TECNICAS DE ANALISIS MULTIVARIANTE EJERCICIOS TEMA 1

franmarq@gmail.com (mailto:franmarq@gmail.com)

2022-10-02

Tarea 1. Abrir el fichero de datos IRIS que estan en R base.

Para hacer la carga del archivo en la sesión de trabajo usamos el comando ‘data’. Hacemos inicialmente un examen visual de la estructura y contenido del mismo utilizando el comando ‘str’

```
data(iris)
str(iris)
```

```
## 'data.frame':    150 obs. of  5 variables:
##  $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
##  $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
##  $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
##  $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
##  $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
```

En resumen, el conjunto de datos está formado por 150 observaciones de flores de la planta iris, 5 Variables de las cuales 4 son numericas y se refieren a distintas mediciones acerca de las plantas y una variable cualtitativa/Factor, la cual presenta valores distintos e indica en tipo de planta a la cual se refiere la observacion: virginica, setosa y versicolor.

Tarea 2. Obtener estadisticos descriptivos de cada variable.

Para la obtención de los estadísticos principales, usaremos el comando ‘summary’

```
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
##
##
##
```

Obtenemos valores de rango y posicion, como el promedio y la mediana de cada una de las variables de largo y ancho tanto de los sépalos como de los pétalos de las plantas. Podemos comprobar como cada una de las especiales cuenta con 50 observaciones. En esta parte del análisis es importante tambien determinar si tenemos valores perdidos. Para ello usamos el siguiente comando

```
colSums(is.na(iris))
```

```
## Sepal.Length  Sepal.Width Petal.Length  Petal.Width      Species
##            0            0            0            0            0
```

El resultado indica que no tenemos valores perdidos, por lo que podremos usar el 100 del archivo.

Tarea 3. Obtener estadisticos descriptivos de cada variable segun la especie.

Para esta actividad, primero se van tres conjuntos de datos, uno por cada especie a partir del archivo 'Iris'. Luego aplicaremos la función 'summary' en cada uno de ellos y así obtendremos los estadísticos descriptivos.

```
irisVer <- subset(iris, Species == "versicolor")
irisSet <- subset(iris, Species == "setosa")
irisVir <- subset(iris, Species == "virginica")

summary(irisVer)
```

```
##   Sepal.Length    Sepal.Width     Petal.Length   Petal.Width          Species
##  Min.   :4.900   Min.   :2.000   Min.   :3.00   Min.   :1.000   setosa    : 0
##  1st Qu.:5.600   1st Qu.:2.525   1st Qu.:4.00   1st Qu.:1.200   versicolor:50
##  Median :5.900   Median :2.800   Median :4.35   Median :1.300   virginica : 0
##  Mean   :5.936   Mean   :2.770   Mean   :4.26   Mean   :1.326
##  3rd Qu.:6.300   3rd Qu.:3.000   3rd Qu.:4.60   3rd Qu.:1.500
##  Max.   :7.000   Max.   :3.400   Max.   :5.10   Max.   :1.800
```

```
summary(irisSet)
```

```
##   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
##  Min.   :4.300   Min.   :2.300   Min.   :1.000   Min.   :0.100
##  1st Qu.:4.800   1st Qu.:3.200   1st Qu.:1.400   1st Qu.:0.200
##  Median :5.000   Median :3.400   Median :1.500   Median :0.200
##  Mean   :5.006   Mean   :3.428   Mean   :1.462   Mean   :0.246
##  3rd Qu.:5.200   3rd Qu.:3.675   3rd Qu.:1.575   3rd Qu.:0.300
##  Max.   :5.800   Max.   :4.400   Max.   :1.900   Max.   :0.600
##        Species
##  setosa    :50
##  versicolor: 0
##  virginica : 0
##
##
##
```

```
summary(irisVir)
```

```
##   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
##  Min.   :4.900   Min.   :2.200   Min.   :4.500   Min.   :1.400
##  1st Qu.:6.225   1st Qu.:2.800   1st Qu.:5.100   1st Qu.:1.800
##  Median :6.500   Median :3.000   Median :5.550   Median :2.000
##  Mean   :6.588   Mean   :2.974   Mean   :5.552   Mean   :2.026
##  3rd Qu.:6.900   3rd Qu.:3.175   3rd Qu.:5.875   3rd Qu.:2.300
##  Max.   :7.900   Max.   :3.800   Max.   :6.900   Max.   :2.500
##        Species
##  setosa    : 0
##  versicolor: 0
##  virginica :50
##
##
##
```

Con base en resultado podemos hacer una comparaciones iniciales entre las mediciones de las tres especies. Sobre la variable 'Sepal.length': vemos como la especie que en promedio registró mayores valores fue Virginica, La de menor valor promedio registrado fue la especie Setosa. Sobre la variable 'Sepal.width': vemos como la especie que en promedio registró mayores valores fue Setosa, La de menor valor promedio registrado fue la especie Versicolor. Sobre la variable 'Petal.length': vemos como la especie que en promedio registró mayores

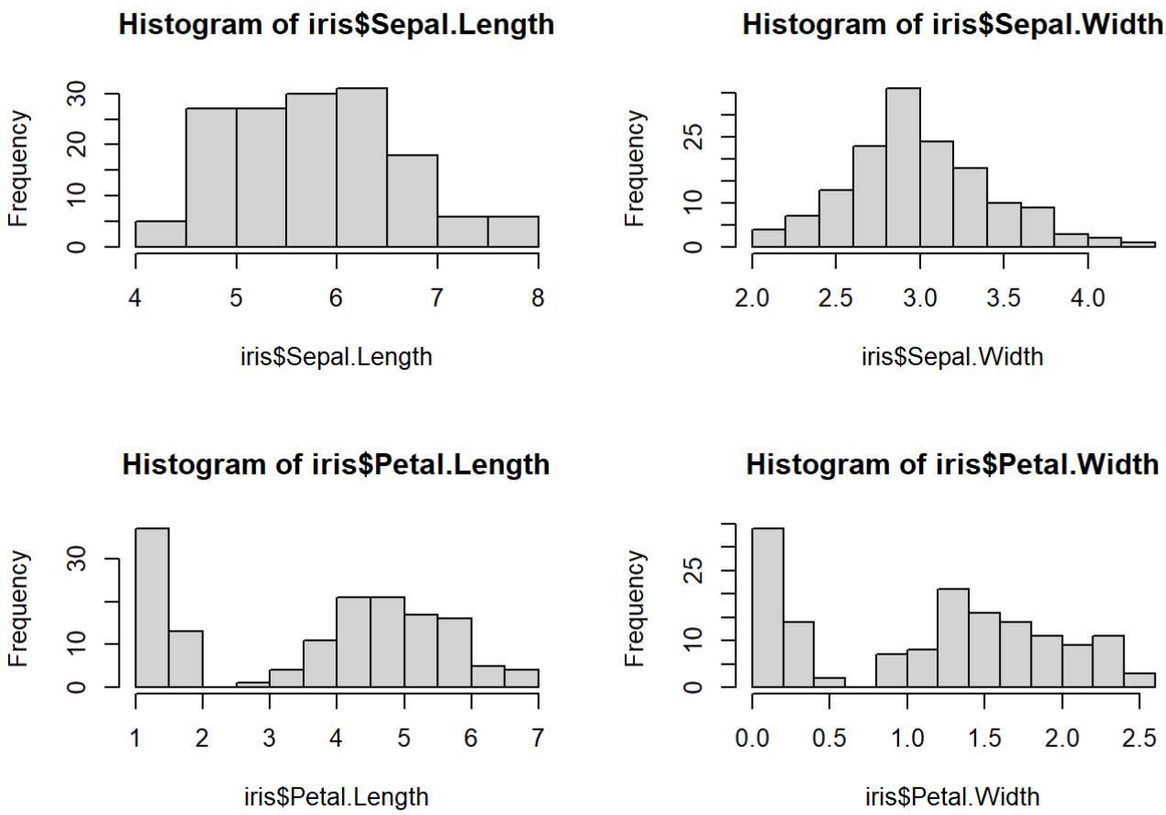
valores fue Virginica, La de menor valor promedio registrado fue la especie Setosa. Sobre la variable ‘Petal.width’: vemos como la especie que en promedio registró mayores valores fue Virginica, La de menor valor promedio registrado fue la especie Setosa.

Tarea 4. Obtener representaciones graficas de cada variable de forma individual y por tipo de planta.

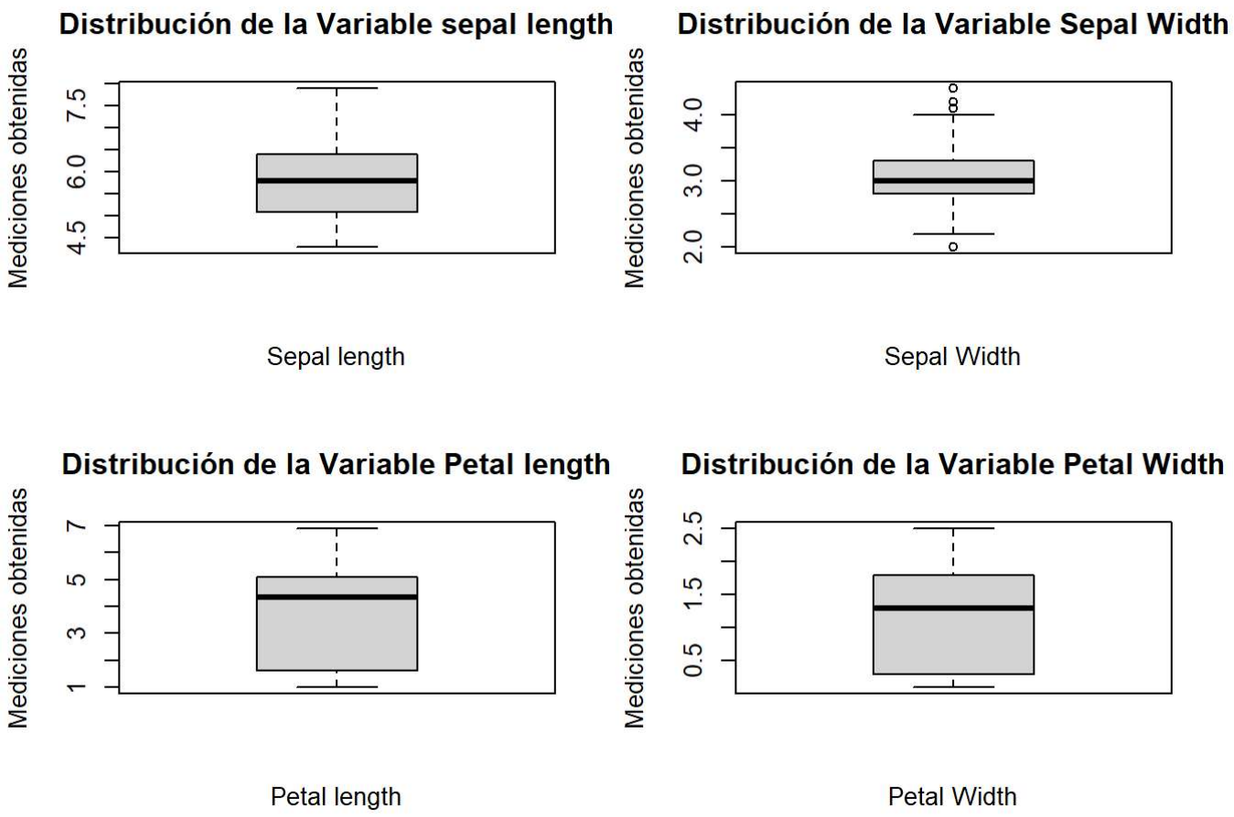
Se usarán Histogramas de frecuencias y gráficos de cajas para las representaciones gráficas individuales y por especies.

Graficos individuales

```
par(mfrow = c(2,2))
hist(iris$Sepal.Length)
hist(iris$Sepal.Width)
hist(iris$Petal.Length)
hist(iris$Petal.Width)
```




```
par(mfrow = c(1,1))

par(mfrow = c(2,2))

boxplot(iris$Sepal.Length,xlab="Sepal length", ylab="Mediciones obtenidas", main="Distribución d
e la Variable sepal length")

boxplot(iris$Sepal.Width,xlab="Sepal Width", ylab="Mediciones obtenidas", main="Distribución de
 la Variable Sepal Width")

boxplot(iris$Petal.Length,xlab="Petal length", ylab="Mediciones obtenidas", main="Distribución d
e la Variable Petal length")

boxplot(iris$Petal.Width,xlab="Petal Width", ylab="Mediciones obtenidas", main="Distribución de
 la Variable Petal Width")
```




```
par(mfrow = c(1,1))
```

Gráficos agrupados por tipo de planta

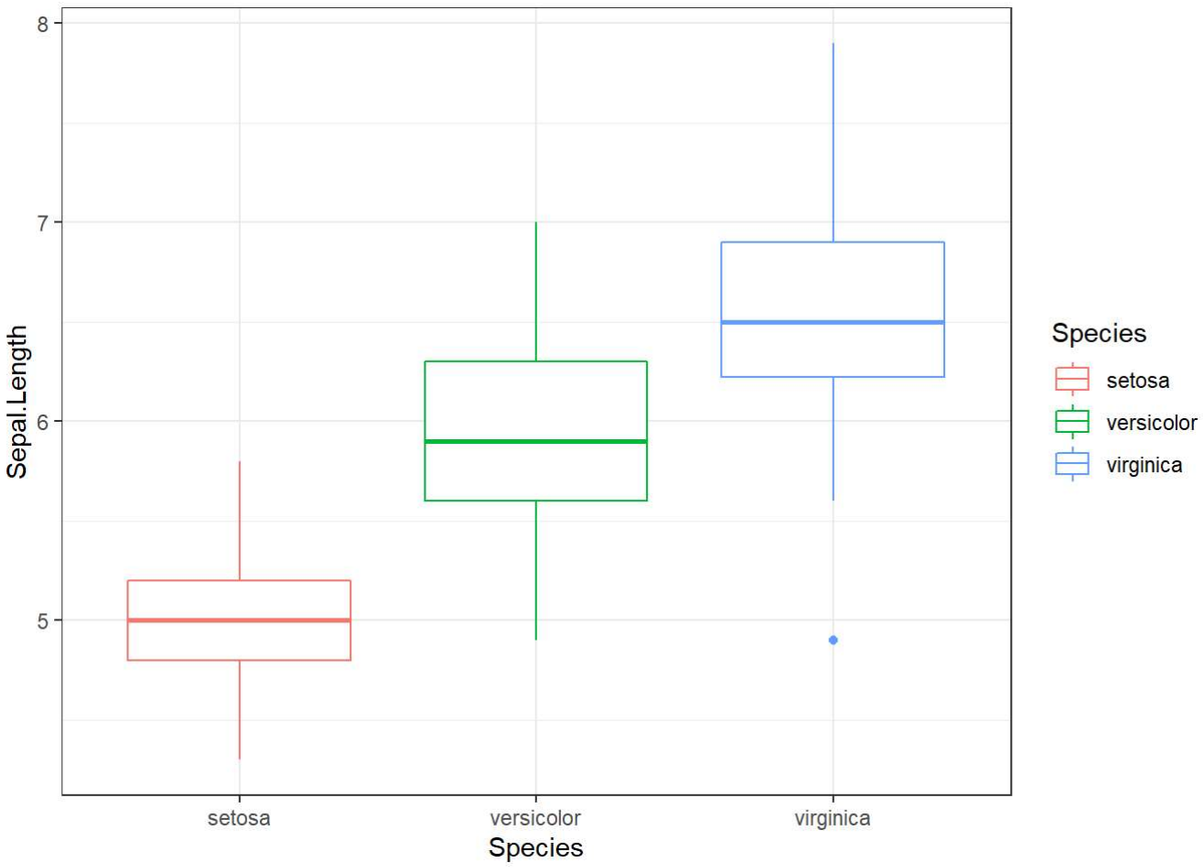
Para los gráficos de Cajas usaremos el paquete ggplot2

```
require(ggplot2)
```

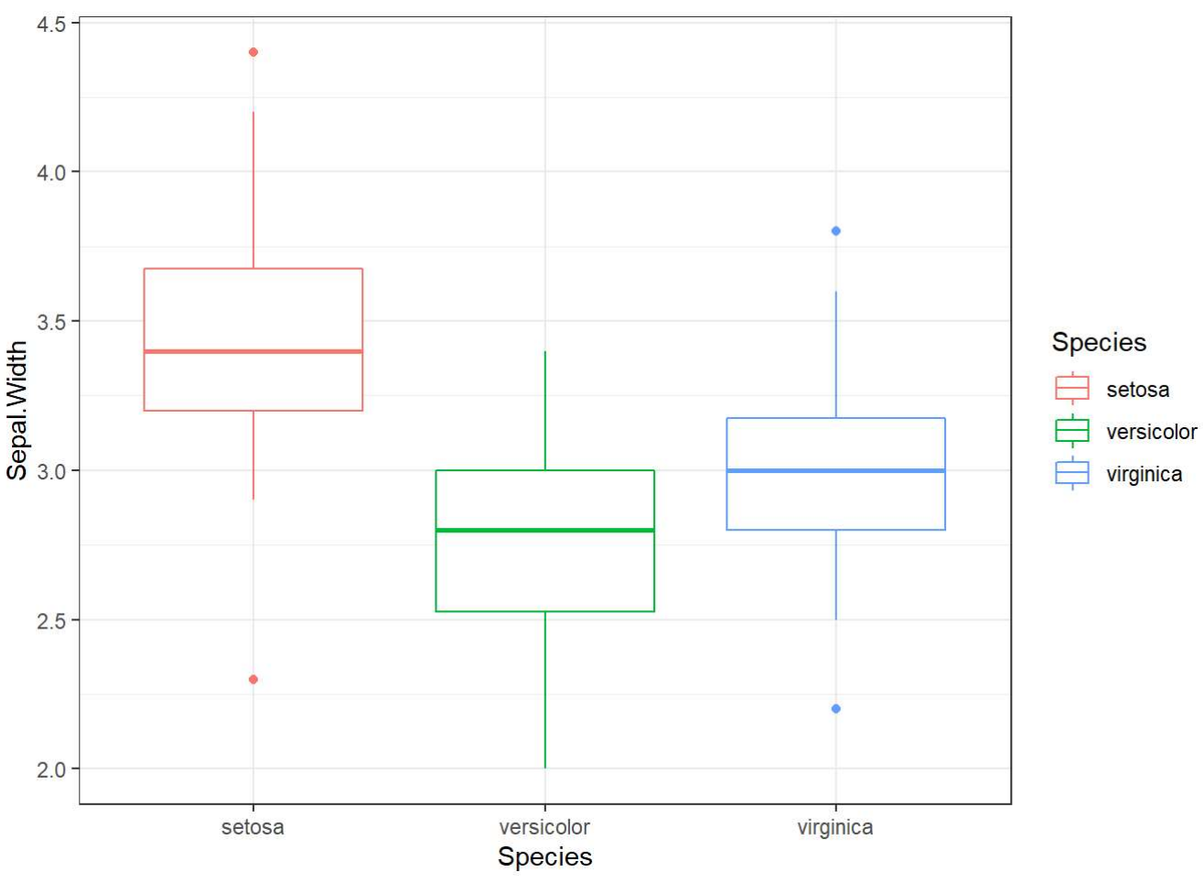
```
## Loading required package: ggplot2
```

```
## Warning: package 'ggplot2' was built under R version 4.1.3
```

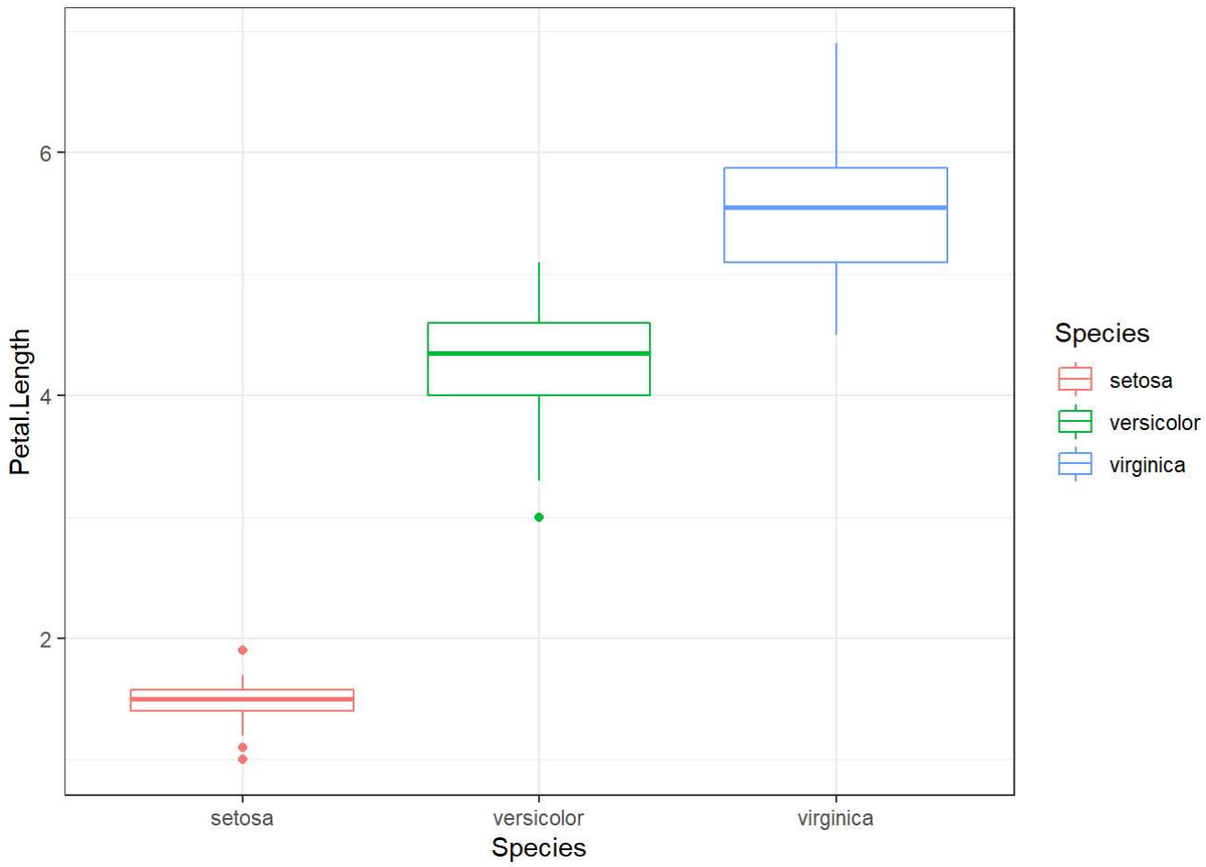
```
par(mfrow = c(2,2))

ggplot(data = iris, aes(x = Species, y = Sepal.Length, color = Species)) +
  geom_boxplot() +
  theme_bw()
```



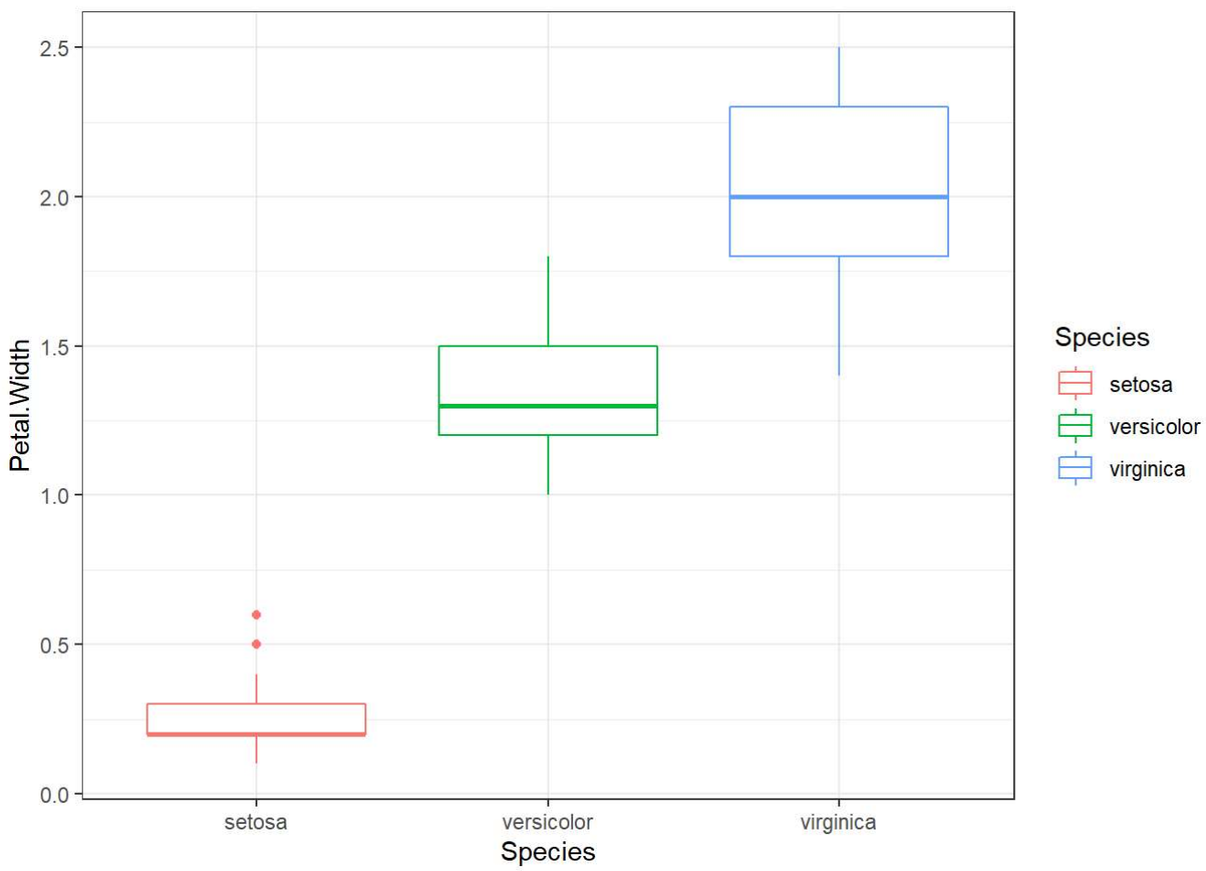

```
ggplot(data = iris, aes(x = Species, y = Sepal.Width, color = Species)) +
  geom_boxplot() +
  theme_bw()
```




```
ggplot(data = iris, aes(x = Species, y = Petal.Length, color = Species)) +
  geom_boxplot() +
  theme_bw()
```




```
ggplot(data = iris, aes(x = Species, y = Petal.Width, color = Species)) +
  geom_boxplot() +
  theme_bw()
```




```
par(mfrow = c(1,1))
```

Tarea 5. Comprobar la hipotesis estudiadas para cada variable y segun el tipo de especie.

Se hará una comprobación de las hipótesis comunes a examinar en el análisis multivariante

Hipotesis 1: Normalidad

Se inicia con pruebas de normalidad individual y luego una comprobación del conjunto completo de variables.

Evaluación individual:

```
require(pastecs)
```

```
## Loading required package: pastecs
```

```
## Warning: package 'pastecs' was built under R version 4.1.3
```

```
round(stat.desc(iris[,c("Sepal.Length","Sepal.Width","Petal.Length","Petal.Width")],basic=FALSE,
norm=TRUE),digits=3)
```

```
##              Sepal.Length Sepal.Width Petal.Length Petal.Width
## median              5.800       3.000        4.350       1.300
## mean                5.843       3.057        3.758       1.199
## SE.mean             0.068       0.036        0.144       0.062
## CI.mean.0.95        0.134       0.070        0.285       0.123
## var                 0.686       0.190        3.116       0.581
## std.dev             0.828       0.436        1.765       0.762
## coef.var            0.142       0.143        0.470       0.636
## skewness            0.309       0.313       -0.269      -0.101
## skew.2SE            0.779       0.789       -0.680      -0.255
## kurtosis           -0.606       0.139       -1.417      -1.358
## kurt.2SE           -0.770       0.176       -1.800      -1.725
## normtest.W          0.976       0.985        0.876       0.902
## normtest.p          0.010       0.101        0.000       0.000
```

En este caso, los valores superiores, de skew.2SE y/o kurt.2se indican cierta proximidad en valor absoluto a 1, acercamiento criterio de normalidad, lo que sugiere que la hipótesis se cumple, excepto para las variables Petal.Length y Petal.Width.

Se usará tambien los gráficos QQ-plot para valorar la normalidad individual.

```
require(car)
```

```
## Loading required package: car
```

```
## Loading required package: carData
```

```
par(mfrow=c(2,2))
qqPlot(iris$Sepal.Length)
```

```
## [1] 132 118
```

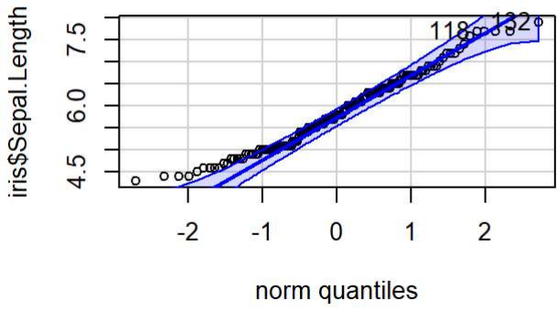
```
qqPlot(iris$Sepal.Width)
```

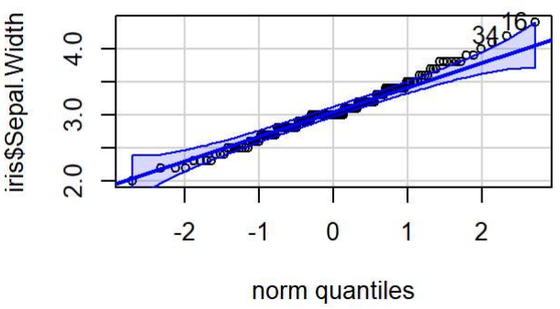
```
## [1] 16 34
```

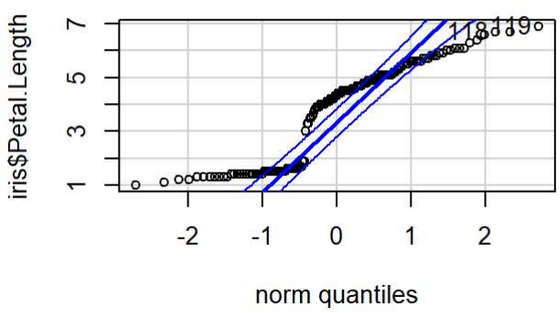
```
qqPlot(iris$Petal.Length)
```

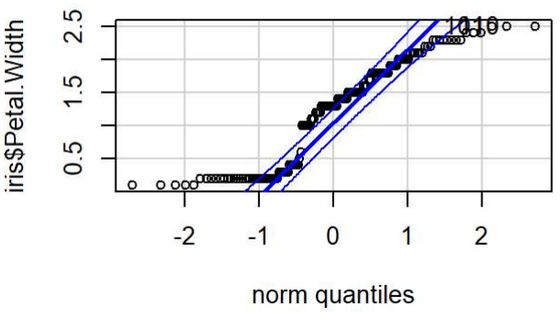
```
## [1] 119 118
```

```
qqPlot(iris$Petal.Width)
```










```
## [1] 101 110
```

```
par(mfrow=c(1,1))
```

La forma de los gráficos corrobora los comportamientos observados en el análisis anterior. Las variables Sepal se comportan de acuerdo a un criterio nromal.

Para evaluar normalidad en el caso multivariante usaremos el test de shapiro-wilk el cual es recomendado en conjuntos de hasta 50 observaciones.

```
shapiro.test(iris$Sepal.Length)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  iris$Sepal.Length
## W = 0.97609, p-value = 0.01018
```

```
shapiro.test(iris$Sepal.Width)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  iris$Sepal.Width
## W = 0.98492, p-value = 0.1012
```

```
shapiro.test(iris$Petal.Length)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  iris$Petal.Length
## W = 0.87627, p-value = 7.412e-10
```

```
shapiro.test(iris$Petal.Width)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  iris$Petal.Width
## W = 0.90183, p-value = 1.68e-08
```

Los resultados del test muestran que sólo para el caso de la variable Sepal.width no podemos rechazar la hipótesis nula: los datos siguen una distribución Normal. Para el resto de variables no se cumple el criterio de normalidad.

Normalidad Multivariante:

Para comprobar la normalidad multivariante usaremos el paquete MNV que nos permita ejecutar el test de Mardia, Henze-Zirkler o Royston. En nuestrao caso aplicaremos el test de Royston

```
require(MVN)
```

```
## Loading required package: MVN
```

```
## Warning: package 'MVN' was built under R version 4.1.3
```

```
mvn(iris[,c("Sepal.Length","Sepal.Width","Petal.Length","Petal.Width")],mvnTest = "royston",univ
ariateTest = "Lillie")
```

```
## $multivariateNormality
##      Test        H      p value MVN
## 1 Royston 50.39667 3.098229e-11  NO
##
## $univariateNormality
##                            Test     Variable Statistic   p value Normality
## 1 Lilliefors (Kolmogorov-Smirnov) Sepal.Length    0.0887  0.0058      NO
## 2 Lilliefors (Kolmogorov-Smirnov) Sepal.Width     0.1057   3e-04      NO
## 3 Lilliefors (Kolmogorov-Smirnov) Petal.Length    0.1982  <0.001      NO
## 4 Lilliefors (Kolmogorov-Smirnov) Petal.Width     0.1728  <0.001      NO
##
## $Descriptives
##                n     Mean   Std.Dev Median Min Max 25th 75th       Skew
## Sepal.Length 150 5.843333 0.8280661   5.80 4.3 7.9  5.1  6.4  0.3086407
## Sepal.Width  150 3.057333 0.4358663   3.00 2.0 4.4  2.8  3.3  0.3126147
## Petal.Length 150 3.758000 1.7652982   4.35 1.0 6.9  1.6  5.1 -0.2694109
## Petal.Width  150 1.199333 0.7622377   1.30 0.1 2.5  0.3  1.8 -0.1009166
##                Kurtosis
## Sepal.Length -0.6058125
## Sepal.Width   0.1387047
## Petal.Length -1.4168574
## Petal.Width  -1.3581792
```

El resultado de la prueba indica que no hay evidencia suficiente para aceptar la Hipótesis de que en conjunto los datos siguen un comportamiento Normal.

Hipótesis 2: Homocedasticidad

Para comprobar esta hipótesis usaremos el test de Levine.

```
require(car)
lapply(iris[,c("Sepal.Length","Sepal.Width","Petal.Length","Petal.Width")],leveneTest,iris$Speci
es)
```

```
## $Sepal.Length
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value   Pr(>F)
## group   2  6.3527 0.002259 **
##       147
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## $Sepal.Width
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   2  0.5902 0.5555
##       147
##
## $Petal.Length
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value    Pr(>F)
## group   2   19.48 3.129e-08 ***
##       147
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## $Petal.Width
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value    Pr(>F)
## group   2  19.892 2.261e-08 ***
##       147
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Con base en el resultado del test, podemos concluir que en los siguientes grupos no se acepta la idea de igualdad de varianzas: “Sepal.Length”,“Petal.Length” y”Petal.Width”.

Hipótesis 3: Linealidad

Evaluaremos este criterio a partir del uso de la siguientes gráficas:

```
require("PerformanceAnalytics")
```

```
## Loading required package: PerformanceAnalytics
```

```
## Warning: package 'PerformanceAnalytics' was built under R version 4.1.3
```

```
## Loading required package: xts
```

```
## Warning: package 'xts' was built under R version 4.1.3
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

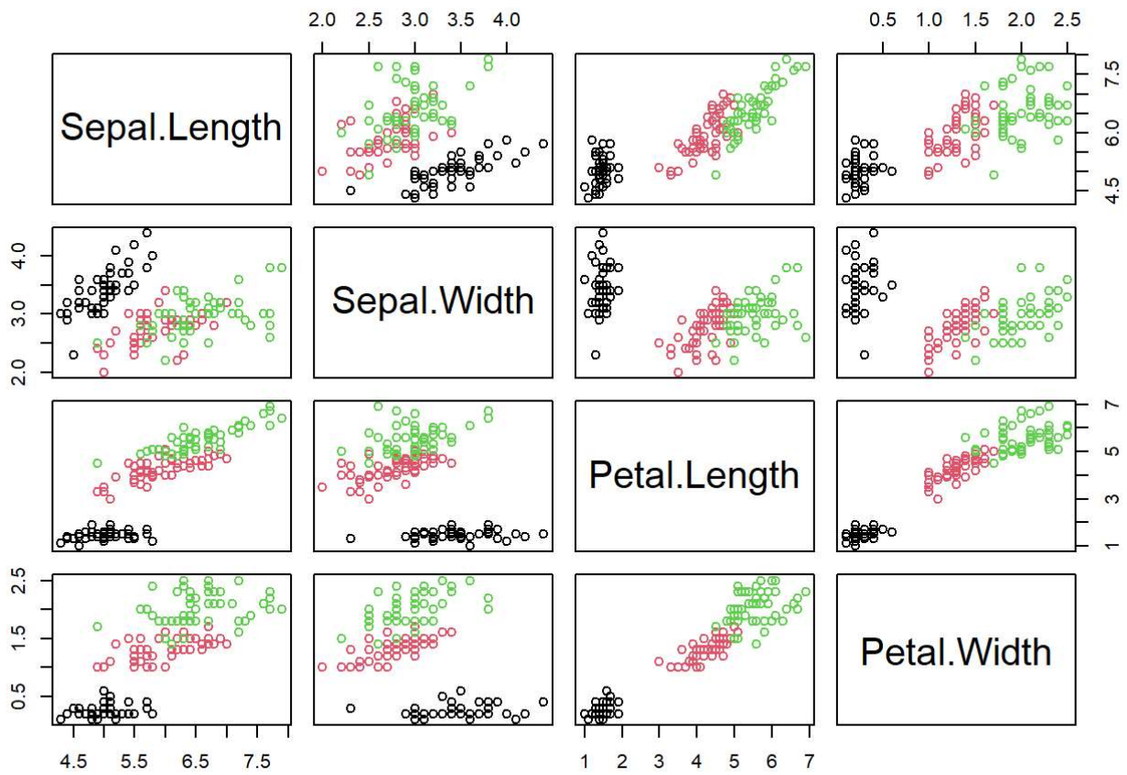
```
##
## Attaching package: 'xts'
```

```
## The following objects are masked from 'package:pastecs':
##
##     first, last
```

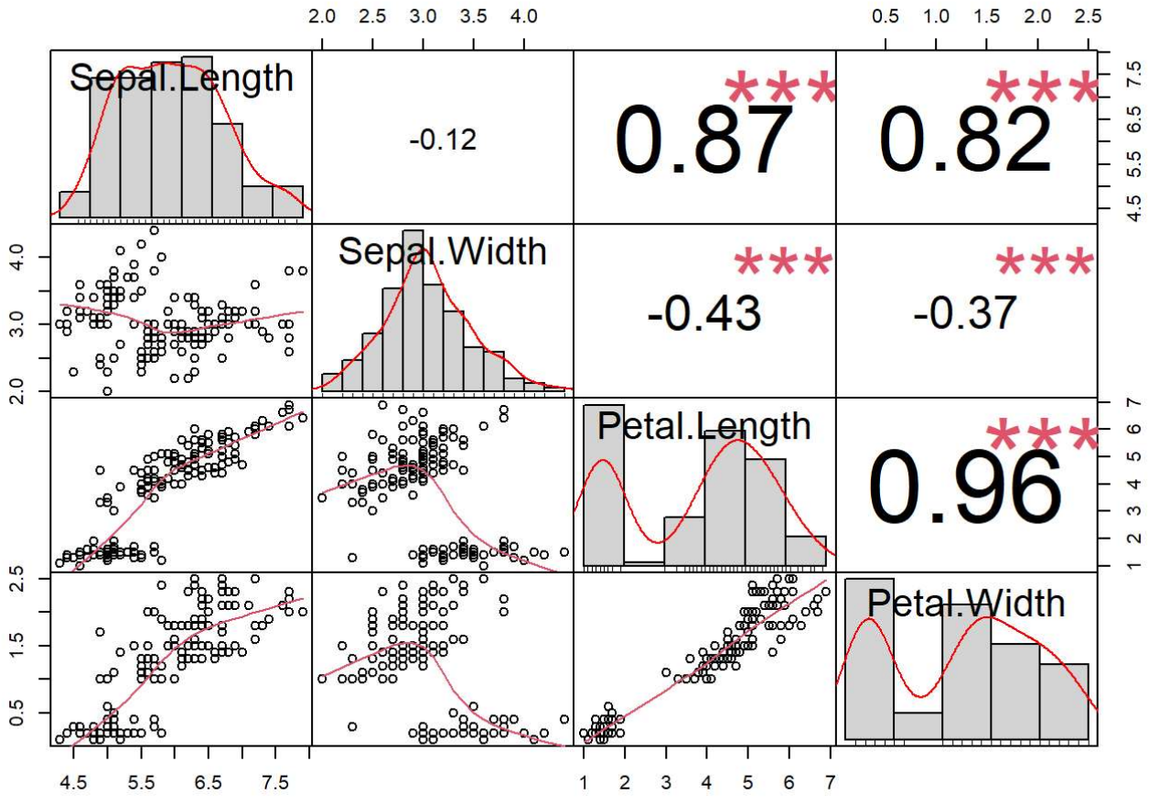
```
##
## Attaching package: 'PerformanceAnalytics'
```

```
## The following object is masked from 'package:graphics':
##
##     legend
```

```
plot(iris[,c("Sepal.Length","Sepal.Width","Petal.Length","Petal.Width")], col=iris$Species)
```




```
chart.Correlation(iris[,c("Sepal.Length","Sepal.Width","Petal.Length","Petal.Width")], histogram
=TRUE,col=iris$Species)
```



El resultado confirma la linealidad para las variables excepto para el caso de las variables Sepal.Length y Sepal.Width