Homework 3

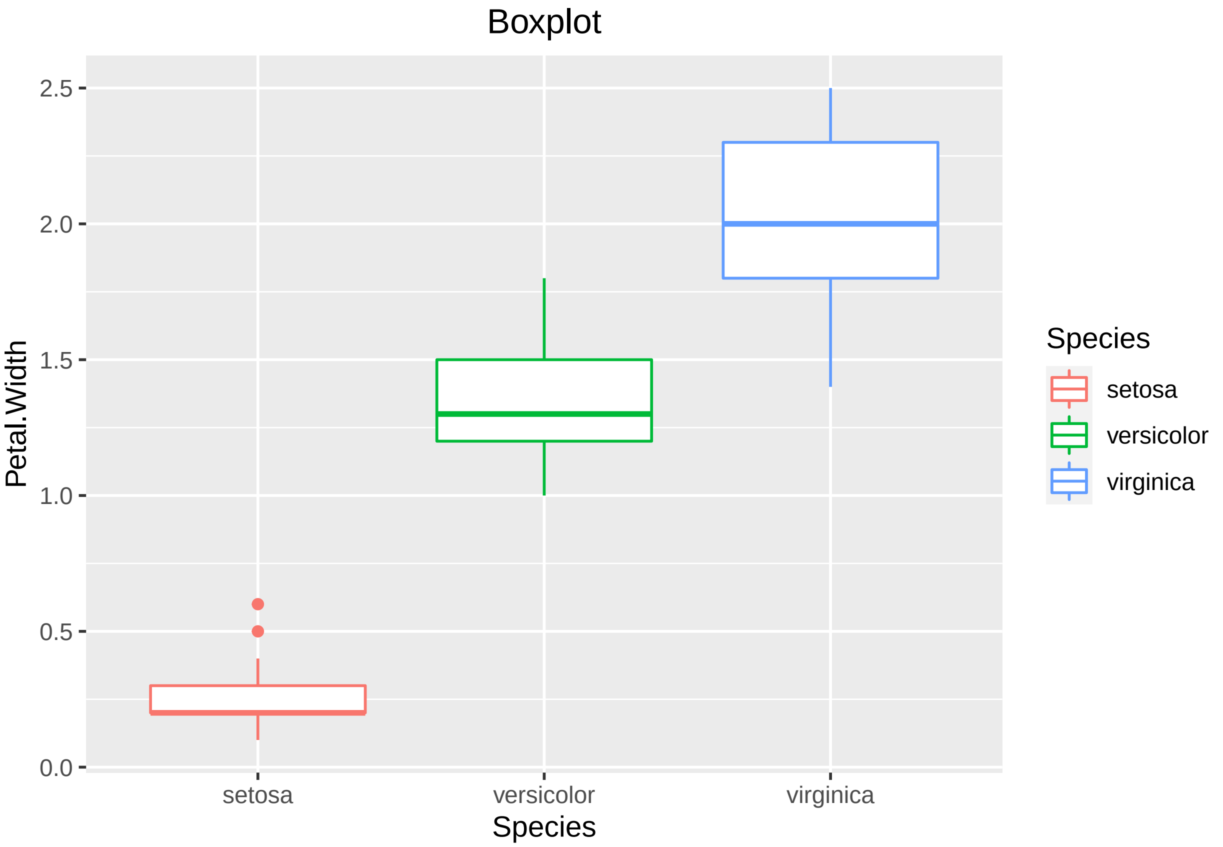
Li Zhang

09/07/2021

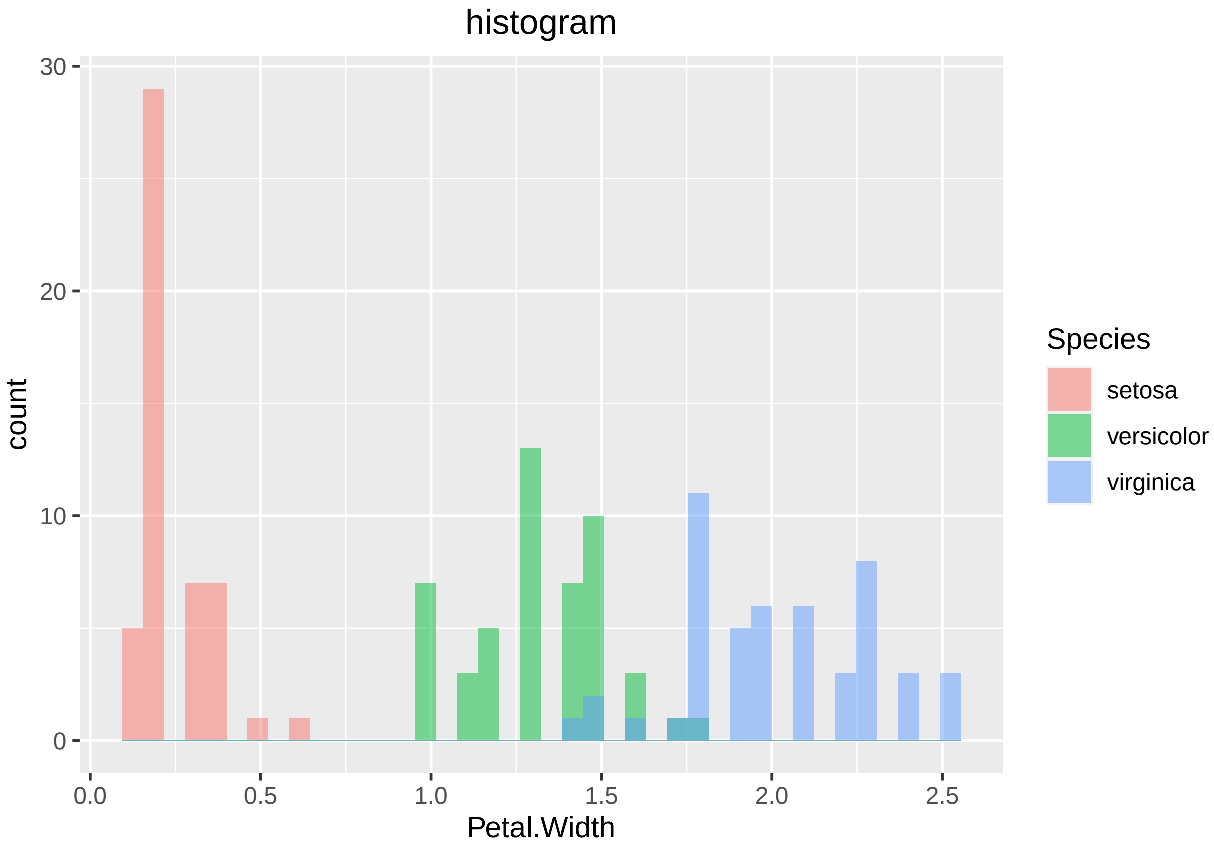
Exercise 2.5

Use boxplot, multiple panel histograms, and density plots to investigate whether petal width is the same among three subspecies.

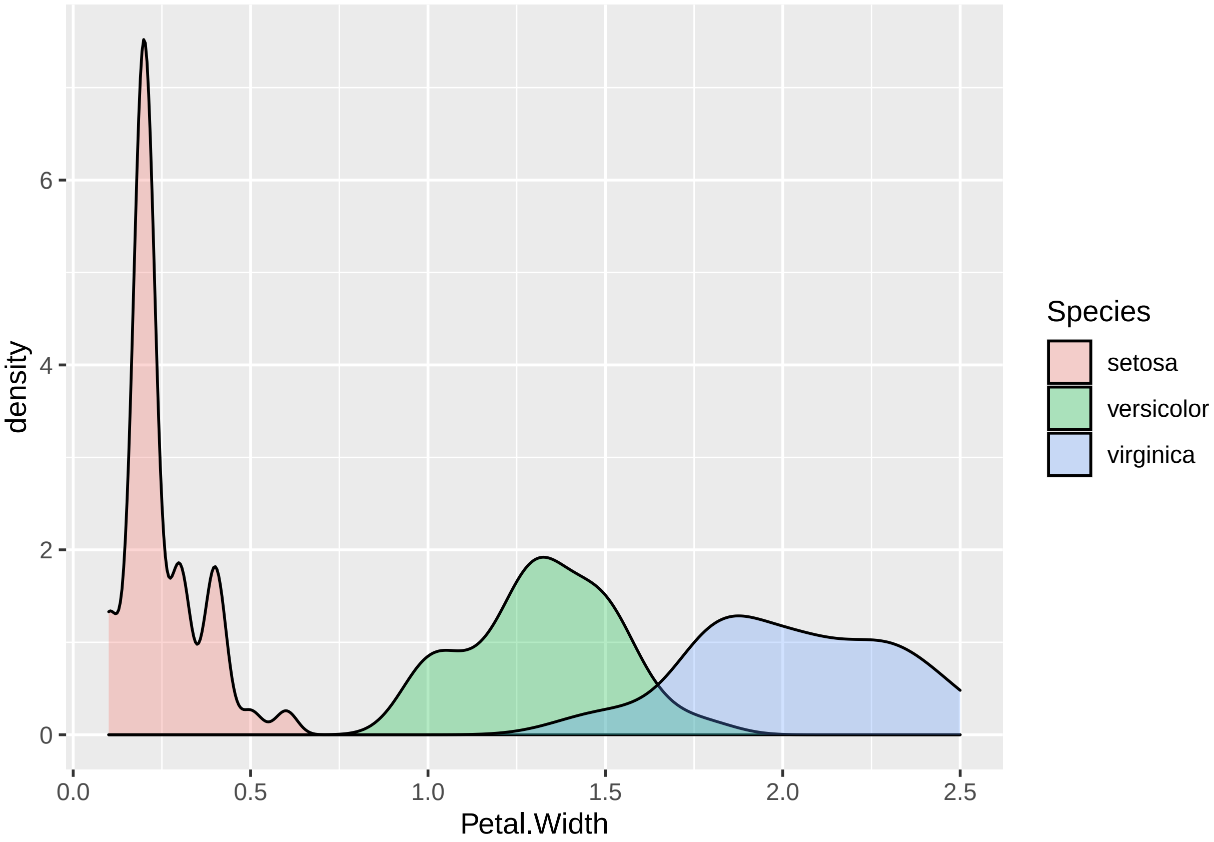
```
#Create a box plot using ggplot2
library(ggplot2)
ggplot(iris) +
  aes(x = Species, y = Petal.Width, color = Species) +
  geom_boxplot() +
  ggtitle("Boxplot")+
  theme(plot.title = element_text(hjust = 0.5))
```



```
#Create multiple panel histograms using ggplot2
ggplot(iris) +
  aes(x = Petal.Width, fill = Species) +
  geom_histogram(alpha = 0.5, position = "identity", bins = 40) +
  ggtitle("histogram")+
  theme(plot.title = element_text(hjust = 0.5))
```



```
#Create density plots using ggplot2
ggplot(iris) +
  aes(x = Petal.Width, fill = Species)+
  geom_density( alpha = 0.3)
```



```
# Based on the above three plots, I can conclude that
# petal.width is not the same among three subspecies.
```

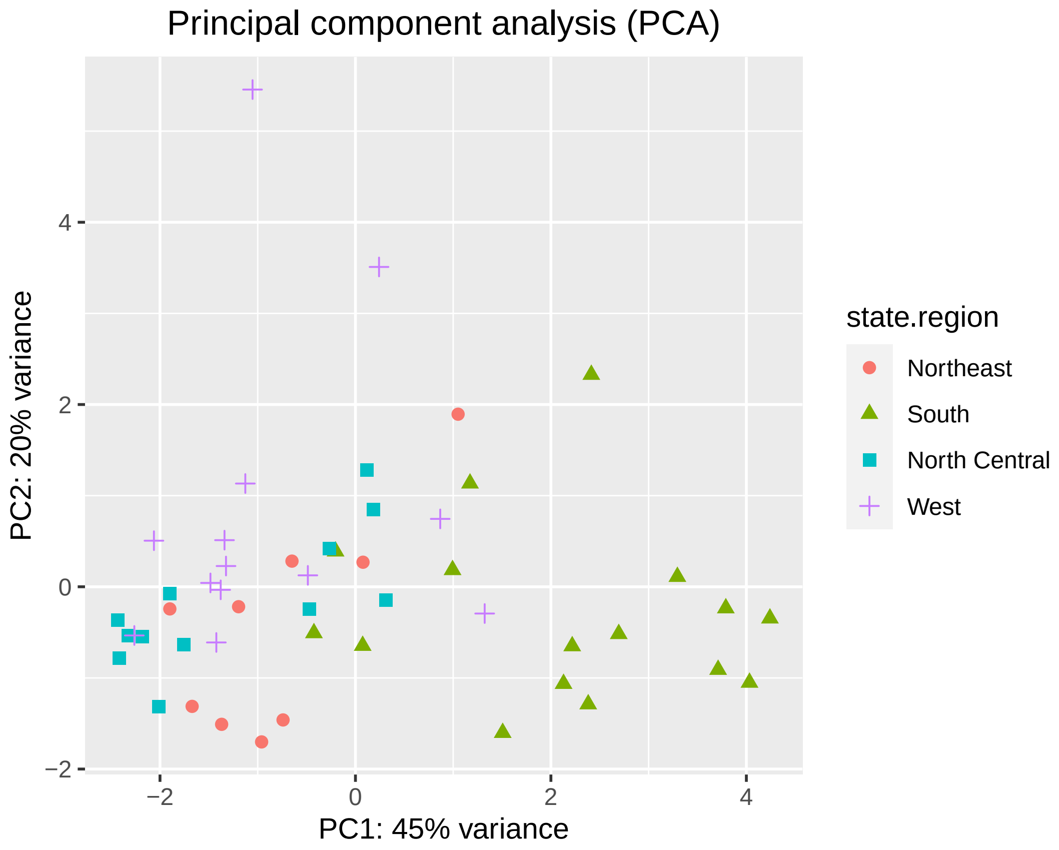
Exercise 2.6

Create PCA plot of the state.x77 data set (convert matrix to data frame). Use the state.region information to color code the states. Interpret your results. Hint: do not forget normalization using the scale option.

```
pca = prcomp(state.x77, scale = TRUE)
# "scale = TRUE" means that
# the data is normalized before conduction PCA.

pcaData <- as.data.frame(pca$x[,1:2])
pcaData <- cbind(pcaData, state.region)
colnames(pcaData) <- c("PC1", "PC2", "state.region")
library(ggplot2)


percentVar <- round(100 * summary(pca)$importance[2, 1:2], 0)
ggplot(pcaData, aes(PC1, PC2, color = state.region, shape = state.region)) +
         geom_point(size = 2) +
         xlab(paste0("PC1: ", percentVar[1], "% variance")) +
         ylab(paste0("PC2: ", percentVar[2], "% variance")) +
         ggtitle("Principal component analysis (PCA)") +
         theme(plot.title = element_text(hjust = 0.5), aspect.ratio = 1)
```



```
# The result is a projection of the 4-dimensional states.x77 data
# on 2-dimensional space using the first two principal components.
# As we can see in the PCA plot,
# It's hard to distinguish all state regions by using
# the first principal component and second principal component.
# We can say if PC1 > 1.75, then South.
# PC2 > 3, then West.
# Another state regions are clustered together in the PCA plot
# so it's hard to distinguish.
```