INFX 573: Problem Set 1 - Exploring Data

Rajendran Seetharaman

Due: Thursday, October 12, 2017

Collaborators:

Instructions:

Before beginning this assignment, please ensure you have access to R and RStudio.

1. Download the `problemset1.Rmd` file from Canvas. 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.
4. Collaboration on problem sets is acceptable, and even encouraged, but each student must turn in an individual write-up in his or her own words and his or her own work. The names of all collaborators must be listed on each assignment. Do not copy-and-paste from other students' responses or code.
5. 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 R Markdown file to `YourLastName_YourFirstName_ps1.Rmd`, knit a PDF and submit the PDF file on Canvas.

stress more visualization, dplyr, less questions/ethics, etc

Setup:

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

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

```
## Warning: package 'tidyverse' was built under R version 3.3.3

## Warning: package 'ggplot2' was built under R version 3.3.3

## Warning: package 'tibble' was built under R version 3.3.3

## Warning: package 'tidyr' was built under R version 3.3.3

## Warning: package 'readr' was built under R version 3.3.3

## Warning: package 'purrr' was built under R version 3.3.3

## Warning: package 'dplyr' was built under R version 3.3.3
```

```r
library("nycflights13")
```

```
## Warning: package 'nycflights13' was built under R version 3.3.3
```

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.

(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.

The data set has 336776 observations of 19 variables

The data set has the following variables- $ year : Integer which represents the year of the flight (one year - 2013) $ month : Integer which represents the month of the flight $ day : Integer which represents the day of the flight $ dep_time : Integer which represents the departure time of the flight $ sched_dep_time: Integer which represents the scheduled departure time of the flight $ dep_delay : Numeric value which represents the departure delay of the flight (Mean - 12.64 mins, 8255 NA values) $ arr_time : Integer which represents the arrival time of the flight $ sched_arr_time: Integer which represents the scheduled arrival time of the flight $ arr_delay : Numeric value which represents the arrival delay of the flight (Mean - 6.89 mins, 9420 NA values) $ carrier : String which represents the carrier code (16 unique carriers)

```r
unique(flights$carrier)
```

```
##  [1] "UA" "AA" "B6" "DL" "EV" "MQ" "US" "WN" "VX" "FL" "AS" "9E" "F9" "HA"
## [15] "YV" "OO"
```

$ flight : Integer which represents the flight number $ tailnum : String which represents the flight tail number $ origin : String which represents origin airport (3 unique values - EWR, JFK, LGA)

```r
unique(flights$origin)
```

```
## [1] "EWR" "LGA" "JFK"
```

$ dest : String which represents destination airport (105 unique destinations)

```r
unique(flights$dest)
```

```
##   [1] "IAH" "MIA" "BQN" "ATL" "ORD" "FLL" "IAD" "MCO" "PBI" "TPA" "LAX"
##  [12] "SFO" "DFW" "BOS" "LAS" "MSP" "DTW" "RSW" "SJU" "PHX" "BWI" "CLT"
##  [23] "BUF" "DEN" "SNA" "MSY" "SLC" "XNA" "MKE" "SEA" "ROC" "SYR" "SRQ"
##  [34] "RDU" "CMH" "JAX" "CHS" "MEM" "PIT" "SAN" "DCA" "CLE" "STL" "MYR"
##  [45] "JAC" "MDW" "HNL" "BNA" "AUS" "BTV" "PHL" "STT" "EGE" "AVL" "PWM"
##  [56] "IND" "SAV" "CAK" "HOU" "LGB" "DAY" "ALB" "BDL" "MHT" "MSN" "GSO"
##  [67] "CVG" "BUR" "RIC" "GSP" "GRR" "MCI" "ORF" "SAT" "SDF" "PDX" "SJC"
##  [78] "OMA" "CRW" "OAK" "SMF" "TUL" "TYS" "OKC" "PVD" "DSM" "PSE" "BHM"
##  [89] "CAE" "HDN" "BZN" "MTJ" "EYW" "PSP" "ACK" "BGR" "ABQ" "ILM" "MVY"
## [100] "SBN" "LEX" "CHO" "TVC" "ANC" "LGA"
```

$ air_time : Numeric value which represents the air time of the flight $ distance : Numeric value which represents the flight distance $ hour : Numeric value which represents the hour of the flight departure $ minute : Numeric value which represents the minute of the flight departure $ time_hour : Timestamp which represents the date and time of the flight

(b) Formulating Questions:

Consider the NYC flights data. Formulate three motivating questions you want to explore using this data and explain why they are of interest.

Q1. Which of the 3 airports has the greatest average departure delay? Which carriers are delayed the most at that airport? - This question is important as delay times are the greatest barriers to operational efficiency of an airport. If the carriers which have high average delays take corrective action to resolve the delays, the overall delay time of flights at the airport can be reduced resulting in higher operational efficiency and revenue for both the carriers and the government.

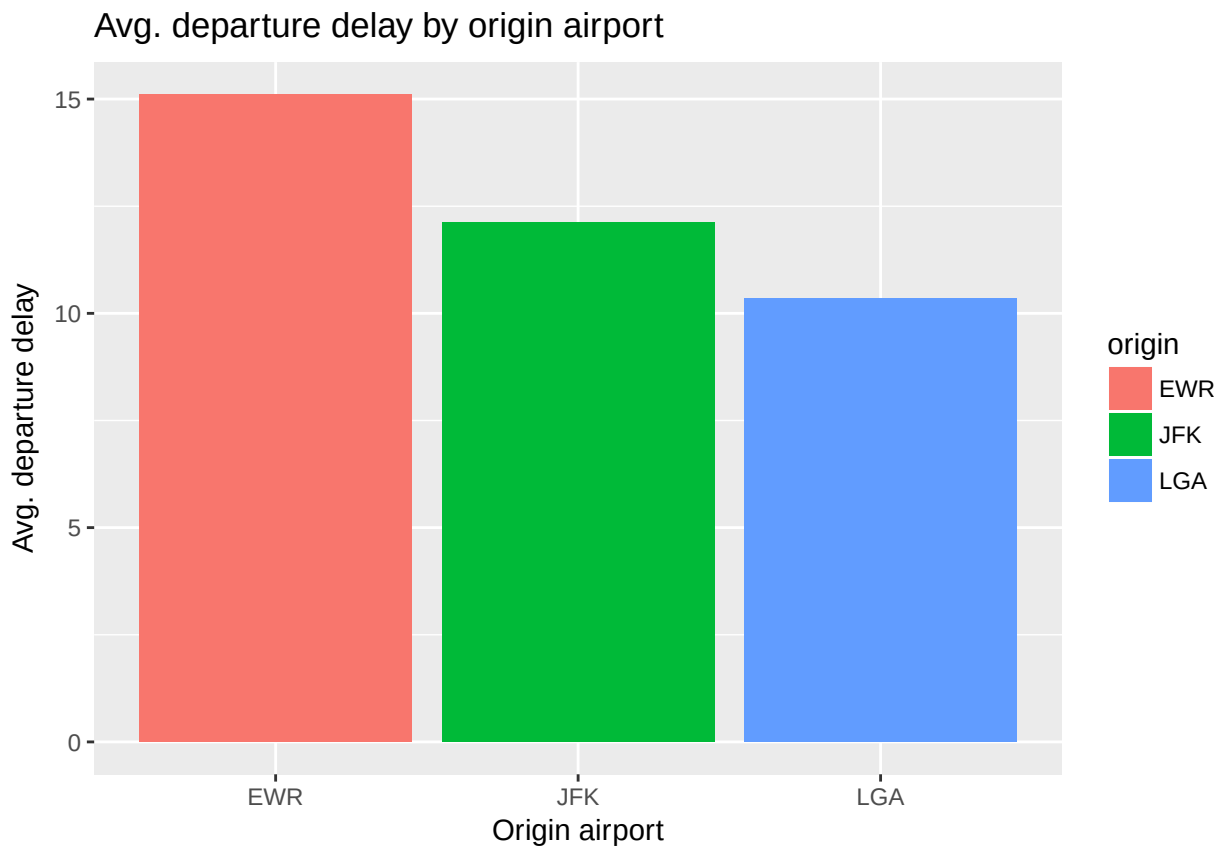
Q2. For each flight having a departure delay, do a majority of the carriers on an average, manage to achieve an on time arrival, despite the delay? For the carriers which managed to achieve this turn around for a large number flights, what were the destination airports that these flights were travelling to? - A lot of carriers try to cover up the departure delay by reducing the flight time so that the flight arrival is on time. Some carriers are better than the others in terms of reducing delays on certain routes. It is important for carriers to understand what routes they are efficient on, so that they can maximize revenue on those routes.

Q3. Is there any seasonality trend in the departure delays by origin airport? Do all three NY airports show a common seasonality trend or is there a variation? Is this trend reflected in the individual carrier delays of the 2 major US airlines - United and Delta? - It is important for airports and carriers to understand the seasonality trend in delays so that they can better plan the flight schedules and take additional corrective to minimize those delays.

(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 related to each question. Be sure to describe what the visuals show and how they speak to your question of interest.

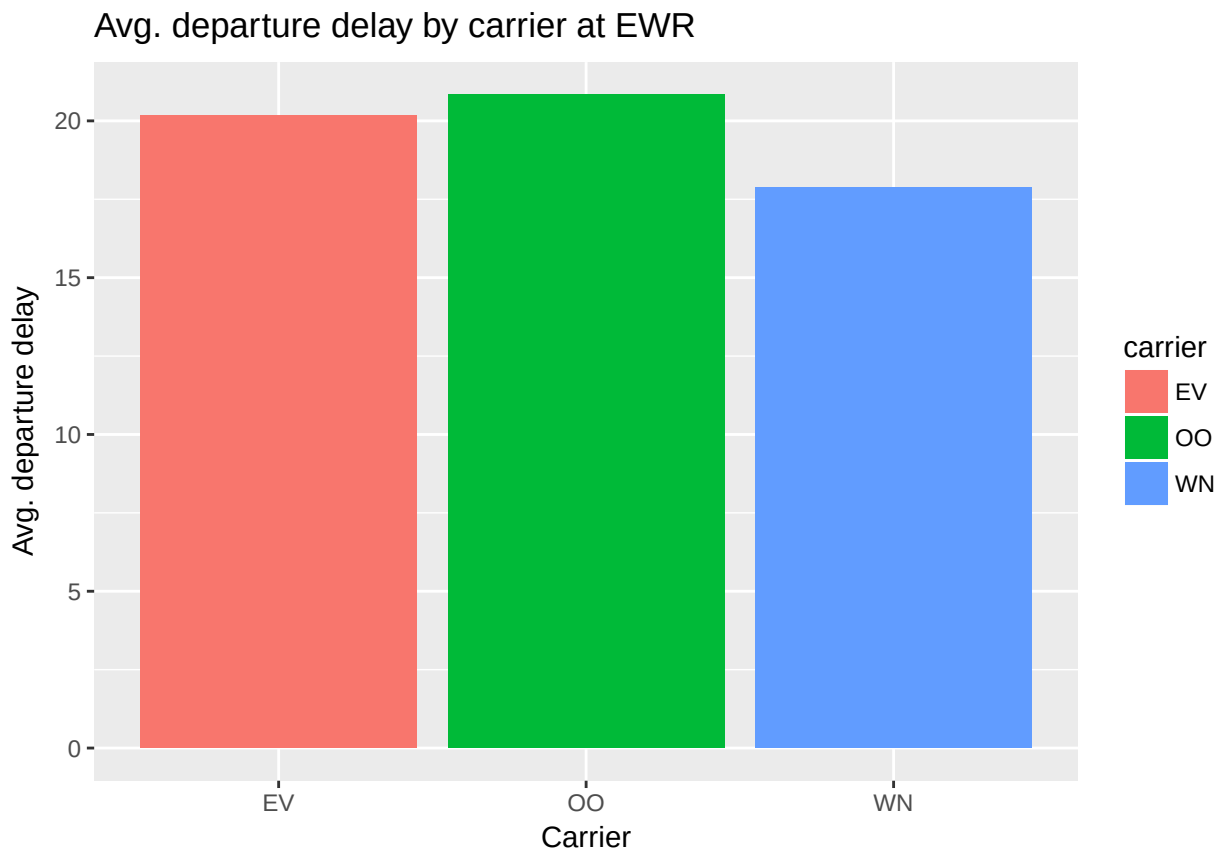
```
# Question 1:
# Answer: The below analysis shows that EWR has the greatest average departure delay (15 mins) out of
# the 3 airports in NY. Within EWR, OO, EV, and WN have the 3 greatest average departure delay times.
# This might indicate that these 3 carriers are responsible for the high average departure delay of EWR
# If these carriers take corrective action, the avg departure delay of EWR can be brought down.

# compute avg departure delay by origin airport
x <- flights %>% group_by(origin) %>% summarise(avg_dep_delay=mean(dep_delay,na.rm=TRUE))

# Plot graph of avg departure delay by airport of origin
ggplot(data=x,mapping=aes(x=origin,y=avg_dep_delay))+geom_bar(mapping=aes(fill=origin), stat="identity"
```



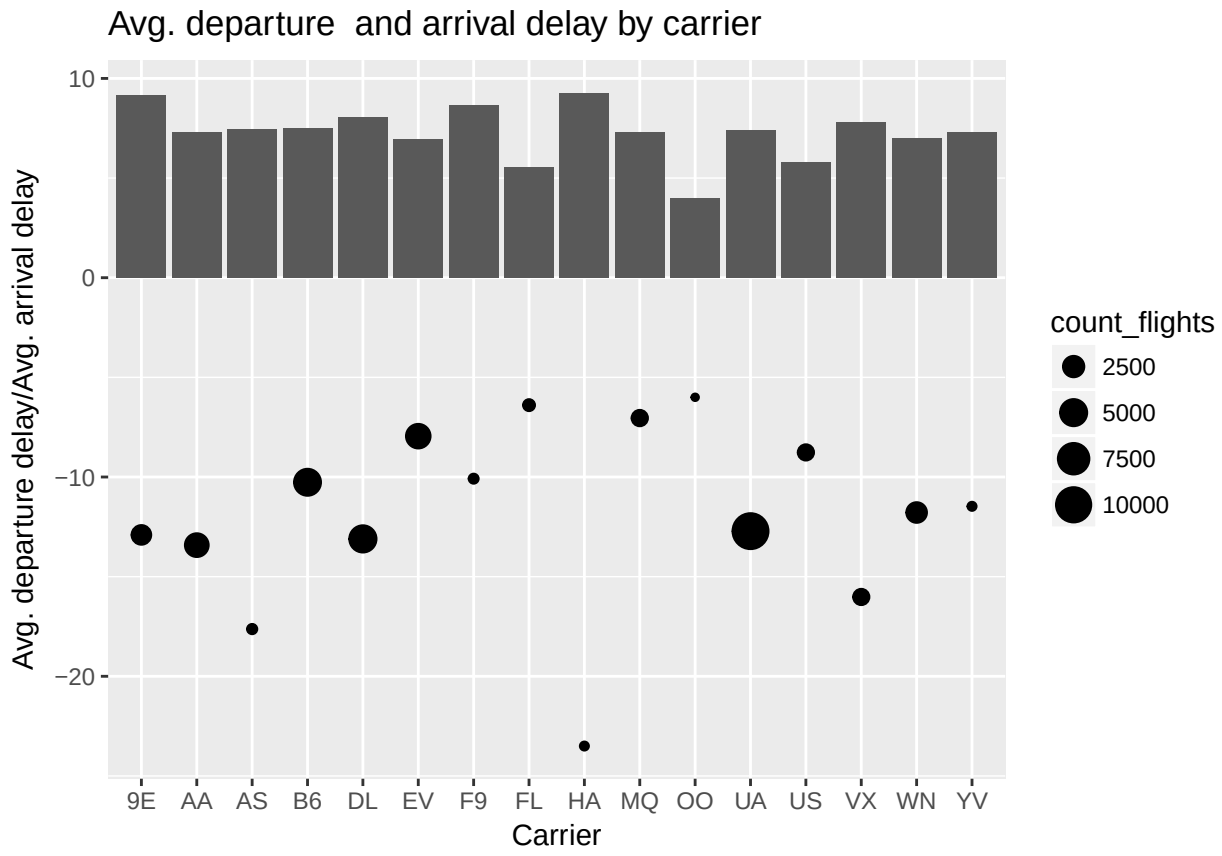

```
# compute avg departure delay by origin airport and carrier
y <- flights %>% group_by(origin,carrier) %>% summarise(avg_dep_delay=mean(dep_delay,na.rm=TRUE))

#Only retain records with origin as EWR
#Only retain top 3 carriers flying out of EWR with the greatest departure delays
z <- y %>% filter(origin=="EWR") %>% arrange(desc(avg_dep_delay)) %>% head(3)
```

```
## Warning: package 'bindrcpp' was built under R version 3.3.3
```

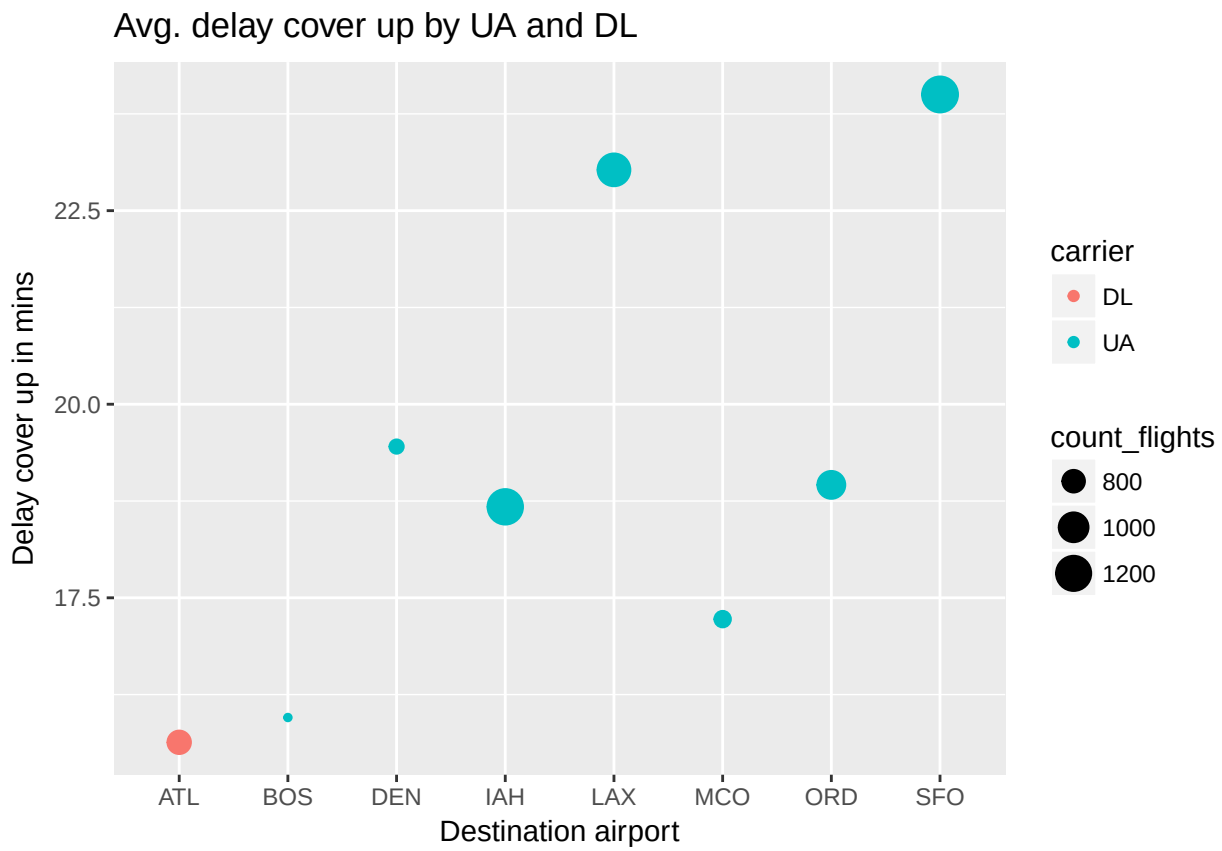
```
#plot graph showing the greatest average departure delays of carriers within EWR
ggplot(data=z,mapping=aes(x=carrier,y=avg_dep_delay))+geom_bar(mapping=aes(fill=carrier),stat="identity
```



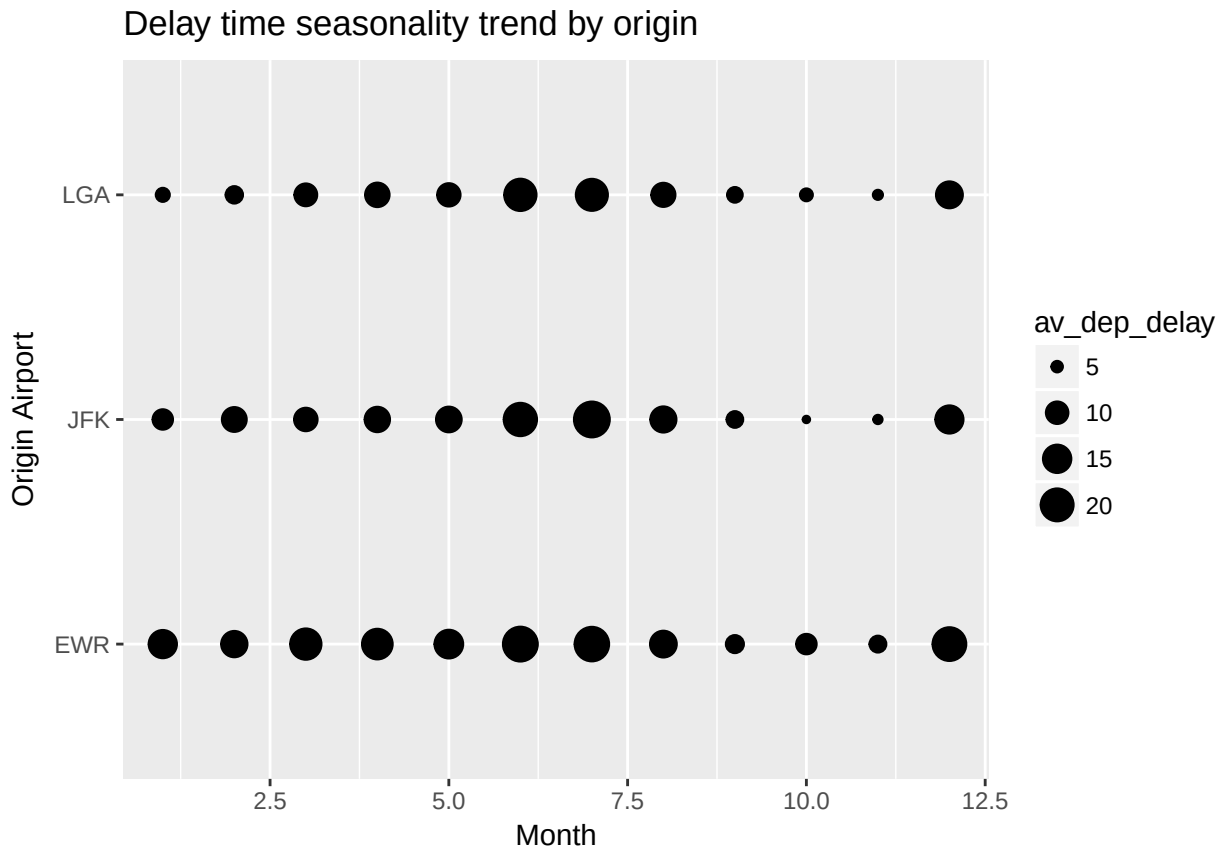

```
# Question 2:
# Answer: Viz 1 shows that most of the carriers managed to achieve an on time arrival
# for atleast a fraction of their flights. UA and DL managed this turn around
# for the greatest proportion of flights.
# UA had the greatest turn around success at - BOS, DEN, IAH, LAX, MCO, ORD, and SFO.
# DL had the greatest turn around success at ATL. (Turn Around flight count greater than 500 flights)

# compute flights having dep delay but arriving on time.
# compute the average dep and arr delay for these flights by carrier and the count of flights
# Arrange by descending count of flights
carrier_cover_up2 <- flights %>% filter(dep_delay>0) %>% filter(arr_delay<=0) %>% group_by(carrier) %>%

# Which of the carrirs managed to achieve a delay turn around? For how many flights?
# What was the average delay turn around time for these carriers?
ggplot(data=carrier_cover_up2,mapping = aes(x=carrier)) + geom_bar(mapping=aes(y=avg_dep_delay),stat="i
```



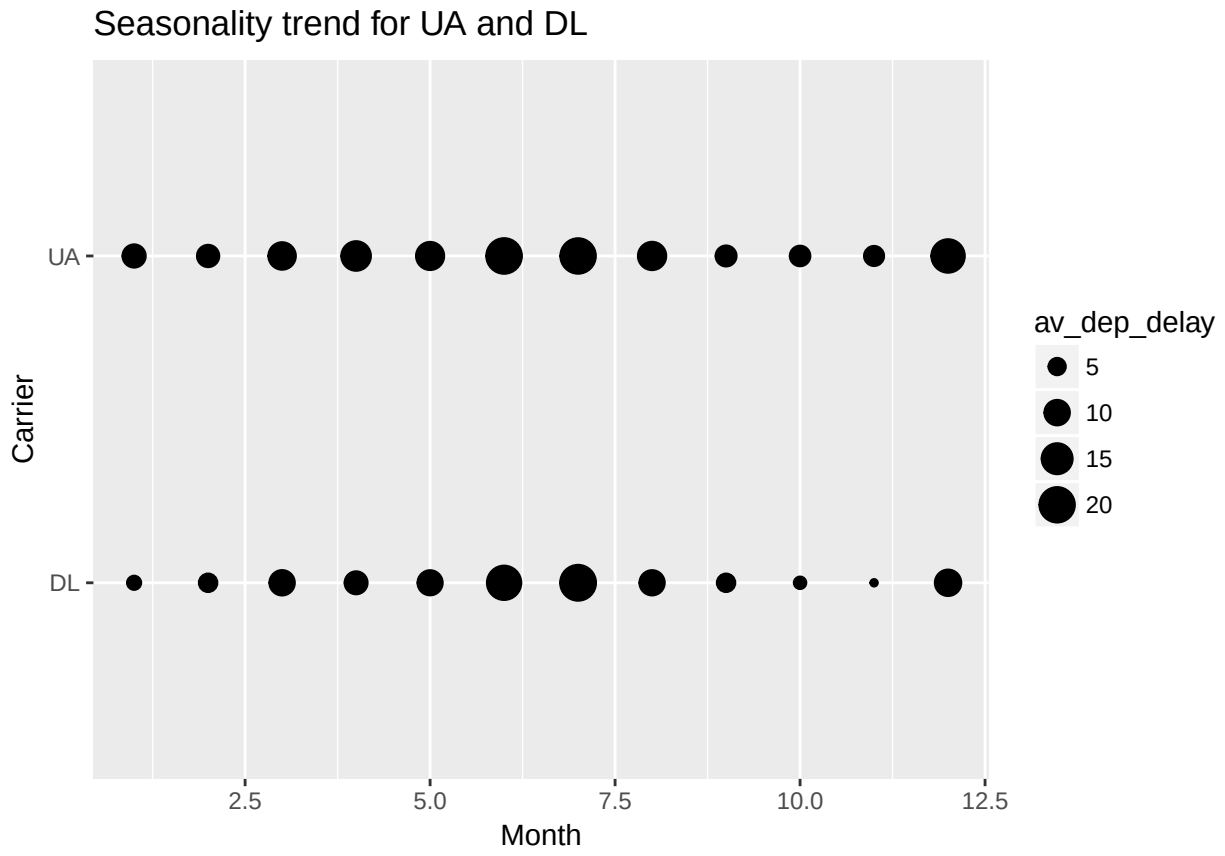

```
# UA and DL managed to achieve a delay turn around for a large proportion of delayed flights
# What were the destinations for which the delay turn around flight count was greater than 500?
carrier_cover_up_dest1 <- flights %>% filter(dep_delay>0) %>% filter(arr_delay<=0) %>% group_by(carrier

# Visualizing the delay cover up by UA and DL. What were the destination airports?
ggplot(data=carrier_cover_up_dest1 %>% filter(carrier %in% c("UA","DL")) %>% filter(count_flights>=500)
```




```
# Question 3
# Answer: Viz 1 shows a clear seasonality trend which is evident across all 3 airports
# The summmer months of June and July have the greatest average departure delay
# There is a period of 3 months - September, October, November with low avg flight delay times.
# December has high avg delay times, followed by a decline in January and february.
# The data indicates that holiday periods (Summer and christmas) might have a correlation
# with Avg. delay times. More people travel during holidays, resulting in more scheduled flights.
# This might result in a greater probability of a flight being delayed.
# EWR has a slight variation from the other airports. It has high avg delay times in January as well.
# This seasonality trend is crearly reflected in the avg delay time trend of United and Delta.

# Visualising the seasonality trend of avg delay times by origin airport
ggplot(data=flights %>% group_by(origin,month) %>% summarize(av_dep_delay=mean(dep_delay,na.rm=TRUE)), r
```




```
# Visualising the seasonality trend of avg delay times by carrier for United and Delta
ggplot(data=flights %>% filter(carrier %in% c("UA","DL")) %>% group_by(carrier,month) %>% summarize(av_
```



(d) Challenge Your Results:

After completing the exploratory analysis from Problem 1c, do you have any concerns about your findings? 1. The high departure delay of the carriers - OO, EV, and WN might indicate that there is a causation effect between those delays and the high average departure delay at Newark airport. However, the effect might merely be a correlation and some other factors might be responsible for the delays, for example high flight traffic, and passenger counts.
2. The destination airport might not be the only factor influncing the delay turn around success of a particular carrier. Other factors like air traffic and experience of the pilot might also heavily influence the delay turn around. 3. The seasonality trend of delays observed in the data could be due to chance; the results might not be statistically significant. Also, not every carrier might show the same seasonality effect like United and Delta as United and Delta are one of the biggest carriers in the US and their trends are likely to follow the general trend.