Assignment 2 - INFX 573

Prem Shah

1/29/2018

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

NOTE: Since the data is only for the year 2013, for the question which specifically ask insights about 2013, we will not be filtering based on the year, since there is only ONE year in the dataset. Also, all flights are from NYC. Hence, no filtering for source is required.

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

```
data(flights)
q1a <- count(flights)
pander(as.integer(q1a))
```

336776

336776 flights departed or arrived NYC in 2013

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

```
q1b <- count(filter(flights, dest == 'SEA'))
pander(as.integer(q1b))
```

3923

3923 flights flew from NYC to SEA in 2013

c. How many airlines fly from NYC to Seattle?

```
q1c <- filter(flights, dest == 'SEA' )
temp <- length(unique(q1c$carrier))
pander(temp)
```

5

5 airlines flew from NYC to SEA in 2013.

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

```
q1d <- filter(flights, dest == 'SEA' )
q1d <- mean(q1d$arr_delay, na.rm = T)
pander(q1d)
```

-1.099

The average arrival delay was -1.0990991 minutes for flights that flew from NYC to SEA. The negative value signifies that flights usually arrived early to SEA from NYC.

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?

```
q2a <- mean(flights$arr_delay, na.rm = T)
pander(q2a)
```

6.895

```
q2a_2 <- median(flights$arr_delay, na.rm = T)
pander(q2a_2)
```

-5

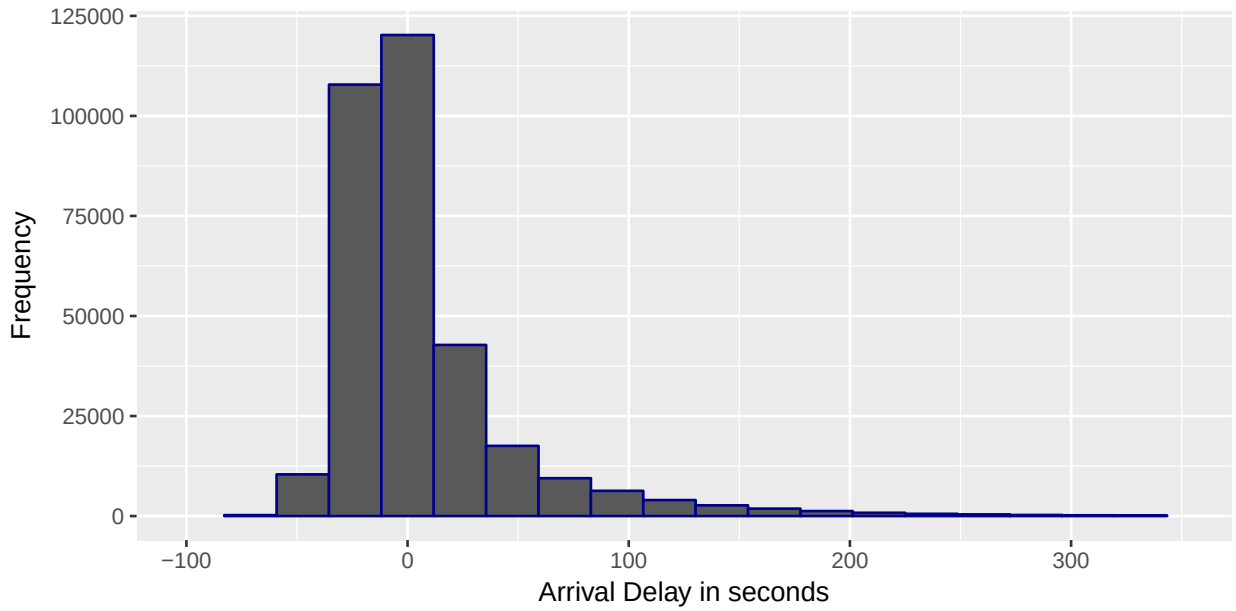
The mean arrival time is 6.8953768 seconds while the median arrival time is -5.

b. What does a negative arrival delay mean?

A negative arrival delay means that the flight landed before its scheduled arrival time

c. Histogram of arrival delay times

```
ggplot(flights, aes(flights$arr_delay)) +
  geom_histogram(bins = 20, color="darkblue") + xlab('Arrival Delay in seconds') +
  ylab('Frequency') +xlim(-100,350)
```



The mean found in (a) seems to correspond with this histogram. Because of the skewness, the mean line deviates from the center.

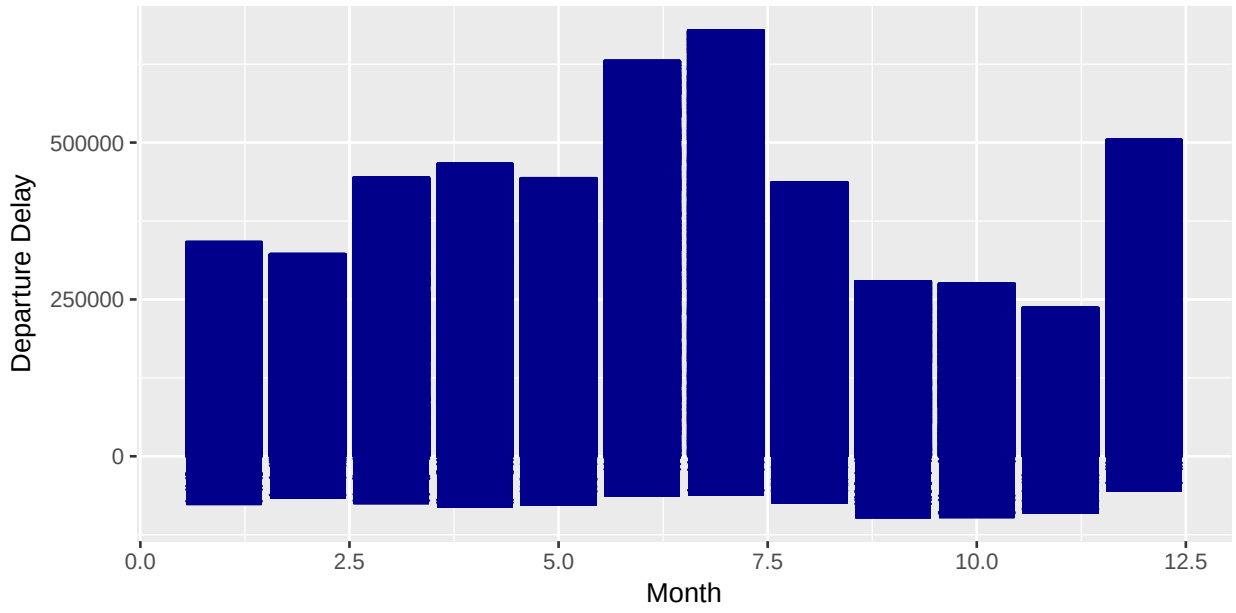
d. Is there seasonality in departure delays?

```
q2d <- by(flights$dep_delay, flights$month, function(x) mean(x, na.rm=T))
pander(q2d)
```

- **1**: *10.04*
- **2**: *10.82*
- **3**: *13.23*
- **4**: *13.94*
- **5**: *12.99*
- **6**: *20.85*
- **7**: *21.73*
- **8**: *12.61*
- **9**: *6.722*
- **10**: *6.244*
- **11**: *5.435*
- **12**: *16.58*

Yes, as we can see from the means from the months, there is seasonality in the dearture delays. It is the lowest during the months of September, October & November which serve as the best months to leave New York since they have the least delays. The worst month to leave would be June since it has the highest average delay. We can also see the same in the graph below.

```
ggplot(flights, aes(x = flights$month, y = flights$dep_delay)) +
    geom_bar(color="darkblue", stat = "identity") + xlab('Month') + ylab('Departure Delay')
```

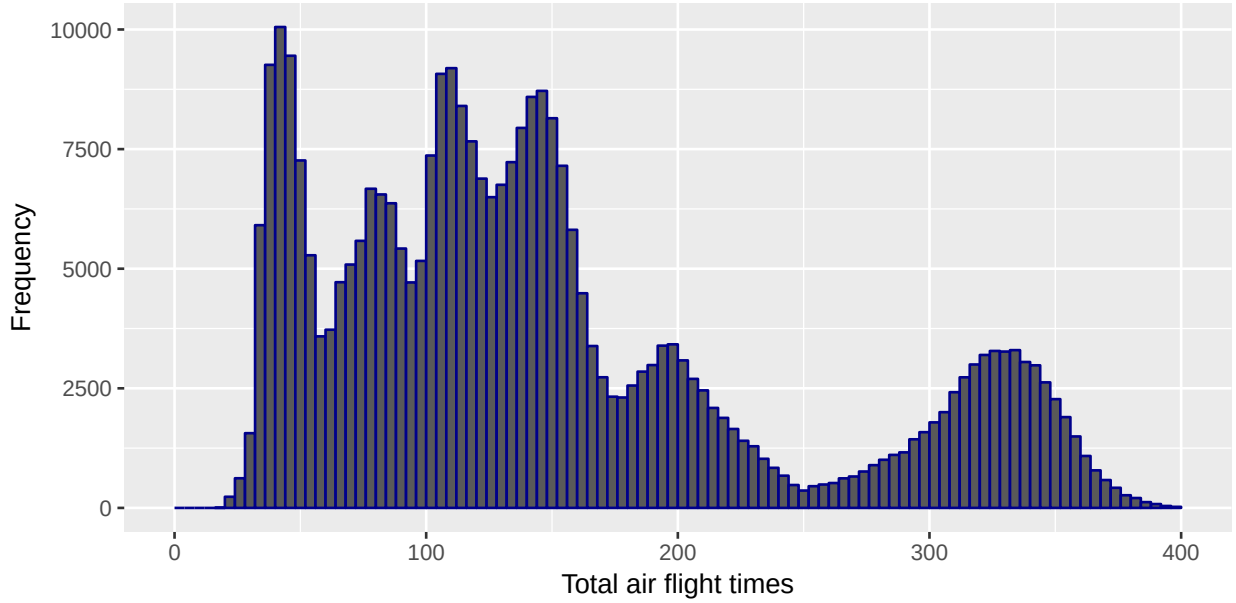


3. Exploratory Data Analysis

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?

```
ggplot(flights, aes(flights$air_time)) +
  geom_histogram(breaks = seq(0,400,by=4), color="darkblue") +
```

```
  xlab('Total air flight times') +
  ylab('Frequency')
```

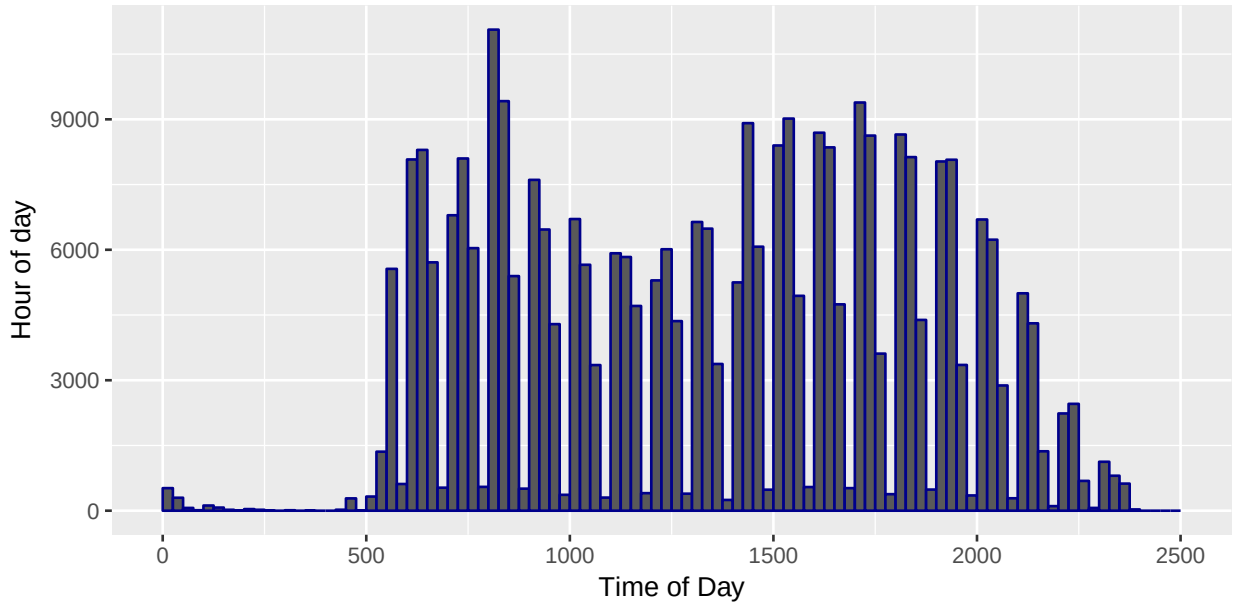


I can see 6 peaks in the distribution and one explanation for this might be that the peaks signify the most common flight routes and hence have more number of flights which justify the higher frequency of air times.

Another inference might be because there are more flights with shorter air times, people might prefer short routes.

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

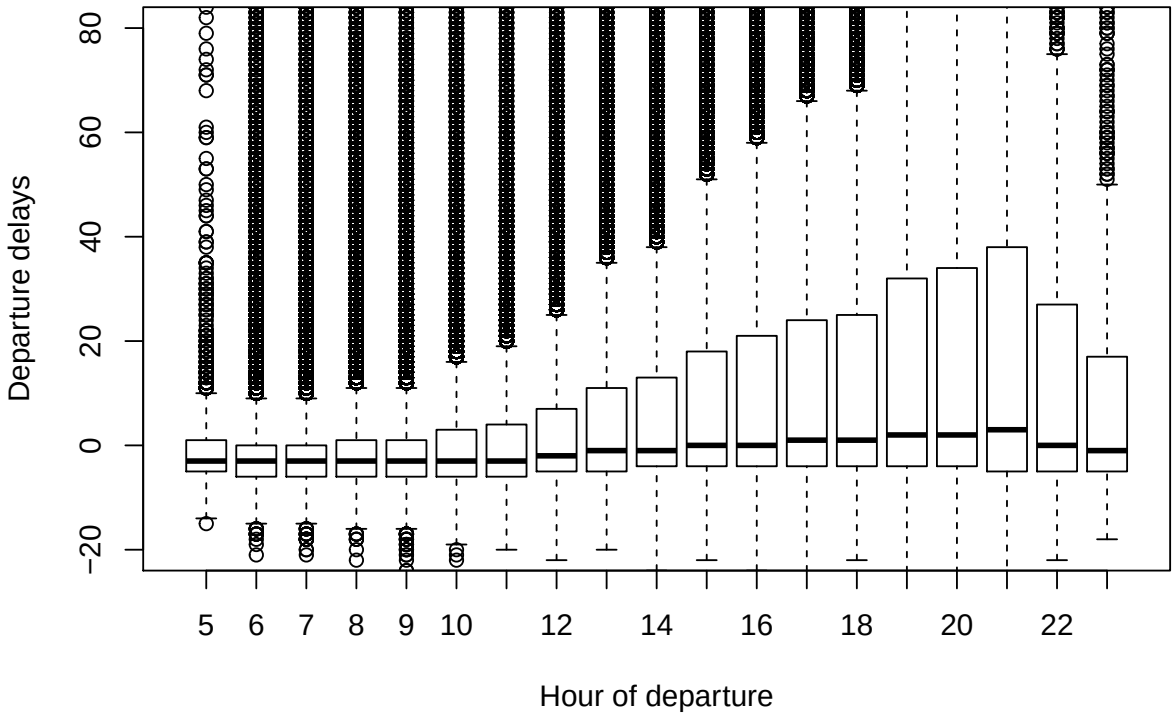
```
ggplot(flights, aes(flights$dep_time)) +
  geom_histogram(breaks = seq(0,2500,by=25), color="darkblue") + xlab('Time of Day') + ylab('Hour of da
```



There can be two popular times of day to depart because it represents after or before office hours. Early mornings and late evenings as wen can see in the graph above. Most common times seem to be from 8-10 AM and 4-7 PM approximately. THis might signify that a lot of people travel for business and they might want to go back the same day after work.

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, range =1.5, ylim=c(-20,80), ylab="Departure delays", xlab='Hour
```



This boxplot suggests that the Q3 range goes on increasing as the hours go from the start till 8 pm and then it decreases which shows that the delay of flights is higher in the night than in the morning

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.

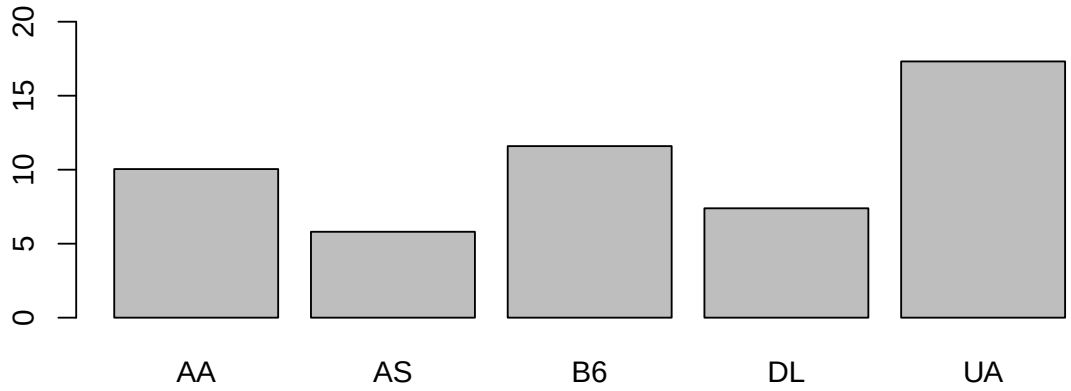
For the flights from New York to Seattle, which airline has the best performance in terms of delays?

This question helps people who do not like flight delays and who want the best flight in terms of the least arrival and departure delay.

```
library(dplyr)
q4 <- filter(flights, dest == 'SEA')
x1 <- by(q4$dep_delay, q4$carrier, function(x) mean(x, na.rm=T))
pander(x1)
```

- **AA**: *10.04*
- **AS**: *5.805*
- **B6**: *11.59*
- **DL**: *7.391*
- **UA**: *17.32*

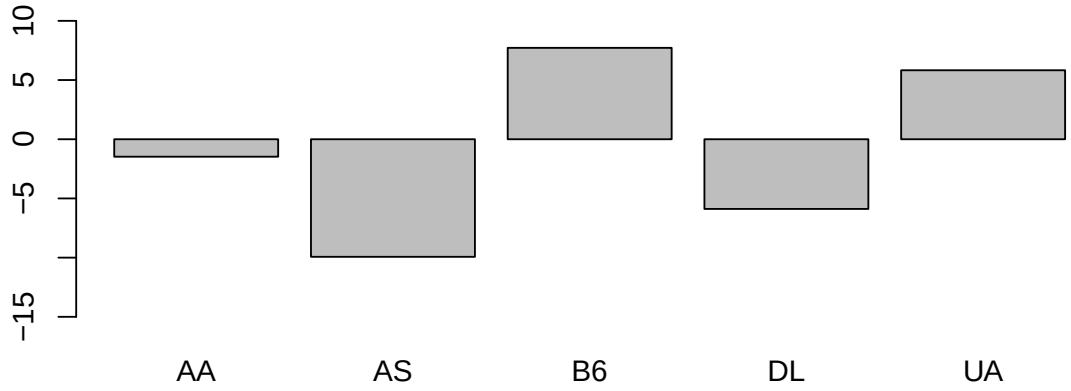
```
barplot(x1, ylim=c(0,20))
```




```
x2 <- by(q4$arr_delay, q4$carrier, function(x) mean(x, na.rm=T))
pander(x2)
```

- **AA**: *-1.475*
- **AS**: *-9.931*
- **B6**: *7.721*
- **DL**: *-5.886*
- **UA**: *5.827*

```
barplot(x2, ylim=c(-15,10))
```



As you can see from the above plots, AS (Alaska Airlines) has the least average departure delay and the least

average arrival delay as well while United Airlines (UA) has the highest average departure delay for flights to Seattle. B6 has the highest average arrival delay for flights from NYC to Seattle.

Hence, people who want departures & arrivals on time should choose Alaska Airlines. This inference might have one problem which we did not consider is that since Alaska Airlines is based out of Seattle, hence might have more flights in this route which might have resulted in the lower departure delay. But as we can see from the below table, that is not the case. Hence we can safely say that Alaska Airlines has a good track record of arriving and departing on time.

```
pander(table(q4$carrier))
```

AA	AS	B6	DL	UA
365	714	514	1213	1117