Homework2

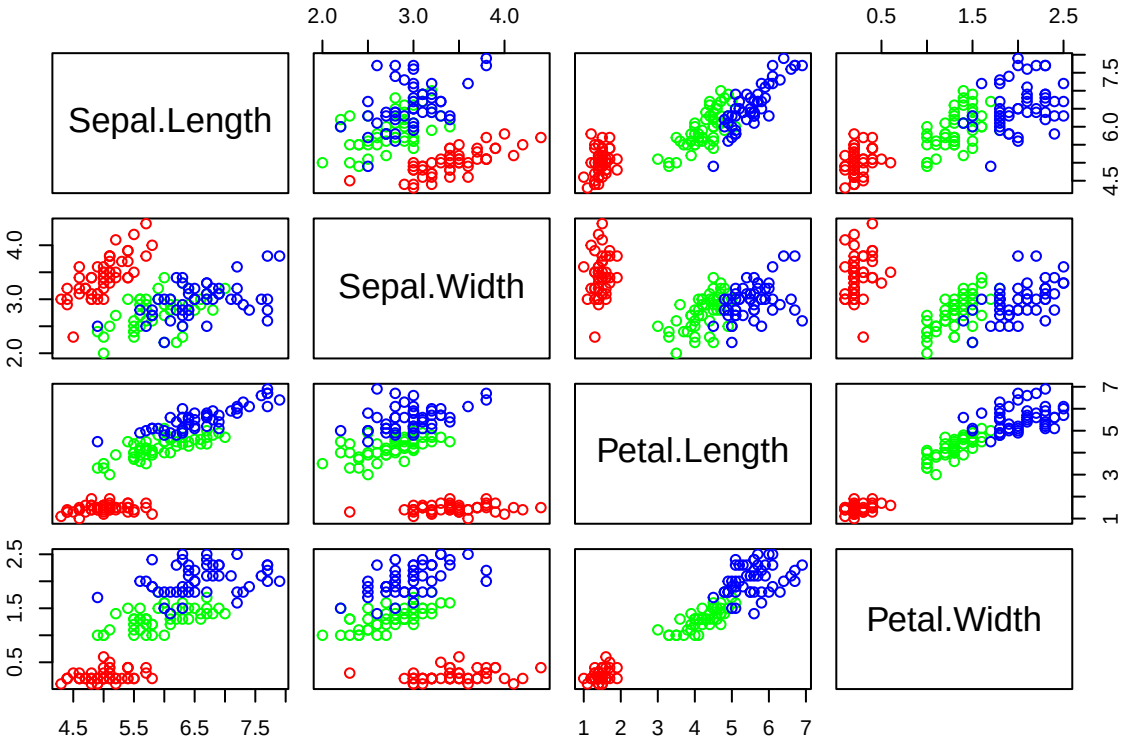
Li Zhang

08/28/2021

Exercise 2.1

Look at this large plot for a moment. What do you see? Provide interpretation of these scatter plots.

```
ma<- as.matrix(iris[ 1:4])
pairs(ma, col = rainbow(3)[iris$Species])
```



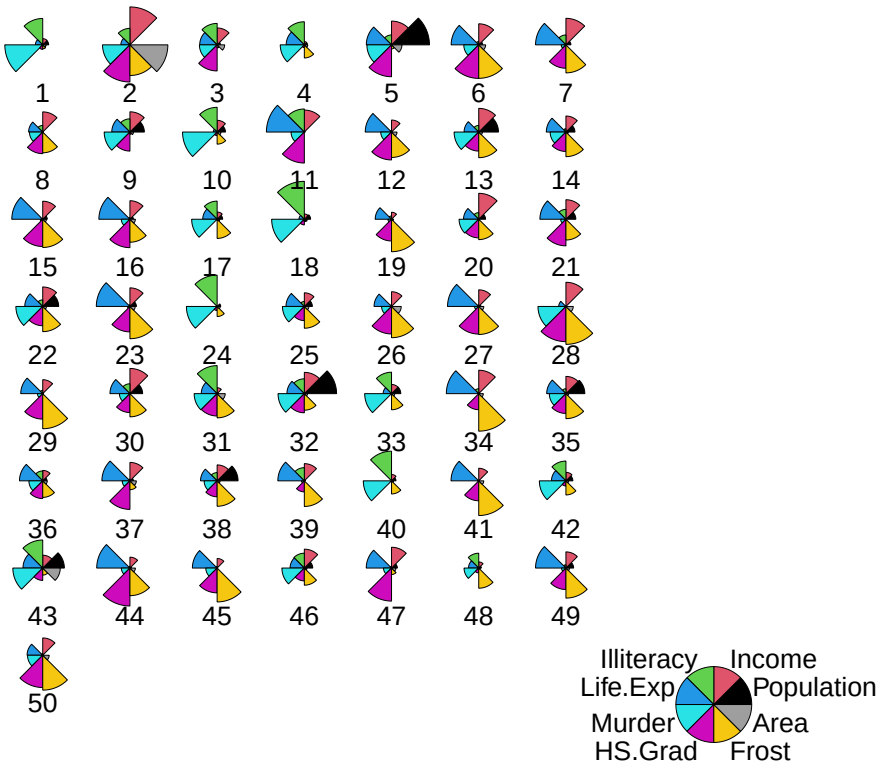

```
### Interpreation:

#Assume Sepal.Length is represented as SL;
#Sepal.Width is represented as SW;
#Petal.Length is represented as PL;
#Petal.With is represented as PW.

#For species type 1 setosa:
#there is a strong relationship between SL and SW.
#(As SW increases, SL also increases.)
#There is no apparent relationship between
#SL and PL, SL and PW, SW and PL, SW and PW, PL and PW.
```

```
#For Species type 2 versicolor:
#there is a week relationship between SL and SW, SL and PL, SW and PL.
#There is no apparent relationship between SL and PW, SW and PW, PL and PW.

#For Species type 3 virginica:
#There is a strong relationship between SL and PL;
#There is a weak relationship between SL and SW, SW and PL.
#There is no apparent relationship between SL and PW, SW and PW, PL and PW.

#Since it is based on observation, the interpretation is very subjective.
```

Exercise 2.3

Produce the segments diagram of the state data (state.x77) and offer some interpretation regarding South Dakota compared with other states. Hints: Convert the matrix to data frame using df.state.x77 <- as.data.frame(state.x77), then attach df.state.x77.

```
#View(state.x77)
#nrow(state.x77.df)
state.x77.df <- data.frame(state.x77)
rownames(state.x77.df) <- 1: nrow(state.x77.df)
stars(state.x77.df,key.loc = c(20, 1.0), draw.segments = TRUE)
```




```
### Interpretation:

#South Dakota is represented as 41.
#As we can see 41 in the segments diagram,
#frost in SD is the higher than almost all other states except North Dakota.
#life.Exp in SD is also higher than almost all the states
#Income and HS.grad are average compared to other states.
```

```
#Population, illiteracy and murder are lower
#than almost all other states,
#Area is not very big compared to the states that have large area.
#However, it is not very small compared to the states that have small area.
```

Exercise 2.4

Create a scatter plot of sepal length vs sepal width, change colors and shapes with species, and add trend line.

```
# Calculate Pearson's correlation coefficient for SL and SW.
SL <- iris$Sepal.Length
SW <- iris$Sepal.Width
PCC <- cor(SW, SL)  # Pearson's correlation coefficient
PCC <- round(PCC, 2) # round to the 2nd place after decimal point.

#Create a scatter plot using ggplot2
library(ggplot2)
ggplot(iris) +
  aes(x = Sepal.Length, y = Sepal.Width) +
  geom_point(aes(color = Species, shape = Species)) +
  geom_smooth(formula = y ~ x, method = lm) +
  annotate("text", x = 7 , y = 0.5 , label = paste("R = ", PCC) ) +
  xlab("Sepal length (cm) ") +
  ylab("Sepal width (cm) ") +
  ggtitle("Correlation between Sepal length and width")
```