Demonstration of reproducible research using Docker, knitr and rmarkdown

Shripad Sinari
Statistics Consulting Laboratory
BIO5, The University of Arizona
shripad@statlab.bio5.org

2019-06-12

Introduction

This document is used in a demonstration of reproducible research and some functionality from the *smisc* package.

If the docker environment setup has gone correctly and you regenerate this report in that environment then your should get a PDF file with the same content as the one in results folder named "demo-original-output.pdf". The only difference will be the date.

We will use the *iris* dataset in this example. The smisc package has isolated functions from *Hmisc* that provides a descriptive summary of each column in a dataset. A biplot using ggplot2 package is also implemented. For more information on smisc, see the github repo.

Analysis

We will explain each code chunk here:

1. Load the libraries required for the analysis

```r
library(tidyverse)
library(smisc)
```

2. Load the data set stored in a local file.

```r
iris_local <- read.csv("../data/iris_local.csv")
```

3. Produce the summary of the data.

```r
summarize(iris_local, "iris")
```

Summary of the iris Dataset
6 Variables 150 Observations

ID

```
  n  missing  distinct  Info  Mean    Gmd   .05    .10    .25    .50     .75     .90     .95
150        0       150     1  75.5  50.33  8.45  15.90  38.25  75.50  112.75  135.10  142.55

lowest :   1   2   3   4   5, highest: 146 147 148 149 150
```

Sepal.Length

```
  n  missing  distinct   Info   Mean     Gmd    .05    .10    .25    .50    .75    .90    .95
150        0        35  0.998  5.843  0.9462  4.600  4.800  5.100  5.800  6.400  6.900  7.255

lowest : 4.3 4.4 4.5 4.6 4.7, highest: 7.3 7.4 7.6 7.7 7.9
```

Sepal.Width

```
  n  missing  distinct   Info   Mean     Gmd    .05    .10    .25    .50    .75    .90    .95
150        0        23  0.992  3.057  0.4872  2.345  2.500  2.800  3.000  3.300  3.610  3.800

lowest : 2.0 2.2 2.3 2.4 2.5, highest: 3.9 4.0 4.1 4.2 4.4
```

Petal.Length

```
  n  missing  distinct   Info   Mean    Gmd   .05   .10   .25   .50   .75   .90   .95
150        0        43  0.998  3.758  1.979  1.30  1.40  1.60  4.35  5.10  5.80  6.10

lowest : 1.0 1.1 1.2 1.3 1.4, highest: 6.3 6.4 6.6 6.7 6.9
```

Petal.Width

```
  n  missing  distinct  Info   Mean     Gmd  .05  .10  .25  .50  .75  .90  .95
150        0        22  0.99  1.199  0.8676  0.2  0.2  0.3  1.3  1.8  2.2  2.3

lowest : 0.1 0.2 0.3 0.4 0.5, highest: 2.1 2.2 2.3 2.4 2.5
```

Species

```
  n  missing  distinct
150        0         3

Value          setosa versicolor  virginica
Frequency          50         50         50
Proportion      0.333      0.333      0.333
```

4. Analyse which variables provide the most distinguishing features of the different species of iris.

```
pca <- iris_local %>%
    dplyr::select(-c(ID,Species)) %>%
    as.matrix() %>%
    prcomp()

biplot <- PCbiplot(PC = pca
       , d = iris_local
       , colors = c("#fc8d59","#ffffbf","#91bfdb")
       , legend_t = "Species"
       , varnames = colnames(iris_local)[-grep("ID|Species", colnames(iris_local))]
       , labels = F
       , title = "")
```

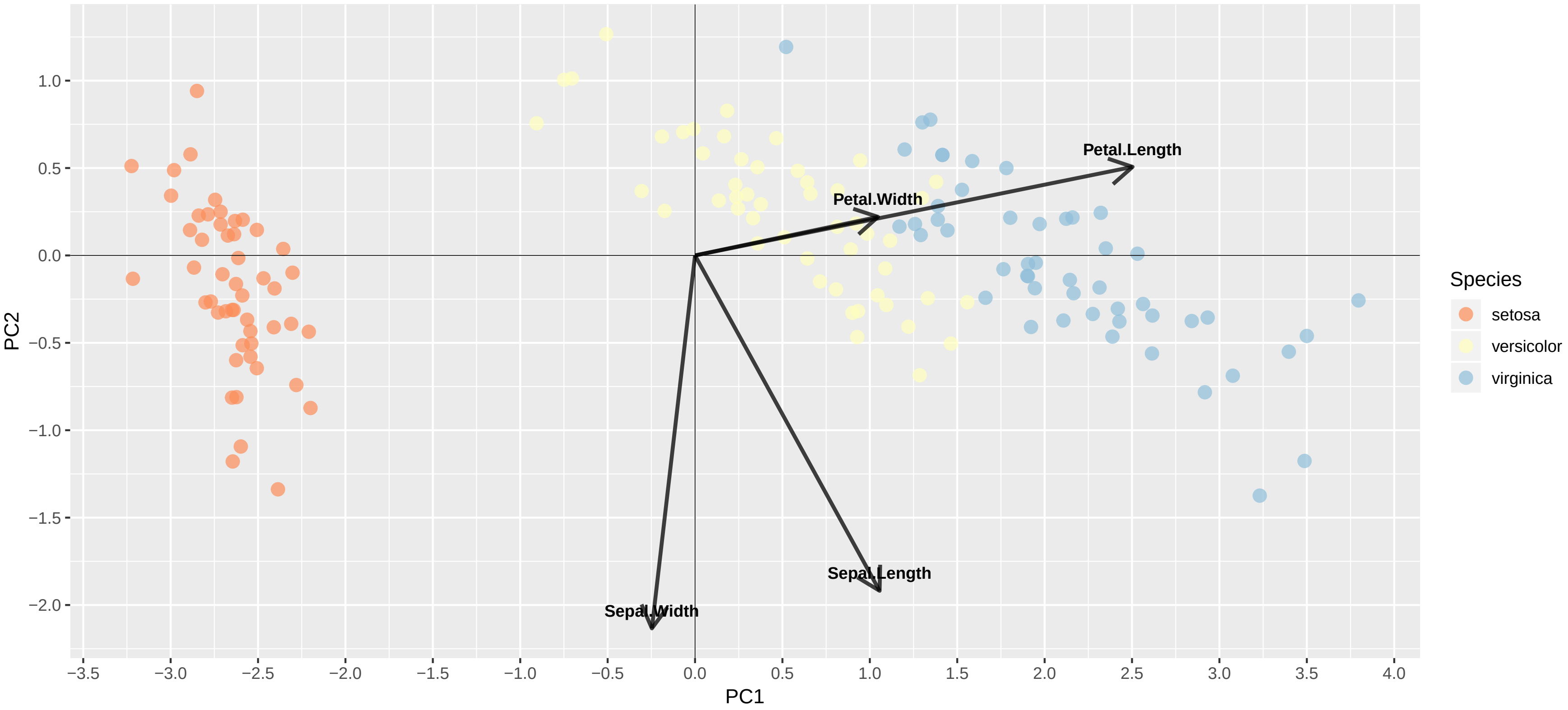

Figure 1: Biplot of the iris data. Petals (length or width) is the distinguishing feature of an iris species rather than its sepal.