   3rd Qu.:17.00
##                     Max.   :661.0   Max.   :4983   Max.   :23.00
##      minute        time_hour
##  Min.   : 0.00   Min.   :2013-12-01 05:00:00
##  1st Qu.: 9.00   1st Qu.:2013-12-08 15:00:00
##  Median :29.00   Median :2013-12-16 15:00:00
##  Mean   :26.27   Mean   :2013-12-16 10:07:49
##  3rd Qu.:45.00   3rd Qu.:2013-12-23 23:00:00
##  Max.   :59.00   Max.   :2013-12-31 23:00:00
```

```r
#Average departure delays in July are 16.48
#Thus using these visualizations and tables we can get better data on the departure delays and factors
```

(d) Challenge Your Results

After completing the exploratory analyses from Problem 1c, do you have any concerns about your findings? How well defined was your original question? Do you still believe this question can be answered using this dataset? Comment on any ethical and/or privacy concerns you have with your analysis.

#After completing teh EDA in 1c I do have the following concerns, firstly the visualizations for the airports with the maximum delays are conflicting. I also do think that the mean is not as reliable a measure as the

median,but in this case, due to outliers skewing the data we are sticking to the mean. Even though we know which among the 3 airports faces the maximum delays, the delay in departure may have nothing to do with the airport and maybe dependent on other factors like the airline carrier etc that we have not accounted for. #In the second question, even though all the visualizations are consistent, it is still not completely clear if the delays are dependent on the weather or season conditions in those months, the air traffic or other unaccounted for factors. #The original question when I look back was not as well defined. While it did cover what needs to be found, it did not set specific conditions limiting the vagueness that we are facing now post analysis. When finding the airport that has the maximum delays it is also important to account for other factors and see if they are correlated. # As we can see here, that the arrival delays and departure delays are correlated, these factors need to be accounted for in the question. #The ethical and privacy concerns surrounding the analysis include the fact that we are monitoring such sensitive data with such deep specifications and hence maybe violating the privacy of passengers as well the airlines. Such data is critical and can be misused if it somehow reaches the hands of someone not responsible. Ethically analysis about finding which carrier had the most delays just from such a dataset, without accounting for real conditions and other factors, could tranish the reputation of that carrier or airport as well as create prejudice in the mind of passengers. One year of data is not enough to make such conclusions and these trends have to observed over a larger window of time.

```
plot(flights$dep_delay, flights$arr_delay)
```