Yash Jain

Code

Hide

```
library(ggplot2)
library(dplyr)
```

```
package <U+393C><U+3E31>dplyr<U+393C><U+3E32> was built under R version 3.5.3
Attaching package: <U+393C><U+3E31>dplyr<U+393C><U+3E32>

The following objects are masked from <U+393C><U+3E31>package:stats<U+393C><U+3E32>:

    filter, lag

The following objects are masked from <U+393C><U+3E31>package:base<U+393C><U+3E32>:

    intersect, setdiff, setequal, union
```

Hide

```
library(magrittr)
```

```
package <U+393C><U+3E31>magrittr<U+393C><U+3E32> was built under R version 3.5.3
```

Hide

```
library(reshape2)
air_data <- read.csv("airline_delay.csv",na.strings="",stringsAsFactors=FALSE)
#Data Exploration
print("Summary of air_data")
```

```
[1] "Summary of air_data"
```

Hide

```
summary(air_data)
```

```
     year          month          carrier          carrier_name
 Min.   :2015   Min.   : 1.000   Length:58494       Length:58494
 1st Qu.:2016   1st Qu.: 4.000   Class :character   Class :character
 Median :2017   Median : 7.000   Mode  :character   Mode  :character
 Mean   :2017   Mean   : 6.509
 3rd Qu.:2018   3rd Qu.:10.000
 Max.   :2018   Max.   :12.000

   airport          airport_name         arr_delay      carrier_delay
 Length:58494       Length:58494       Min.   :     0   Min.   :     0
 Class :character   Class :character   1st Qu.:   443   1st Qu.:   143
 Mode  :character   Mode  :character   Median :  1233   Median :   441
                                       Mean   :  4721   Mean   :  1481
                                       3rd Qu.:  3351   3rd Qu.:  1215
                                       Max.   :429194   Max.   :196944
                                       NA's   :58       NA's   :58
 weather_delay       nas_delay      security_delay     late_aircraft_delay
 Min.   :    0.0   Min.   :     0   Min.   :   0.000   Min.   :     0
 1st Qu.:    0.0   1st Qu.:    56   1st Qu.:   0.000   1st Qu.:   116
 Median :   23.0   Median :   189   Median :   0.000   Median :   430
 Mean   :  229.7   Mean   :  1138   Mean   :   6.565   Mean   :  1866
 3rd Qu.:  162.0   3rd Qu.:   584   3rd Qu.:   0.000   3rd Qu.:  1296
 Max.   :31960.0   Max.   :112018   Max.   :2897.000   Max.   :147167
 NA's   :58        NA's   :58       NA's   :58         NA's   :58
```

Hide

```
#Removing rows with NA
air_data2 <- na.omit(air_data)
print("-----------------------------------------------------------")
```

```
[1] "-----------------------------------------------------"
```

Hide

```
print("Comparing dimensions before and afeter cleaning")
```

```
[1] "Comparing dimensions before and afeter cleaning"
```

Hide

```
cat("\n")
```

Hide

```
cat("air_data:",dim(air_data))
```

```
air_data: 58494 12
```

Hide

```
cat("\n")
```

Hide

```
cat("air_data2:",dim(air_data2))
```

```
air_data2: 58436 12
```

Hide

```
cat("\n")
```

Hide

```
print("-----------------------------------------------------")
```

```
[1] "-----------------------------------------------------"
```

Hide

```
print("Summary of air_data2")
```

```
[1] "Summary of air_data2"
```

Hide

```
summary(air_data2) # to check if null
```

```
     year          month          carrier          carrier_name      
 Min.   :2015   Min.   : 1.000   Length:58436       Length:58436      
 1st Qu.:2016   1st Qu.: 4.000   Class :character   Class :character  
 Median :2017   Median : 7.000   Mode  :character   Mode  :character  
 Mean   :2017   Mean   : 6.508                                        
 3rd Qu.:2018   3rd Qu.:10.000                                        
 Max.   :2018   Max.   :12.000                                        
   airport          airport_name         arr_delay      carrier_delay    
 Length:58436       Length:58436       Min.   :     0   Min.   :     0  
 Class :character   Class :character   1st Qu.:   443   1st Qu.:   143  
 Mode  :character   Mode  :character   Median :  1233   Median :   441  
                                       Mean   :  4721   Mean   :  1481  
                                       3rd Qu.:  3351   3rd Qu.:  1215  
                                       Max.   :429194   Max.   :196944  
 weather_delay       nas_delay      security_delay     late_aircraft_delay
 Min.   :    0.0   Min.   :     0   Min.   :   0.000   Min.   :     0     
 1st Qu.:    0.0   1st Qu.:    56   1st Qu.:   0.000   1st Qu.:   116     
 Median :   23.0   Median :   189   Median :   0.000   Median :   430     
 Mean   :  229.7   Mean   :  1138   Mean   :   6.565   Mean   :  1866     
 3rd Qu.:  162.0   3rd Qu.:   584   3rd Qu.:   0.000   3rd Qu.:  1296     
 Max.   :31960.0   Max.   :112018   Max.   :2897.000   Max.   :147167     
```

Hide

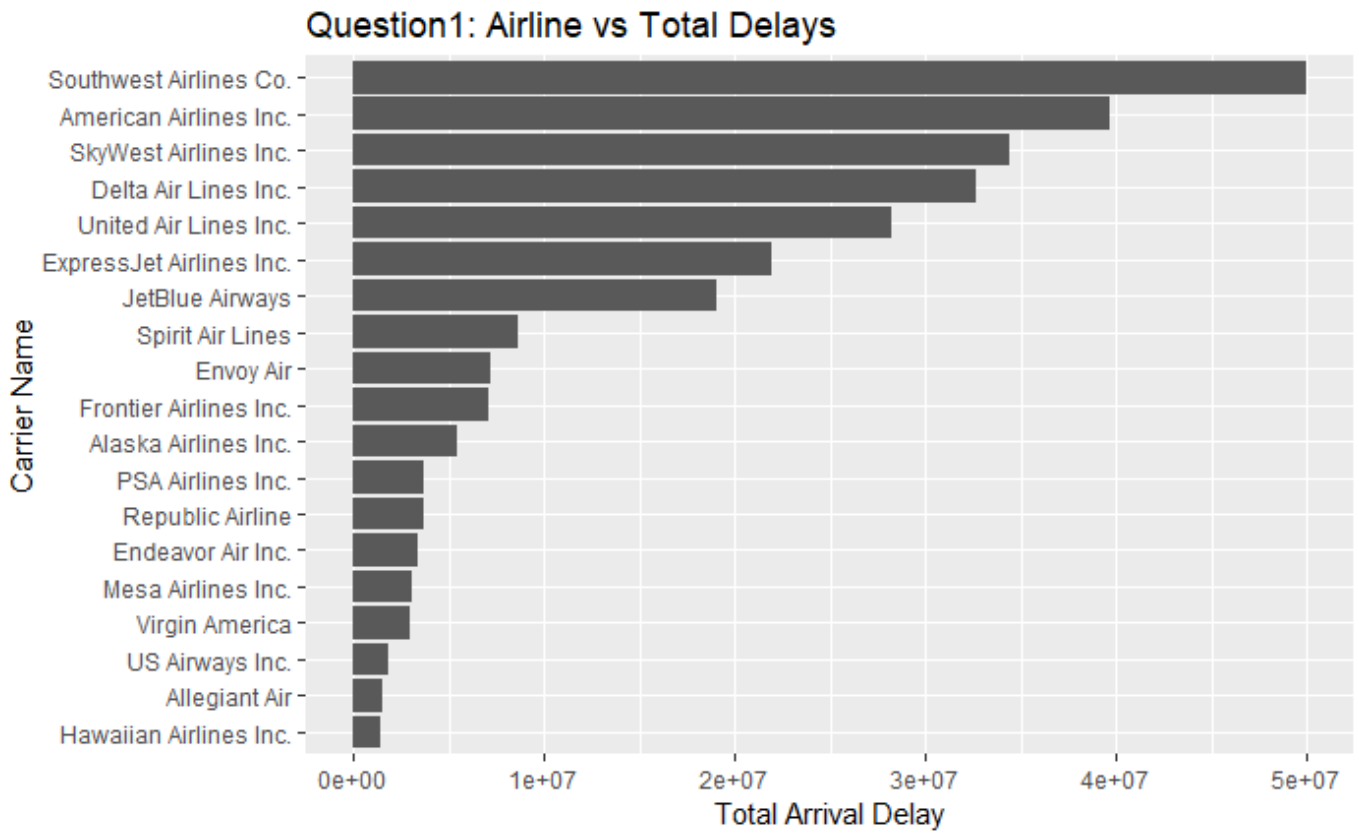
```
#Question 1
df_1 <- air_data2 %>%
  select(carrier_name, arr_delay) %>%
  group_by(carrier_name) %>%
  summarise(total_arrival_delay = sum(arr_delay)) %>%
  arrange(desc(total_arrival_delay))
head(df_1, 19)
```

carrier_name <chr>	total_arrival_delay <int>
Southwest Airlines Co.	49980502
American Airlines Inc.	39646209
SkyWest Airlines Inc.	34412580
Delta Air Lines Inc.	32676571
United Air Lines Inc.	28214925
ExpressJet Airlines Inc.	21922288
JetBlue Airways	19001712
Spirit Air Lines	8681183
Envoy Air	7234283
Frontier Airlines Inc.	7059606

1-10 of 19 rows Previous **1** 2 Next

Hide

```
ggplot(df_1, aes(y=total_arrival_delay, x=reorder(carrier_name, total_arrival_delay))) +
  geom_bar(stat="identity")+
  labs(y="Total Arrival Delay",
       x="Carrier Name",
       fill = "Year",
       title="Question1: Airline vs Total Delays" ) +
  coord_flip()
```



Hide

```
NA
```

Hide

```
#Question 2
df_2 <- air_data2 %>%
  select(carrier_name, arr_delay,year) %>%
  group_by(carrier_name,year) %>%
  summarise(total_arrival_delay = sum(arr_delay)) %>%
  arrange(desc(total_arrival_delay))
dim(df_2)
```

```
[1] 56  3
```

Hide

```
head(df_2, 56)
```

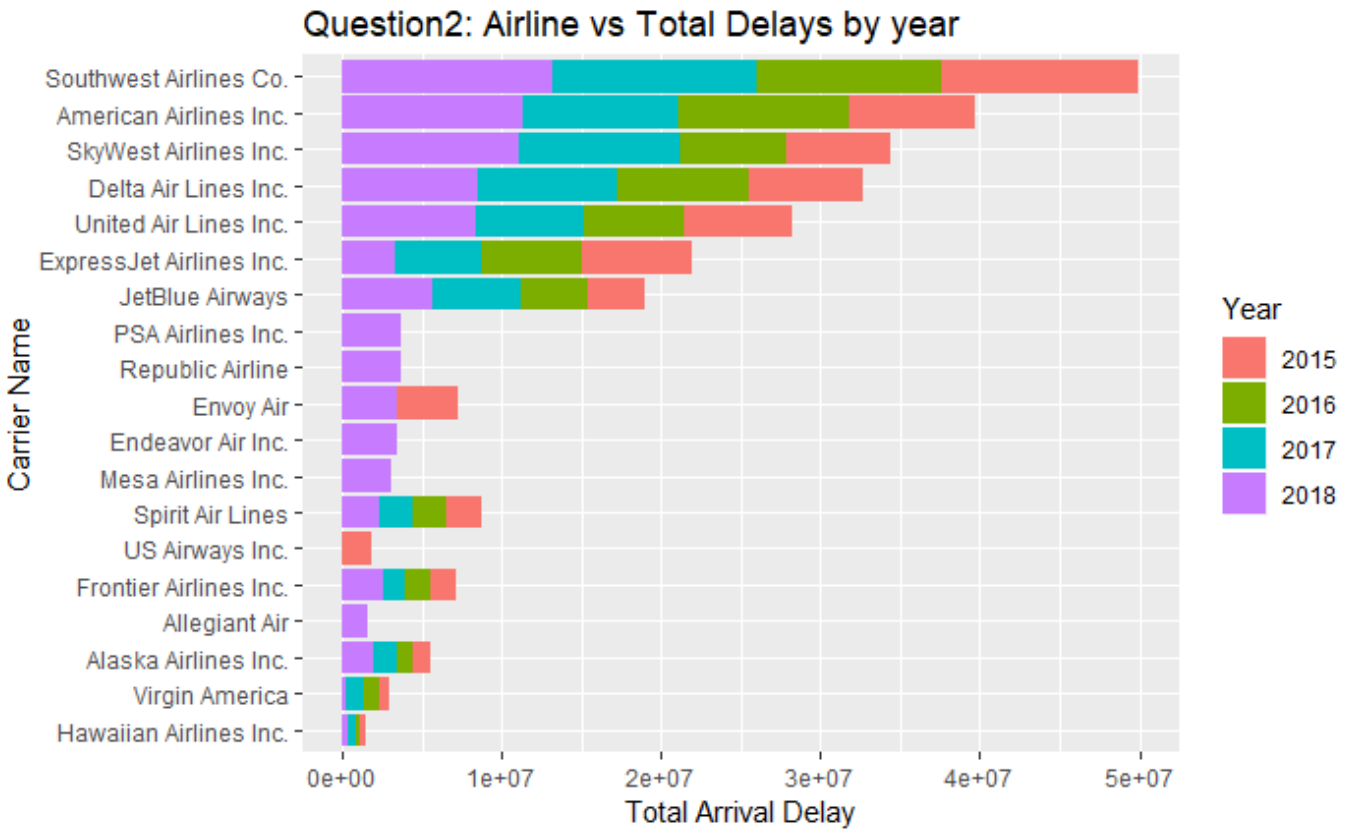
carrier_name <chr>	year <int>	total_arrival_delay <int>
Southwest Airlines Co.	2018	13190774
Southwest Airlines Co.	2017	12858246
Southwest Airlines Co.	2015	12371384
Southwest Airlines Co.	2016	11560098
American Airlines Inc.	2018	11336457

carrier_name <chr>	year <int>	total_arrival_delay <int>
SkyWest Airlines Inc.	2018	11077383
American Airlines Inc.	2016	10709365
SkyWest Airlines Inc.	2017	10080739
American Airlines Inc.	2017	9768952
Delta Air Lines Inc.	2017	8782255

1-10 of 56 rows Previous **1** 2 3 4 5 6 Next

Hide

```
ggplot(df_2, aes(y=total_arrival_delay,x=reorder(carrier_name,total_arrival_delay),
                                                fill=factor(year))) +
  geom_bar(stat="identity")+
  labs(y="Total Arrival Delay",
       x="Carrier Name",
       fill = "Year",
       title="Question2: Airline vs Total Delays by year") +
  coord_flip()
```



Hide

```
#Question 3
df_3 <- air_data2 %>%
  filter((airport == "SFO") | (airport == "ORD") | (airport == "LGA") |
           (airport == "LAX") | (airport == "JFK") | (airport == "EWR")|
           (airport == "DFW") | (airport == "DEN") | (airport == "BOS")|
           (airport == "ATL")) %>%
  select(airport, arr_delay,year) %>%
  group_by(airport,year) %>%
  summarise(total_arrival_delay = sum(arr_delay)) %>%
  arrange(desc(total_arrival_delay))
dim(df_3)
```

```
[1] 40  3
```

Hide

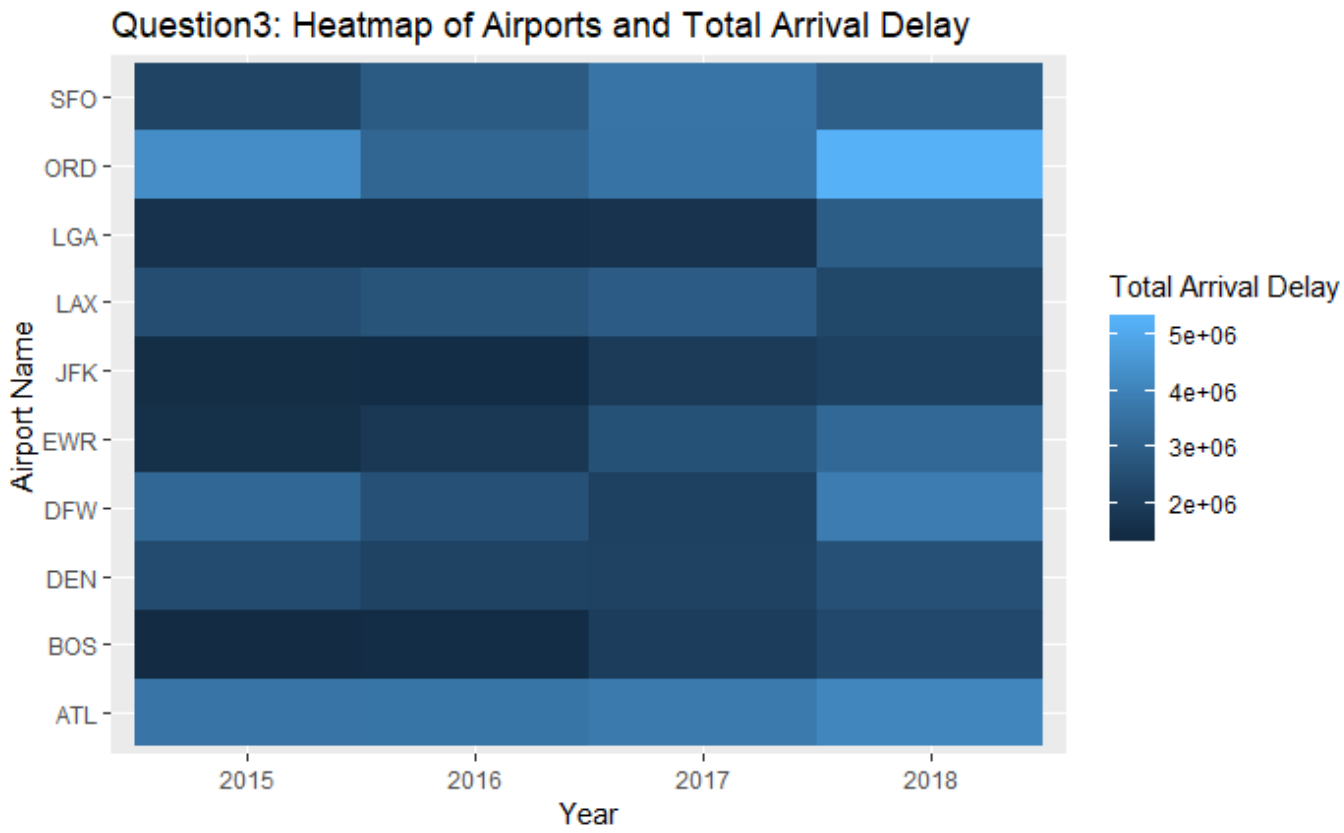
```
head(df_3, 56)
```

airport <chr>	year <int>	total_arrival_delay <int>
ORD	2018	5234855
ORD	2015	4280327
ATL	2018	4125306
DFW	2018	3835082
ATL	2017	3776552
ATL	2016	3648623
SFO	2017	3621342
ATL	2015	3612426
ORD	2017	3599611
EWR	2018	3271729

1-10 of 40 rows Previous **1** 2 3 4 Next

Hide

```
ggplot(df_3, aes(x=factor(year), y=factor(airport), fill=total_arrival_delay)) +
  geom_tile() +
  labs(x="Year",
       y="Airport Name",
       title="Question3: Heatmap of Airports and Total Arrival Delay",
       fill="Total Arrival Delay")
```



Hide

```
#Question 4
df_4 <- air_data2 %>%
  select(year, late_aircraft_delay, carrier_delay) %>%
  group_by(year) %>%
  summarise(carrier_delayy = sum(carrier_delay),
            late_aircraft_delayy = sum(late_aircraft_delay))
head(df_4)
```

year <int>	carrier_delayy <int>	late_aircraft_delayy <int>
2015	20172956	24961931
2016	19533337	23458398
2017	20516702	25905070
2018	26316981	34689058

4 rows

Hide

```
test_4 <- melt(data=df_4, id="year")
head(test_4, 10)
```

	year <int>	variable <fctr>	value <int>
1	2015	carrier_delayy	20172956
2	2016	carrier_delayy	19533337

	year <int>	variable <fctr>	value <int>
3	2017	carrier_delayy	20516702
4	2018	carrier_delayy	26316981
5	2015	late_aircraft_delayy	24961931
6	2016	late_aircraft_delayy	23458398
7	2017	late_aircraft_delayy	25905070
8	2018	late_aircraft_delayy	34689058

8 rows

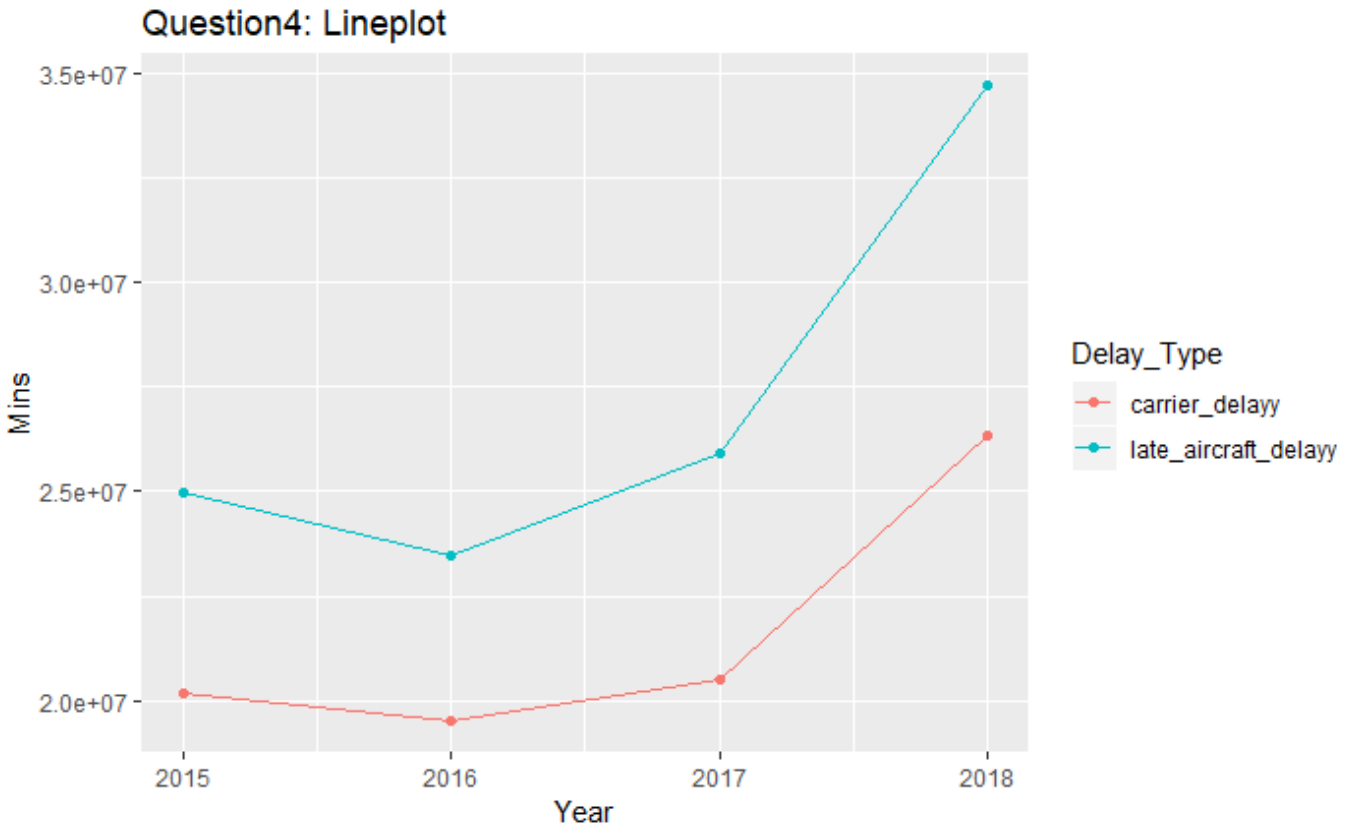
Hide

```
colnames(test_4)
```

```
[1] "year"     "variable" "value"
```

Hide

```
ggplot(test_4, aes(x=year, y= value, color=variable))+
  geom_line()+
  geom_point()+
  labs(x="Year",
       y="Mins",
       color="Delay_Type",
       title="Question4: Lineplot")
```



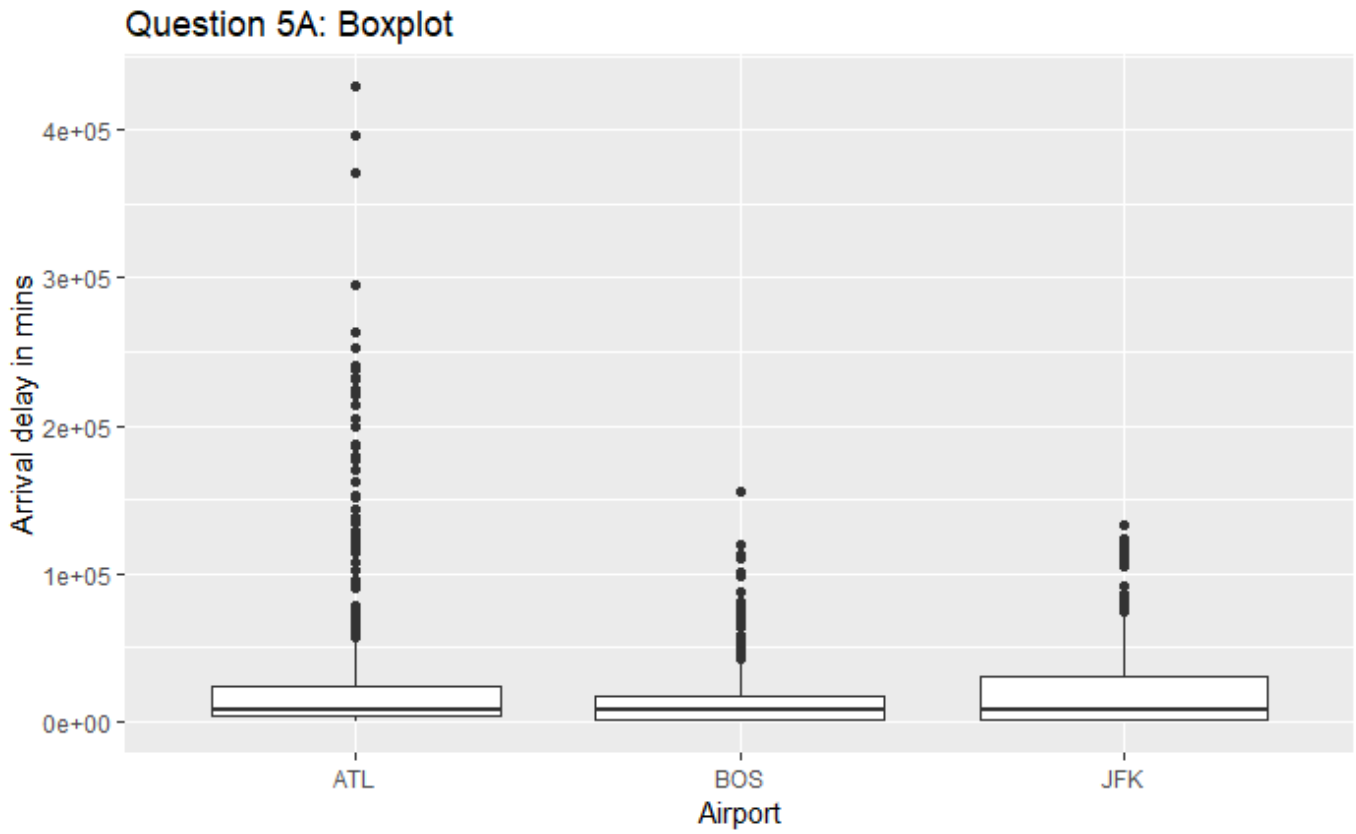
Hide

```
#Question 5
df_5 <- air_data2 %>%
  filter(airport=="ATL" | airport=="BOS" | airport=="JFK")%>%
  select(airport, arr_delay)
dim(df_5)
```

```
[1] 1391    2
```

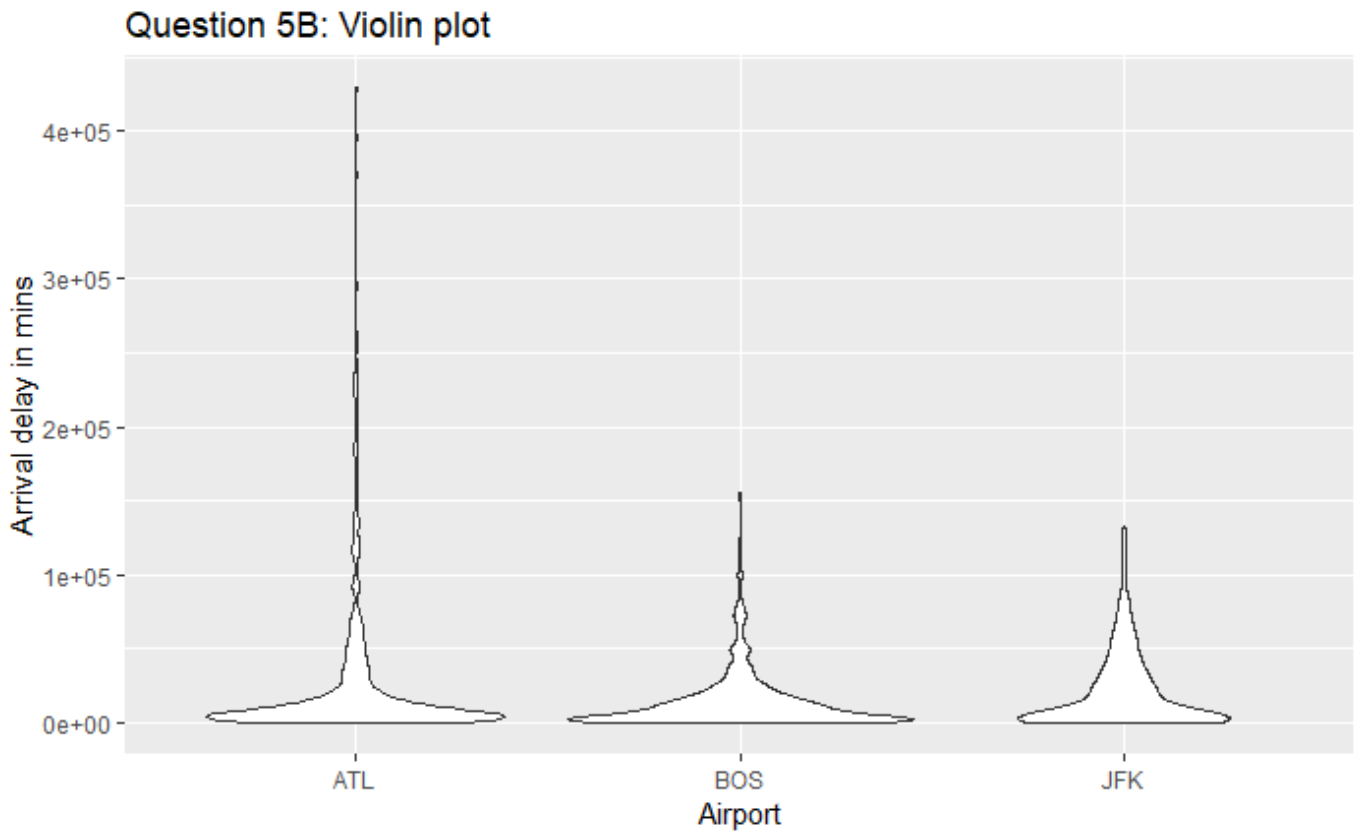
Hide

```
ggplot(df_5, aes(x=airport, y=arr_delay))+
  geom_boxplot()+
  labs(x="Airport",
       y="Arrival delay in mins",
       title="Question 5A: Boxplot")
```



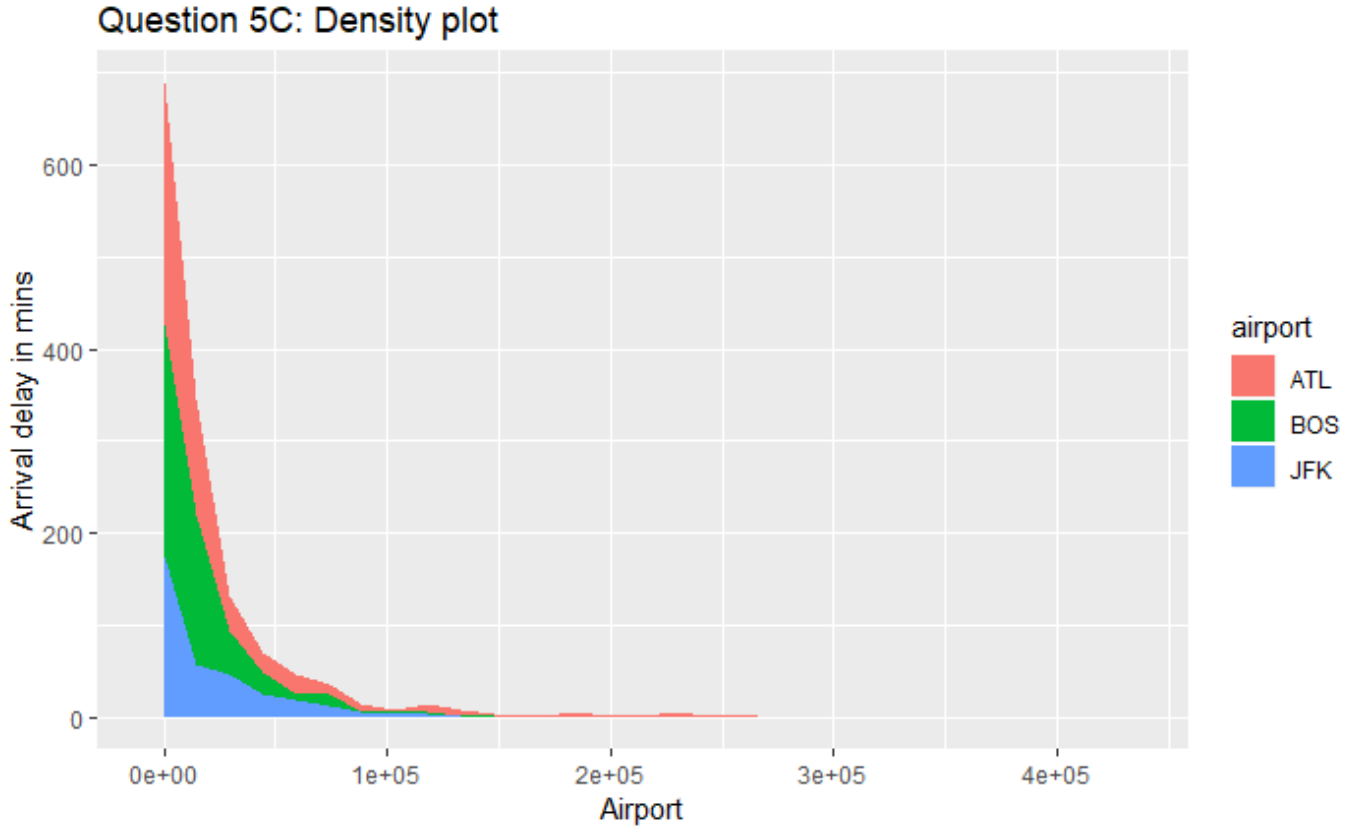
Hide

```
ggplot(df_5, aes(x=airport, y=arr_delay))+
  geom_violin()+
  labs(x="Airport",
       y="Arrival delay in mins",
       title="Question 5B: Violin plot")
```



Hide

```
ggplot(df_5, aes(x=arr_delay, fill=airport))+
  geom_area(stat='bin')+
  labs(x="Airport",
       y="Arrival delay in mins",
       title="Question 5C: Density plot")
```



Hide

```
#Question 6:
#Import data
enplanement_df <- read.csv("enplanement_2017_csv.csv")
summary(enplanement_df)
```

```
   ï..Rank            RO            ST          Locid            City
 Min.   :  1.0   AL     : 84   AK     : 84   0AK    :  1           :  6
 1st Qu.:128.5   SO     : 79   CA     : 27   125    :  1   Columbus :  4
 Median :256.0   GL     : 71   TX     : 24   16A    :  1   Anchorage:  3
 Mean   :265.2   NM     : 63   FL     : 21   1G4    :  1   Jackson  :  3
 3rd Qu.:401.5   WP     : 57   NY     : 18   255    :  1   Albany   :  2
 Max.   :555.0   EA     : 53   MI     : 16   2A3    :  1   Atlanta  :  2
 NA's   :6       (Other):110   (Other):327   (Other):511   (Other)  :497
                  Airport.Name S.L        Hub      CY.17.Enplanements
 Tri-Cities             :  2     :  6       :  6           :  6
 Aberdeen Regional      :  1   CS:125   L   : 30   12,735  :  2
 Abilene Regional       :  1   P :386   M   : 31   1,00,133:  1
 Abraham Lincoln Capital:  1            N   :255   1,02,988:  1
 Adirondack Regional    :  1            None:125   1,03,547:  1
 Akiachak               :  1            S   : 70   1,03,569:  1
 (Other)                :510                       (Other) :505
 CY.16.Enplanements   X..Change
         :  6       9.50%  :  4
 16,822  :  2       -0.21% :  3
 45,300  :  2       3.06%  :  3
 5,442   :  2       4.03%  :  3
 1,00,433:  1              :  2
 1,01,115:  1       -0.29% :  2
 (Other) :503       (Other):500
```

Hide

```
#Preparing data
#Remove df records
enplanement_df2 <- na.omit(enplanement_df)
#Check dimensions of before and after removing NAs records
dim(enplanement_df)
```

```
[1] 517  11
```

Hide

```
dim(enplanement_df2)
```

```
[1] 511  11
```

Hide

```
#Change column name of dpwnloaded table
colnames(enplanement_df2)[which(names(enplanement_df2) == "Locid")] <- "airport"
#Change from factor to character
enplanement_df2$airport <- as.character(enplanement_df2$airport)
merged_df <- merge(air_data2, enplanement_df2,
                   by.x="airport", by.y = "airport")
df_6 <- merged_df %>%
  select(airport, CY.17.Enplanements, arr_delay) %>%
  group_by(airport, CY.17.Enplanements) %>%
  summarise(sum(arr_delayy = arr_delay))
summary(df_6)
```

```
   airport          CY.17.Enplanements sum(arr_delayy = arr_delay)
 Length:347         1,00,133:  1       Min.   :     119
 Class :character   1,02,988:  1       1st Qu.:   31548
 Mode  :character   1,03,547:  1       Median :  110307
                    1,03,569:  1       Mean   :  793925
                    1,03,679:  1       3rd Qu.:  405290
                    1,03,724:  1       Max.   :16340155
                    (Other) :341
```

Hide

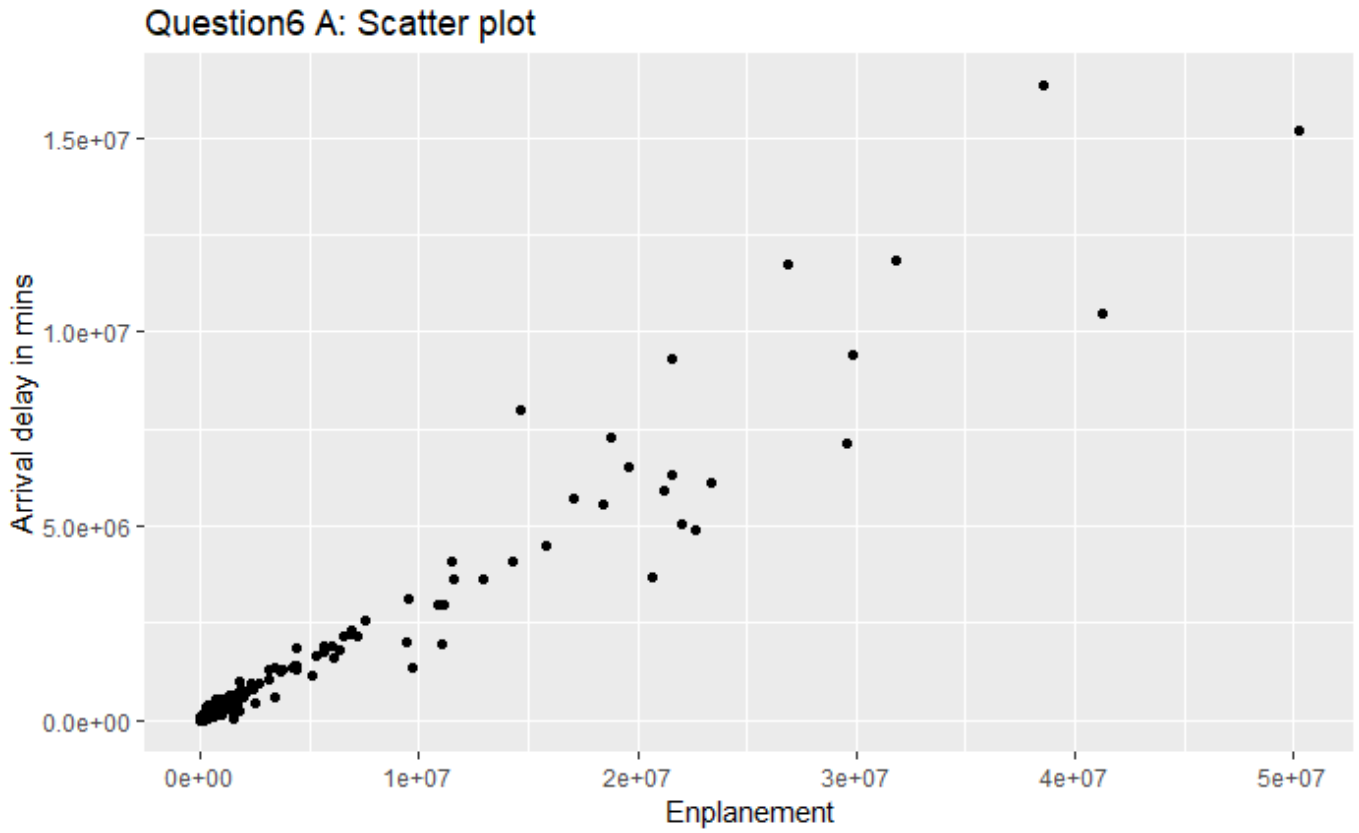
```
#Rename columns
colnames(df_6) <-c("airport", "Enplanement","Arrival_Delay")
#remove comma
df_6$Enplanement <- as.numeric(gsub(",","",df_6$Enplanement))
head(df_6,20)
```

airport <chr>	Enplanement <dbl>	Arrival_Delay <int>
ABE	328914	128836
ABI	85085	45985
ABQ	2412328	829002
ABR	27635	25727
ABY	37920	49994
ACK	113009	37920
ACT	58888	56762
ACV	65932	84935
ACY	552690	217012
ADQ	83577	10718

1-10 of 20 rows Previous **1** 2 Next

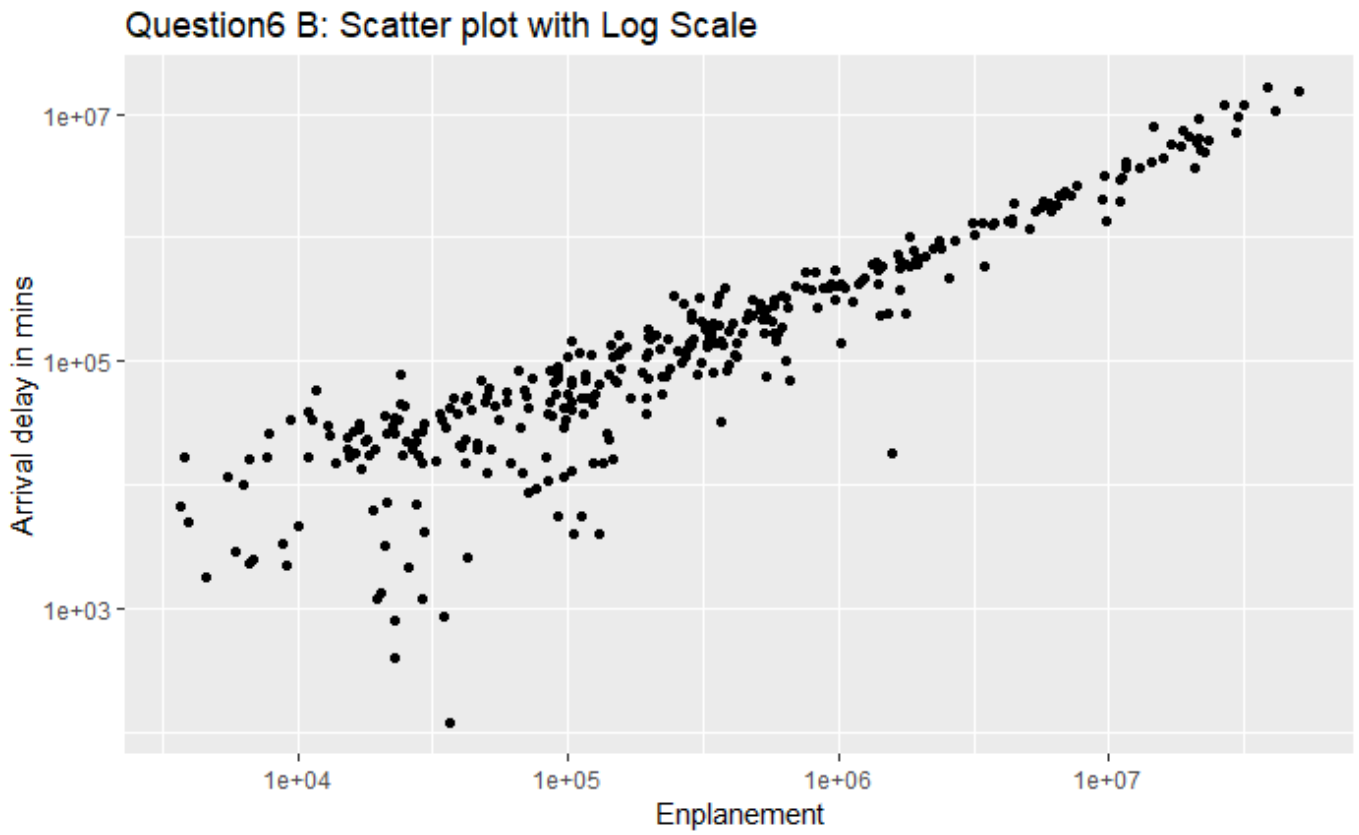
Hide

```
#Question 6 (continued)
#Scatter plot
ggplot(df_6, aes(x=Enplanement, y=Arrival_Delay))+
  geom_point()+
  labs(x="Enplanement",
       y="Arrival delay in mins",
       title="Question6 A: Scatter plot")
```



Hide

```
#Scatter Plot on log scale
ggplot(df_6, aes(x=Enplanement, y=Arrival_Delay))+
  geom_point()+
  scale_x_continuous(trans='log10')+
  scale_y_continuous(trans='log10')+
  labs(x="Enplanement",
       y="Arrival delay in mins",
       title="Question6 B: Scatter plot with Log Scale ")
```



Hide

```
#Remove all free dataframes from working memory in the end
rm(enplanement_df, enplanement_df2, merged_df,air_data)
```