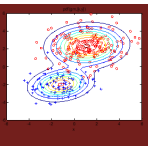


STAT 445/645 Assignment Cover Page

Student Name _____

SFU Student Number _____

SFU email address _____

Assignment Number _____

Due Date _____

Provide references for any data sets used in this assignment

List software used in this assignment.

List **ALL** resources used to complete this assignment, including books, internet sources and people.

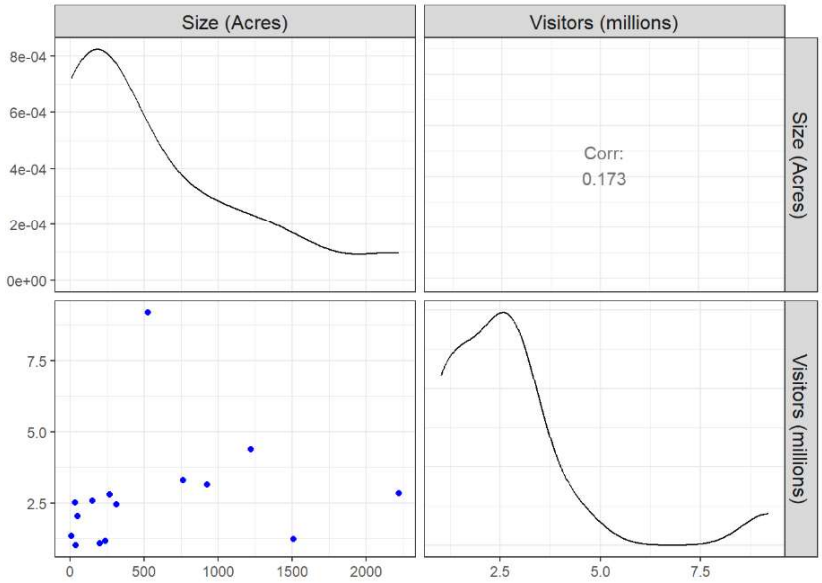
☐ I personally completed the computations and wrote the solutions submitted in this document.

Question1

a)

a. Present a matrix scatter plot.

```
ggpairs(X[,-1], lower=list(continuous = wrap("points", color = "blue")))+
  theme_bw()+
  theme(strip.text=element_text(size=12))
```



b)

b. Compute and display the sample correlation matrix.

```
numericData <- X[sapply(X, is.numeric)]

cor(numericData, use = "complete.obs")
```

```
##                    Size (Acres) Visitors (millions)
## Size (Acres)          1.0000000           0.1725274
## Visitors (millions)   0.1725274           1.0000000
```

The sample correlation matrix is $\begin{bmatrix} 1.000 & 0.173 \\ 0.173 & 1.000 \end{bmatrix}$

c) An unusual park was Great Smoky because of the much higher number of visitors compared to other national parks. Specifically, the Great Smoky park had 9.19 millions visitors, which significantly deviated from the pattern observed in the rest of the data.

.

d)

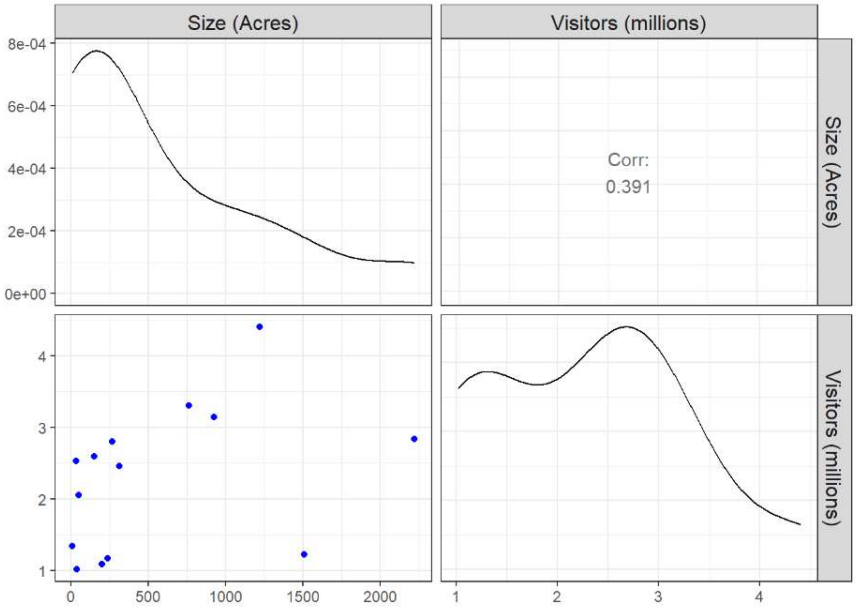
d. Produce a new data set by removing the data for the unusual park. Present a matrix scatter plot from the new data

```
new_X <- X[-c(7), ]

new_X
```

```
## # A tibble: 14 × 3
##    Park            `Size (Acres)` `Visitors (millions)`
##    <chr>                    <dbl>                 <dbl>
##  1 Arcadia                   47.4                  2.05
##  2 Bryce Canyon              35.8                  1.02
##  3 Cuyahoga Valley           32.9                  2.53
##  4 Everglades              1508.                   1.23
##  5 Grand Canyon            1217.                   4.4
##  6 Grand Teton              310                    2.46
##  7 Hot Springs                5.6                  1.34
##  8 Olympic                  923.                   3.14
##  9 Mount Rainier            236.                   1.17
## 10 Rocky Mountain           266.                   2.8
## 11 Shenandoah               199                    1.09
## 12 Yellowstone             2220.                   2.84
## 13 Yosemite                 761.                   3.3
## 14 Zion                     147.                   2.59
```

```
ggpairs(new_X[,-1], lower=list(continuous = wrap("points", color = "blue")))+
  theme_bw()+
  theme(strip.text=element_text(size=12))
```



e)

e. Compute the sample correlation matrix for the new data set

```
new_numericData <- new_X[sapply(new_X, is.numeric)]

cor(new_numericData, use = "complete.obs")
```

```
##                     Size (Acres) Visitors (millions)
## Size (Acres)           1.0000000           0.3907829
## Visitors (millions)    0.3907829           1.0000000
```

The sample correlation matrix for the new data set is $\begin{bmatrix} 1.000 & 0.392 \\ 0.391 & 1.000 \end{bmatrix}$

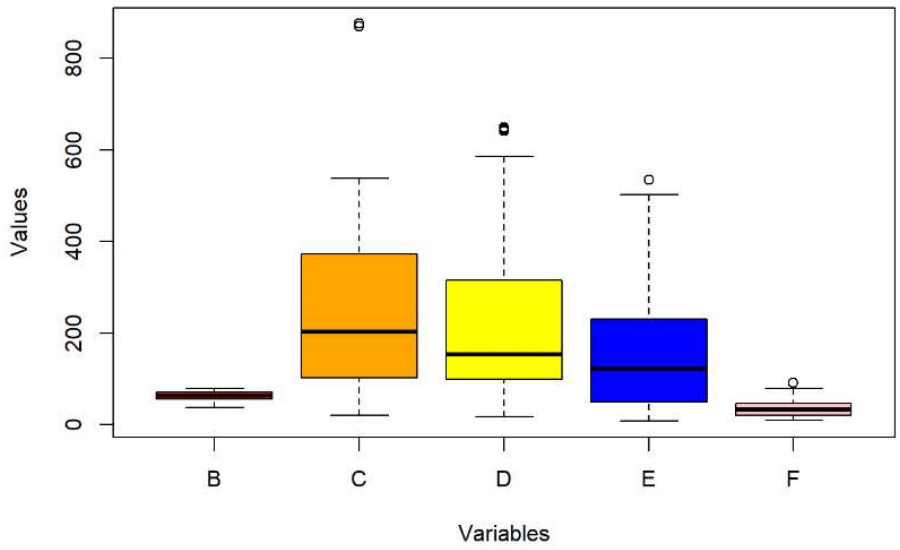
Initially, the correlation was 0.173. It indicates a weak positive relationship. After remove unusual park, the correlation increased to 0.391, indicating a stronger relationship.

This means the remove of the unusual park resulted in an increase in the correlation by 0.218. This difference quantifies the impact on the relationship between park size and visitor numbers, demonstrating that the outlier was masking a stronger positive relationship between these two variables.

Question2

a)

```
boxplot(data[, -1], col = c("red","orange","yellow","blue","pink"),
        xlab="Variables", ylab="Values", names=c("B", "C", "D","E","F"))
```



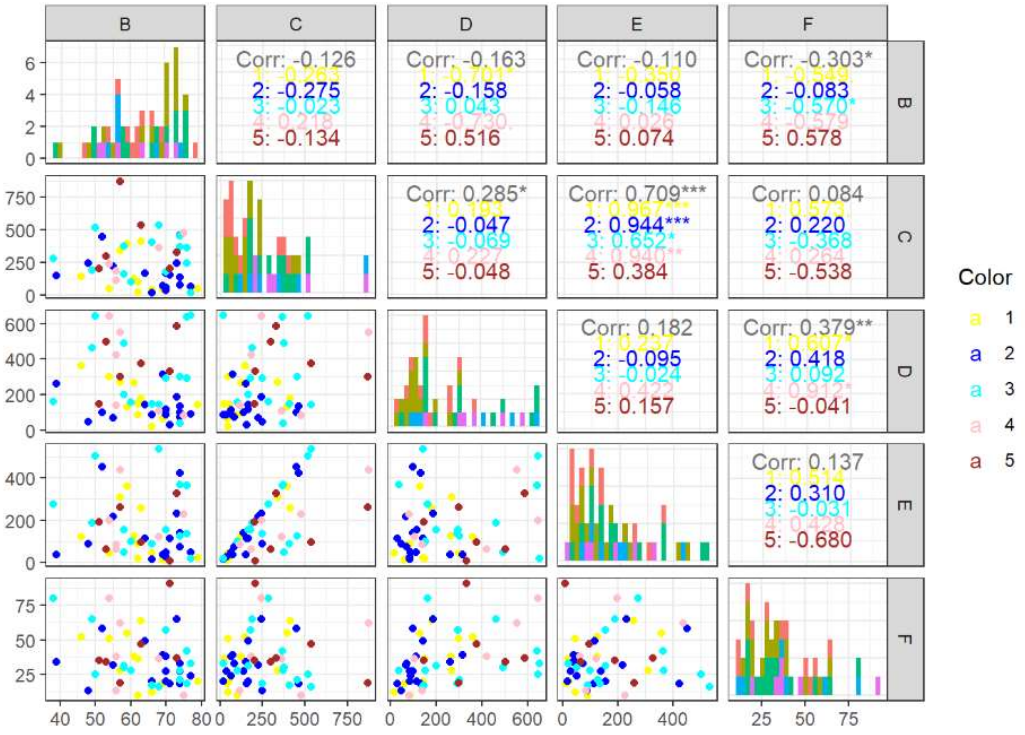
Yes, there is potential for scale and outlier issues because the range of variables C, D, and E are much larger than variables B and F. Also, there are visible outliers in variables C, D, E, and F, indicating outlier issues. This means that the data point is significantly different from other observations.

b)

b. Display a matrix scatter plot

```
Color<-as.factor(data$Group)
colors <- c("yellow","blue","cyan","pink","brown")

data %>%
  mutate(Color = factor(Color)) %>%
  ggpairs(columns = 2:6, aes(color = Color),legend=4,
          diag = list(continuous = "barDiag")) +
  theme_bw() +
  theme(plot.title=element_text(hjust=0.5))+
  scale_color_manual(values = colors, labels = c("1", "2", "3", "4", "5"),
                     name = "Color")
```



c)

c. Compute the sample mean vector and standard deviations across all groups, display

```
data %>% group_by(Group) %>%
  summarise(mean_B=mean(B), sd_B=sd(B), mean_C=mean(C), sd_C=sd(C),
            mean_D=mean(D), sd_D=sd(D), mean_E=mean(E), sd_E=sd(E),
            mean_F=mean(F), sd_F=sd(F))
```

```
## # A tibble: 5 × 11
##   Group mean_B  sd_B mean_C  sd_C mean_D  sd_D mean_E  sd_E mean_F  sd_F
##   <dbl>  <dbl> <dbl>  <dbl> <dbl>  <dbl> <dbl>  <dbl> <dbl>  <dbl> <dbl>
## 1     1   61.6  8.42   157.  151.   178. 108.    128   124.   34.9  18.7
## 2     2   65.4 11.0    171.  131.   122.  75.4   139.  133.   32.2  14.8
## 3     3   63.9 12.1    301.  157.   330. 210.    219.  164.   35.1  19.4
## 4     4   61    8.49   376.  277.   324. 247.    188.  137.   39.8  27.3
## 5     5   61.3  9.24   407.  258.   373. 154.    157.  122.   43.8  24.8
```

d)

d. Compute the matrix of sample means within groups using the standardized values, display

```
stand <- function(x) {
  (x - mean(x)) / sd(x)
}

mean_value<-data %>%
  mutate(across(B:F, stand)) %>%
  group_by(Group) %>%
  summarise(mean_B=mean(B), mean_C=mean(C),
            mean_D=mean(D), mean_E=mean(E),
            mean_F=mean(F))
mean_value
```

```
## # A tibble: 5 × 6
##   Group  mean_B mean_C mean_D  mean_E  mean_F
##   <dbl>   <dbl>  <dbl>  <dbl>   <dbl>   <dbl>
## 1     1 -0.158  -0.486 -0.339 -0.269  -0.0411
## 2     2  0.210  -0.415 -0.648 -0.190  -0.183
## 3     3  0.0602  0.248  0.494  0.382  -0.0326
## 4     4 -0.221   0.631  0.464  0.165   0.215
## 5     5 -0.188   0.789  0.733 -0.0582  0.423
```

The matrix of sample means within groups using the standardized values is

$$\begin{bmatrix} -0.158 & -0.486 & -0.339 & -0.269 & -0.041 \\ 0.210 & -0.415 & -0.648 & -0.190 & -0.183 \\ 0.060 & 0.248 & 0.494 & 0.382 & -0.033 \\ -0.221 & 0.631 & 0.464 & 0.165 & 0.215 \\ -0.188 & 0.789 & 0.733 & -0.058 & 0.423 \end{bmatrix}$$

e)

e. Compute the distance matrix between the groups using the standardized values, display.

```
distance_matrix <- as.dist(dist(mean_value[,-1]))

as.matrix(distance_matrix)
```

```
##           1         2         3         4         5
## 1 0.0000000 0.5127254 1.3055503 1.4662787 1.7423809
## 2 0.5127254 0.0000000 1.4551391 1.6738889 1.9754213
## 3 1.3055503 1.4551391 0.0000000 0.5795336 0.9022303
## 4 1.4662787 1.6738889 0.5795336 0.0000000 0.4374654
## 5 1.7423809 1.9754213 0.9022303 0.4374654 0.0000000
```

The distance matrix between the groups is

$$\begin{bmatrix} 0.000 & 0.513 & 1.306 & 1.466 & 1.742 \\ 0.513 & 0.000 & 1.455 & 1.674 & 1.975 \\ 1.306 & 1.455 & 0.000 & 0.580 & 0.902 \\ 1.466 & 1.674 & 0.580 & 0.000 & 0.437 \\ 1.742 & 1.975 & 0.902 & 0.437 & 0.000 \end{bmatrix}$$

f) Group 4 and Group 5 are relatively close to each other, with a distance of 0.437, suggesting similar characteristics between these groups. However, the distance between Group 2 and Group 5 is 1.975, indicating these groups are distinct from each other. And Group 5 is significantly further away from all other groups, which can mean unique characteristics that distinguish them.

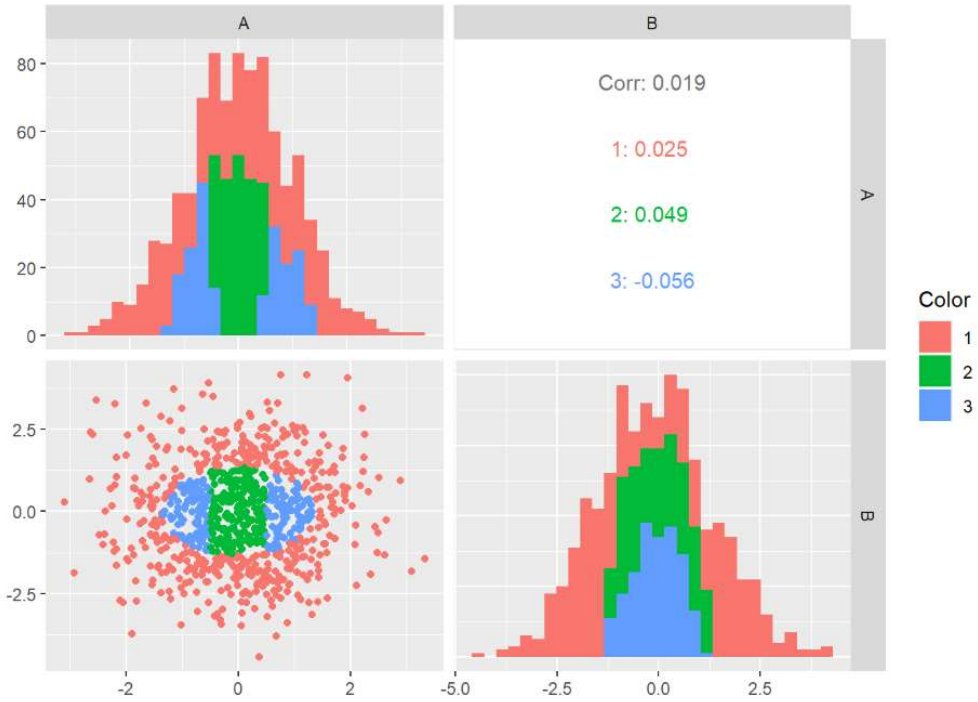
Question3

a)

a. Display a matrix scatter plot, indicating data from different groups with different color symbols

```
Color<-as.factor(data$Group)

ggpairs(data[,c("A","B")],mapping = aes(color = Color),legend=4,
        diag = list(continuous = "barDiag"))+
  theme(plot.title=element_text(hjust=0.5))
```



b)

b. Compute the matrix of sample means within groups, display

```
mean_value<-data %>%
  group_by(Group) %>%
  summarise(mean_A=mean(A), mean_B=mean(B))
print(mean_value)
```

```
## # A tibble: 3 × 3
##   Group   mean_A  mean_B
##   <int>    <dbl>   <dbl>
## 1     1  0.00500 -0.133
## 2     2 -0.00482 -0.0177
## 3     3  0.00641 -0.0560
```

The matrix of sample mean within groups is $\begin{bmatrix} 0.005 & -0.133 \\ -0.005 & -0.018 \\ 0.006 & -0.056 \end{bmatrix}$

c)

c. Compute the distance matrix between the groups, display

```
distance_matrix <- as.dist(dist(mean_value[,-1]))

as.matrix(distance_matrix)
```

```
##           1          2          3
## 1 0.00000000 0.11534319 0.07662038
## 2 0.11534319 0.00000000 0.03993015
## 3 0.07662038 0.03993015 0.00000000
```

The distance matrix between the groups is $\begin{bmatrix} 0.000 & 0.115 & 0.077 \\ 0.115 & 0.000 & 0.040 \\ 0.077 & 0.040 & 0.000 \end{bmatrix}$

d) I conclude that the groups of data are fairly close because the distance between group1 and group2 is 0.115, between group1 and group3 is 0.077 and between group2 and group3 is 0.04. These results show that the groups of data are fairly close together and share relatively similar characteristics with no significant differences.