CSE – 3020

Data Visualization

Lab DA – 2

Name	:	Anish Desai
Reg. No.	:	20BCE0461
Slot	:	L39 + L40
Guided by	:	Prof. Jyotismita Chaki

VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

Loading Required Libraries and Dataset:

Data Viz Lab DA - 2

Multivariate Analysis and PCA

Anish Desai
20BCE0461

```
# Load Libraries
library(nycflights13)
library(ggplot2)
library(tidyverse)
# Loading the dataset
data(flights)
# Viewing the dataset
View(flights)
# Storing thedataset in another variable
data_flights <- flights
```

```
> #Load Libraries
> library(nycflights13)
> library(ggplot2)
> library(tidyverse)
-- Attaching packages --------------------------------------------------------- tidyverse 1.3.1 --
v tibble  3.1.6     v dplyr   1.0.7
v tidyr   1.1.4     v stringr 1.4.0
v readr   2.1.1     v forcats 0.5.1
v purrr   0.3.4
-- Conflicts ------------------------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
> #Loading the dataset
> data(flights)
> #Viewing the dataset
> View(flights)
> #Storing the dataset in another variable
> data_flights <- flights
```

Code Execution

LabDA2.R × flights ×

Filter

	year	month	day	dep_time	sched_dep_time	dep_delay	arr_time	sched_arr_time	arr_delay	carrier	fligl
439	2013	1	1	1456	1500	-4	1649	1632	17	UA	
440	2013	1	1	1456	1455	1	1830	1813	17	UA	
441	2013	1	1	1457	1500	-3	1758	1815	-17	UA	
442	2013	1	1	1457	1500	-3	1652	1656	-4	US	
443	2013	1	1	1458	1500	-2	1658	1655	3	MQ	
444	2013	1	1	1459	1501	-2	1651	1651	0	EV	
445	2013	1	1	1459	1454	5	1750	1751	-1	UA	
446	2013	1	1	1500	1459	1	1809	1806	3	B6	
447	2013	1	1	1502	1500	2	1802	1806	-4	UA	
448	2013	1	1	1505	1310	115	1638	1431	127	EV	
449	2013	1	1	1505	1510	-5	1654	1655	-1	MQ	
450	2013	1	1	1506	1505	1	1838	1820	18	AA	
451	2013	1	1	1506	1512	-6	1723	1741	-18	UA	
452	2013	1	1	1507	1515	-8	1651	1656	-5	9E	
453	2013	1	1	1507	1510	-3	1748	1745	3	MQ	
454	2013	1	1	1508	1450	18	1813	1747	26	UA	
455	2013	1	1	1510	1517	-7	1811	1811	0	B6	

Showing 438 to 455 of 336,776 entries, 19 total columns

Dataset

Question 1 :

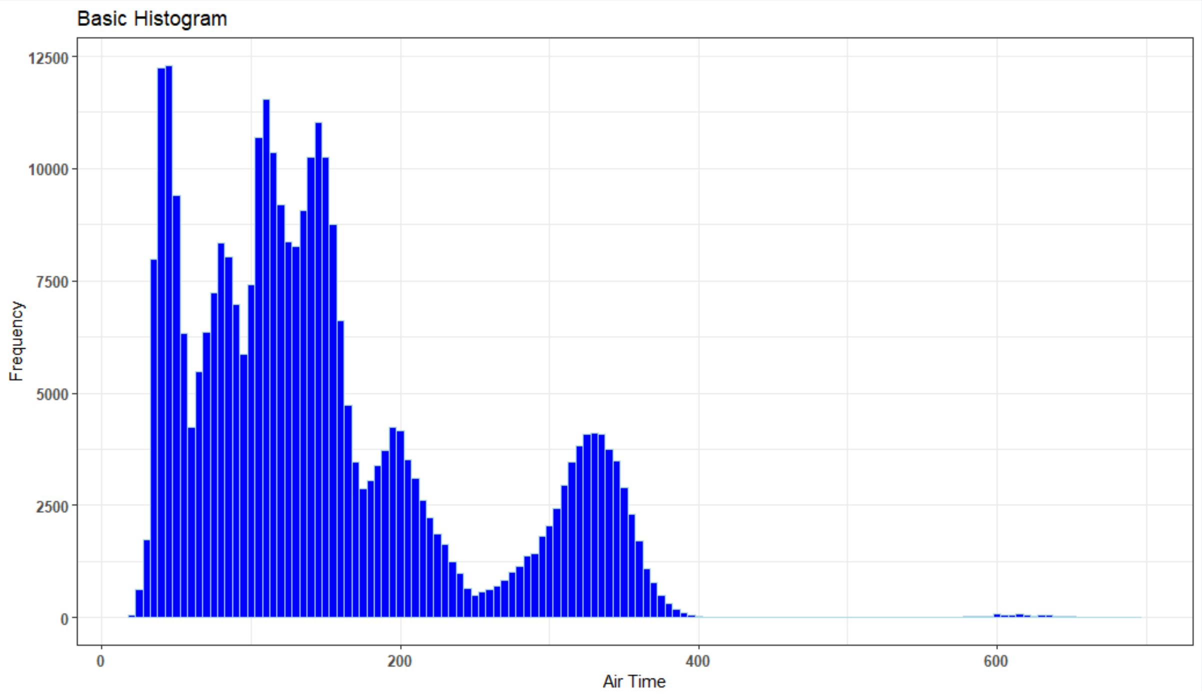
Perform six different types of multivariate analysis. Write the respective interpretations.

Multi-variate Analysis (MVA)

MVA – 1 :

```
#Q1
#MVA-1
#Histogram
ggplot(data_flights)+
  geom_histogram(aes(x = air_time), fill = 'blue',
                 color = "lightblue", binwidth = 5)+
  ggtitle("Basic Histogram") +
  xlab("Air Time") +
  ylab("Frequency") +
  theme_bw() +
  theme(axis.text.x = element_text(face = 'bold', size=10),
        axis.text.y = element_text(face='bold', size=10))
```

```
> #MVA-1
> #histogram
> ggplot(data_flights) +
+   geom_histogram(aes(x = air_time), fill = 'blue',
+                  color = "lightblue", binwidth = 5)+
+   ggtitle("Basic Histogram") +
+   xlab("Air Time") +
+   ylab("Frequency") +
+   theme_bw() +
+   theme(axis.text.x = element_text(face = 'bold', size = 10),
+         axis.text.y = element_text(face = 'bold', size = 10))
```



Interpretation #MVA-1.

Using Histogram plot, we have found out the frequency of Air Time of the flights.
From the plot, we can infer that :

1. The Air Time 45-50 mins has the highest number of observations, closely followed by 40-45 mins.
   Both have nearly 12300 observations each.

2. Most of the flights have Air Time in the range of either 30 - 400 mins or very few having in the range 575 - 650 mins.

3. There are hardly any flights which have an air time in the range 400 ~ 575 mins.

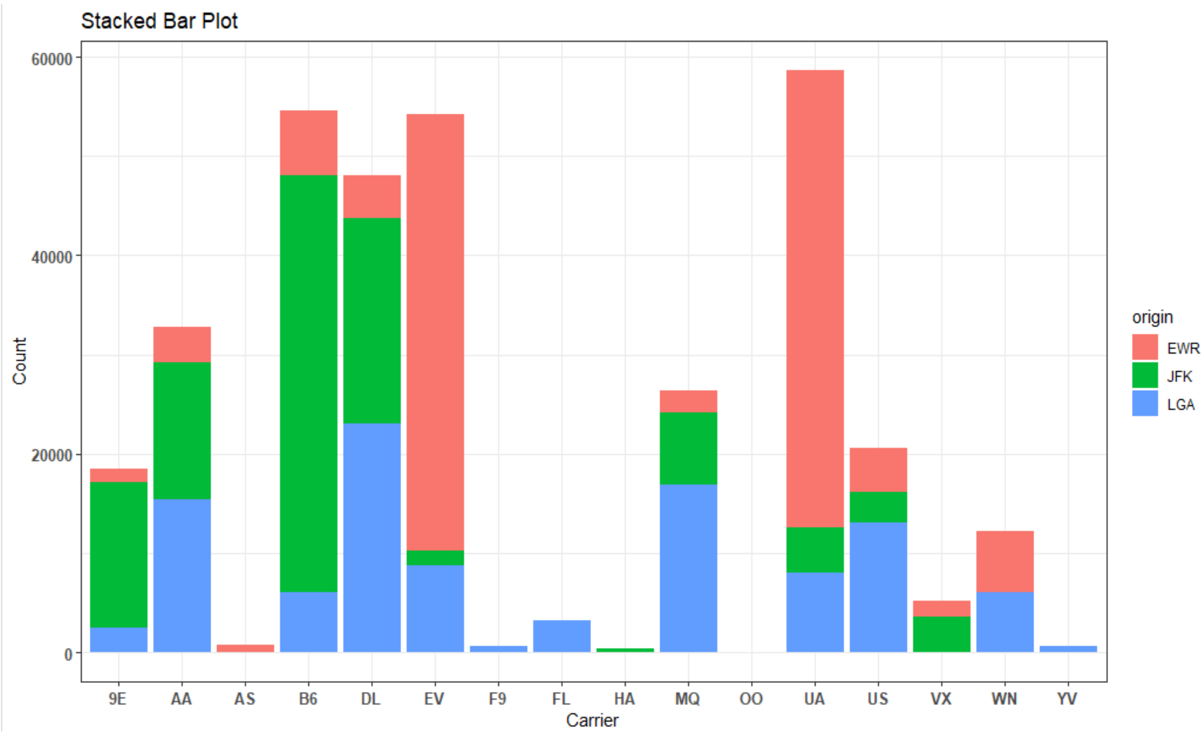
4. Most of the flights are concentrated in the range of 30 - 200 mins.

[Air Time of the flight ≈ Travelling Time]

MVA – 2 :

```
#MVA-2
#Stacked Bar Plot
data_flights %>%
    ggplot(aes(x= carrier, fill = origin))+
    geom_bar() +
    ggtitle("Stacked Bar Plot") +
    xlab("Carrier") +
    ylab("Count") +
    theme_bw() +
   theme(axis.text.x = element_text(face = 'bold', size=10),
         axis.text.y= element_text(face= 'bold', size = 10))
```

```
> #MVA-2
> #Stacked Bar Plot
> data_flights %>%
+   ggplot(aes(x = carrier, fill = origin)) +
+   geom_bar() +
+   ggtitle("Stacked Bar Plot") +
+   xlab("Carrier") +
+   ylab("Count") +
+   theme_bw() +
+   theme(axis.text.x = element_text(face = 'bold', size = 10),
+         axis.text.y = element_text(face = 'bold', size = 10))
```



Interpretation #MVA-2

Using Stacked Bar Plot, we have plotted 'Carrier' against Count, filled with 'Origin' as the parameter.

From the plot, we can infer that :

1. The Carrier with most number of flights is the carrier UA with almost 59000~60000 flights in the year 2013.

   The Carrier OO has so less number of flights that it is negligibly shown in the graph and hardly noticeable

2. Most of the flights from the year 2013 for the carriers AS, EV, UA and WN originated in EWR.

The origin of most of the flights (highest) for the carriers 9E, B6, HA, VX is JFK

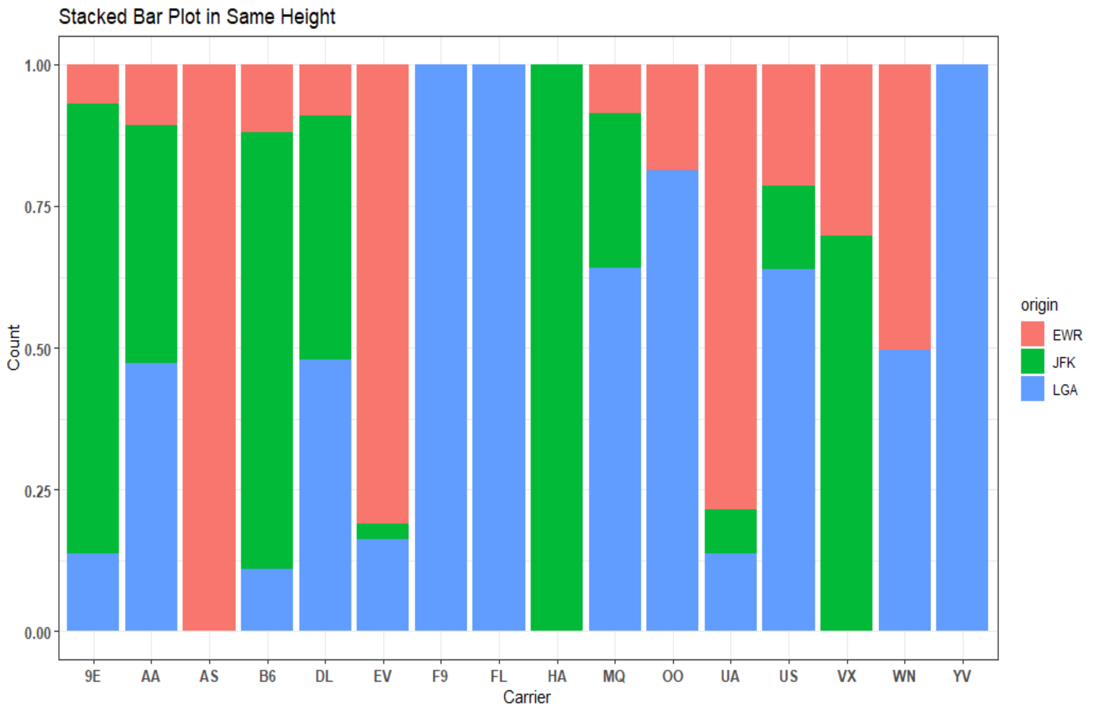
The rest of the carriers have their maximum number of flights originated from LGA.

The graph for OO is negligible, thus we cannot estimate the origin of the flights.

MVA – 3 :

```
# MVA-3
# Stacked Bar Plot in same height
ggplot(data_flights) +
    geom_bar(aes(x = carrier, fill = origin),
                                  position = 'fill') +
    ggtitle("Stacked Bar Plot in Same Height") +
    xlab("Carrier") +
    ylab("Count") +
    theme_bw() +
    theme(axis.text.x = element_text(face = 'bold', size = 10),
          axis.text.y = element_text(face = 'bold', size = 10))
```

```
> #MVA-3
> #Stacked Bar Plot in same height
> ggplot(data_flights) +
+   geom_bar(aes(x = carrier, fill = origin), position = 'fill') +
+   ggtitle("Stacked Bar Plot in Same Height") +
+   xlab("Carrier") +
+   ylab("Count") +
+   theme_bw() +
+   theme(axis.text.x = element_text(face = 'bold', size = 10),
+         axis.text.y = element_text(face = 'bold', size = 10))
```



Interpretation #MVA-3

In this, we have plotted a variation of Stacked Bar Plot which normalizes the count on Y-axis in between 0 and 1, thus called Stacked Bar Plot in same height.

From the plot, we can infer that :

1. Almost all the flights of the carriers F9, FL and YV originated at LGA. Similarly, all the flights of the carrier AS originated at EWR, while JFK was the origin of all the flights of the carrier HA.

2. The most useful feature of this graph is that, when the observations for any parameter is negligible, it doesn't get plotted in Stacked Bar plot, whereas in this, we can clearly see the trend of observations irrespective of its number

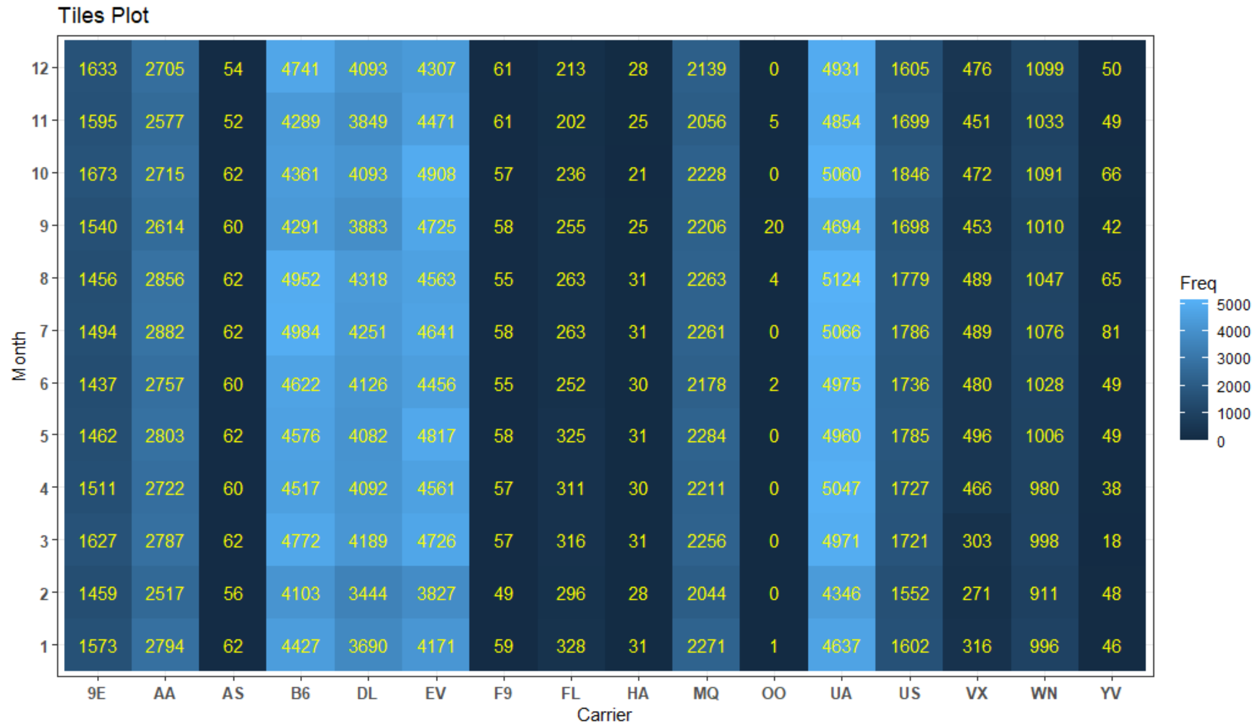
For example, in the previous plot, carrier OO is negligible. In this, we can see that out of all the flights of OO, irrespective of its number, majority (~80%) originated at LGA and the rest at EWR.

For the rest of the carriers too, it shows the proportion of each of the origins of the flights.

MVA – 4 :

```
# MVA-4
# Tiles Plot
ggplot (as.data.frame(table (data_flights $carrier,
                              data_flights $ month))) +
  geom_tile (aes (x= var1, y= var2, fill = Freq)) +
  geom_text(aes(x=var1, y=var 2, label= Freq),
                  color = "yellow") +
  ggtitle ("Tiles Plot") +
  xlab ("Carrier") +
  ylab ("Month") +
  theme_bw() +
  theme(axis.text.x = element_text (face='bold', size =10),
        axis.text.y = element_text(face='bold', size =10))
```

```
> #MVA-4
> #Tiles plot
> ggplot(as.data.frame(table(data_flights$carrier,
+                          data_flights$month))) +
+   geom_tile(aes(x = Var1, y = Var2, fill = Freq)) +
+   geom_text(aes(x = Var1, y = Var2, label = Freq),
+             color = "yellow") +
+   ggtitle("Tiles Plot") +
+   xlab("Carrier") +
+   ylab("Month") +
+   theme_bw() +
+   theme(axis.text.x = element_text(face = 'bold', size = 10),
+         axis.text.y = element_text(face = 'bold', size = 10))
```



Interpretation - # MVA-4

Using Tiles Plot, we have plotted Carrier against Month.

Using the plot, we can infer that:

1. Obtain the number of flights of each carrier per month of 2013.

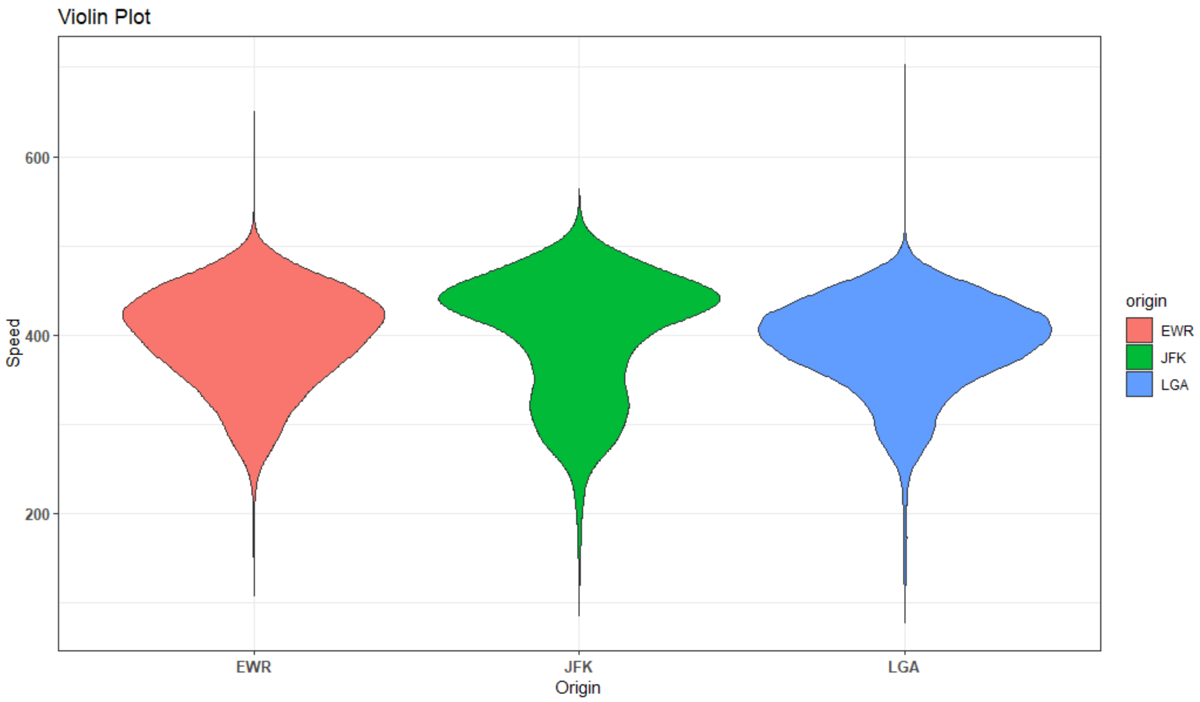
2. The highest number of flights in a month is of carrier. UA in the month of August(8) in which it had 5124 flights.
The least number of flights is of carrier OO in the months Feb, March, Apr, May, July, Oct and December in which it had 0 flights.

3. From the colour scheme, we can notice that the UA has been marked the lightest, implying it had maximum number of flights when compared to other carriers.
On the other hand, the carrier OO column is marked the darkest, implying it had the least number of flights.

MVA – 5 :

```
# MVA - 5
# Violin Plot
data_flights <- data_flights %>%
        mutate(speed = distance / air_time * 60)
ggplot (data_flights) +
   geom_violin (aes(x=origin, y=speed, fill = origin))+
   ggtitle ("Violin Plot") +
   xlab ("Origin") +
   ylab ("Speed") +
   theme_bw() +
  theme(axis.text.x = element_text(face='bold', size=10),
     axis.text.y = element_text(face = 'bold', size = 10))
```

```
> #MVA-5
> #violin plot
> data_flights <- data_flights %>%
+   mutate(speed = distance / air_time * 60)
> ggplot(data_flights) +
+   geom_violin(aes(x = origin , y = speed, fill = origin)) +
+   ggtitle("Violin Plot") +
+   xlab("Origin") +
+   ylab("Speed") +
+   theme_bw() +
+   theme(axis.text.x = element_text(face = 'bold', size = 10),
+         axis.text.y = element_text(face = 'bold', size = 10))
```



Interpretation # MVA-5

From the available features of the dataset, I have obtained a new feature 'Speed' using distance and air_time.

In this plot, we can see the speed comparisons of the flights originating at the airports EWR, JFK and LGA.

From the plot, we can infer that :

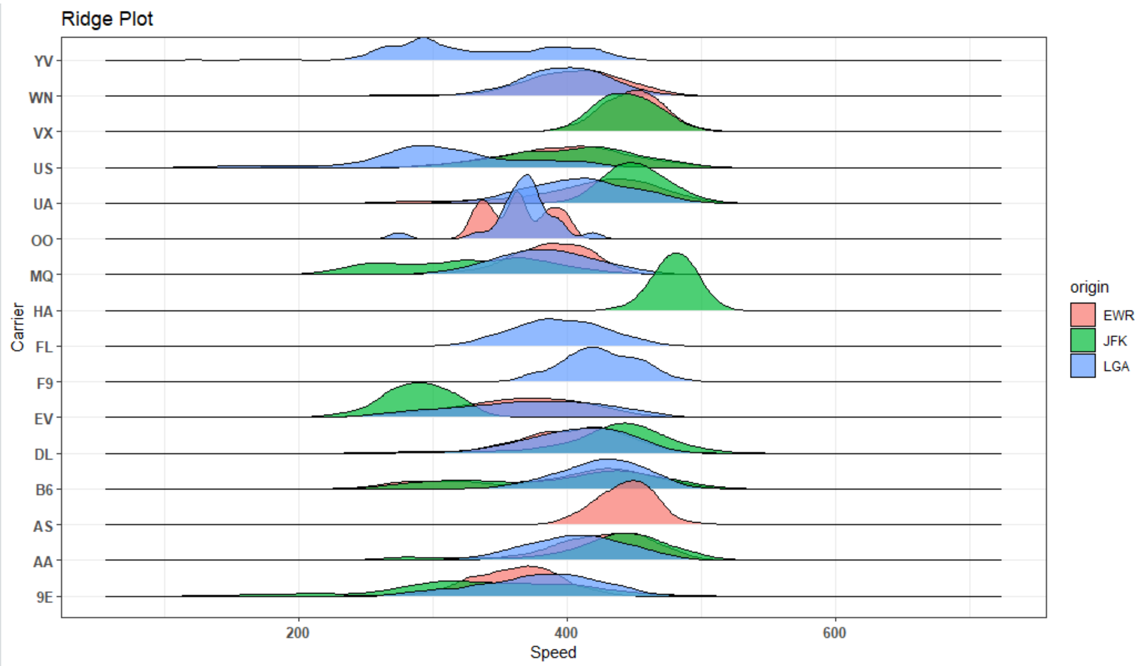
1. The speed of maximum flights originating at EWR is in the range 400 - 450, while for JFK flights, it is in the range of around 450s and the maximum number of flights from LGA had speed in the range of 400s.

2. The tips of LGA in the violin plot extends beyond the other graphs' tips implying the range of speed of flights originating at LGA is more wide as compared to flights of other origins (ranging from as low as 100 to as high as 700).

MVA – 6 :

```
# MVA-6
# Ridge Plot
library(ggridges)
ggplot(data_flights) +
    geom_density_ridges(aes(x=speed , y= carrier,
                        fill = origin), alpha = 0.7) +
    ggtitle("Ridge Plot") +
    xlab("Speed") +
    ylab(" Carrier") +
   theme_bw() +
  theme(axis.text.x = element_text(face='bold', size=10)
    axis.text.y = element_text(face='bold', size=10))
```

```
> #MVA-6
> #Ridge Plot
> library(ggridges)
> ggplot(data_flights) +
+   geom_density_ridges(aes(x = speed , y = carrier,
+                           fill = origin), alpha = 0.7) +
+   ggtitle("Ridge Plot") +
+   xlab("Speed") +
+   ylab("Carrier") +
+   theme_bw() +
+   theme(axis.text.x = element_text(face = 'bold', size = 10),
+         axis.text.y = element_text(face = 'bold', size = 10))
Picking joint bandwidth of 6.77
```



Interpretation #MVA-6

In the Ridge plot, we have used 3 parameters : Speed on X axis, Carrier on y axis and Origin as fill.

In the previous Violin plot, we could only see the speed distribution of all the flights across all origins without any info on the different types of carriers. In this Ridge plot, we broke down the Speed vs Origin comparison Carrier-wise.

We can infer that :

1. In carrier YV, maximum flights had the speed of around 300, which originates at LGA.
In case of carrier WN, the peak is at 400 and blue in colour, implying maximum no. of flights had the speed of around 400 and these flights originated at LGA.
Similarly, for all other carriers

2. The peak of the carrier HA occurs the most rightward, implying the maximum number of flights of HA had the speed which is more than the peak speed of other carriers.

The peak gives us an idea about the speed possessed by maximum no. of flights from a specific origin

Thus, 6 different types of Multi-variate analysis (MVA) have been performed.

The codes, output and interpretation of respective output plots have been included.

---------------------------------------End of Qsn 1---
---------------------------------------Start of Qsn 2---

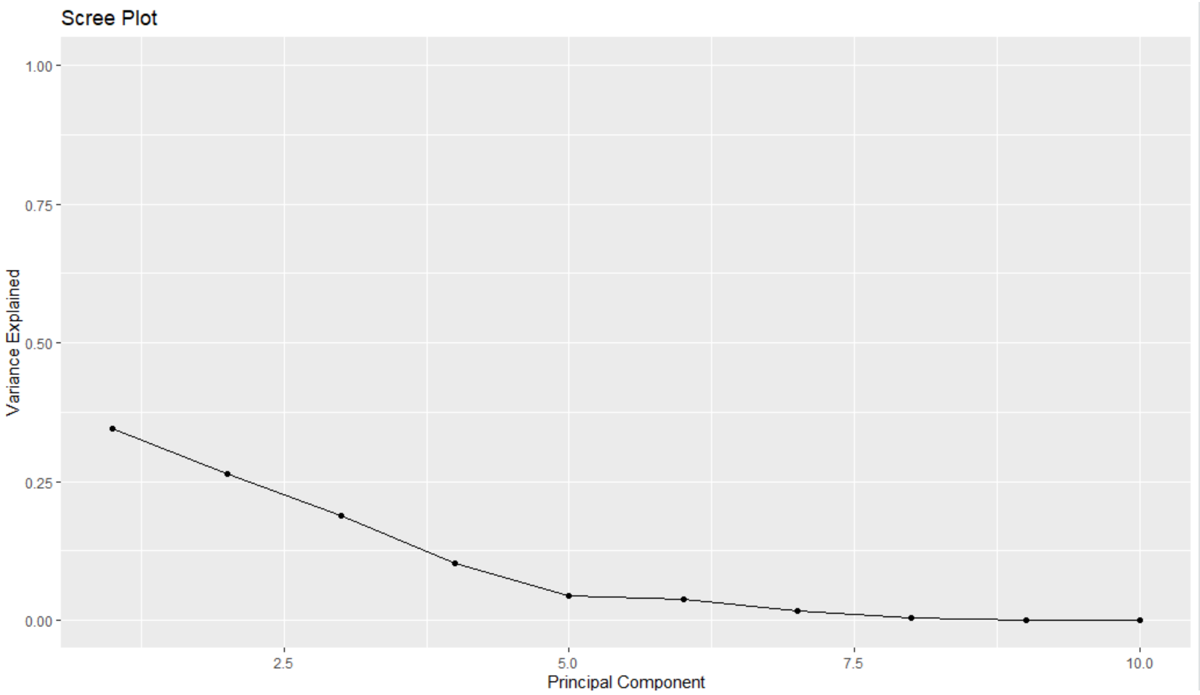
Question 2 :

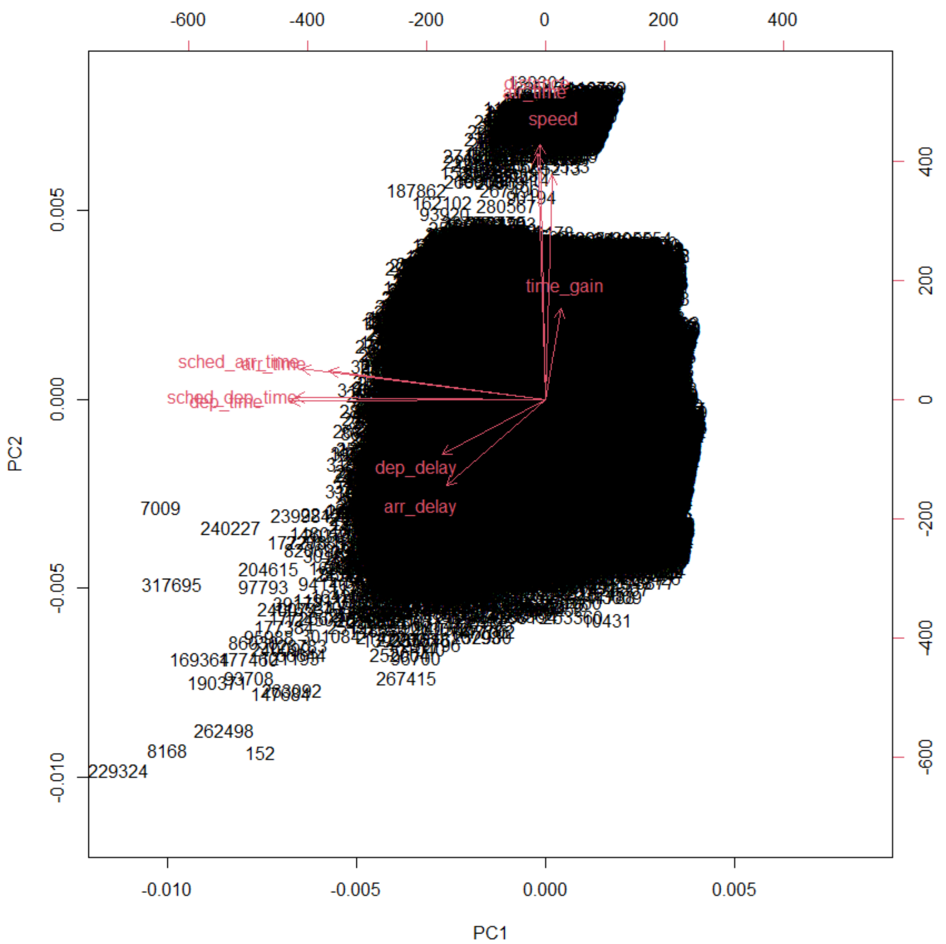
Perform and plot PCA using at least 8 features (think before you select the effective 8 features so that you can clearly interpret the data) and write the proper interpretation.

```
# Q2
# Adding Space and Time gain
data_flights <- data_flights %>%
        mutate(time_gain = dep_delay - arr_delay,
               speed = distance / air_time * 60)

# Extracting certain features of the dataset
df <- data_flights %>% select(air_time,
             dep_time, arr_time, sched_dep_time,
             sched_arr_time, dep_delay, arr_delay,
             distance, speed, time_gain)

# View the new dataset
View(df)
df <- na.omit(df) # Omit NA
# Re-numbering the rows
rownames(df) <- NULL
# Obtaining Principal Components
pca.fit <- prcomp(df, scale. = TRUE)

# Storing the variance result
var_explained = pca.fit$dev^2 / sum(pca.fit$dev^2)
```

```
# Create scree plot
# Ten features used
qplot(c(1:10), var_explained) +
      geom_line() +
      xlab("Principal Component") +
      ylab("Variance explained") +
      ggtitle("Scree Plot") +
      ylim(0, 1)

# Plot using PC1 and PC2
biplot(pca.fit, choices = c(1,2))
```

```
> #Q2
> #Adding Speed and Time gain feature
> data_flights <- data_flights %>%
+   mutate(time_gain = dep_delay - arr_delay ,
+          speed = distance / air_time * 60)
> #Extracting certain features of the dataset
> df <- data_flights %>% select(air_time,dep_time,arr_time,
+                               sched_dep_time,sched_arr_time,dep_delay,
+                               arr_delay,distance,speed,time_gain)
> #View the new dataset
> View(df)
> df <- na.omit(df) # Omit NA
> #R does not automatically re-number the rows when we drop those with NA values,
> #we can force re-numbering
> rownames(df) <- NULL
> #Obtaining Principal components
> pca.fit <- prcomp(df, scale. = TRUE)
> #Calculate total variance explained by each principal component
> #sdev : standard deviation
> pca.fit$sdev^2 / sum(pca.fit$sdev^2)
 [1] 3.444389e-01 2.635184e-01 1.883227e-01 1.016416e-01 4.309536e-02 3.709071e-02 1.722721e-02
 [8] 4.227611e-03 4.374781e-04 4.036456e-29
> #Storing the variance result
> var_explained = pca.fit$sdev^2 / sum(pca.fit$sdev^2)
> #Create scree plot
> #Ten columns used
> qplot(c(1:10), var_explained) +
+   geom_line() +
+   xlab("Principal Component") +
+   ylab("Variance Explained") +
+   ggtitle("Scree Plot") +
+   ylim(0, 1)
> #Plot using PC1 and PC2
> biplot(pca.fit,choices=c(1,2))
```


Scree Plot
Variance Explained
1.00
0.75
0.50
0.25
0.00
2.5
5.0
7.5
10.0
Principal Component


-600
-400
-200
0
200
400
speed
time_gain
sched_arr_time
arr_time
sched_dep_time
dep_time
dep_delay
arr_delay
PC1
PC2
-0.010
-0.005
0.000
0.005

To obtain the eigen-vector values for each feature, use the command

pca.fit$rotation

The rotation column of pca.fit (in which the principal components are stored) gives the values of eigen vectors for each feature principal component-wise.

```
> pca.fit$rotation
                       PC1          PC2         PC3         PC4          PC5         PC6
air_time       -0.01520842  0.555153865  0.21071427 -0.22682662  0.013116985 -0.36339438
dep_time       -0.50518001 -0.004035085 -0.08141312  0.03646922 -0.424098450 -0.05258409
arr_time       -0.42879796  0.064386778 -0.21141127 -0.14707618  0.675542914  0.13197639
sched_dep_time -0.49520833  0.004524592 -0.11965147  0.02104310 -0.490718217 -0.05040042
sched_arr_time -0.48346037  0.069024385 -0.13625292 -0.04759473  0.260371337  0.02362020
dep_delay      -0.20468137 -0.123466315  0.57962703  0.41056068  0.142159635 -0.08333261
arr_delay      -0.19634898 -0.193676963  0.63670291  0.03633712  0.076281937  0.05182487
distance       -0.01166887  0.575417612  0.19602178 -0.15523580 -0.007961166 -0.27291777
speed           0.01310841  0.506952803  0.09337631  0.23154246 -0.113134679  0.81390016
time_gain       0.03120221  0.204929597 -0.28791245  0.82176011  0.126970419 -0.31323447
                        PC7          PC8          PC9          PC10
air_time       -0.018085652 -0.007853467  0.680518750  0.000000e+00
dep_time       -0.198577083 -0.717457009 -0.004096274 -1.674881e-14
arr_time       -0.518756351  0.052672875 -0.001394991  9.951665e-15
sched_dep_time -0.128876451  0.692651211  0.016417824  1.369028e-14
sched_arr_time  0.819344202 -0.029279694 -0.011756549 -7.431903e-15
dep_delay      -0.031726230  0.030735483  0.005790594 -6.396840e-01
arr_delay      -0.016481239  0.024643851 -0.005375406  7.126098e-01
distance       -0.031391116  0.012401620 -0.728387964  2.080293e-15
speed           0.007307328 -0.006117728  0.071918179 -2.097711e-16
time_gain      -0.029679212  0.007287997  0.026154699  2.880827e-01
```

Interpretation #82

PCA helps to capture most of the variance in the data.

We have reduced the dataset to 10 features so that we can use some meaningful strictly numeric data.

Then, we obtain the Principal components. The number of PCs obtained is equal to the number of features.

The variance explained by each PC is obtained.

The maximum variance is explained by PC1 and PC2. This can be visualized using Scree Plot.

Thus, we plot between PC1 and PC2. We get two clusters and few outliers.

To interpret each PC, examine the magnitude and direction of the features. The larger the absolute value of eigen vector, the more important the corresponding feature is in calculating the PC.

The first PC has large negative associations with dep_time, sched_dep_time, arr_time and sched_arr_time, thus it is mainly concerned with the Arrival and Departure of flights.

The second PC is mainly concerned with air_time, distance and speed since it has large positive associations with them.

The third PC is concerned with arr_delay and dep_delay, thus focussing on delays by virtue of its large positive associations with them. And so on.

Principal Component Analysis (using 10 features) and Interpretation has been done.

--Thank you---