ADS501-01: Assignment 1.2 Dataset Visualization

Gabi Rivera

10Sep2022

```
#Library(s)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.8     v dplyr   1.0.10
## v tidyr   1.2.1     v stringr 1.4.1
## v readr   2.1.2     v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(skimr)
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
library(ggplot2)
library(ggpubr)

##Part 1
#Create a scatter plot using the Iris Dataset comparing
  #Petal Length to Petal Width by Species or
  #Sepal Length to Sepal Width by species.

#Import data
iris = read.csv("Iris.csv", header = TRUE, sep = ",")
head(iris)
```

```
##   Id SepalLengthCm SepalWidthCm PetalLengthCm PetalWidthCm     Species
## 1  1          5.1          3.5          1.4         0.2 Iris-setosa
## 2  2          4.9          3.0          1.4         0.2 Iris-setosa
## 3  3          4.7          3.2          1.3         0.2 Iris-setosa
## 4  4          4.6          3.1          1.5         0.2 Iris-setosa
## 5  5          5.0          3.6          1.4         0.2 Iris-setosa
## 6  6          5.4          3.9          1.7         0.4 Iris-setosa
```

```
#General informations
skim(iris)
```

Data summary

Name	iris
Number of rows	150
Number of columns	6

Column type frequency:

character	1
numeric	5

Group variables: None

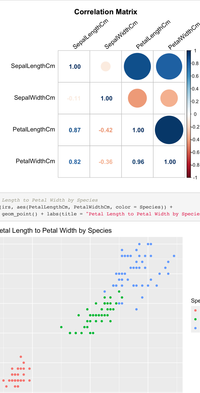
Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
Species	0	1	11	15	0	3	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
Id	0	1	75.50	43.45	1.0	38.25	75.50	112.75	150.0	▇▇▇▇▇
SepalLengthCm	0	1	5.84	0.83	4.3	5.10	5.80	6.40	7.9	▆▇▅▃▂
SepalWidthCm	0	1	3.05	0.43	2.0	2.80	3.00	3.30	4.4	▁▆▇▂▁
PetalLengthCm	0	1	3.76	1.76	1.0	1.60	4.35	5.10	6.9	▇▁▂▅▃
PetalWidthCm	0	1	1.20	0.76	0.1	0.30	1.30	1.80	2.5	▇▁▃▅▃

```
#Correlation
n_iris = subset(iris, select = c("SepalLengthCm", "SepalWidthCm",
                                 "PetalLengthCm", "PetalWidthCm"))
c_iris = cor(n_iris, y = NULL, use = "complete.obs", method = "pearson")
corrplot.mixed(c_iris, title = "Correlation Matrix",
               tl.col = "black", tl.srt =45, mar=c(1,1,2,1),
               tl.pos = 'lt', diag = 'l')
```

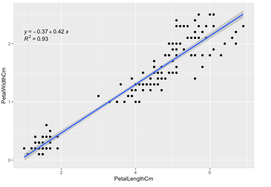



```
#Petal Length to Petal Width by Species
ggplot(iris, aes(PetalLengthCm, PetalWidthCm, color = Species)) +
  geom_point() + labs(title = "Petal Length to Petal Width by Species")
```



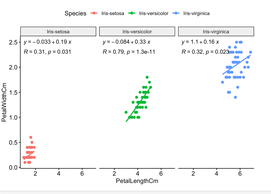
```
#Petal cumulative linear regression
ggplot(iris, aes(PetalLengthCm, PetalWidthCm)) +
  geom_point() + stat_smooth(method = 'lm') +
  stat_regline_equation(label.y = 2.2, aes(label = ..eq.label..)) +
  stat_regline_equation(label.y = 2.1, aes(label = ..rr.label..))
```

```
## `geom_smooth()` using formula 'y ~ x'
```

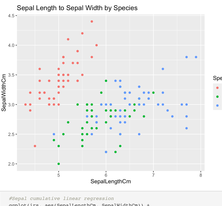



```
#Petal species linear regression
ggscatter(iris, x = "PetalLengthCm", y = "PetalWidthCm",color = "Species",
  add = "reg.line") + facet_wrap(~Species) + stat_cor(label.y = 2.3) +
  stat_regline_equation(label.y = 2.5)
```

```
## `geom_smooth()` using formula 'y ~ x'
```

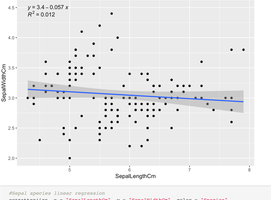



```
#Sepal Length to Sepal Width by species
ggplot(iris, aes(SepalLengthCm, SepalWidthCm, color = Species)) +
  geom_point() + labs(title = "Sepal Length to Sepal Width by Species")
```



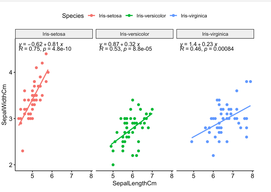

```
#Sepal cumulative linear regression
ggplot(iris, aes(SepalLengthCm, SepalWidthCm)) +
  geom_point() + stat_smooth(method = 'lm') +
  stat_regline_equation(label.y = 4.5, aes(label = ..eq.label..)) +
  stat_regline_equation(label.y = 4.4, aes(label = ..rr.label..))
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
#Sepal species linear regression
ggscatter(iris, x = "SepalLengthCm", y = "SepalWidthCm", color = "Species",
  add = "reg.line") + facet_wrap(~Species) + stat_cor(label.y = 4.5) +
  stat_regline_equation(label.y = 4.4)
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
##Part II
#Interpret your scatter plot.
  #At a minimum your interpretation should describe
  #if there is any relationship between the variables and amongst the species.
  #Identify how strong the relationship appears.
  #If there are any major outliers, and if there are any notable findings.

#Petal Length to Petal Width by Species:
  #Petal length and petal width both measured in centimeters have a correlation
  #score of 96%. The r-squared of petal length against petal width is at 93%.
  #All these suggest a strong positive correlation between the two variables.
  #The cumulative linear regression visually supports the calculated relationship
  #as well. The individual data are close to regression line. As a whole, petal
  #length has the propensity to predict the petal width.
  #Looking at individual species however, it seems that only Iris-versicolor
  #score a high positive correlation at 79% r-squared with 5.5 alpha. The other
  #two species are at less than 33% r-squared. Iris-setosa and Iris-virginica
  #seeme to have a weak correlation pattern at greater than 30 observations.
  #Finally, the samples of both species have wider spread along
  #each of their regression line compared to Iris-versicolor. But noticeably,
  #each cluster among their own species precisely. The clustering can be used
  #for categorical differentiation across or within species.

#Sepal Length to Sepal Width by Species:
  #Sepal length and sepal width both measured in centimeters have a negative
  #correlation score of -11%. The r-squared of sepal length against sepal width
  #is at 1.2%. All these suggest a weak negative correlation between the two
  #variables.The cumulative linear regression visually supports the calculated
  #relationship with the individual data widely spread along the regression line.
  #There doesn't seem to be any pattern as a whole.
  #Within species however, shows relatively stronger relationships between sepal
  #length and sepal width compared to the cumulative observation. Iris-setosa
  #has 75% r-squared followed by Iris-versicolor at 53% then of Iris-virginica
  #at 46%. Just like in petals, the sepal clustering within species can be used
  #to differentiate each species using the two variables.
```