```
#install.packages("nycflights13")
library(nycflights13)
if (!require('rmarkdown'))
{
  install.packages('rmarkdown');
  library(rmarkdown);
}
```

```
## Loading required package: rmarkdown
```

```
data(flights)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
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 05:02:36
##  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
##
```

```
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 variable
s:
##  $ year          : int  2013 2013 2013 2013 2013 2013 2013 2013 2013 201
3 ...
##  $ month         : int  1 1 1 1 1 1 1 1 1 ...
##  $ day           : int  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" ...
```

1 a) How many flights were there to and from NYC in 2013?

```
nrow(flights)
```

```
## [1] 336776
```

```
df.toNYC<-flights %>%
  select(flight,dest)%>%
  filter(dest=="EWR" | dest=="JFK" | dest=="LGA")
df.toNYC
```

```
## # A tibble: 1 × 2
##   flight  dest
##    <int> <chr>
## 1   1632   LGA
```

Using the above code, we find that all the flights departed from NYC and one flight arrived at LGA airport. So, total number of flights to and from NYC= 336776

b. How many flights were there from NYC airports to Seattle (SEA) in 2013?

```
df<-flights %>%
  select(flight, origin, dest) %>%
  filter(dest=="SEA") %>%
  summarise(count=n())
df
```

```
## # A tibble: 1 × 1
##   count
##   <int>
## 1  3923
```

No. of flights from NYC to Seattle is 3923

c. How many airlines fly from NYC to Seattle?

```
df.unique<-flights %>%
  select(flight, carrier, origin, dest) %>%
  filter(dest=="SEA") %>%
  distinct(carrier)%>%
  summarise(count=n())
df.unique
```

```
## # A tibble: 1 × 1
##   count
##   <int>
## 1     5
```

```
df.unique<-flights %>%
  select(flight, carrier, origin, dest) %>%
  filter(dest=="SEA") %>%
  distinct(carrier)
df.unique
```

```
## # A tibble: 5 × 1
##   carrier
##     <chr>
## 1      AS
## 2      DL
## 3      UA
## 4      B6
## 5      AA
```

There are 5 unique airlines to Seattle. The carriers are AS,DL,UA,B6,AA

d. What is the average arrival delay for flights from NYC to Seattle?

```
df.delay<-flights %>%
  filter(dest=="SEA") %>%
  summarise(mean(arr_delay, na.rm=T))
df.delay
```

```
## # A tibble: 1 × 1
##   `mean(arr_delay, na.rm = T)`
##                          <dbl>
## 1                    -1.099099
```

The average arrival delay for flights from NYC to Seattle is -1.099099

2 a) What is the mean arrival delay time? What is the median arrival delay time?

```
fl.nyc<-data.frame(flights)
arrdelay.avg<-mean(fl.nyc$arr_delay, na.rm=T)
arrdelay.avg
```

```
## [1] 6.895377
```

The mean arrival delay for all flights is 6.895377

```
arrdelay.med<-median(fl.nyc$arr_delay, na.rm=T)
arrdelay.med
```

```
## [1] -5
```

The median arrival delay for all flights is -5

b. What does a negative arrival delay mean?

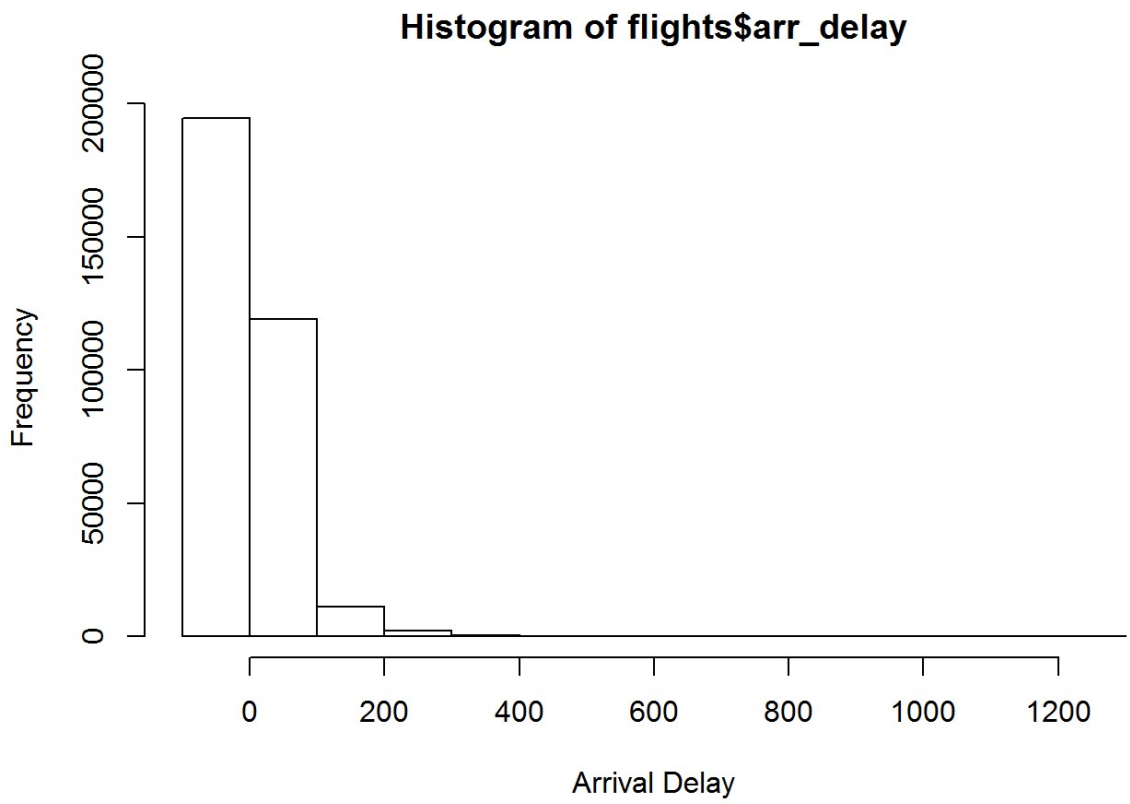
```
neg<-flights %>%
  select(flight, origin, dest, arr_delay) %>%
  filter(arr_delay<0)
neg
```

```
## # A tibble: 188,933 × 4
##    flight origin  dest arr_delay
##     <int>  <chr> <chr>     <dbl>
## 1     725    JFK   BQN       -18
## 2     461    LGA   ATL       -25
## 3    5708    LGA   IAD       -14
## 4      79    JFK   MCO        -8
## 5      49    JFK   PBI        -2
## 6      71    JFK   TPA        -3
## 7    1124    EWR   SFO       -14
## 8    1806    JFK   BOS        -4
## 9    1187    EWR   LAS        -8
## 10    371    LGA   FLL        -7
## # ... with 188,923 more rows
```

A negative arrival delay means that the flight arrived before the scheduled time of arrival. A positive arrival delay implies that the flight arrived after the scheduled time of arrival.

c. Plot a histogram of arrival delay times. Does the answers you obtained in (a) consistent with the shape of the delay time distribution?

```
hist(flights$arr_delay, xlab="Arrival Delay", breaks=10)
abline(v = mean(flights$arr_delay),col = "blue",lwd = 2)
abline(v = median(flights$arr_delay),col = "red",lwd = 2)
```



The shape of the histogram is consistent with the answers we derived from a. The peak of the histogram lies before Zero. This signifies a high concentration of data before Zero, which implies that the median is negative and the mean arrival delay is 6.89

d. Is there seasonality in departure delays? Try and describe what patterns you see. Is there a best month to leave New York? A worst? Why might this be

```
by(flights$dep_delay, flights$month, function(x) mean(x, na.rm=T))
```

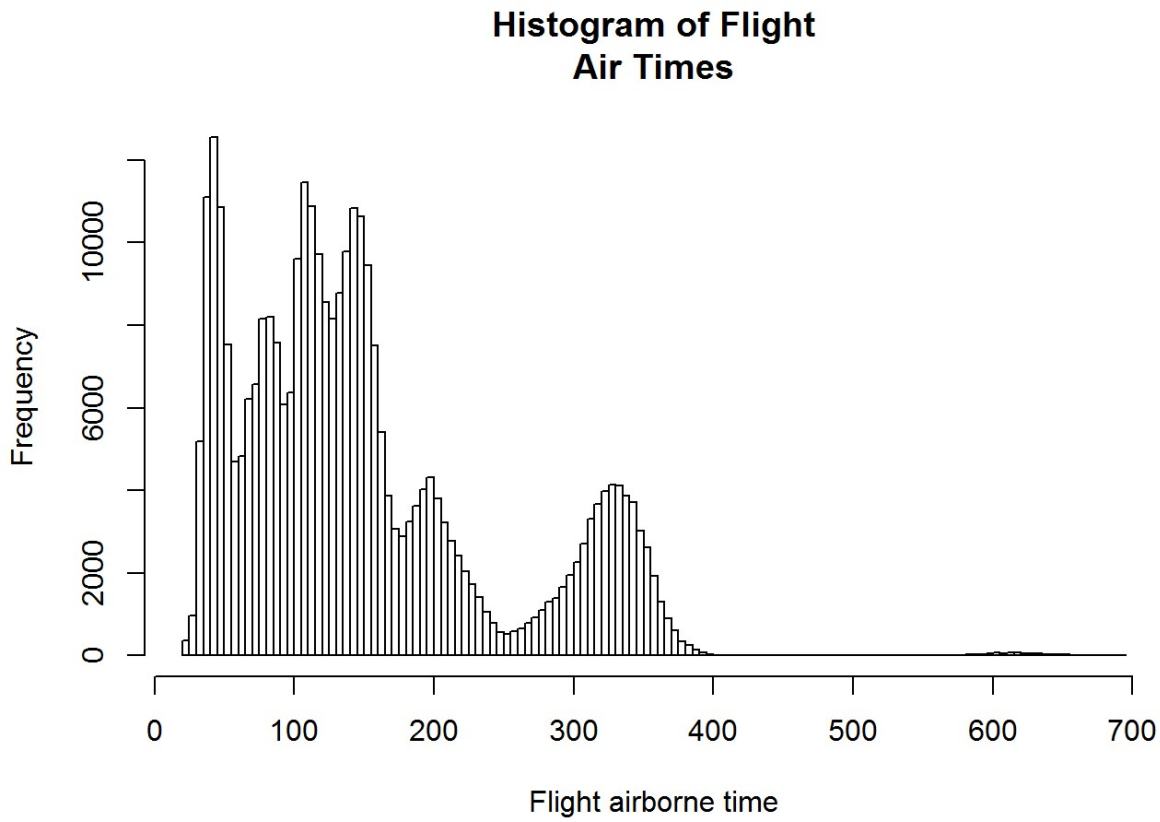
```
## flights$month: 1
## [1] 10.03667
## ------------------------------------------------------------
## flights$month: 2
## [1] 10.81684
## ------------------------------------------------------------
## flights$month: 3
## [1] 13.22708
## ------------------------------------------------------------
## flights$month: 4
## [1] 13.93804
## ------------------------------------------------------------
## flights$month: 5
## [1] 12.98686
## ------------------------------------------------------------
## flights$month: 6
## [1] 20.84633
## ------------------------------------------------------------
## flights$month: 7
## [1] 21.72779
## ------------------------------------------------------------
## flights$month: 8
## [1] 12.61104
## ------------------------------------------------------------
## flights$month: 9
## [1] 6.722476
## ------------------------------------------------------------
## flights$month: 10
## [1] 6.243988
## ------------------------------------------------------------
## flights$month: 11
## [1] 5.435362
## ------------------------------------------------------------
## flights$month: 12
## [1] 16.57669
```

It appears that the mean departure delay time varies across the months. In Spring(March-May), the mean departure delay is around 13. In Summer(June-Aug), the average departure delay is around 17. In Fall(Sep-Nov), the average departure delay is around 5.5. In Winter(Dec-Feb), the average departure delay is 12. To conclude, the departure delay increases from Winter and drops at Fall. The best month to leave NYC is November and the worst month to leave NYC is July. This might be

because during Fall, the skies are clear and it's easier for an airline to take off. During July, there might be a lot of passenger traffic, who board flights to Europe for summer vacations. NYC is the central hub for these airlines and that may account for the departure delay.

3a) Plot a histogram of the total air flight time with 100 breaks. (look at the help for hist()). How many peaks do you see in this distribution? What is an explanation for this?

```
hist(flights$air_time, xlab="Flight airborne time", main="Histogram of Fligh
t
Air Times", breaks=100)
```



The number of peaks are 6. The flight air time variation can be attributed to two factors. Firstly, there are more short-distance flights than long-distance flights. Short distance flights take shorter air time which causes a high density in the left of the histogram around zero.Secondly, the flight time can also vary due to an atmospheric phenomena called Jet stream. The jet stream is a very high altitude wind which blows from West to East over the Atlantic Ocean. The jet stream might be responsible for the varying flight times. Source: http://curious.astro.cornell.edu/about-us/40-our-solar-system/ (http://curious.astro.cornell.edu/about-us/40-our-solar-system/) the-earth/climate-and-weather/68-why-do-airplanes-take-longer-to-fly-west-than-east-intermediate

b. What time of day do flights most commonly depart? Why might there be two most popular times of day to depart?

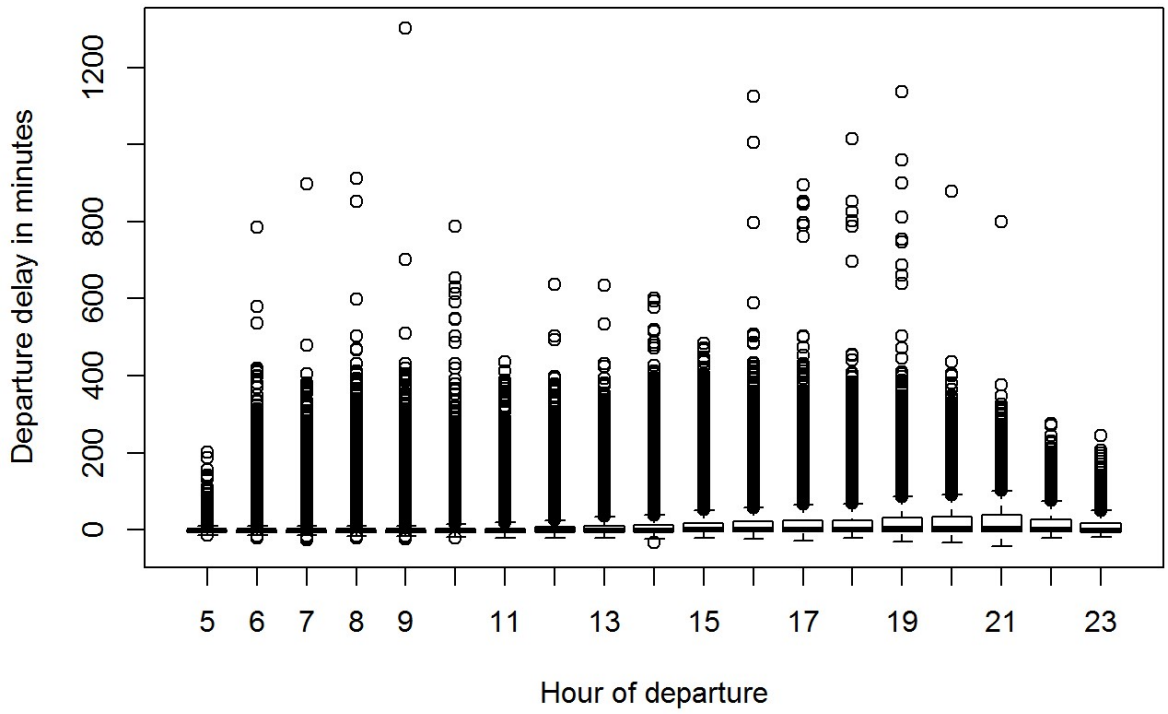
```
fl.time<-flights %>%
select(dep_time, hour, minute) %>%
count(dep_time, sort=TRUE)
head(fl.time)
```

```
## # A tibble: 6 × 2
##   dep_time     n
##      <int> <int>
## 1       NA  8255
## 2      555   834
## 3      755   820
## 4      556   818
## 5      557   799
## 6      655   798
```

The time of day when flights most commonly depart is 5:55AM. There are a total of 834 flights that depart at 5:55AM. The 2nd most popular time of day is 7:55AM where 820 flights depart. The timings of the flights suggest that early morning flights are mostly preferred by passengers.This may be due to the fact that air traffic is low in the morning, because of which flights depart on time and there is very low chances of delay. As the day progresses, the airspace gets more crowded and air traffic controllers delay flight departures.

c. Plot a box plot of departure delays and hour of departure. What pattern do you see? What is an explanation for this?

```
boxplot(flights$dep_delay ~ flights$hour, xlab="Hour of departure", ylab="De
parture delay in minutes")
```



From the box plot, we find that most of the values are concentrated around zero. The departure hour 21:00 has the largest 3rd quartile value. Also, there are a lot of outliers in the distribution. This indicates that some flights had a higher departure delay which can be attributed to high passenger volume during the summer months.

4. Develop one research question you can address using the nycflights2013 dataset. Provide two visualizations to support your exploration of this question. Discuss what you find.

Research Question: Which airline carriers have the best and worst service in terms of the lowest average flight arrival and departure delays in June 2013?

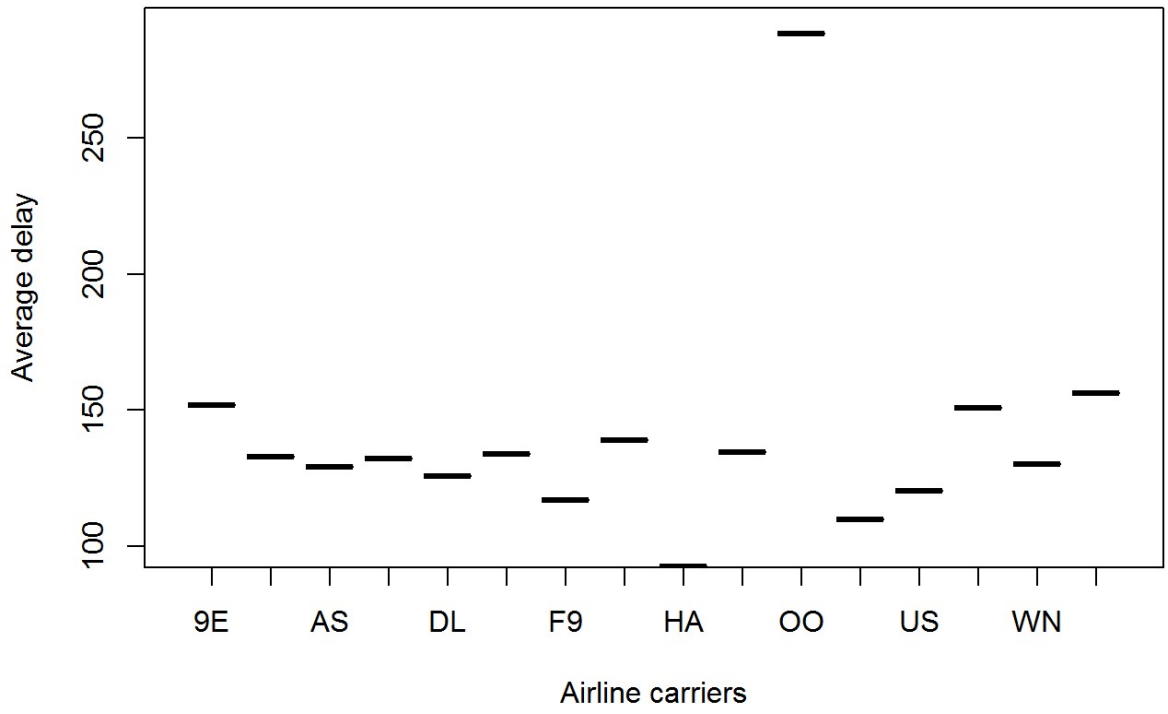
```
fl.carrier<-flights %>%
select(carrier, arr_delay, dep_delay, dest) %>%
filter(arr_delay>0 & dep_delay>0 & flights$month==6) %>% #only positive valu
es have been considered since they signify actual delays.
group_by(carrier) %>%
summarise(avg_arr_delay=mean(arr_delay, na.rm=TRUE)+mean(dep_delay, na.rm=TR
UE))%>%
arrange(desc(avg_arr_delay))

fl.carrier
```

```
## # A tibble: 16 × 2
##    carrier avg_arr_delay
##      <chr>         <dbl>
## 1       OO      288.0000
## 2       YV      156.1304
## 3       9E      151.8828
## 4       VX      150.7396
## 5       FL      138.9510
## 6       MQ      134.4397
## 7       EV      134.0200
## 8       AA      132.7273
## 9       B6      132.0277
## 10      WN      130.3163
## 11      AS      129.2143
## 12      DL      125.8116
## 13      US      120.5193
## 14      F9      116.8929
## 15      UA      109.9387
## 16      HA       92.7500
```

Thus, we find that the carrier OO or SkyWest Airlines has the worst average arrival and departure delays for the month of June 2013. The average delay for Skywest airlines is 288 minutes. The best airline carrier in terms of average arrival delay is Hawaiian airlines. The average arrival and departure delay for Hawaiian airlines is 92.75 minutes for the month of June 2013.

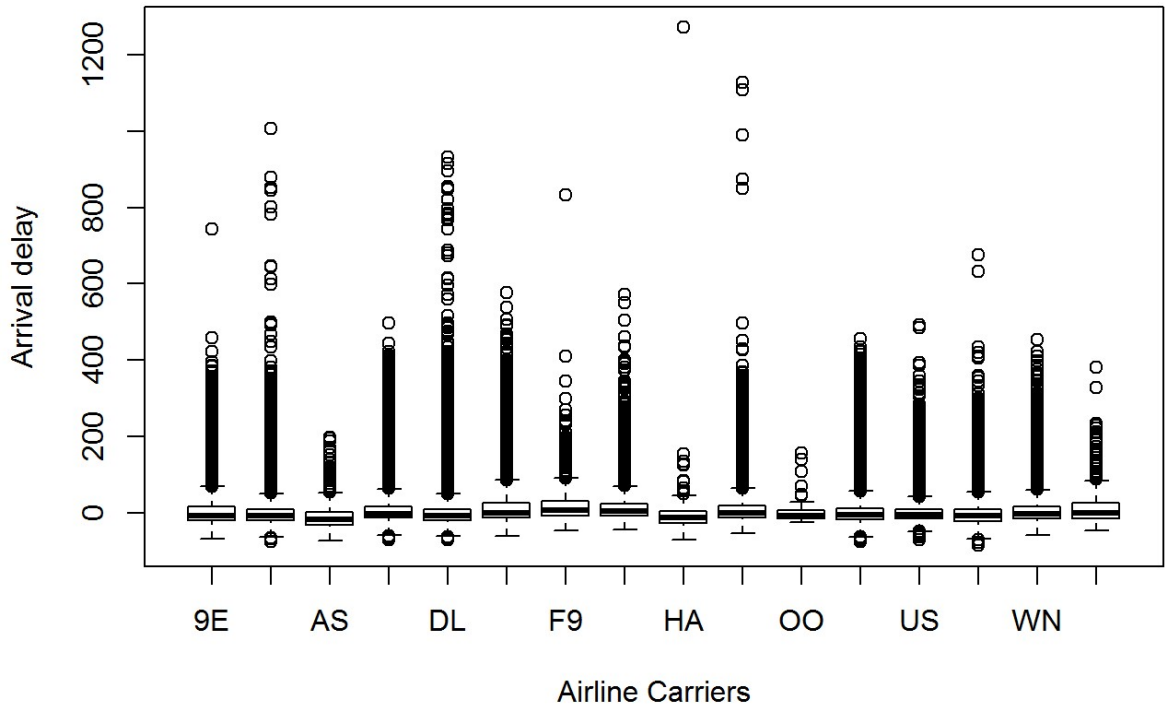
The distribution of average flight arrival and departure delay with respect to airline carriers is as follows:

```
boxplot(fl.carrier$avg_arr_delay ~ fl.carrier$carrier, ylim=c(100,290), xlab
="Airline carriers", ylab="Average delay")
```



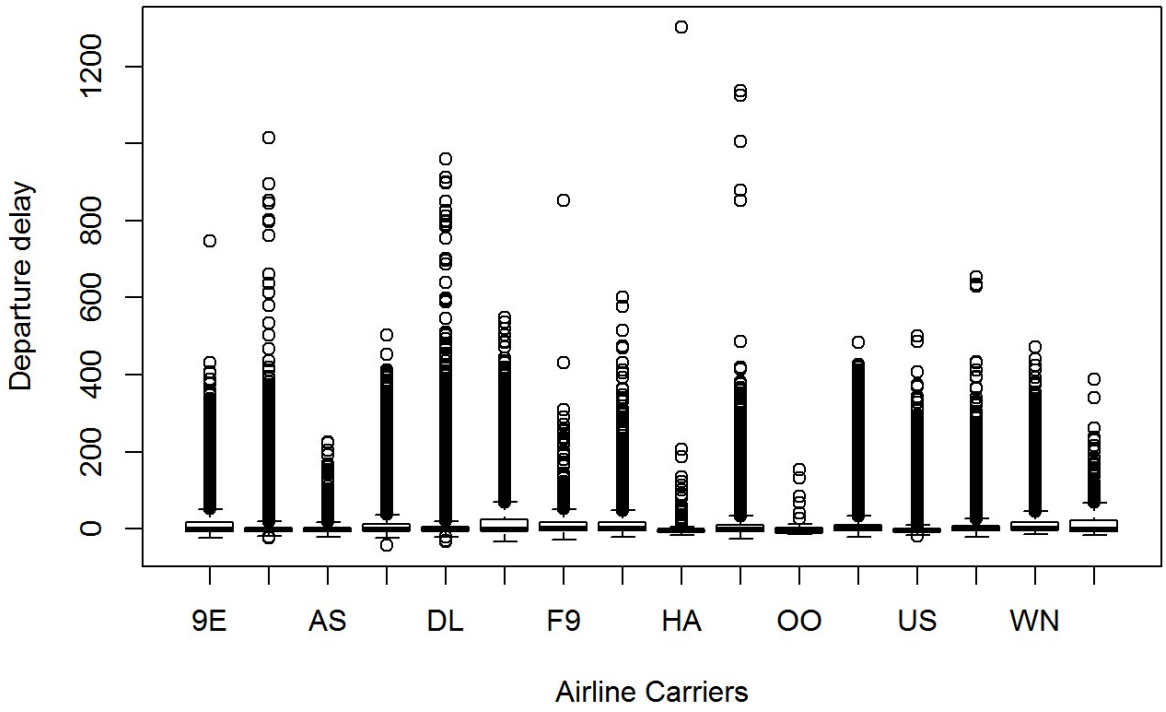
The distribution of arrival delay and airline carriers are as follows:

```
boxplot(flights$arr_delay ~ flights$carrier, xlab="Airline Carriers", ylab
="Arrival delay")
```



The distribution of departure delay and airline carriers are as follows:

```
boxplot(flights$dep_delay ~ flights$carrier, xlab="Airline Carriers", ylab
="Departure delay")
```


Departure delay
0
200
400
600
800
1200
9E
AS
DL
F9
HA
OO
US
WN
Airline Carriers