Assignment_2

Abhinav Garg

1/25/2017

Loading the Data

```
library(nycflights13) # load library
```

```
## Warning: package 'nycflights13' was built under R version 3.3.2
```

```
data(flights) # load data on flights
```

1.

Let’s explore flights from NYC to Seattle. Use the flights dataset to answer the following questions

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

```
f <- flights
nyc_dest <- f[f$dest %in% c('LGA','JFK','EWR'),] # FLights to NYC airport
a<- nrow(f) # all Flights originate from NYC Airports
b<- nrow(nyc_dest)

# Number of flights from and to NYC
a+b
```

```
## [1] 336777
```

b. How many flights were from NYC Airports to Seattle in 2013 ?

```
nyc_sea <- f[(f$dest %in% c('SEA') & f$origin %in% c('LGA','EWR','JFK')),] # Or
igin is one of NYC Airports and dest is SEA
nrow(nyc_sea)
```

```
## [1] 3923
```

c. How many airlines fly from NYC to Seattle?

```
nyc_sea <- f[(f$dest %in% c('SEA') & f$origin %in% c('LGA','EWR','JFK')),] # Origin is one of NYC Airports and dest is SEA
nyc_sea_carrier <- unique(nyc_sea$carrier) # Unique carriers that fly from NYC to SEA
length(nyc_sea_carrier) # Number of unique Carriers
```

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

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

```
nyc_sea <- f[(f$dest %in% c('SEA') & f$origin %in% c('LGA','EWR','JFK')),] # Origin is one of NYC Airports and dest is SEA
avgarr_delay <- mean(nyc_sea$arr_delay,na.rm = T)
avgarr_delay
```

```
## [1] -1.099099
```

2.

Flights are often delayed. Consider the following questions exploring delay patterns.

a. What is the mean arrival delay time? What is the median arrival delay time?

```
mean_arrdelay <- mean(f$arr_delay, na.rm = T)
# Mean Arrival Delay Time
mean_arrdelay
```

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

```
median_arrdelay <- median(f$arr_delay, na.rm = T)
# Median Arrival Delay Time
median_arrdelay
```

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

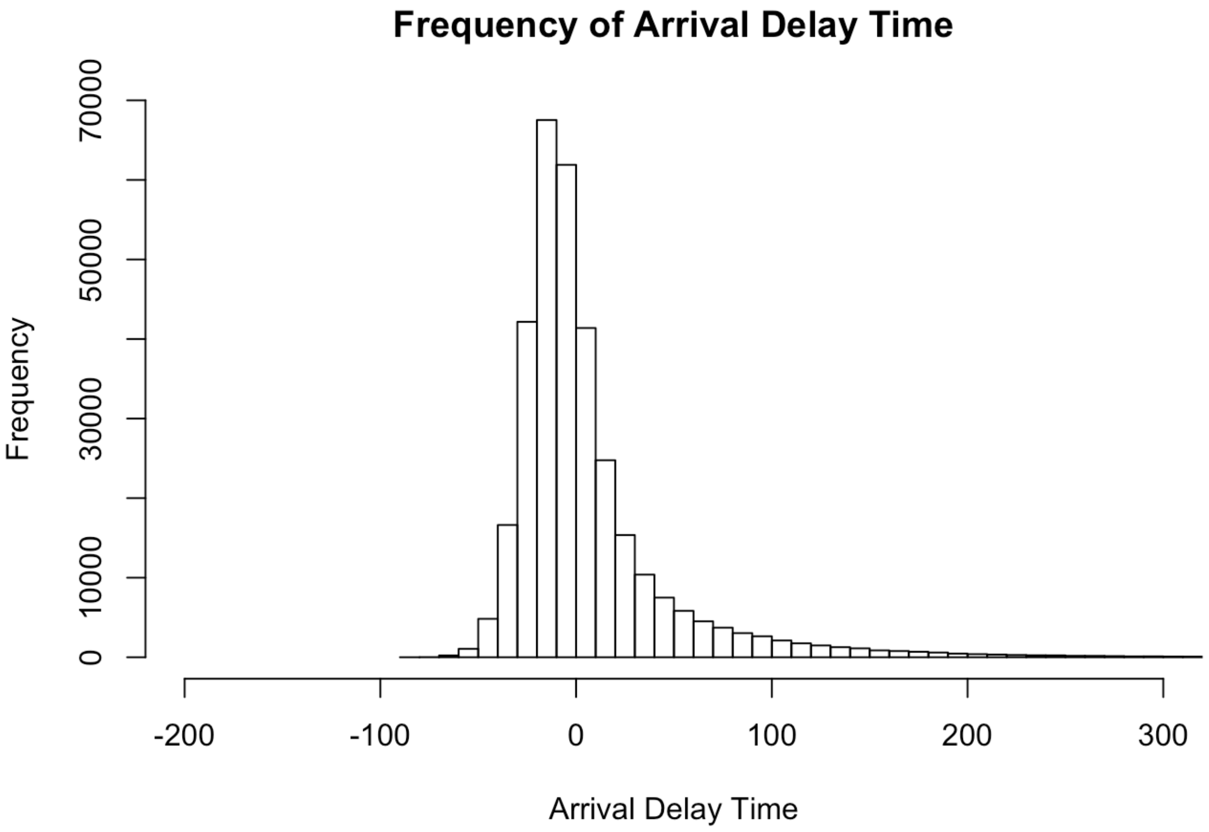
b. What does a negative arrival delay mean ?

```
?flights # Gives description about variables
```

Negative times represents early arrivals

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(f$arr_delay,xlim = c(-200,300),breaks = 100,xlab = 'Arrival Delay Time',ma
in= 'Frequency of Arrival Delay Time')
```



```
# Yes the answers obtained in (a) are consistent with the shape
```

d. Is there seasonality in departure delays?

```
depdelay_month <- by(flights$dep_delay, flights$month, function(x) mean(x, na.r
m=T))
max(depdelay_month)
```

```
## [1] 21.72779
```

```
# Departure delay is highest in the month of July which could be due to Summer
break and also due to 4th of July
min(depdelay_month)
```

```
## [1] 5.435362
```

```
# Departure delay is lowest in the month of November. Hence this could be the b
est month to leave New York
depdelay_month
```

```
## 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
```

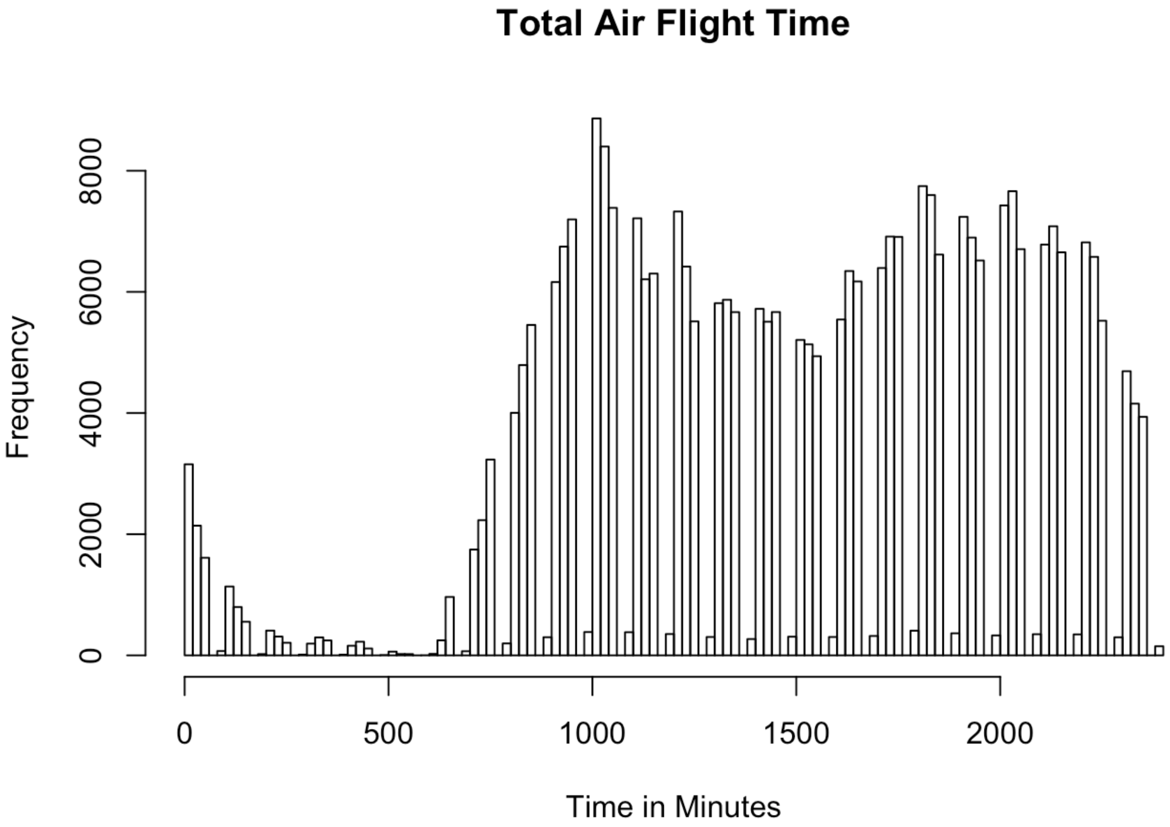
```
# The pattern is that the best time to fly out of NYC is during the months of S
eptember October November
```

3

Exploratory Data Analysis

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

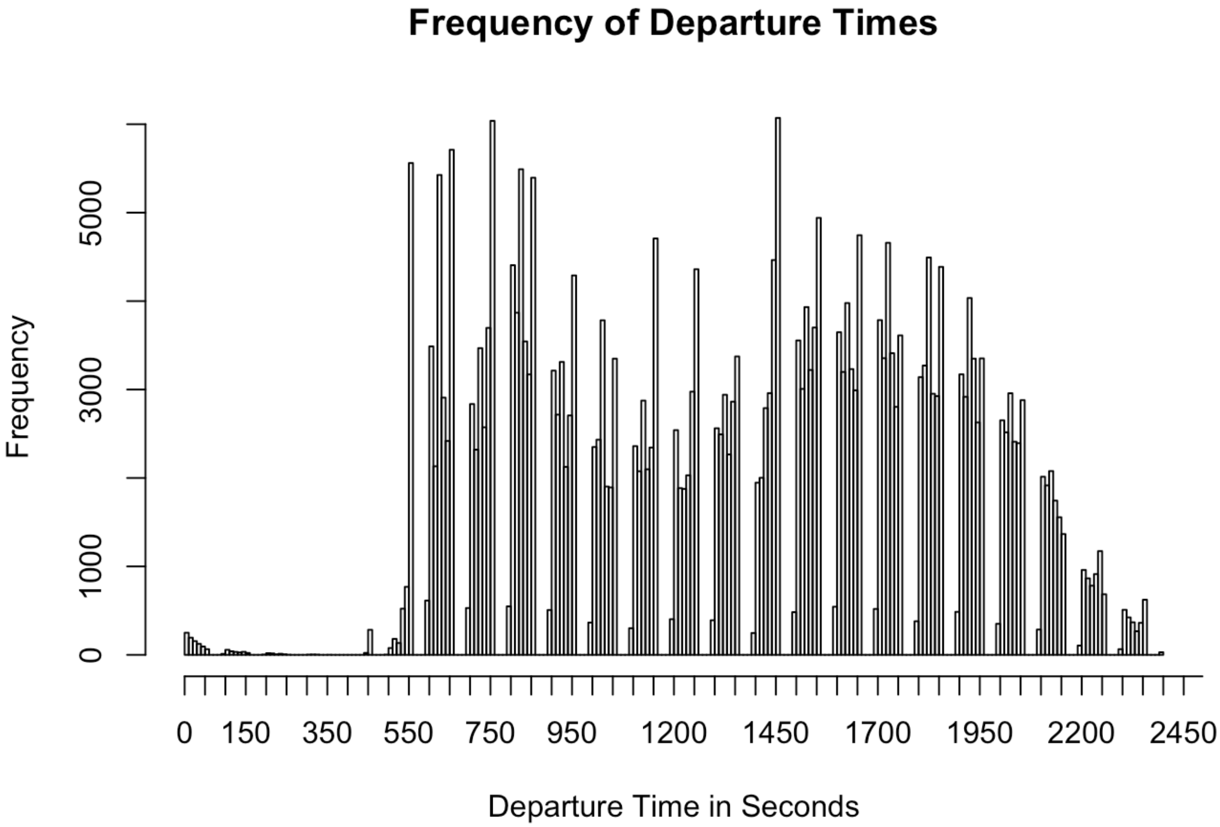
```
hist(f$arr_time, breaks = 100, main = 'Total Air Flight Time', xlab = 'Time in
Minutes')
```



There is one clear peak for this distribution

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

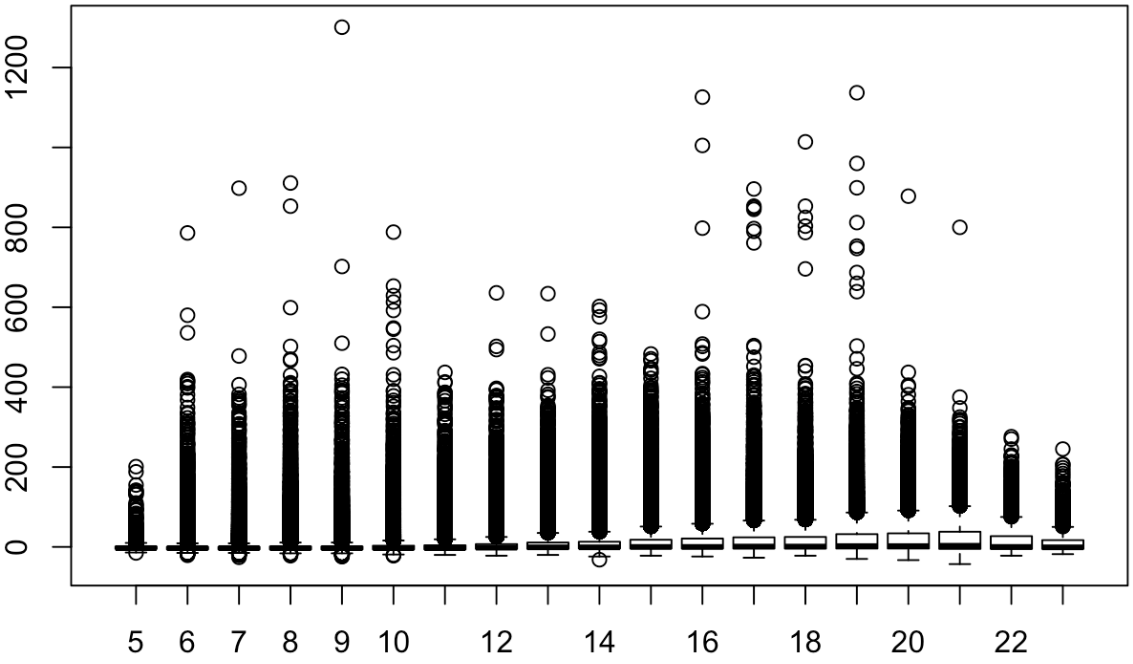
```
time_inhr <- f$dep_time
hist(time_inhr, breaks = 300, main = 'Frequency of Departure Times', xlab = 'De
parture Time in Seconds',xaxt='n')
axis(side=1,at= seq(0,3000,50),  labels=seq(0,3000,50))
```


Frequency of Departure Times

Flights most commonly depart in the afternoon. There could be two popular times of the day to depart because one could be early in the morning for business people flying out to work at around 7.30 am and the other popular time would be for the general public that flies at most time of the day which is during the afternoon somewhere near 2.50 pm

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

```
boxplot(dep_delay ~ hour,breaks = 10,  data = f)
```

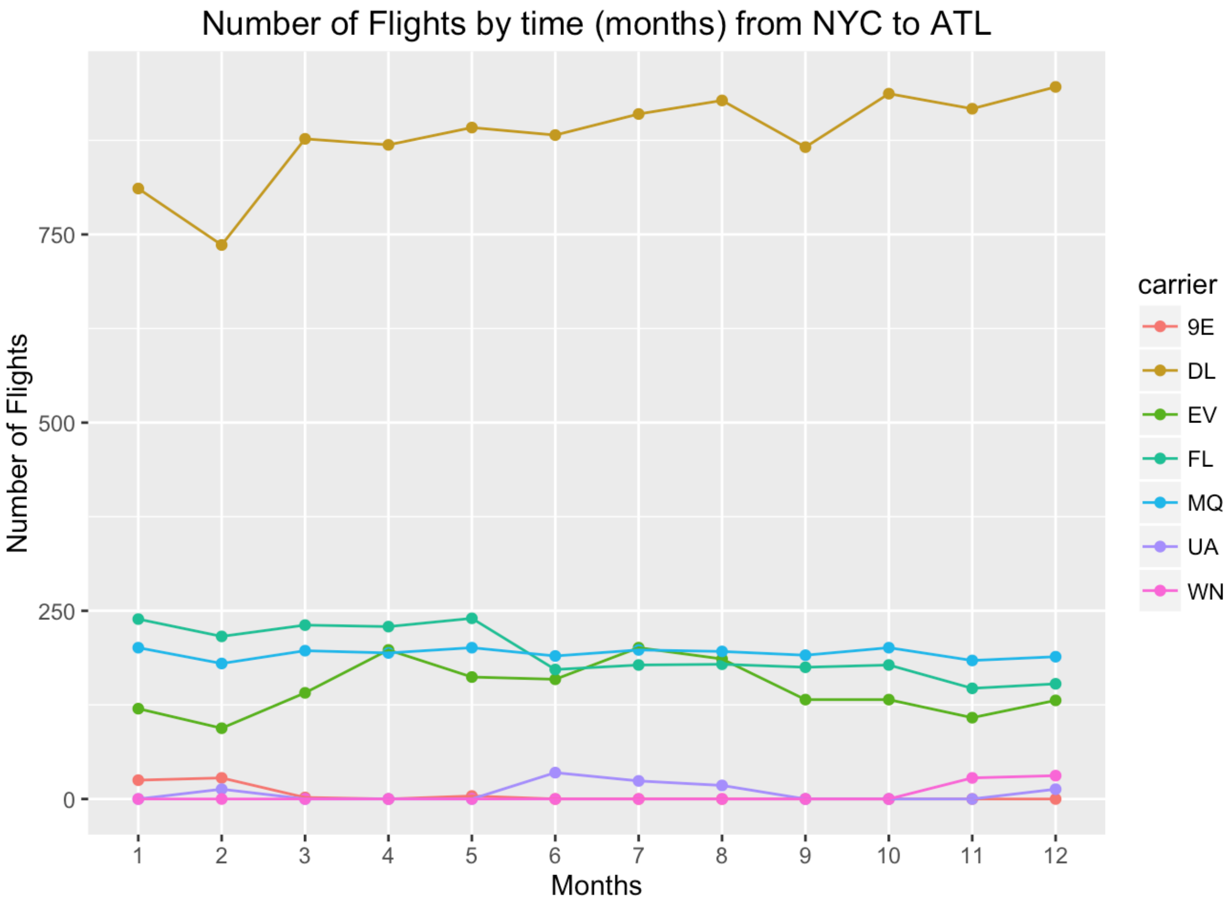


4

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.

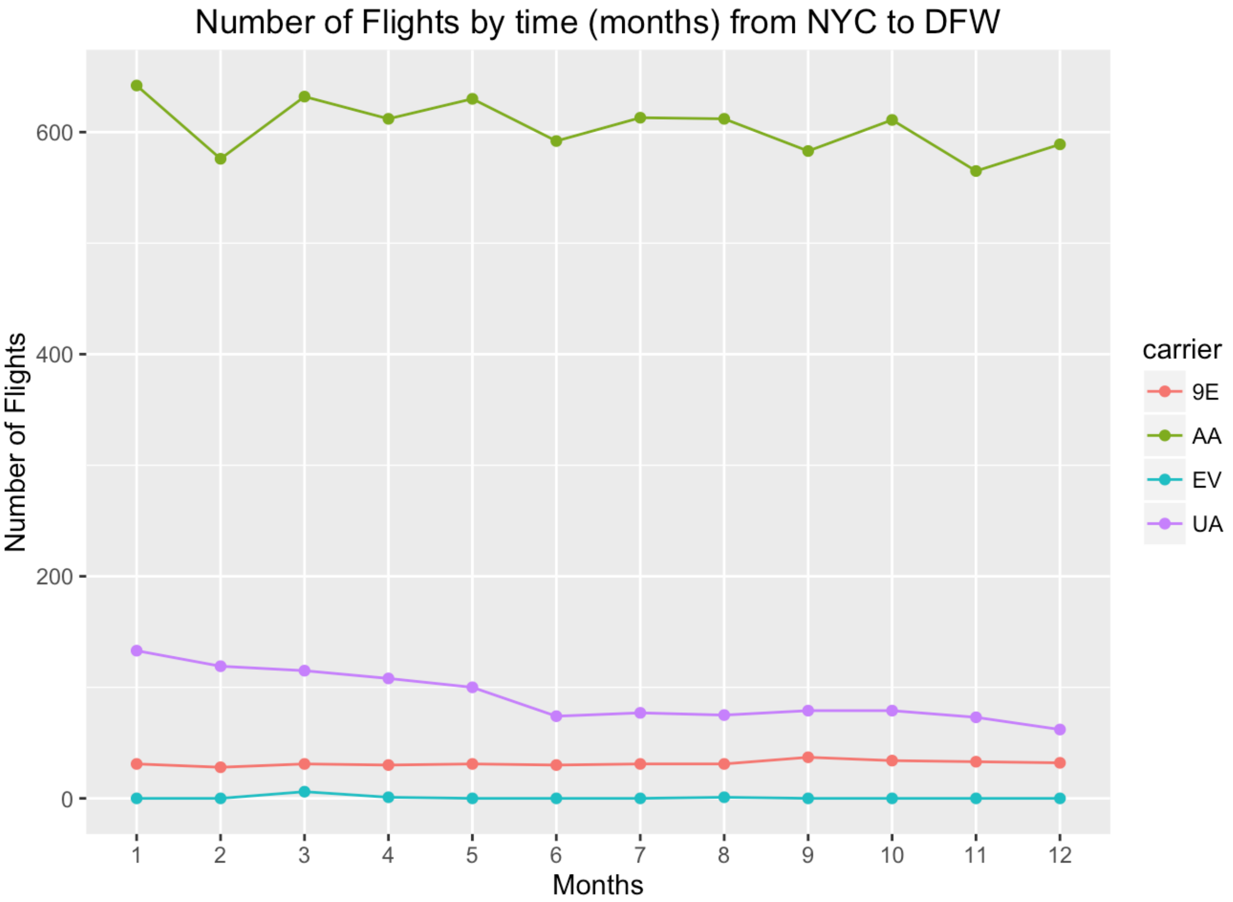
Research Question: Does the airlines HUB airport influence the number of flights to that airport from NYC over the months in the year 2013

```
# Number of flights of the carriers from NYC to Atl over the 12 months
library(ggplot2)
library(plyr)
nyc_atl <- f[(f$dest %in% c('ATL') & f$origin %in% c('LGA','EWR','JFK')),]
frame <- data.frame(table(nyc_atl$month,nyc_atl$carrier))
frame <- rename(frame,c("Var1" = "month","Var2" = "carrier", "Freq"= "frequency
"))
ggplot(frame,mapping = aes(x = frame$month, y = frequency,color = carrier, grou
p = carrier)) +
geom_point() + geom_line() + xlab('Months') + ylab('Number of Flights')  + labs
(title = 'Number of Flights by time (months) from NYC to ATL')
```



As we can observe in the plot above there are 5 carriers that fly from NYC to Atlanta. Delta has the highest total number of flights to ATL from NYC airports over all months in 2013 and there is a huge gap between Delta and other carriers. Let us now take a look at flights to Fort Worth, TX which is the hub for American Airlines

```
library(ggplot2)
library(plyr)
labmon =c("JAN","FEB","MAR","APR","MAY","JUN","JUL","AUG","SEP","OCT","NOV","DE
C")
nyc_dfw <- f[(f$dest %in% c('DFW') & f$origin %in% c('LGA','EWR','JFK')),]
frame <- data.frame(table(nyc_dfw$month,nyc_dfw$carrier))
frame <- rename(frame,c("Var1" = "month","Var2" = "carrier", "Freq"= "frequency
"))
ggplot(frame,mapping = aes(x = frame$month, y = frequency,color = carrier, grou
p = carrier)) +
geom_point() + geom_line() + xlab('Months') + ylab('Number of Flights') + labs(
title = 'Number of Flights by time (months) from NYC to DFW')
```



It is interesting to see how there is such a stark difference between the total number of flights of other carriers to DFW and the total number of flights of American Airlines to DFW. So our research question is in some way answered that the HUB of the airline influences the number of flights that Airline has from NYC airports