# Principal Component Analysis

Dennis Do, Abbey Rosario, Sierra Stein

12/7/2021

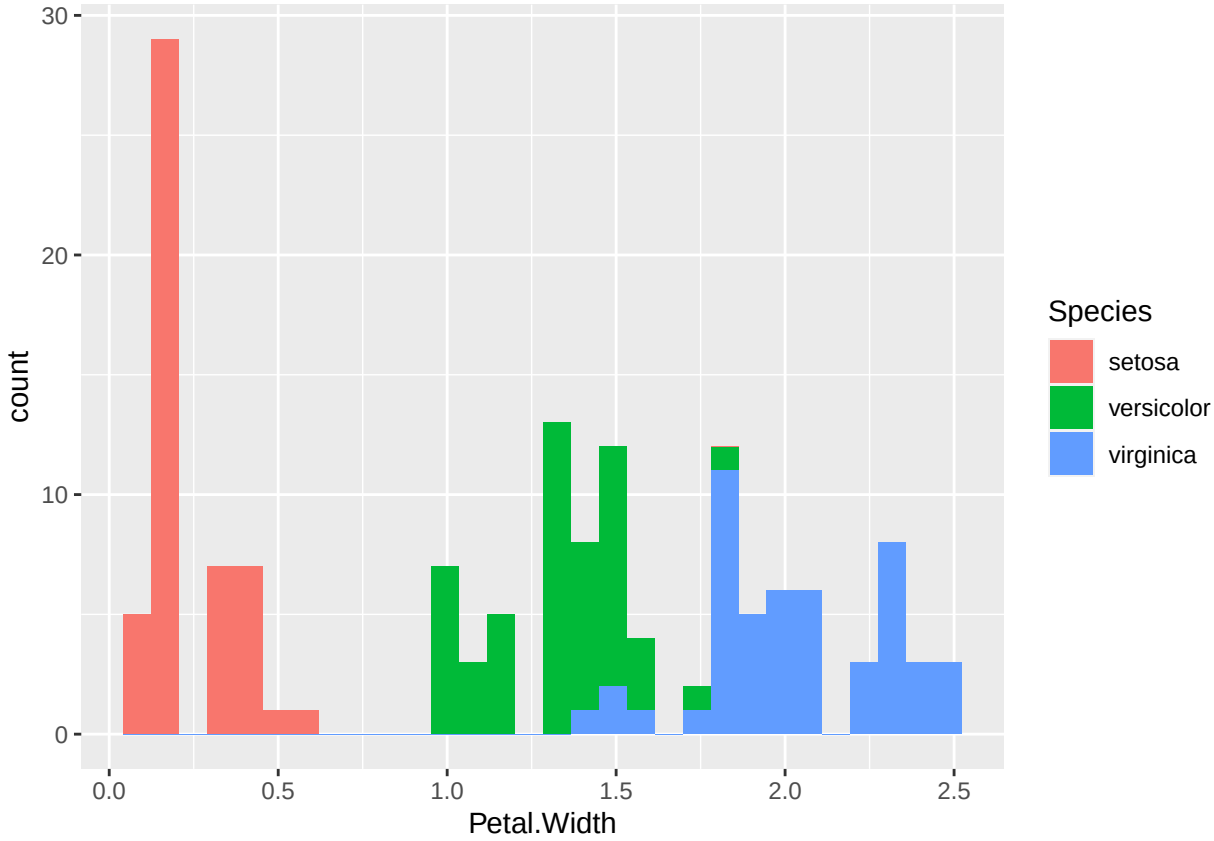
#EDA: Iris Dataset

```
summary(iris)
```

```
##   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
##  Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
##  1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
##  Median :5.800   Median :3.000   Median :4.350   Median :1.300
##  Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
##  3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
##  Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
##        Species
##  setosa    :50
##  versicolor:50
##  virginica :50
##
##
##
```

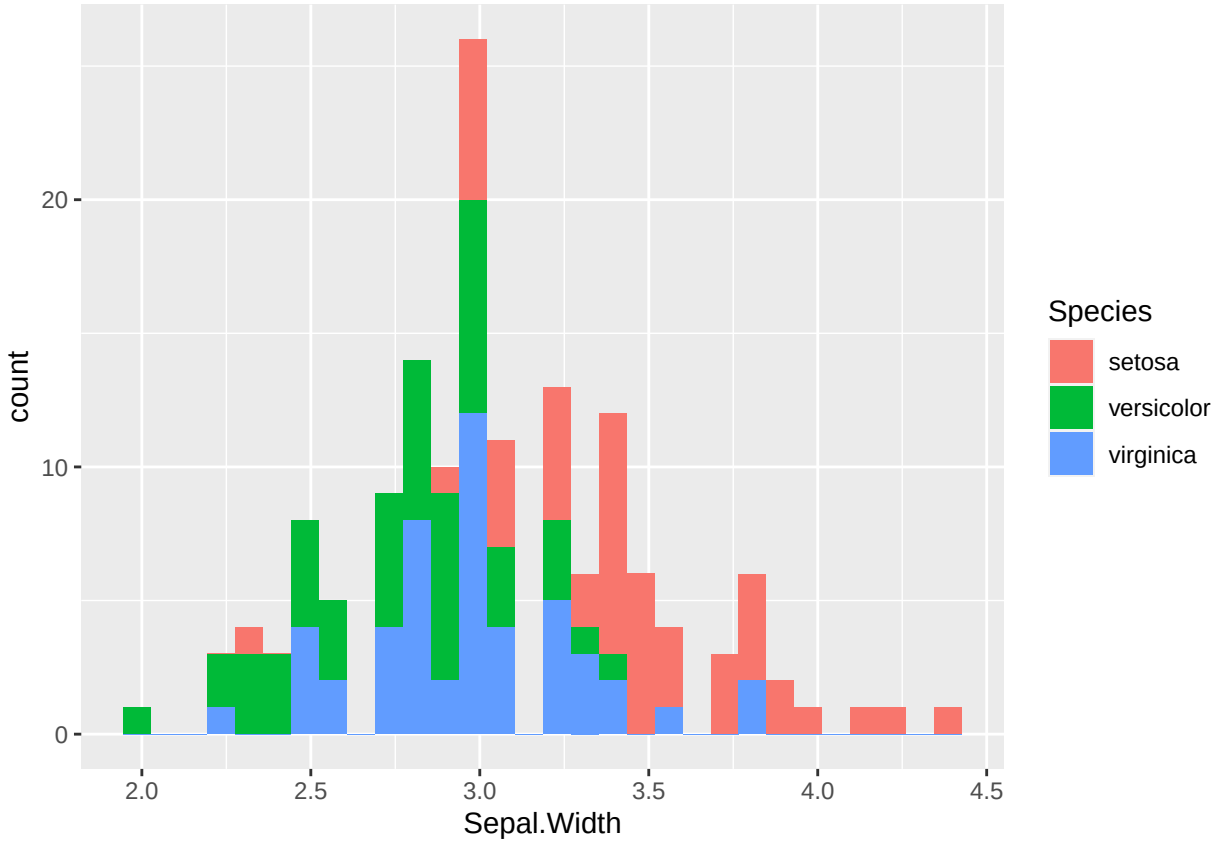
```
iris%>%
    ggplot(aes(x=Petal.Width, fill=Species))+
    geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



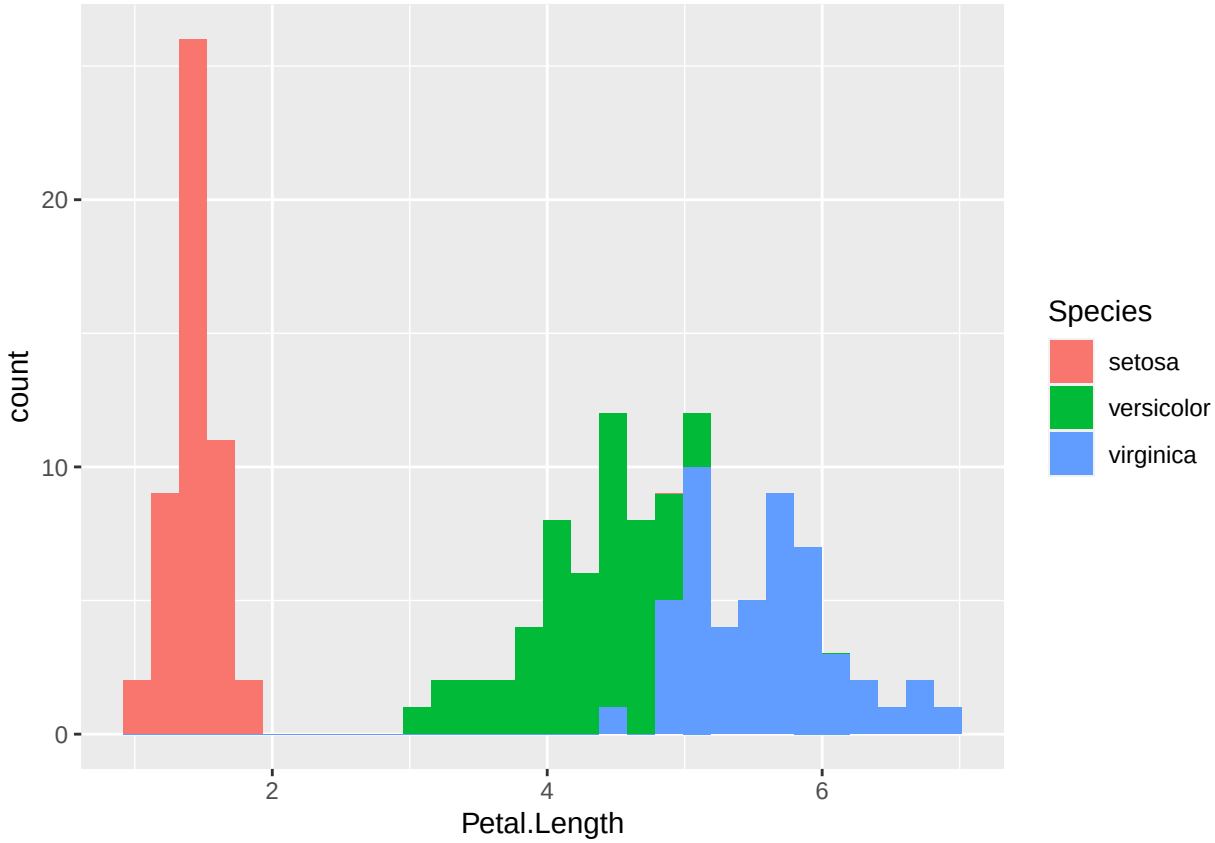

```
iris%>%
    ggplot(aes(x=Sepal.Width, fill=Species))+
    geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



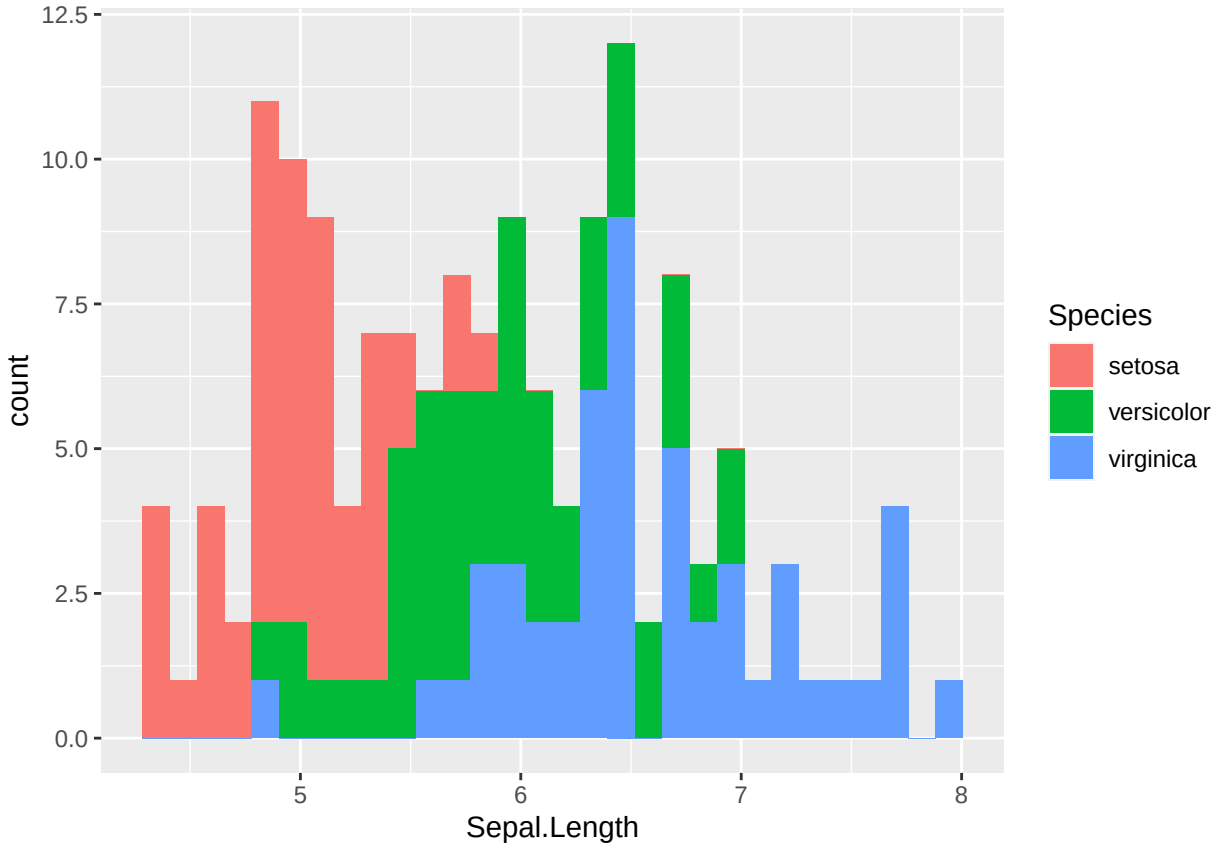

```
iris%>%
    ggplot(aes(x=Petal.Length, fill=Species))+
    geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

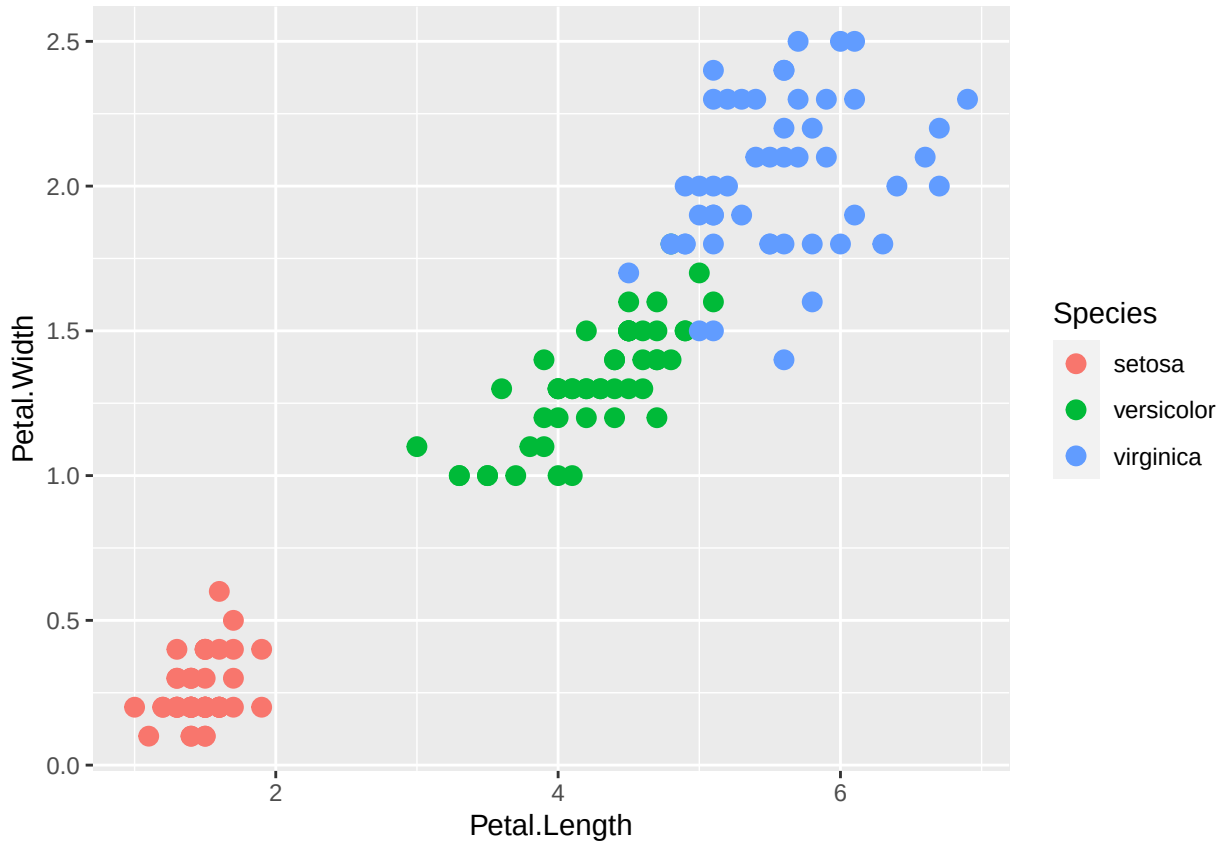



```
iris%>%
    ggplot(aes(x=Sepal.Length, fill=Species))+
    geom_histogram()
```

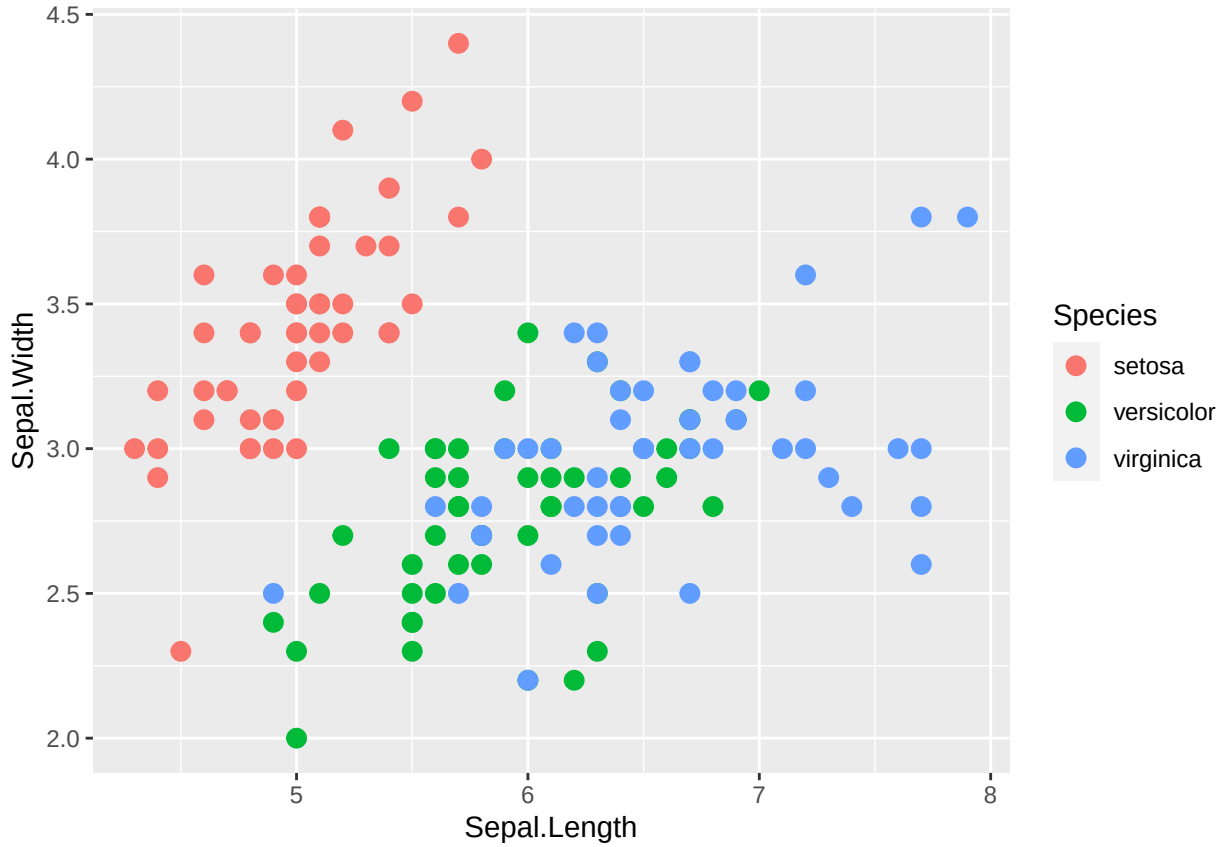
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
iris%>%
    ggplot(aes(x = Petal.Length, y = Petal.Width, colour = Species))+
    geom_point(size = 3)
```




```
iris%>%
    ggplot(aes(x = Sepal.Length, y = Sepal.Width, colour = Species))+
    geom_point(size = 3)
```




#Splitting Data

```
dt = sort(sample(nrow(iris), nrow(iris)*.7))
train<-iris[dt,]
test<-iris[-dt,]

summary(train)
```

```
##   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
##  Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
##  1st Qu.:5.200   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
##  Median :5.800   Median :3.000   Median :4.400   Median :1.400
##  Mean   :5.881   Mean   :3.053   Mean   :3.825   Mean   :1.223
##  3rd Qu.:6.500   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
##  Max.   :7.900   Max.   :4.400   Max.   :6.700   Max.   :2.500
##        Species
##  setosa    :32
##  versicolor:38
##  virginica :35
##
##
##
```

```
summary(test)
```

```
##   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
```

```
##  Min.   :4.500   Min.   :2.200   Min.   :1.200   Min.   :0.100
##  1st Qu.:5.000   1st Qu.:2.800   1st Qu.:1.500   1st Qu.:0.300
##  Median :5.600   Median :3.100   Median :4.000   Median :1.300
##  Mean   :5.756   Mean   :3.067   Mean   :3.602   Mean   :1.144
##  3rd Qu.:6.300   3rd Qu.:3.400   3rd Qu.:5.100   3rd Qu.:1.800
##  Max.   :7.700   Max.   :4.100   Max.   :6.900   Max.   :2.500
##        Species
##  setosa    :18
##  versicolor:12
##  virginica :15
##
##
##
```

#PCA

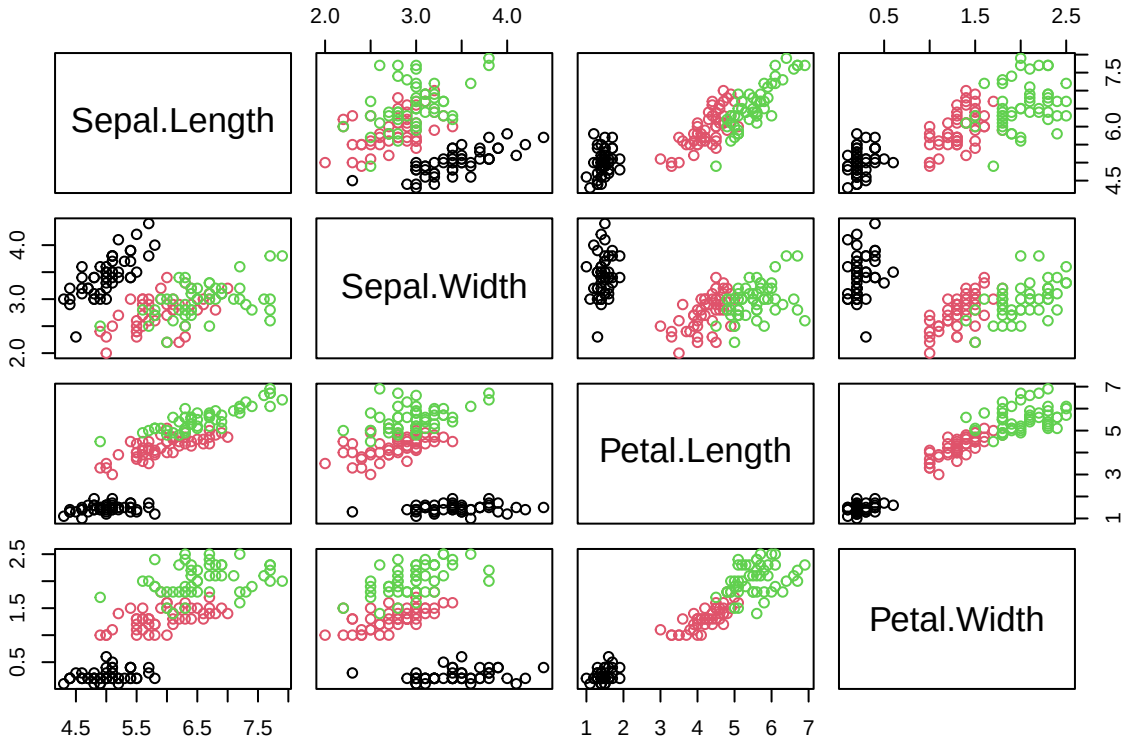
```
pc <- prcomp(train[,-5],center = T,scale. = T)
pc
```

```
## Standard deviations (1, .., p=4):
## [1] 1.7067707 0.9537263 0.3943775 0.1476700
##
## Rotation (n x k) = (4 x 4):
##                     PC1         PC2        PC3        PC4
## Sepal.Length  0.5112271 -0.41809359  0.7097605 -0.2451216
## Sepal.Width  -0.2878382 -0.90574388 -0.2779918  0.1396343
## Petal.Length  0.5806955 -0.01493703 -0.1508142  0.7998904
## Petal.Width   0.5644366 -0.06784295 -0.6294566 -0.5297103
```

```
summary(pc)
```

```
## Importance of components:
##                           PC1    PC2     PC3     PC4
## Standard deviation     1.7068 0.9537 0.39438 0.14767
## Proportion of Variance 0.7283 0.2274 0.03888 0.00545
## Cumulative Proportion  0.7283 0.9557 0.99455 1.00000
```

```
pairs(iris[,1:4],col=iris[,5])
```




```
pred <- predict(pc,train)
train_1 <- data.frame(pred,train[5])
pred1 <- predict(pc,test)
test_1 <- data.frame(pred1,test[5])

library(nnet)
set.seed(100)
mymodel <- multinom(Species~PC1 +PC2,data = train_1)
```

```
## # weights:  12 (6 variable)
## initial  value 115.354290
## iter  10 value 21.028921
## iter  20 value 19.989171
## iter  30 value 19.899086
## final  value 19.898081
## converged
```

```
summary(mymodel)
```

```
## Call:
## multinom(formula = Species ~ PC1 + PC2, data = train_1)
##
## Coefficients:
##            (Intercept)       PC1       PC2
```

```
## versicolor    7.805943 10.04322 4.282975
## virginica     2.049043 15.37869 4.916588
##
## Std. Errors:
##            (Intercept)      PC1      PC2
## versicolor    62.23453 52.19538 77.39694
## virginica     62.25177 52.21185 77.39917
##
## Residual Deviance: 39.79616
## AIC: 51.79616
```

#Confusion matrix

```
library(caret) #install caret package
```

```
## Warning: package 'caret' was built under R version 4.1.2
```

```
## Loading required package: lattice
```

```
prd <- predict(mymodel,train_1)
confusionMatrix(prd,train_1$Species)
```

```
## Confusion Matrix and Statistics
##
##             Reference
## Prediction   setosa versicolor virginica
##   setosa         32          0         0
##   versicolor      0         32         4
##   virginica       0          6        31
##
## Overall Statistics
##
##                Accuracy : 0.9048
##                  95% CI : (0.8318, 0.9534)
##     No Information Rate : 0.3619
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.8569
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: setosa Class: versicolor Class: virginica
## Sensitivity                 1.0000            0.8421           0.8857
## Specificity                 1.0000            0.9403           0.9143
## Pos Pred Value              1.0000            0.8889           0.8378
## Neg Pred Value              1.0000            0.9130           0.9412
## Prevalence                  0.3048            0.3619           0.3333
## Detection Rate              0.3048            0.3048           0.2952
## Detection Prevalence        0.3048            0.3429           0.3524
## Balanced Accuracy           1.0000            0.8912           0.9000
```

```r
prt <- predict(mymodel,test_1)
confusionMatrix(prt,test_1$Species)
```

```
## Confusion Matrix and Statistics
##
##             Reference
## Prediction   setosa versicolor virginica
##   setosa         18          0         0
##   versicolor      0         11         1
##   virginica       0          1        14
##
## Overall Statistics
##
##                Accuracy : 0.9556
##                  95% CI : (0.8485, 0.9946)
##     No Information Rate : 0.4
##     P-Value [Acc > NIR] : 2.842e-15
##
##                   Kappa : 0.9324
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: setosa Class: versicolor Class: virginica
## Sensitivity                    1.0            0.9167           0.9333
## Specificity                    1.0            0.9697           0.9667
## Pos Pred Value                 1.0            0.9167           0.9333
## Neg Pred Value                 1.0            0.9697           0.9667
## Prevalence                     0.4            0.2667           0.3333
## Detection Rate                 0.4            0.2444           0.3111
## Detection Prevalence           0.4            0.2667           0.3333
## Balanced Accuracy              1.0            0.9432           0.9500
```