Iris Analysis Report

DimitriMo

2025-11-19

Introduction

Analyse du dataset Iris pour un mini projet Data Analyst.
Objectifs : - Statistiques descriptives - Visualisations - Tests statistiques - Régression linéaire

Analyse et visualisations

```
# --------------------
# Charger le dataset
# --------------------
data(iris)

# Aperçu
head(iris)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

```
summary(iris)
```

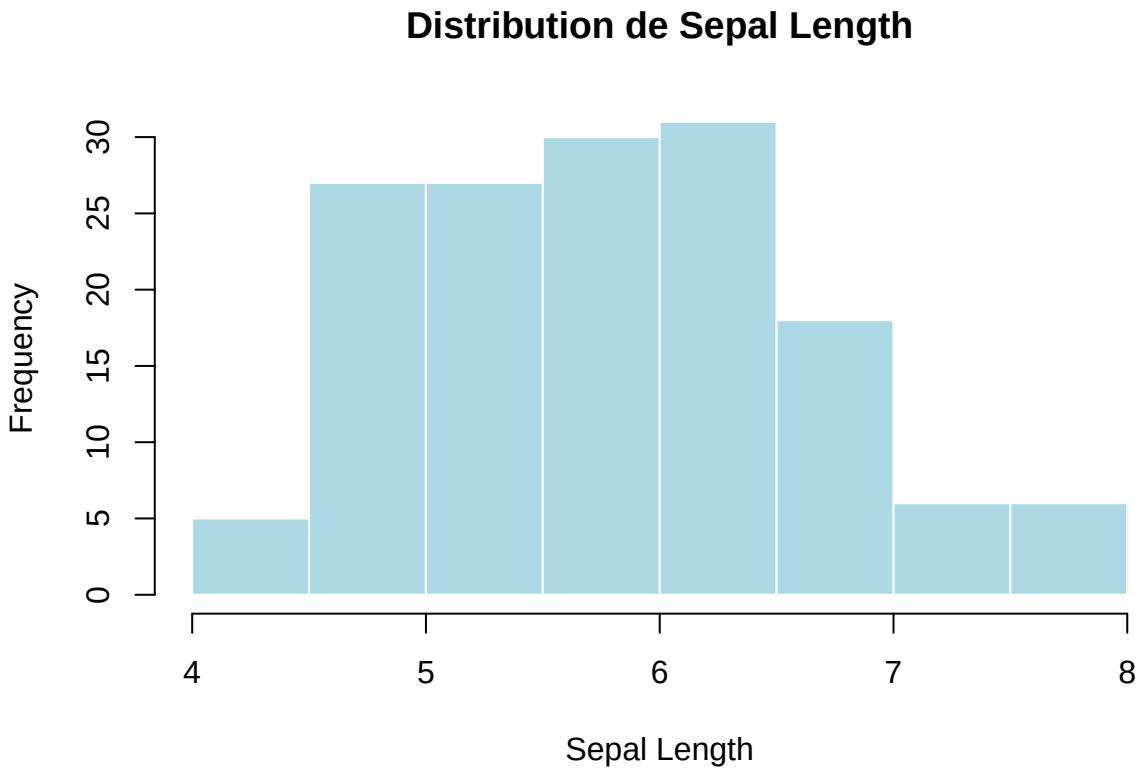
```
##   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
##  Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
##  1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
##  Median :5.800   Median :3.000   Median :4.350   Median :1.300
##  Mean   :5.843   Mean   :3.057   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
##
##
##
```

```
str(iris)
```

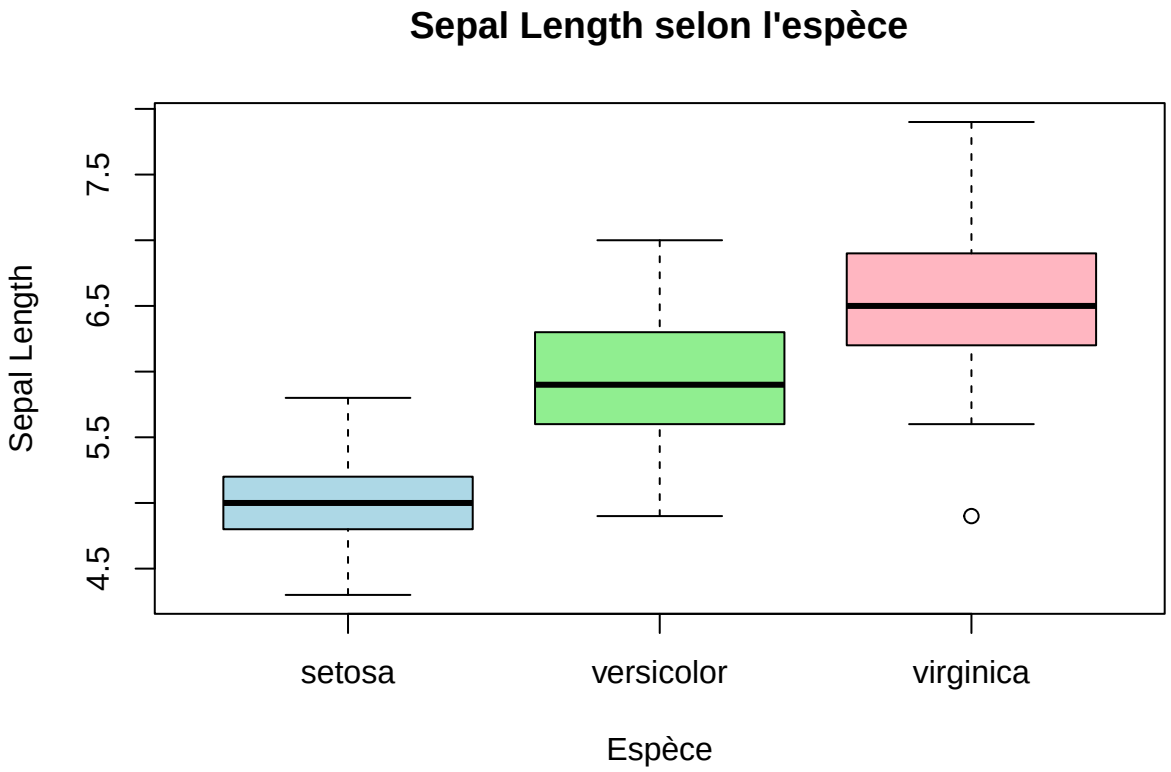
```
## 'data.frame':    150 obs. of  5 variables:
##  $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
```

```
##  $ 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 ...
```

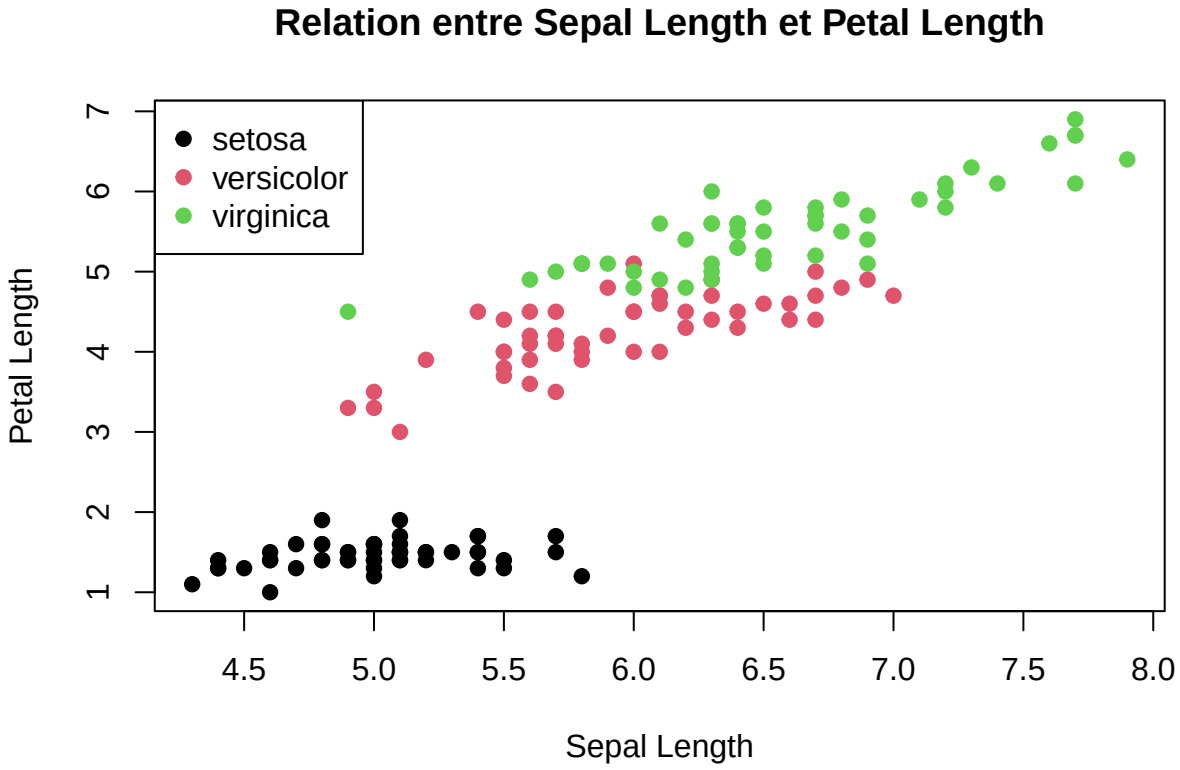
```r
# --------------------
# Visualisations
# --------------------

# Histogramme
hist(
  iris$Sepal.Length,
  main = "Distribution de Sepal Length",
  xlab = "Sepal Length",
  col = "lightblue",
  border = "white"
)
```




```r
# Boxplot
boxplot(
  Sepal.Length ~ Species,
  data = iris,
  main = "Sepal Length selon l'espèce",
  xlab = "Espèce",
  ylab = "Sepal Length",
  col = c("lightblue", "lightgreen", "lightpink")
)
```




```r
# Scatter plot
plot(
  iris$Sepal.Length,
  iris$Petal.Length,
  main = "Relation entre Sepal Length et Petal Length",
  xlab = "Sepal Length",
  ylab = "Petal Length",
  pch = 19,
  col = as.numeric(iris$Species)
)
legend(
  "topleft",
  legend = levels(iris$Species),
  col = 1:3,
  pch = 19
)
```




```
# --------------------
# Test ANOVA et Tukey HSD
# --------------------
anova_model <- aov(Sepal.Length ~ Species, data = iris)
summary(anova_model)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Species       2  63.21  31.606   119.3 <2e-16 ***
## Residuals   147  38.96   0.265
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
tukey_result <- TukeyHSD(anova_model)
tukey_result
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Sepal.Length ~ Species, data = iris)
##
## $Species
##                        diff       lwr       upr p adj
## versicolor-setosa    0.930 0.6862273 1.1737727     0
## virginica-setosa     1.582 1.3382273 1.8257727     0
## virginica-versicolor 0.652 0.4082273 0.8957727     0
```

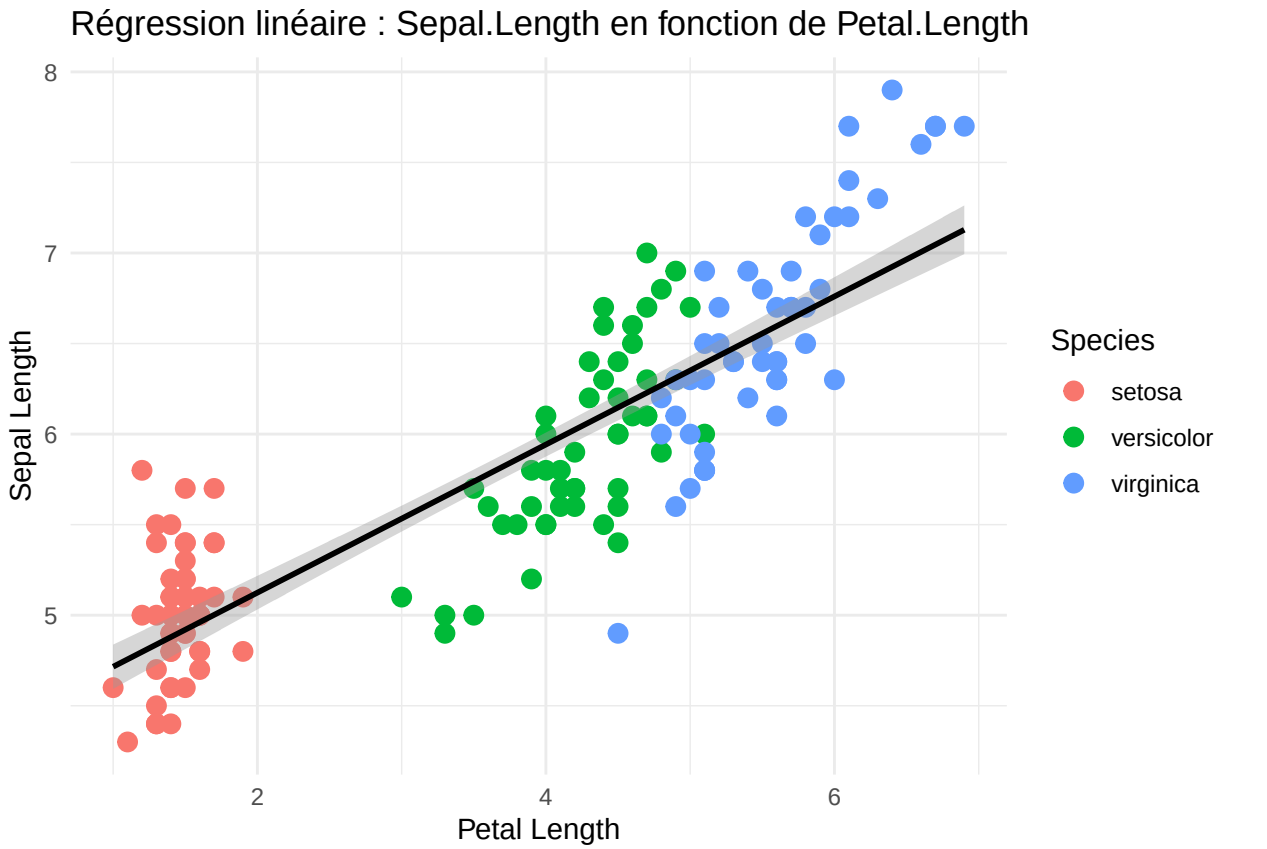
```
# Ici, toutes les p-values ajustées = 0,
# donc chaque paire d'espèces présente une différence significative
# de longueur de sépales.
```

```r
# --------------------
# Régression linéaire et plot ggplot2
# --------------------
linear_model <- lm(Sepal.Length ~ Petal.Length, data = iris)
summary(linear_model)
```

```
##
## Call:
## lm(formula = Sepal.Length ~ Petal.Length, data = iris)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.24675 -0.29657 -0.01515  0.27676  1.00269
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.30660    0.07839   54.94   <2e-16 ***
## Petal.Length  0.40892    0.01889   21.65   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4071 on 148 degrees of freedom
## Multiple R-squared:   0.76,  Adjusted R-squared:  0.7583
## F-statistic: 468.6 on 1 and 148 DF,  p-value: < 2.2e-16
```

```r
library(ggplot2)
ggplot(iris, aes(x = Petal.Length, y = Sepal.Length)) +
  geom_point(aes(color = Species), size = 3) +
  geom_smooth(method = "lm", se = TRUE, color = "black") +
  labs(
    title = "Régression linéaire : Sepal.Length en fonction de Petal.Length",
    x = "Petal Length",
    y = "Sepal Length"
  ) +
  theme_minimal()
```

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




```
#   - Coefficient positif et significatif : Sepal.Length augmente avec Petal.Length
#   - R-squared = 0.76 : modèle explique 76% de la variance
```

Conclusion

- Les analyses montrent que **chaque espèce d'iris a des sépales de longueur moyenne différente**.
- La **régression linéaire** confirme que **Petal.Length est un bon prédicteur de Sepal.Length**.