NAME – SAARA ANAND

REG NO – 21BCE8156

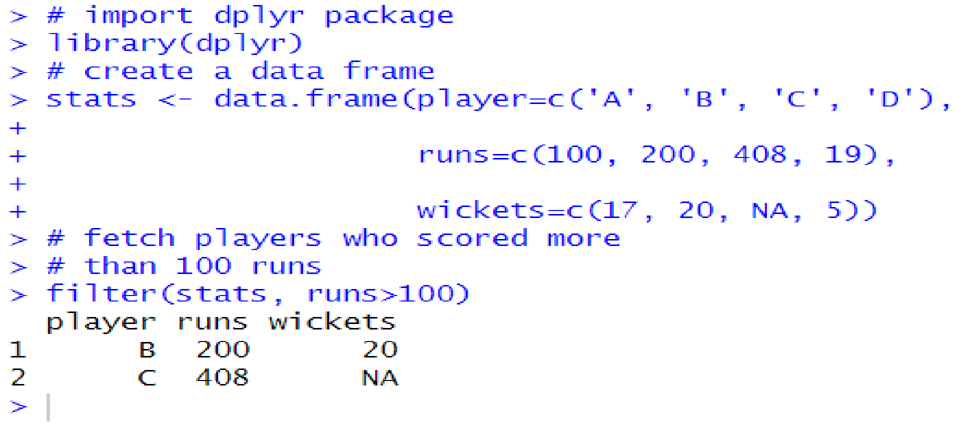
SLOT – L55+L56

FDA LAB ASSIGNMENT 6-

Install the dplyr package and perform the following operations:

filter() method:

```
# import dplyr package
library(dplyr)
# create a data frame
stats <- data.frame(player=c('A', 'B', 'C', 'D'),

                    runs=c(100, 200, 408, 19),

                    wickets=c(17, 20, NA, 5))
# fetch players who scored more
# than 100 runs
filter(stats, runs>100)
```



```
> # import dplyr package
> library(dplyr)
> # create a data frame
> stats <- data.frame(player=c('A', 'B', 'C', 'D'),
+ 
+                     runs=c(100, 200, 408, 19),
+ 
+                     wickets=c(17, 20, NA, 5))
> # fetch players who scored more
> # than 100 runs
> filter(stats, runs>100)
  player runs wickets
1      B  200      20
2      C  408      NA
> 
```

distinct() method:

import dplyr package

library(dplyr)

create a data frame

stats <- data.frame(player=c('A', 'B', 'C', 'D', 'A', 'A'),

runs=c(100, 200, 408, 19, 56, 100),

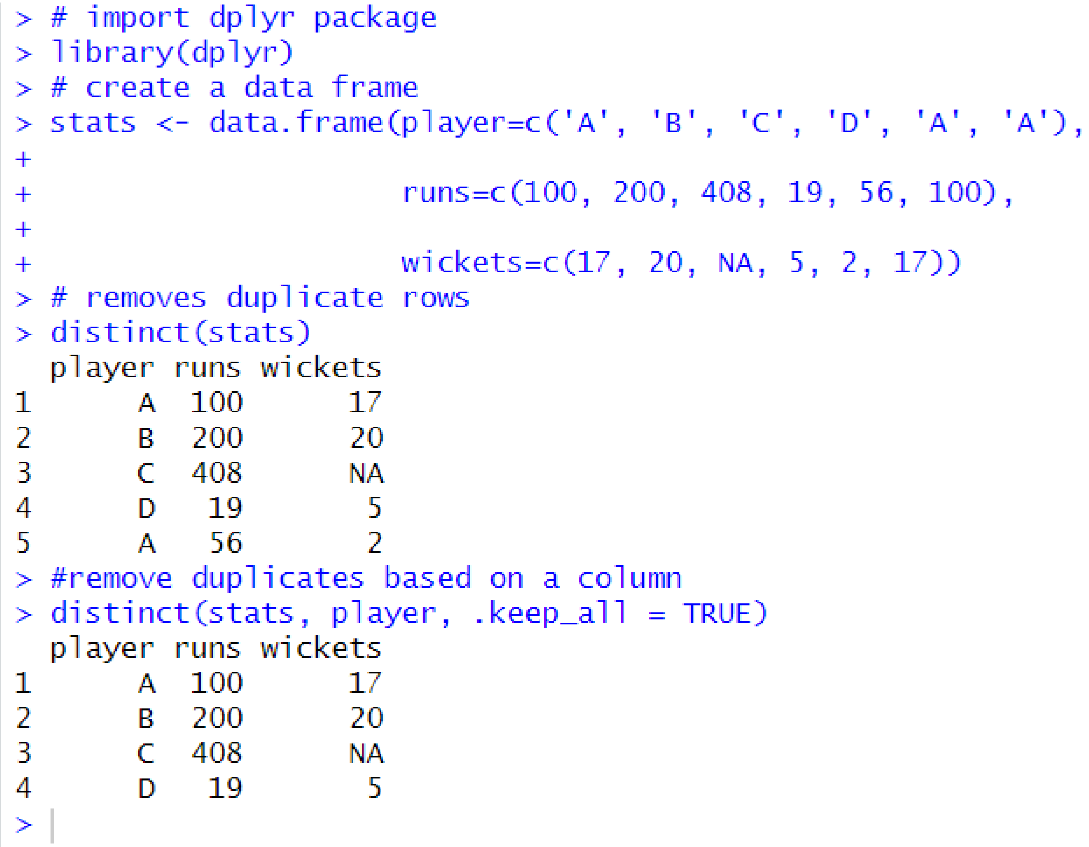
wickets=c(17, 20, NA, 5, 2, 17))

removes duplicate rows

distinct(stats)

#remove duplicates based on a column

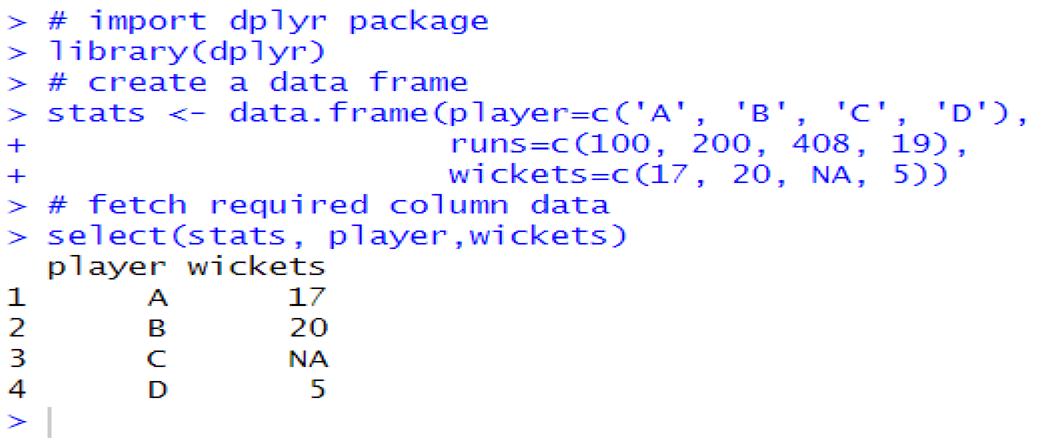
distinct(stats, player, .keep_all = TRUE)



```
> # import dplyr package
> library(dplyr)
> # create a data frame
> stats <- data.frame(player=c('A', 'B', 'C', 'D', 'A', 'A'),
+
+                       runs=c(100, 200, 408, 19, 56, 100),
+
+                       wickets=c(17, 20, NA, 5, 2, 17))
> # removes duplicate rows
> distinct(stats)
  player runs wickets
1      A  100      17
2      B  200      20
3      C  408      NA
4      D   19       5
5      A   56       2
> #remove duplicates based on a column
> distinct(stats, player, .keep_all = TRUE)
  player runs wickets
1      A  100      17
2      B  200      20
3      C  408      NA
4      D   19       5
> 
```

select() method:

```
# import dplyr package

library(dplyr)

# create a data frame

stats <- data.frame(player=c('A', 'B', 'C', 'D'),

                    runs=c(100, 200, 408, 19),

                    wickets=c(17, 20, NA, 5))

# fetch required column data

select(stats, player,wickets)
```

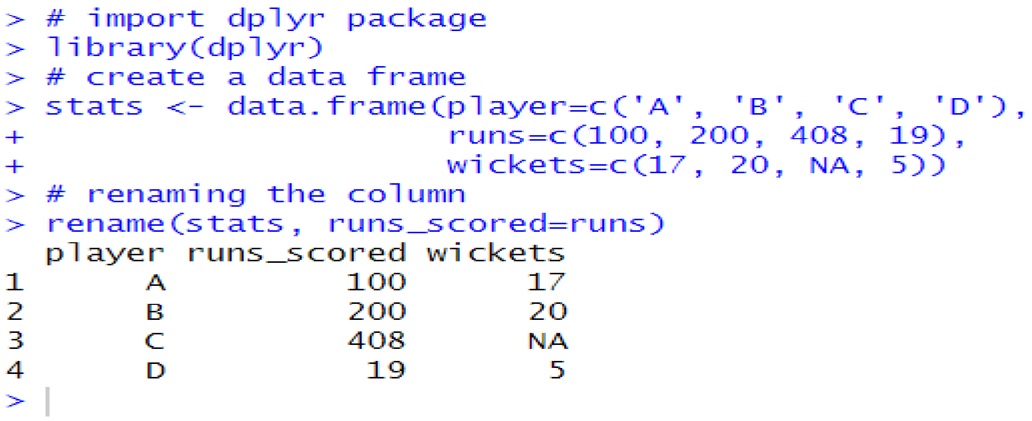


```
> # import dplyr package
> library(dplyr)
> # create a data frame
> stats <- data.frame(player=c('A', 'B', 'C', 'D'),
+                     runs=c(100, 200, 408, 19),
+                     wickets=c(17, 20, NA, 5))
> # fetch required column data
> select(stats, player,wickets)
  player wickets
1      A      17
2      B      20
3      C      NA
4      D       5
> 
```

rename() method:

```
# import dplyr package

library(dplyr)

# create a data frame

stats <- data.frame(player=c('A', 'B', 'C', 'D'),

                    runs=c(100, 200, 408, 19),

                    wickets=c(17, 20, NA, 5))

# renaming the column
```

rename(stats, runs_scored=runs)



```
> # import dplyr package
> library(dplyr)
> # create a data frame
> stats <- data.frame(player=c('A', 'B', 'C', 'D'),
+                     runs=c(100, 200, 408, 19),
+                     wickets=c(17, 20, NA, 5))
> # renaming the column
> rename(stats, runs_scored=runs)
  player runs_scored wickets
1      A         100      17
2      B         200      20
3      C         408      NA
4      D          19       5
> 
```

mutate() and transmutate() method:

import dplyr package

library(dplyr)

create a data frame

stats <- data.frame(player=c('A', 'B', 'C', 'D'),

                    runs=c(100, 200, 408, 19),

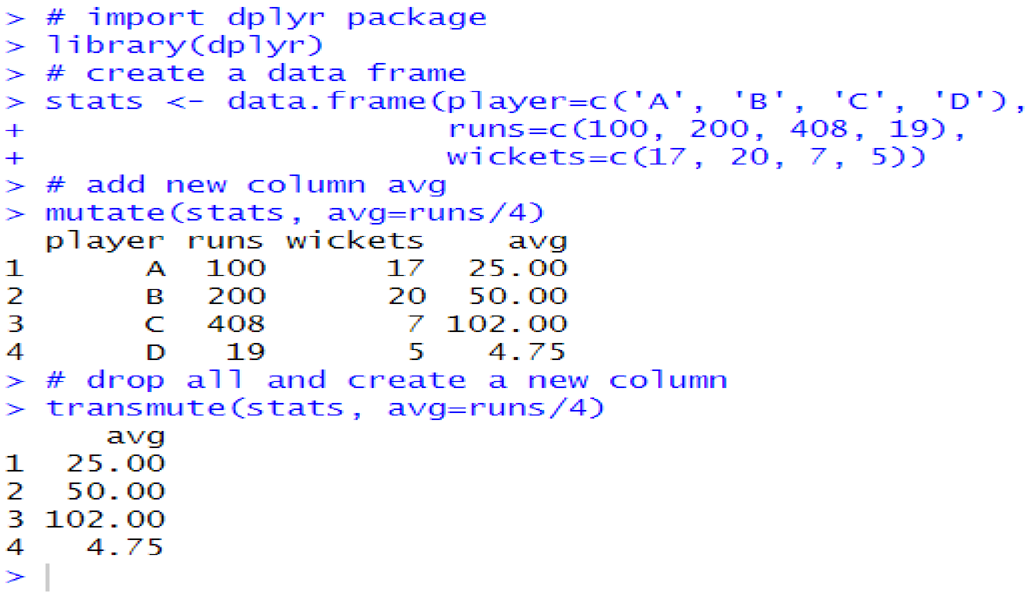
                    wickets=c(17, 20, 7, 5))

# add new column avg

mutate(stats, avg=runs/4)

# drop all and create a new column

transmute(stats, avg=runs/4)



```
> # import dplyr package
> library(dplyr)
> # create a data frame
> stats <- data.frame(player=c('A', 'B', 'C', 'D'),
+                     runs=c(100, 200, 408, 19),
+                     wickets=c(17, 20, 7, 5))
> # add new column avg
> mutate(stats, avg=runs/4)
  player runs wickets    avg
1      A  100      17  25.00
2      B  200      20  50.00
3      C  408       7 102.00
4      D   19       5   4.75
> # drop all and create a new column
> transmute(stats, avg=runs/4)
     avg
1  25.00
2  50.00
3 102.00
4   4.75
> 
```

summarize() method:

import dplyr package

library(dplyr)

create a data frame

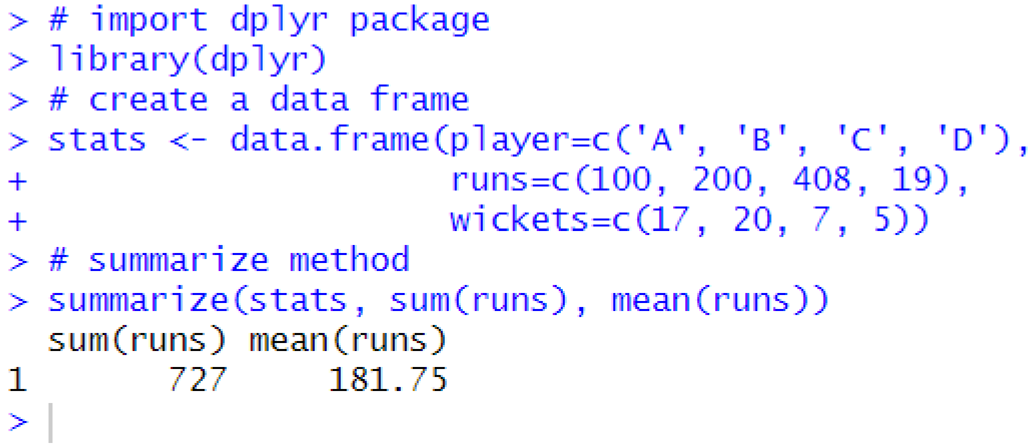
stats <- data.frame(player=c('A', 'B', 'C', 'D'),

runs=c(100, 200, 408, 19),

wickets=c(17, 20, 7, 5))

summarize method

summarize(stats, sum(runs), mean(runs))



```
> # import dplyr package
> library(dplyr)
> # create a data frame
> stats <- data.frame(player=c('A', 'B', 'C', 'D'),
+                     runs=c(100, 200, 408, 19),
+                     wickets=c(17, 20, 7, 5))
> # summarize method
> summarize(stats, sum(runs), mean(runs))
  sum(runs) mean(runs)
1       727     181.75
> 
```

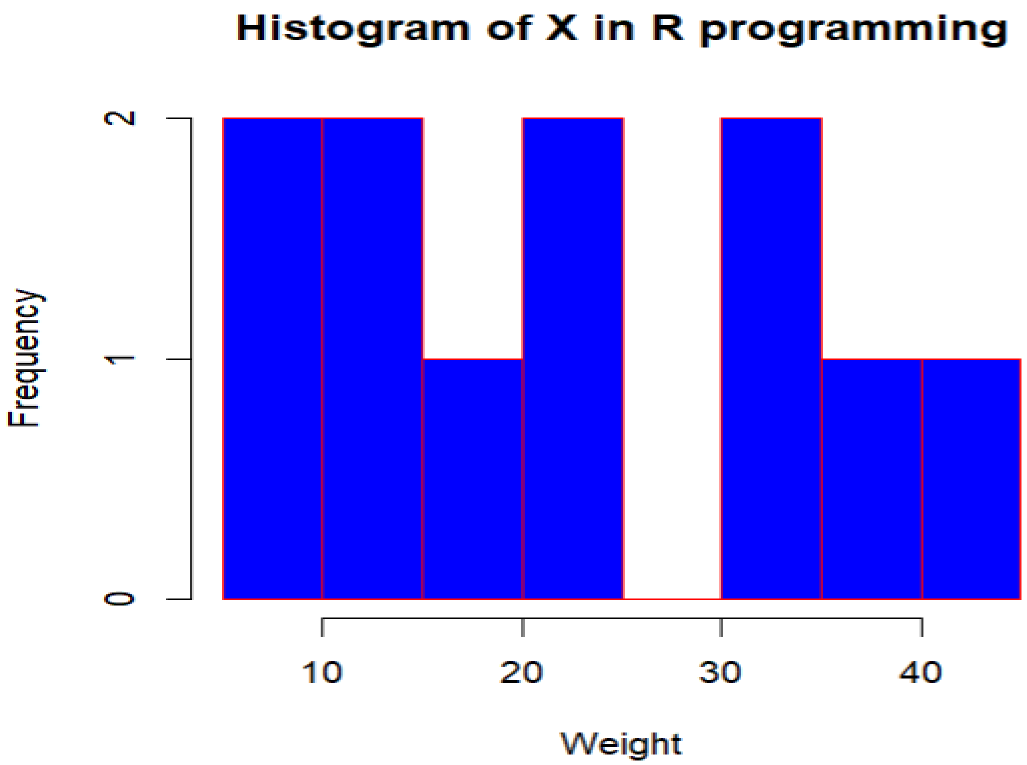
Descriptive Statistics:

Histogram:

```
x=c(9, 13, 21, 8, 36, 22, 12, 41, 31, 33, 19)
```

```
hist(x, col ="blue", border="red", xlab="Weight", main = "Histogram of X in R programming")
```

output

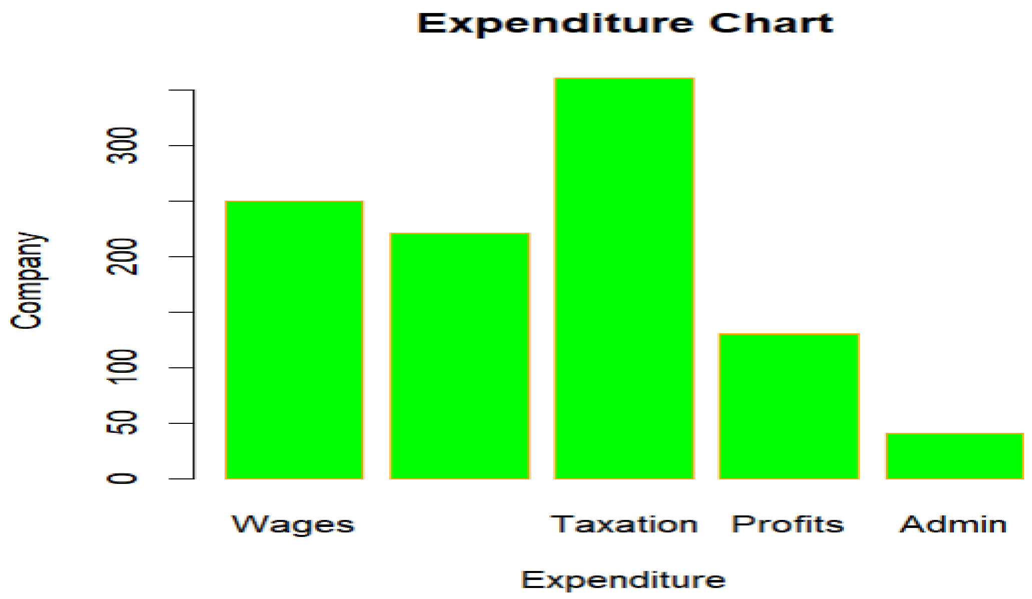


Bar Diagram:

```
x=c(250, 220, 360, 130, 40)
```

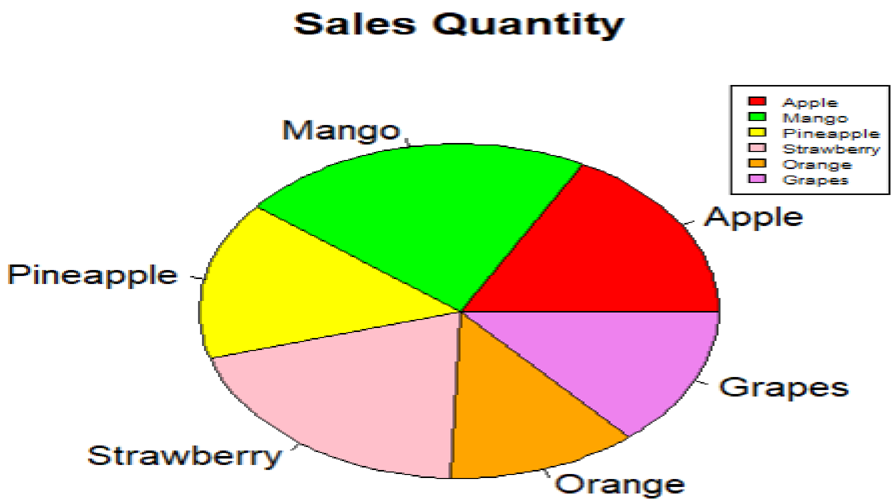
```
y=c("Wages", "Materials", "Taxation", "Profits", "Admin")
```

```
barplot(x, names.arg=y, xlab="Expenditure", ylab="Company", col="green", border="orange", main="Expenditure Chart")
```



Pie Chart:

```
x=c(500, 650, 450, 580, 350, 400)

labels=c("Apple", "Mango", "Pineapple", "Strawberry", "Orange",
"Grapes")

cols=c("red", "green", "yellow", "pink", "orange", "violet", "blue")

pie(x, labels, col=cols, main="Sales Quantity")

legend("topright", c("Apple", "Mango", "Pineapple", "Strawberry",
"Orange", "Grapes"), cex=0.5, fill=cols)
```



Create different data frames (STUDENT DATA, ORGANIZATION DATA)

Do the following task

1. **Ordering of data**

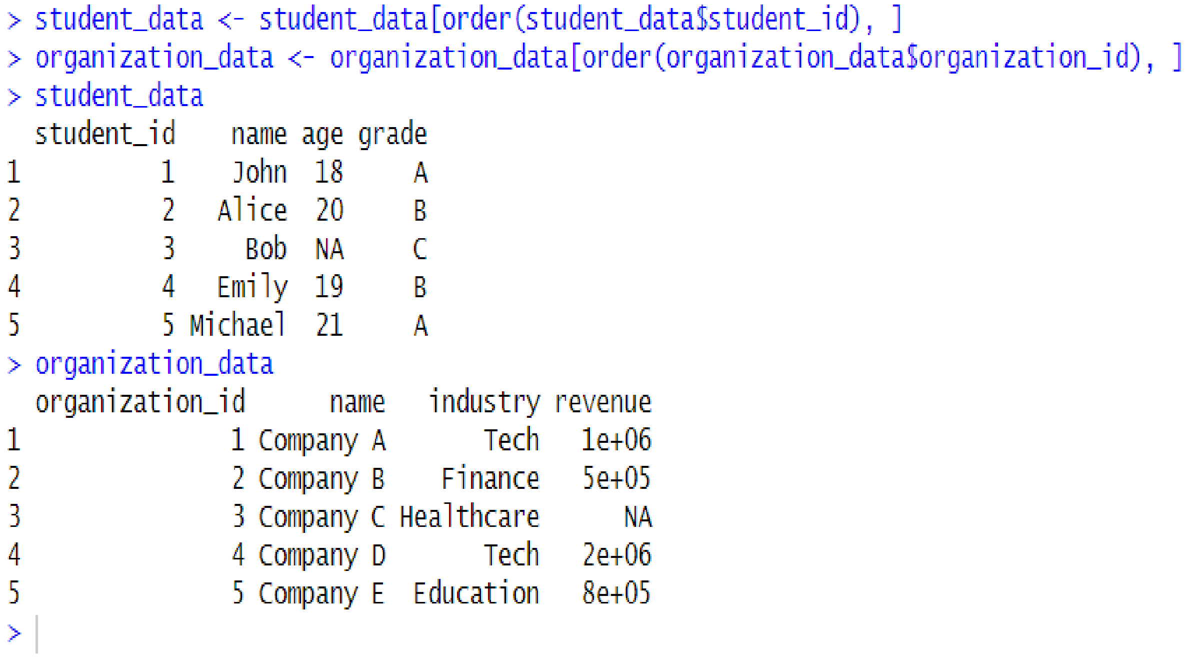
```
# Create STUDENT DATA data frame

student_data <- data.frame(

  student_id = c(1, 2, 3, 4, 5),

  name = c("John", "Alice", "Bob", "Emily", "Michael"),

  age = c(18, 20, NA, 19, 21),

  grade = c("A", "B", "C", "B", "A")

)

# Create ORGANIZATION DATA data frame

organization_data <- data.frame(

  organization_id = c(1, 2, 3, 4, 5),

  name = c("Company A", "Company B", "Company C", "Company D", "Company E"),

  industry = c("Tech", "Finance", "Healthcare", "Tech", "Education"),

  revenue = c(1000000, 500000, NA, 2000000, 800000)

)
```

Do the following task

1. **Ordering of data**

```
student_data <- student_data[order(student_data$student_id), ]

organization_data <- organization_data[order(organization_data$organization_id), ]

student_data
```

organization_data



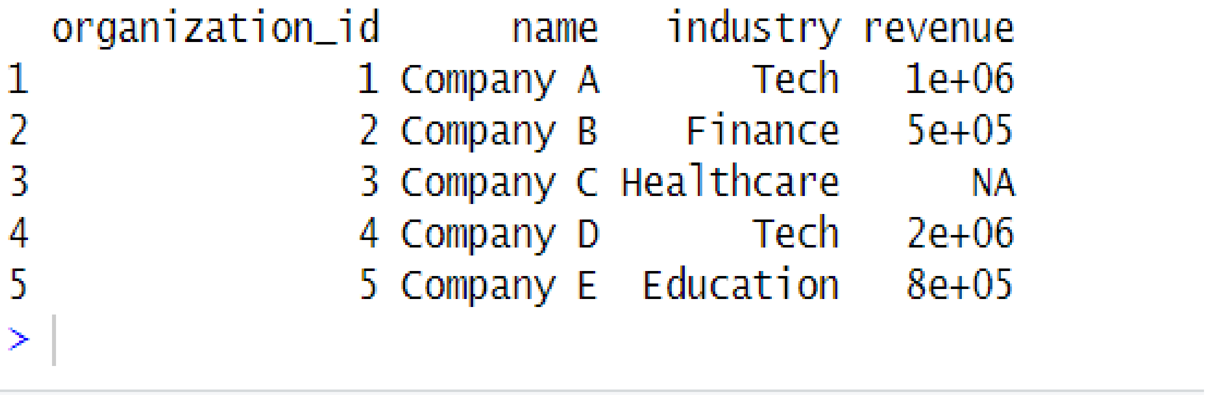
```
> student_data <- student_data[order(student_data$student_id), ]
> organization_data <- organization_data[order(organization_data$organization_id), ]
> student_data
  student_id    name age grade
1          1    John  18     A
2          2   Alice  20     B
3          3     Bob  NA     C
4          4   Emily  19     B
5          5 Michael  21     A
> organization_data
  organization_id      name   industry revenue
1               1 Company A       Tech   1e+06
2               2 Company B    Finance   5e+05
3               3 Company C Healthcare      NA
4               4 Company D       Tech   2e+06
5               5 Company E  Education   8e+05
> 
```

2. Finding and removing duplicate data

student_data <- unique(student_data)

organization_data <- unique(organization_data)

organization_data

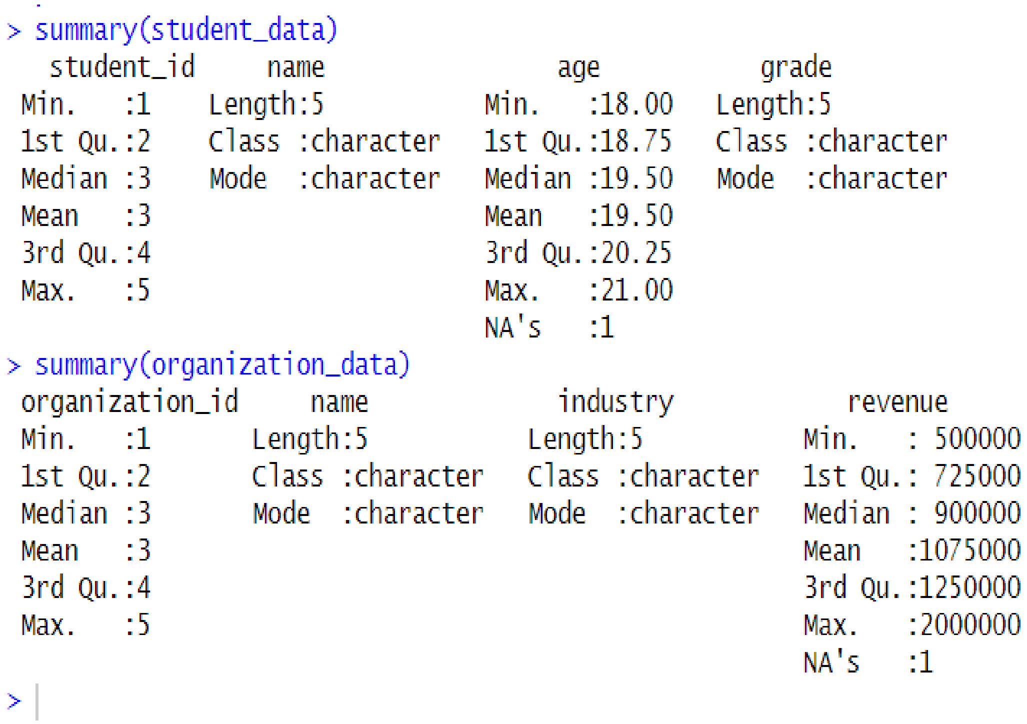


```
  organization_id      name   industry revenue
1               1 Company A       Tech   1e+06
2               2 Company B    Finance   5e+05
3               3 Company C Healthcare      NA
4               4 Company D       Tech   2e+06
5               5 Company E  Education   8e+05
> 
```

3. Handling missing values and perform summarize.

summary(student_data)

summary(organization_data)

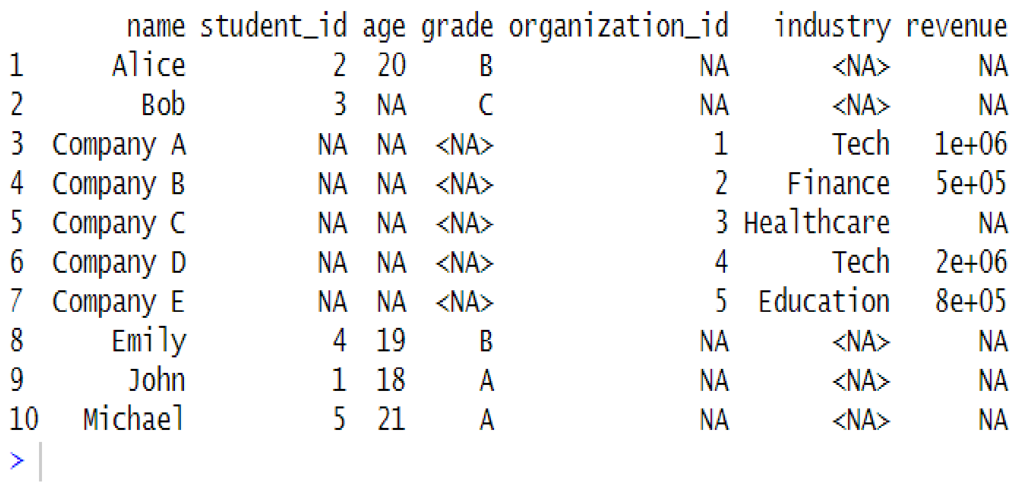


```
> summary(student_data)
   student_id      name                age            grade          
 Min.   :1    Length:5           Min.   :18.00   Length:5          
 1st Qu.:2    Class :character   1st Qu.:18.75   Class :character  
 Median :3    Mode  :character   Median :19.50   Mode  :character  
 Mean   :3                       Mean   :19.50                     
 3rd Qu.:4                       3rd Qu.:20.25                     
 Max.   :5                       Max.   :21.00                     
                                 NA's   :1                         
> summary(organization_data)
 organization_id     name             industry            revenue       
 Min.   :1       Length:5           Length:5           Min.   : 500000  
 1st Qu.:2       Class :character   Class :character   1st Qu.: 725000  
 Median :3       Mode  :character   Mode  :character   Median : 900000  
 Mean   :3                                             Mean   :1075000  
 3rd Qu.:4                                             3rd Qu.:1250000  
 Max.   :5                                             Max.   :2000000  
                                                       NA's   :1        
> 
```

4. Do the merging operations.

merged_data <- merge(student_data, organization_data, by = "name", all = TRUE)

merged_data



```
        name student_id age grade organization_id   industry revenue
1      Alice          2  20     B              NA       <NA>      NA
2        Bob          3  NA     C              NA       <NA>      NA
3  Company A         NA  NA  <NA>               1       Tech   1e+06
4  Company B         NA  NA  <NA>               2    Finance   5e+05
5  Company C         NA  NA  <NA>               3 Healthcare      NA
6  Company D         NA  NA  <NA>               4       Tech   2e+06
7  Company E         NA  NA  <NA>               5  Education   8e+05
8      Emily          4  19     B              NA       <NA>      NA
9       John          1  18     A              NA       <NA>      NA
10   Michael          5  21     A              NA       <NA>      NA
> 
```

# Perform Left Join

```
left_join <- merge(student_data, organization_data, by = "name", all.x = TRUE)
```

```
# Perform Right Join
```

```
right_join <- merge(student_data, organization_data, by = "name", all.y = TRUE)
```

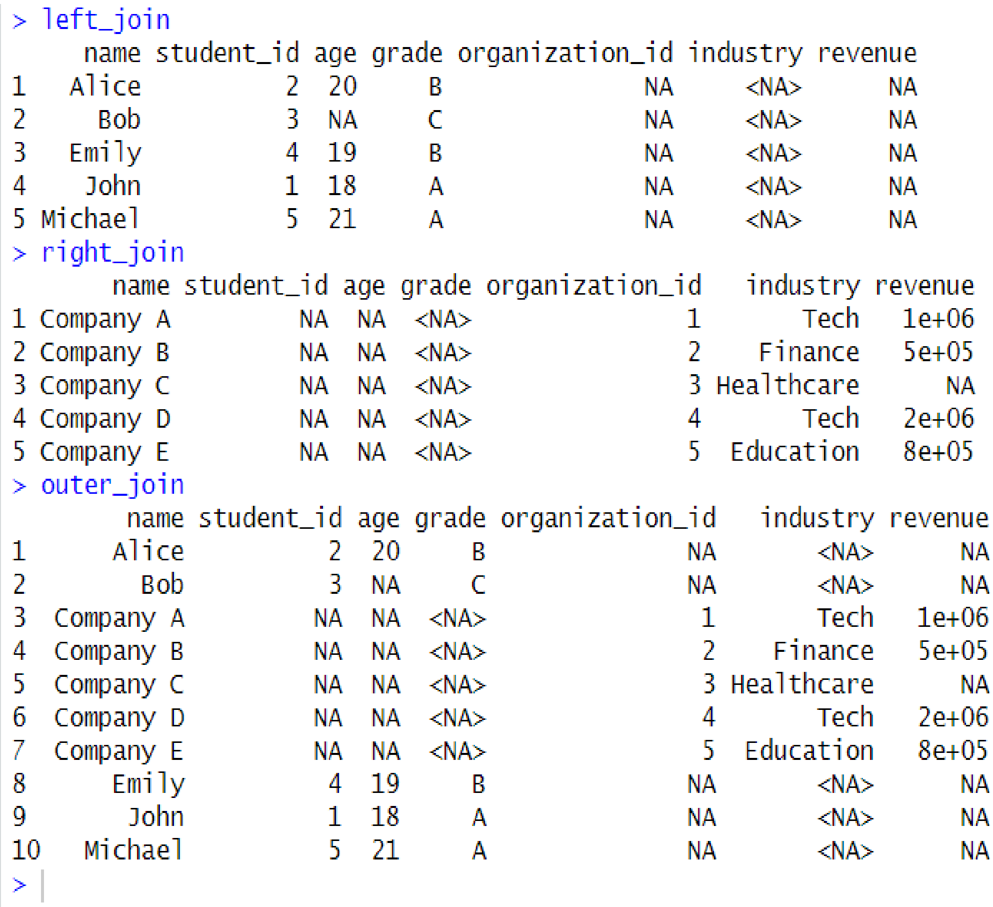
```
# Perform Outer Join
```

```
outer_join <- merge(student_data, organization_data, by = "name", all = TRUE)
```

```
left_join
```

```
right_join
```

```
outer_join
```

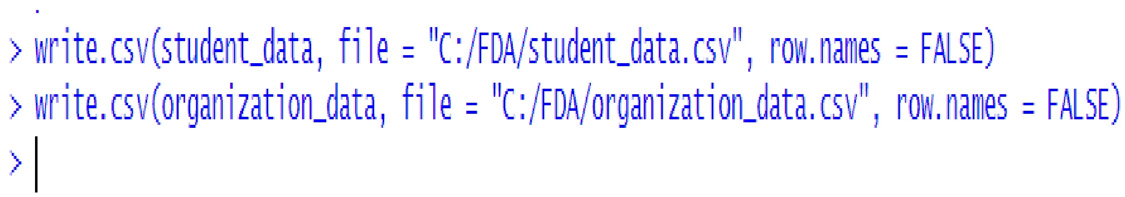


```
> left_join
     name student_id age grade organization_id industry revenue
1   Alice          2  20     B              NA     <NA>      NA
2     Bob          3  NA     C              NA     <NA>      NA
3   Emily          4  19     B              NA     <NA>      NA
4    John          1  18     A              NA     <NA>      NA
5 Michael          5  21     A              NA     <NA>      NA
> right_join
       name student_id age grade organization_id   industry revenue
1 Company A         NA  NA  <NA>               1       Tech   1e+06
2 Company B         NA  NA  <NA>               2    Finance   5e+05
3 Company C         NA  NA  <NA>               3 Healthcare      NA
4 Company D         NA  NA  <NA>               4       Tech   2e+06
5 Company E         NA  NA  <NA>               5  Education   8e+05
> outer_join
        name student_id age grade organization_id   industry revenue
1      Alice          2  20     B              NA       <NA>      NA
2        Bob          3  NA     C              NA       <NA>      NA
3  Company A         NA  NA  <NA>               1       Tech   1e+06
4  Company B         NA  NA  <NA>               2    Finance   5e+05
5  Company C         NA  NA  <NA>               3 Healthcare      NA
6  Company D         NA  NA  <NA>               4       Tech   2e+06
7  Company E         NA  NA  <NA>               5  Education   8e+05
8      Emily          4  19     B              NA       <NA>      NA
9       John          1  18     A              NA       <NA>      NA
10   Michael          5  21     A              NA       <NA>      NA
> 
```

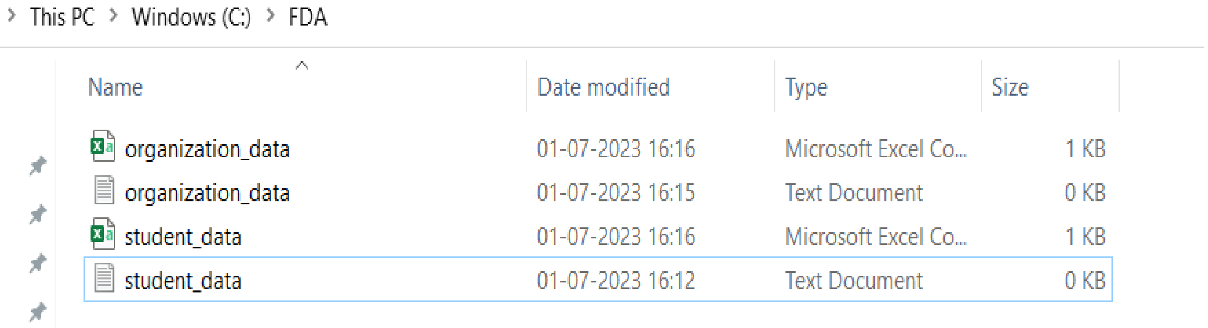
5. Write/Copy the content of data frames to csv/txt files.

write.csv(student_data, file = "C:/FDA/student_data.csv", row.names = FALSE)

write.csv(organization_data, file = "C:/FDA/organization_data.csv", row.names = FALSE)

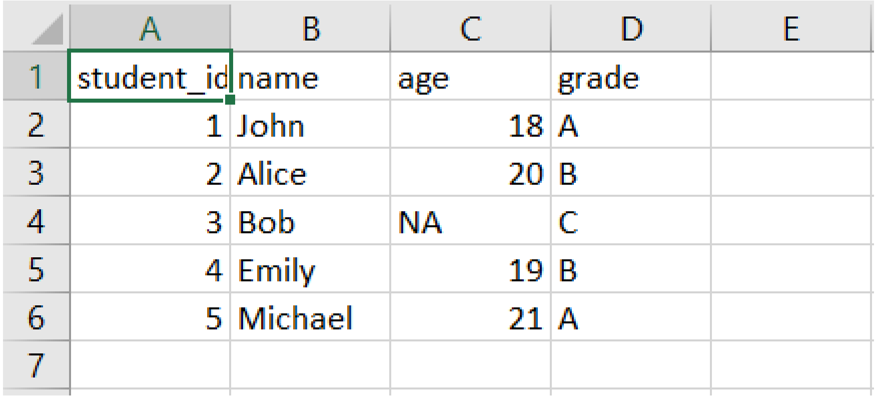


```
> write.csv(student_data, file = "C:/FDA/student_data.csv", row.names = FALSE)
> write.csv(organization_data, file = "C:/FDA/organization_data.csv", row.names = FALSE)
> |
```

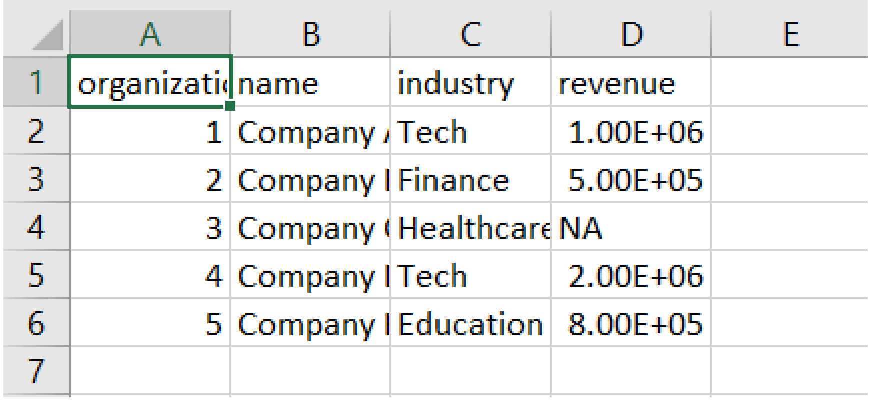


› This PC › Windows (C:) › FDA

Name	Date modified	Type	Size
organization_data	01-07-2023 16:16	Microsoft Excel Co...	1 KB
organization_data	01-07-2023 16:15	Text Document	0 KB
student_data	01-07-2023 16:16	Microsoft Excel Co...	1 KB
student_data	01-07-2023 16:12	Text Document	0 KB



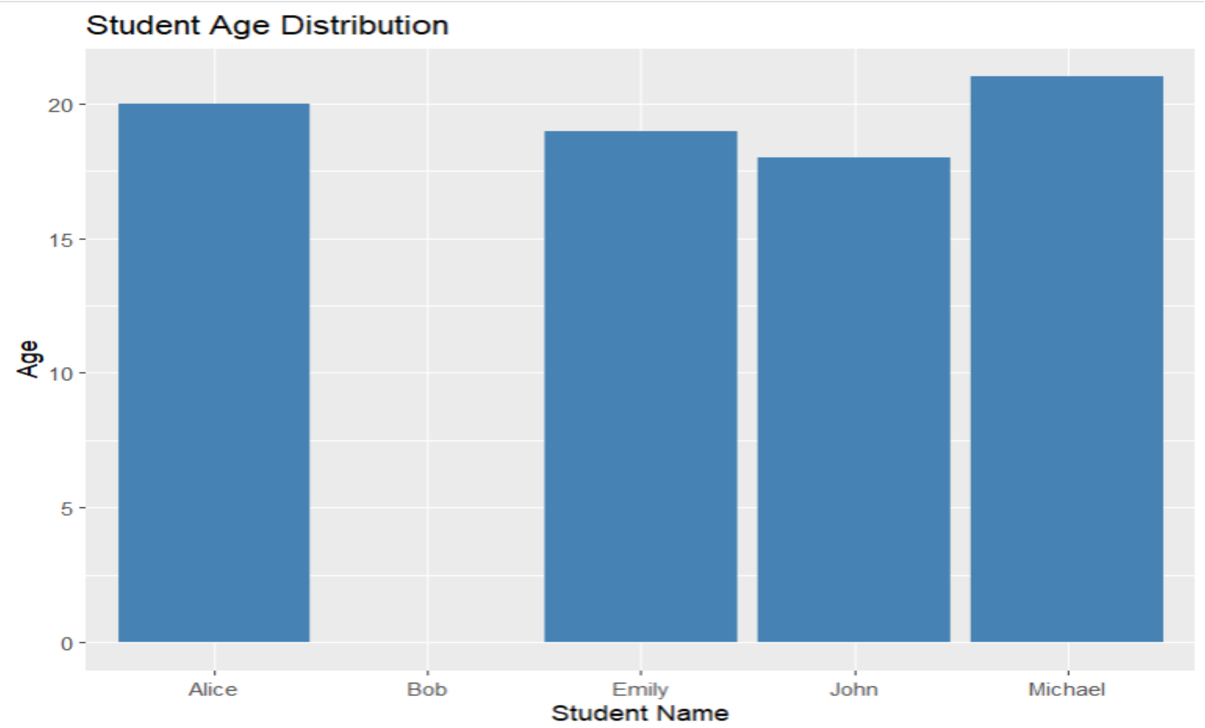
	A	B	C	D	E
1	student_id	name	age	grade	
2	1	John	18	A	
3	2	Alice	20	B	
4	3	Bob	NA	C	
5	4	Emily	19	B	
6	5	Michael	21	A	
7					

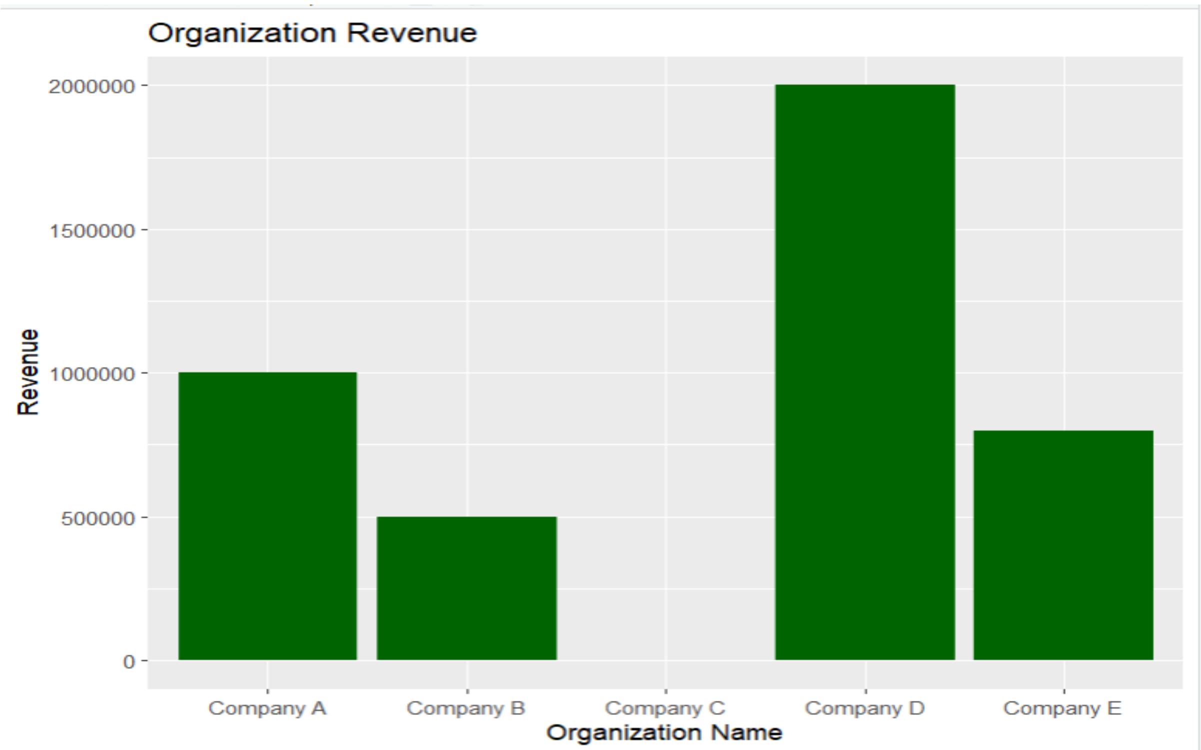


	A	B	C	D	E
1	organizati	name	industry	revenue	
2	1	Company /	Tech	1.00E+06	
3	2	Company I	Finance	5.00E+05	
4	3	Company (	Healthcare	NA	
5	4	Company I	Tech	2.00E+06	
6	5	Company I	Education	8.00E+05	
7					

6. Project the data values using basic plots.

```
library(ggplot2)

ggplot(student_data, aes(x = name, y = age)) +

  geom_bar(stat = "identity", fill = "steelblue") +

  labs(title = "Student Age Distribution", x = "Student Name", y = "Age")

ggplot(organization_data, aes(x = name, y = revenue)) +

  geom_bar(stat = "identity", fill = "darkgreen") +

  labs(title = "Organization Revenue", x = "Organization
Name",y="Revenue")
```




Organization Revenue
2000000
1500000
1000000
500000
0
Revenue
Company A
Company B
Company C
Company D
Company E
Organization Name

---X--

Thank you!