TP1 Statistique

January 12, 2024

0.1 TP1 Statistique descripyive avec R

0.2 EXERCICE 1

1°) Exécutons le code suivant et expliquons ce qu'il fait

[2]:
```
data()  # Charge les ensembles de données disponibles dans R.

x1 = trees  # Sélectionne l'ensemble de données "trees" et le stocke dans la␣
 ↪variable x1.
str(x1)  # Affichez la structure de l'ensemble de données trees.

x1$Girth  # Affiche la colonne "Girth" de l'ensemble de données trees.

x2 = HairEyeColor  # Sélectionne l'ensemble de données "HairEyeColor" et le␣
 ↪stocke dans la variable x2.
str(x2)  # Affiche la structure de l'ensemble de données HairEyeColor.

x2[1,2,2]  # Accéde à l'élément situé à la première ligne, deuxième colonne et␣
 ↪deuxième "profondeur" de l'ensemble de données HairEyeColor.

x2[,2,2]  # Accéde à tous les éléments de la deuxième colonne et deuxième␣
 ↪"profondeur" de l'ensemble de données HairEyeColor.
```

```
'data.frame':   31 obs. of  3 variables:
 $ Girth : num  8.3 8.6 8.8 10.5 10.7 10.8 11 11 11.1 11.2 …
 $ Height: num  70 65 63 72 81 83 66 75 80 75 …
 $ Volume: num  10.3 10.3 10.2 16.4 18.8 19.7 15.6 18.2 22.6 19.9 …
```

1. 8.3 2. 8.6 3. 8.8 4. 10.5 5. 10.7 6. 10.8 7. 11 8. 11 9. 11.1 10. 11.2 11. 11.3 12. 11.4 13. 11.4 14. 11.7 15. 12 16. 12.9 17. 12.9 18. 13.3 19. 13.7 20. 13.8 21. 14 22. 14.2 23. 14.5 24. 16 25. 16.3 26. 17.3 27. 17.5 28. 17.9 29. 18 30. 18 31. 20.6

```
 'table' num [1:4, 1:4, 1:2] 32 53 10 3 11 50 10 30 10 25 …
 - attr(*, "dimnames")=List of 3
  ..$ Hair: chr [1:4] "Black" "Brown" "Red" "Blond"
  ..$ Eye : chr [1:4] "Brown" "Blue" "Hazel" "Green"
  ..$ Sex : chr [1:2] "Male" "Female"
```

9

Black 9 **Brown** 34 **Red** 7 **Blond** 64

```
Data sets in package 'datasets':

AirPassengers           Monthly Airline Passenger Numbers 1949-1960
BJsales                 Sales Data with Leading Indicator
BJsales.lead (BJsales)
                        Sales Data with Leading Indicator
BOD                     Biochemical Oxygen Demand
CO2                     Carbon Dioxide Uptake in Grass Plants
ChickWeight             Weight versus age of chicks on different diets
DNase                   Elisa assay of DNase
EuStockMarkets          Daily Closing Prices of Major European Stock
                        Indices, 1991-1998
Formaldehyde            Determination of Formaldehyde
HairEyeColor            Hair and Eye Color of Statistics Students
Harman23.cor            Harman Example 2.3
Harman74.cor            Harman Example 7.4
Indometh                Pharmacokinetics of Indomethacin
InsectSprays            Effectiveness of Insect Sprays
JohnsonJohnson          Quarterly Earnings per Johnson & Johnson Share
LakeHuron               Level of Lake Huron 1875-1972
LifeCycleSavings        Intercountry Life-Cycle Savings Data
Loblolly                Growth of Loblolly pine trees
Nile                    Flow of the River Nile
Orange                  Growth of Orange Trees
OrchardSprays           Potency of Orchard Sprays
PlantGrowth             Results from an Experiment on Plant Growth
Puromycin               Reaction Velocity of an Enzymatic Reaction
Seatbelts               Road Casualties in Great Britain 1969-84
Theoph                  Pharmacokinetics of Theophylline
Titanic                 Survival of passengers on the Titanic
ToothGrowth             The Effect of Vitamin C on Tooth Growth in
                        Guinea Pigs
UCBAdmissions           Student Admissions at UC Berkeley
UKDriverDeaths          Road Casualties in Great Britain 1969-84
UKgas                   UK Quarterly Gas Consumption
USAccDeaths             Accidental Deaths in the US 1973-1978
USArrests               Violent Crime Rates by US State
USJudgeRatings          Lawyers' Ratings of State Judges in the US
                        Superior Court
USPersonalExpenditure   Personal Expenditure Data
UScitiesD               Distances Between European Cities and Between
                        US Cities
VADeaths                Death Rates in Virginia (1940)
WWWusage                Internet Usage per Minute
WorldPhones             The World's Telephones
ability.cov             Ability and Intelligence Tests
```

airmiles	Passenger Miles on Commercial US Airlines, 1937-1960
airquality	New York Air Quality Measurements
anscombe	Anscombe's Quartet of 'Identical' Simple Linear Regressions
attenu	The Joyner-Boore Attenuation Data
attitude	The Chatterjee-Price Attitude Data
austres	Quarterly Time Series of the Number of Australian Residents
beaver1 (beavers)	Body Temperature Series of Two Beavers
beaver2 (beavers)	Body Temperature Series of Two Beavers
cars	Speed and Stopping Distances of Cars
chickwts	Chicken Weights by Feed Type
co2	Mauna Loa Atmospheric CO2 Concentration
crimtab	Student's 3000 Criminals Data
discoveries	Yearly Numbers of Important Discoveries
esoph	Smoking, Alcohol and (O)esophageal Cancer
euro	Conversion Rates of Euro Currencies
euro.cross (euro)	Conversion Rates of Euro Currencies
eurodist	Distances Between European Cities and Between US Cities
faithful	Old Faithful Geyser Data
fdeaths (UKLungDeaths)	Monthly Deaths from Lung Diseases in the UK
freeny	Freeny's Revenue Data
freeny.x (freeny)	Freeny's Revenue Data
freeny.y (freeny)	Freeny's Revenue Data
infert	Infertility after Spontaneous and Induced Abortion
iris	Edgar Anderson's Iris Data
iris3	Edgar Anderson's Iris Data
islands	Areas of the World's Major Landmasses
ldeaths (UKLungDeaths)	Monthly Deaths from Lung Diseases in the UK
lh	Luteinizing Hormone in Blood Samples
longley	Longley's Economic Regression Data
lynx	Annual Canadian Lynx trappings 1821-1934
mdeaths (UKLungDeaths)	Monthly Deaths from Lung Diseases in the UK
morley	Michelson Speed of Light Data
mtcars	Motor Trend Car Road Tests
nhtemp	Average Yearly Temperatures in New Haven
nottem	Average Monthly Temperatures at Nottingham, 1920-1939
npk	Classical N, P, K Factorial Experiment
occupationalStatus	Occupational Status of Fathers and their Sons
precip	Annual Precipitation in US Cities
presidents	Quarterly Approval Ratings of US Presidents

```
pressure                Vapor Pressure of Mercury as a Function of
                        Temperature
quakes                  Locations of Earthquakes off Fiji
randu                   Random Numbers from Congruential Generator
                        RANDU
rivers                  Lengths of Major North American Rivers
rock                    Measurements on Petroleum Rock Samples
sleep                   Student's Sleep Data
stack.loss (stackloss)
                        Brownlee's Stack Loss Plant Data
stack.x (stackloss)     Brownlee's Stack Loss Plant Data
stackloss               Brownlee's Stack Loss Plant Data
state.abb (state)       US State Facts and Figures
state.area (state)      US State Facts and Figures
state.center (state)    US State Facts and Figures
state.division (state)
                        US State Facts and Figures
state.name (state)      US State Facts and Figures
state.region (state)    US State Facts and Figures
state.x77 (state)       US State Facts and Figures
sunspot.month           Monthly Sunspot Data, from 1749 to "Present"
sunspot.year            Yearly Sunspot Data, 1700-1988
sunspots                Monthly Sunspot Numbers, 1749-1983
swiss                   Swiss Fertility and Socioeconomic Indicators
                        (1888) Data
treering                Yearly Treering Data, -6000-1979
trees                   Diameter, Height and Volume for Black Cherry
                        Trees
uspop                   Populations Recorded by the US Census
volcano                 Topographic Information on Auckland's Maunga
                        Whau Volcano
warpbreaks              The Number of Breaks in Yarn during Weaving
women                   Average Heights and Weights for American Women

Use ‘data(package = .packages(all.available = TRUE))’
to list the data sets in all *available* packages.
```

0.2.1 2°) Que contiennent les variables x1 et x2? Comparons et discutons les commandes str versus summary.

[3]:
```r
str(x1)
```

```
'data.frame':   31 obs. of  3 variables:
 $ Girth : num  8.3 8.6 8.8 10.5 10.7 10.8 11 11 11.1 11.2 …
 $ Height: num  70 65 63 72 81 83 66 75 80 75 …
 $ Volume: num  10.3 10.3 10.2 16.4 18.8 19.7 15.6 18.2 22.6 19.9 …
```

[12]:
```r
summary(x1)
```

```
     Girth           Height       Volume     
 Min.   : 8.30   Min.   :63   Min.   :10.20  
 1st Qu.:11.05   1st Qu.:72   1st Qu.:19.40  
 Median :12.90   Median :76   Median :24.20  
 Mean   :13.25   Mean   :76   Mean   :30.17  
 3rd Qu.:15.25   3rd Qu.:80   3rd Qu.:37.30  
 Max.   :20.60   Max.   :87   Max.   :77.00  
```

[4]:
```
str(x2)
```

```
 'table' num [1:4, 1:4, 1:2] 32 53 10 3 11 50 10 30 10 25 …
 - attr(*, "dimnames")=List of 3
  ..$ Hair: chr [1:4] "Black" "Brown" "Red" "Blond"
  ..$ Eye : chr [1:4] "Brown" "Blue" "Hazel" "Green"
  ..$ Sex : chr [1:2] "Male" "Female"
```

[13]:
```
summary(x2)
```

```
Number of cases in table: 592 
Number of factors: 3 
Test for independence of all factors:
	Chisq = 164.92, df = 24, p-value = 5.321e-23
	Chi-squared approximation may be incorrect
```

Les variables x1 et x2 contiennent les ensembles de données "trees" et "HairEyeColor"

Les commandes `str` et `summary` sont deux commandes couramment utilisées dans R pour obtenir des informations sur la structure et le résumé statistique d'un objet de données. Comparons ces deux commandes :

0.2.2 `str()` (Structure)

La fonction `str` (structure) est utilisée pour afficher la structure interne d'un objet R. Elle est particulièrement utile pour explorer la structure des listes, des data frames, des matrices, et d'autres objets complexes. Elle fournit des informations sur le type de données, la structure et les premières valeurs des composants.

Avantages de `str` : - Donne une vue détaillée de la structure de l'objet. - Utile pour explorer des objets complexes comme les listes et les data frames.

Limitations de `str` : - Ne fournit pas de statistiques descriptives comme la moyenne, la médiane, etc.

0.2.3 `summary()` (Résumé)

La fonction `summary` est utilisée pour obtenir un résumé statistique des objets R. Elle est couramment utilisée avec des data frames, des vecteurs, ou d'autres objets pour obtenir des statistiques descriptives telles que la moyenne, la médiane, les quartiles, etc.

Avantages de `summary` : - Fournit des statistiques descriptives utiles. - Utile pour obtenir un aperçu rapide des caractéristiques importantes d'un ensemble de données.

Limitations de `summary` : - Moins détaillé sur la structure interne de l'objet par rapport à `str`. - Moins approprié pour explorer des objets complexes comme les listes.

0.2.4 Conclusion :

En général, `str` est plus approprié pour explorer la structure interne des objets, tandis que `summary` est plus approprié pour obtenir un résumé statistique rapide. En pratique, il est souvent judicieux d'utiliser les deux en combinaison pour obtenir une compréhension complète d'un ensemble de données ou d'un objet complexe.

0.2.5 3°) Quels types de graphes nous semblent appropriés pour résumer graphiquement l'information contenue dans le jeu de données x1? Justifions. Même question pour x2.

Pour résumer graphiquement l'information contenue dans le jeu de données `x1`, plusieurs types de graphiques peuvent être appropriés:

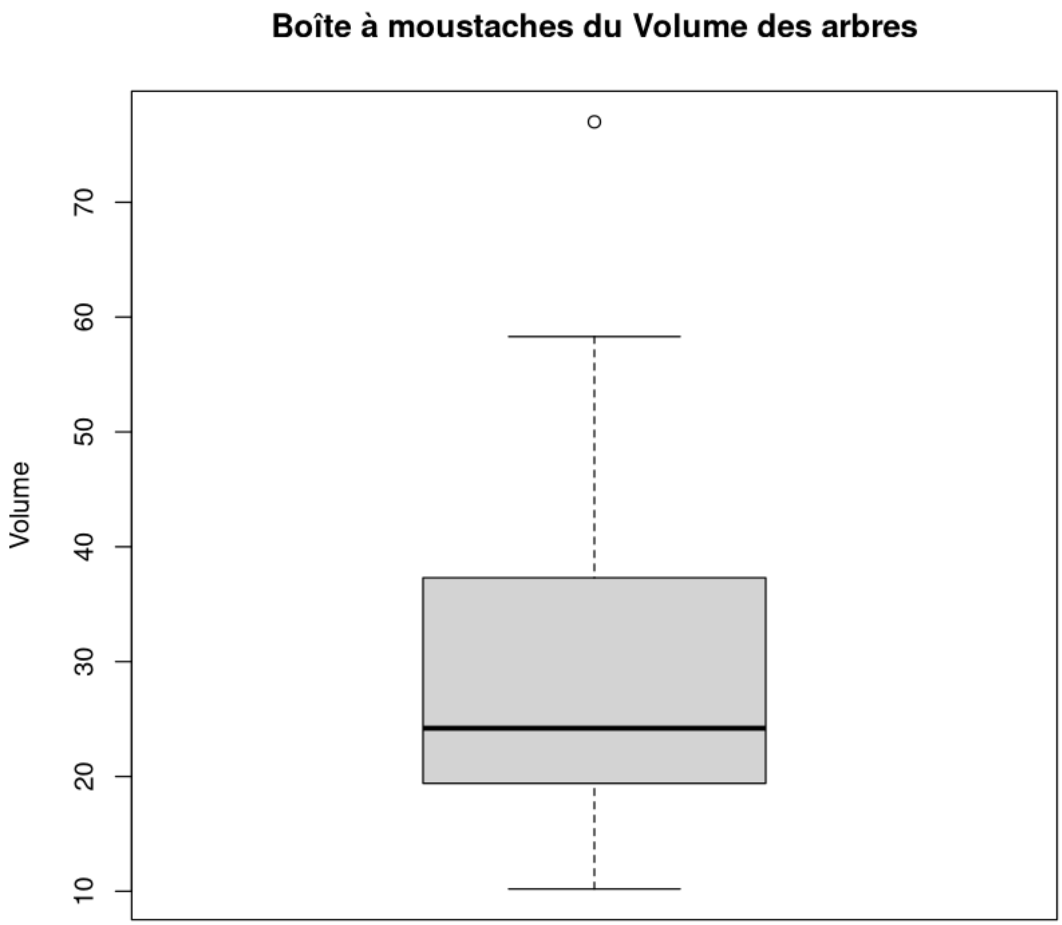
1. **Histogramme :**
   - Un histogramme peut être utile pour visualiser la distribution des valeurs dans chaque variable. Cela pourrait être appliqué à des variables comme la circonférence, la hauteur et le volume pour comprendre la répartition des données.
2. **Boîte à moustaches (Boxplot) :**
   - Les boîtes à moustaches peuvent aider à visualiser la dispersion des données, les valeurs aberrantes éventuelles, et les mesures de tendance centrale. Chaque variable pourrait avoir son propre boxplot pour une comparaison visuelle.

Pour le jeu de données `x2`, qui contient des informations sur la fréquence des combinaisons de couleurs de cheveux, de yeux et de sexe, voici quelques types de graphiques appropriés pour résumer graphiquement l'information :

1. **Diagramme en barres (Bar Chart) :**
   - Un diagramme en barres peut être utilisé pour visualiser la fréquence des combinaisons de couleurs de cheveux, de yeux et de sexe. Chaque barre représenterait une combinaison, et la hauteur de la barre serait proportionnelle à la fréquence.
2. **Diagramme en mosaïque (Mosaic Plot) :**
   - Un diagramme en mosaïque est adapté pour représenter la relation entre deux variables catégoriques. Cela peut être utilisé pour visualiser la distribution des couleurs de cheveux et de yeux par sexe.
3. **Diagramme de secteurs (Pie Chart) :**
   - Si on veut représenter la répartition des couleurs de cheveux, de yeux et de sexe en pourcentage du total, un diagramme de secteurs pourrait être approprié.
4. **Heatmap :**
   - Une heatmap peut être utilisée pour visualiser la fréquence des combinaisons de couleurs sous forme de tableau coloré. Les couleurs indiqueraient la fréquence relative.

0.2.6 4°) Traçons la boîte à moustaches des valeurs du volume des arbres contenues dans x1.

[5]:
```
boxplot(x1$Volume, main = "Boîte à moustaches du Volume des arbres", ylab =␣
↪"Volume")
```



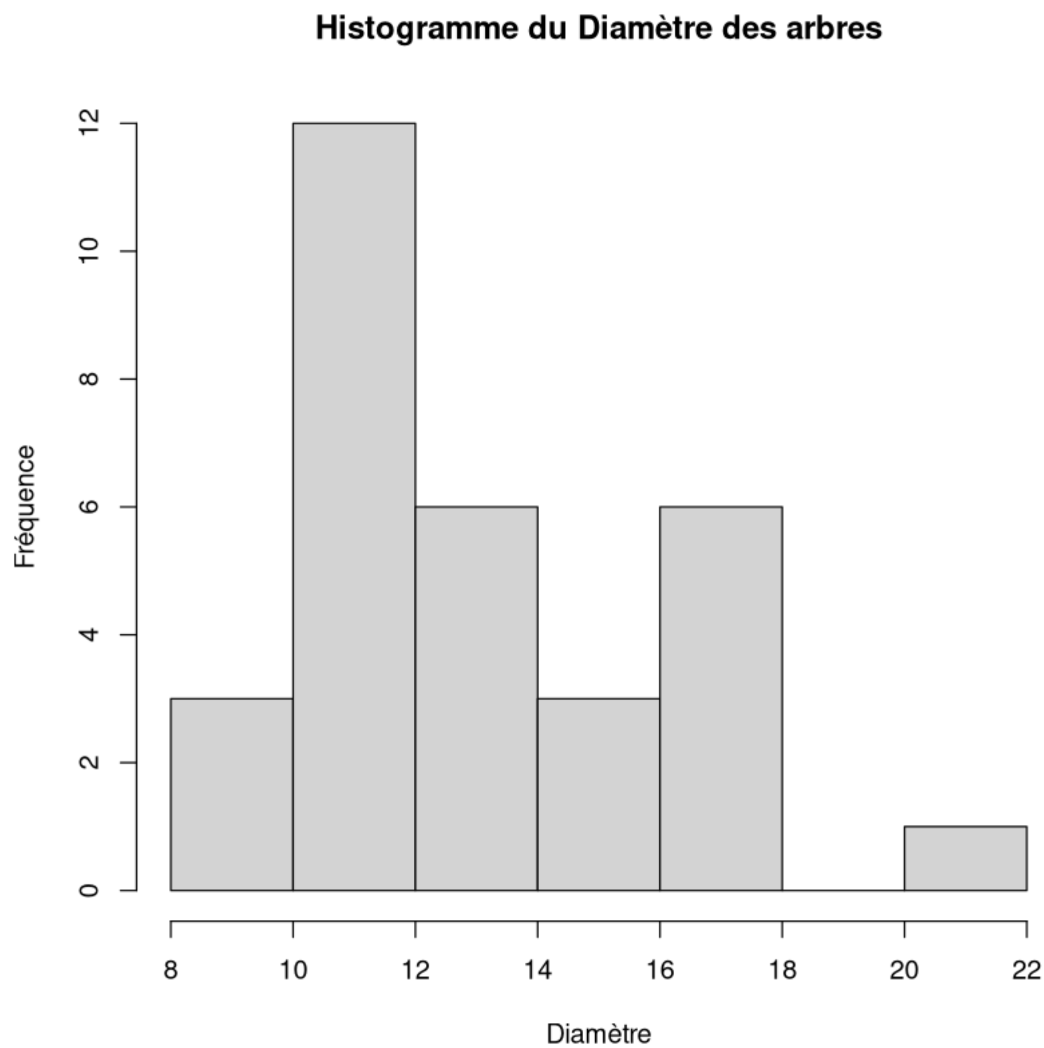
0.2.7 COMMENTAIRE

Le graphe obtenu à partir de la boîte à moustaches du volume des arbres donne une représentation visuelle des principales caractéristiques statistiques de la distribution du volume. L'objectif de ce graphique est de fournir une vue rapide de la distribution des valeurs du volume des arbres, en mettant en évidence les mesures de tendance centrale (médiane) et de dispersion (étendue interquartile). Les points aberrants peuvent également attirer l'attention sur des observations inhabituelles dans les données.

En examinant ce graphique, nous pouvons avoir une idée de la variabilité du volume des arbres et identifier s'il y a des valeurs aberrantes.

0.2.8 5°) Traçons l'histogramme des valeurs du diamètre des arbres contenues dans x1

[6]:
```
hist(x1$Girth, main = "Histogramme du Diamètre des arbres", xlab = "Diamètre",␣
↪ylab = "Fréquence")
```

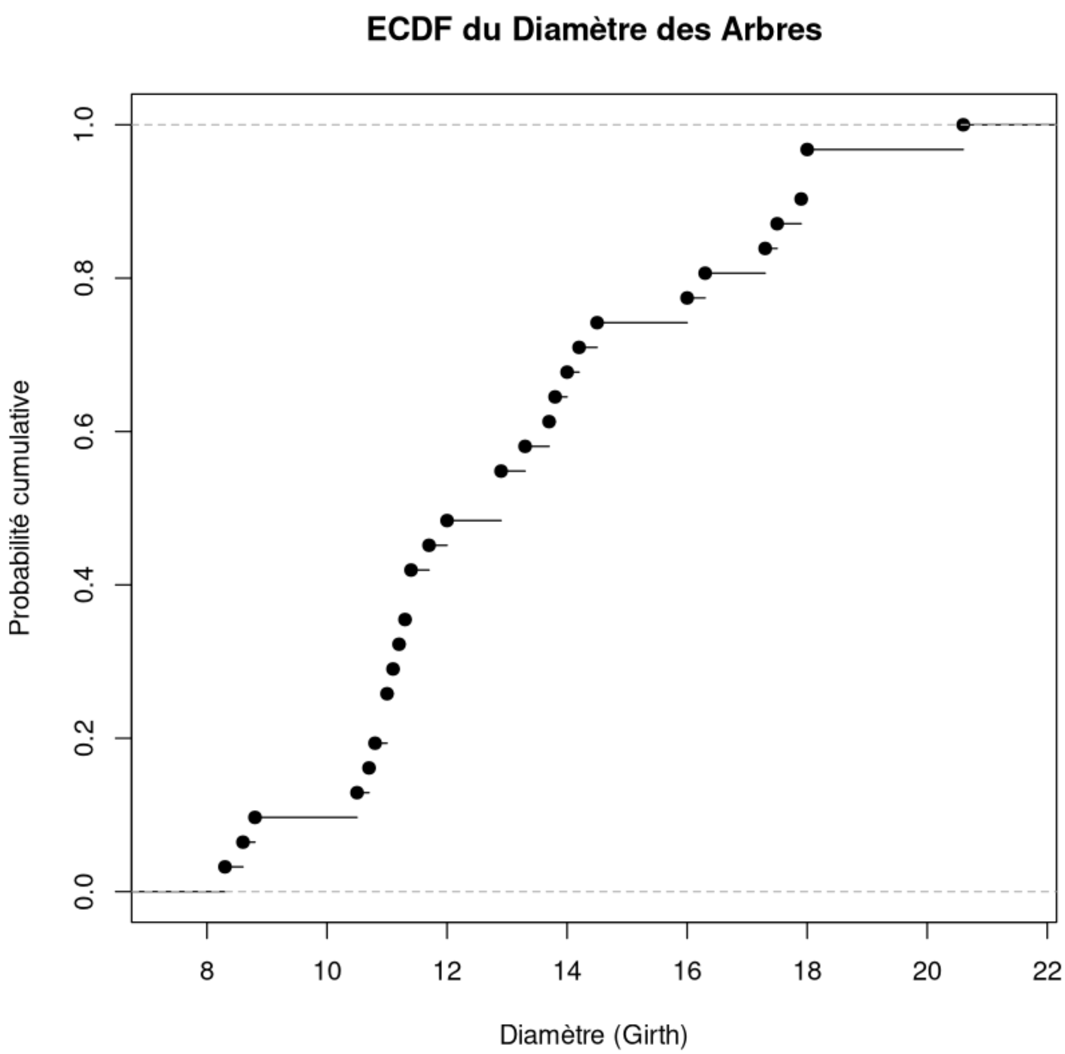


0.2.9 COMMENTAIRE

L'objectif de l'histogramme est de fournir un aperçu visuel de la distribution des valeurs du diamètre des arbres. Nous pouvons observer la forme générale de la distribution, identifier les plages de valeurs les plus fréquentes et évaluer la dispersion des données. Les histogrammes sont particulièrement utiles pour détecter des tendances ou des structures dans les données, comme la normalité, la présence de pics, ou la symétrie. Dans notre cas, l'histogramme est asymétrique alors cela indique une distribution biaisée.

0.2.10 6°) Traçons la fonction de répartition empirique des valeurs du diamètre des arbres contenues dans x1.

[7]:
```
plot(ecdf(x1$Girth), main = "ECDF du Diamètre des Arbres", xlab = "Diamètre␣
↪(Girth)", ylab = "Probabilité cumulative")
```



0.2.11 COMMENTAIRE

L'ECDF est particulièrement utile pour comprendre la répartition cumulative des données sans faire d'hypothèses sur la forme de la distribution. En examinant la courbe, nous pouvons rapidement estimer la probabilité d'observer des valeurs de diamètre inférieures ou égales à une certaine valeur.

0.2.12 7°)Construisons un noyau de jeu de données x3 contenant x1, et dans lequel la dernière valeur de diamètre (soit 20.6 pouces) a malencontreusement été modifiée en 206 pouces.

[8]:
```r
# Copier x1 pour créer x3
x3 <- x1

# Modifie la dernière valeur de diamètre dans x3
x3$Girth[length(x3$Girth)] <- 206

# Affiche les premières lignes de x3 pour vérification
head(x3)
```

A data.frame: 6 × 3

	Girth <dbl>	Height <dbl>	Volume <dbl>
1	8.3	70	10.3
2	8.6	65	10.3
3	8.8	63	10.2
4	10.5	72	16.4
5	10.7	81	18.8
6	10.8	83	19.7

La modification de la dernière valeur de diamètre dans le jeu de données `x3` (passant de 20.6 pouces à 206 pouces) aura un impact sur les différentes représentations numériques et graphiques que nous avons discutées précédemment. Voici comment cela pourrait affecter chaque représentation :

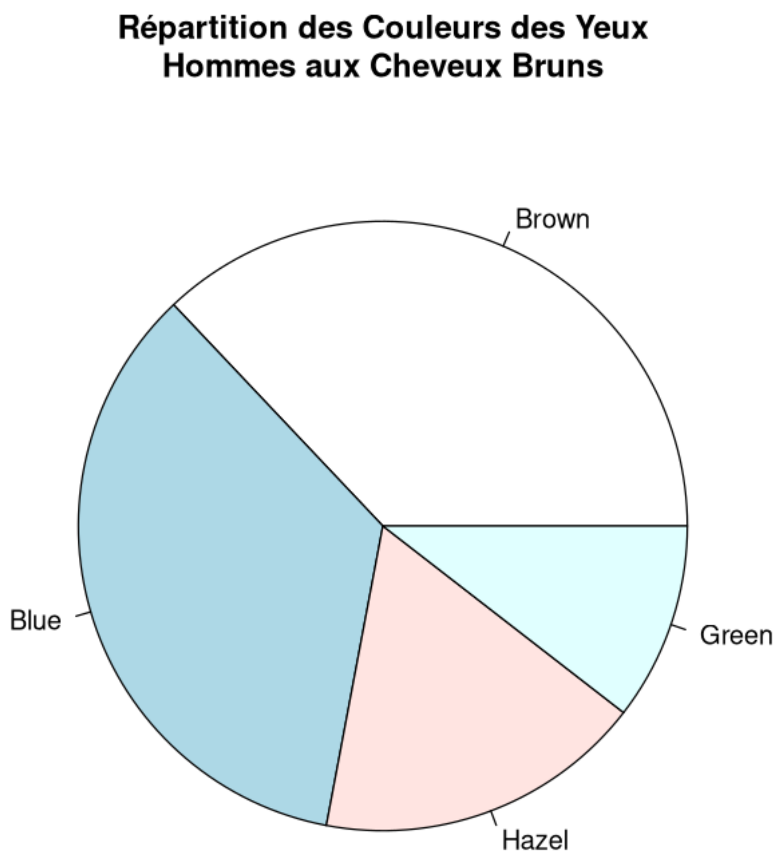
1. **Boxplot du Volume**:
   - La modification du diamètre n'affectera pas directement le boxplot du volume, car le volume n'est pas directement dépendant du diamètre. Cependant, si le volume est calculé en utilisant le diamètre, cela peut influencer le volume en fonction de la méthode de calcul.
2. **Histogramme du Diamètre**:
   - L'histogramme du diamètre sera influencé par la modification de la dernière valeur. La classe correspondant à la valeur modifiée (206 pouces) pourrait être la seule classe visible dans l'histogramme, donnant une fausse impression de la distribution réelle des diamètres.
3. **ECDF du Diamètre**:
   - La fonction de répartition empirique (ECDF) sera également affectée. La courbe montrera une probabilité cumulée significativement élevée à la nouvelle valeur de diamètre modifiée.
4. **Résumé Statistique du Diamètre**:
   - Les mesures statistiques telles que la moyenne et l'écart-type du diamètre seront influ-

encées par la nouvelle valeur aberrante.

5. **Affichage des Premières Lignes de x3**:
   - Les premières lignes de `x3` montreront la modification du diamètre dans la dernière observation.

0.2.13 8°)A l'aide de la fonction pie, traçons pour le jeu de données x2 un diagramme en camembert de la répartition des couleurs des yeux chez les hommes aux cheveux bruns.

[9]:
```
# Sélectionne les données pour les hommes aux cheveux bruns
data_subset <- x2["Brown", , "Male"]

# Crée un diagramme en camembert
pie(data_subset, main = "Répartition des Couleurs des Yeux\nHommes aux Cheveux␣
↪Bruns")
```

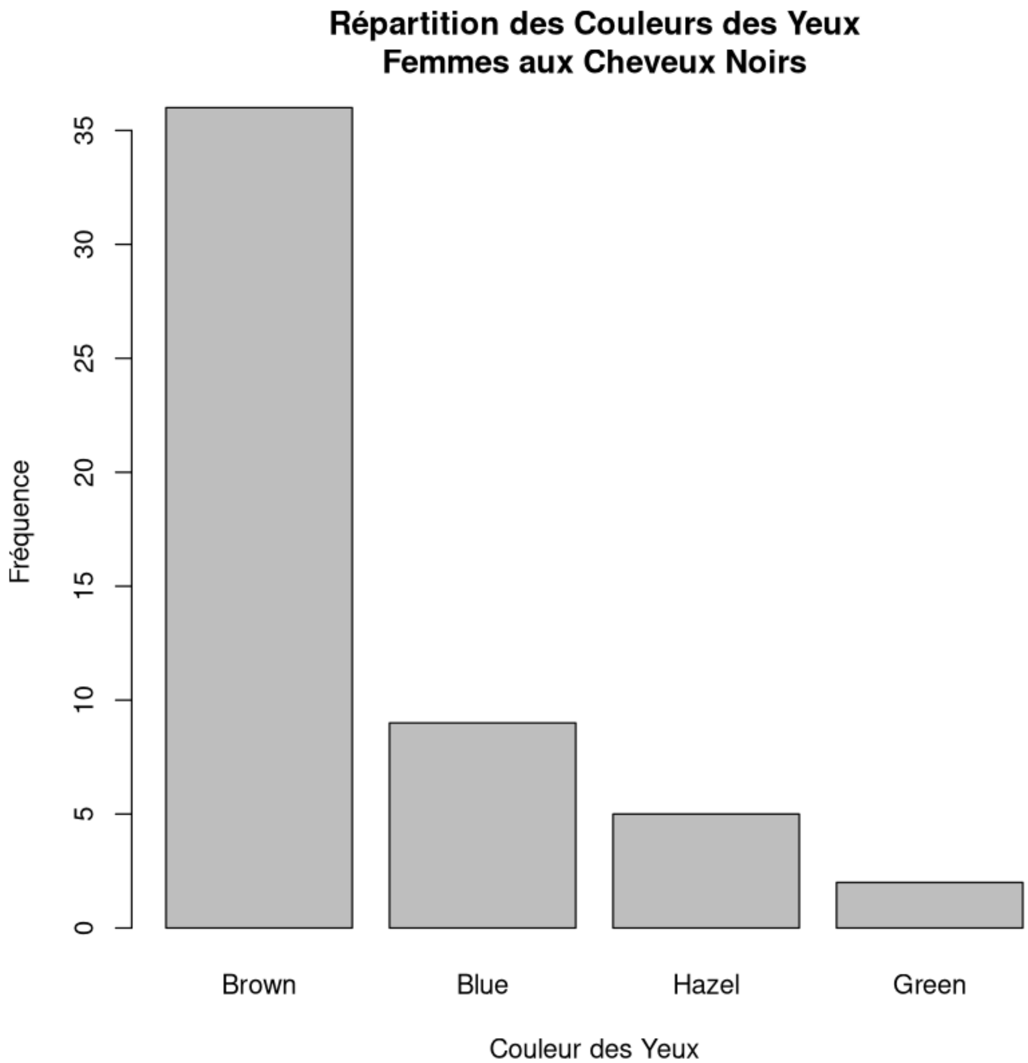


0.2.14 COMMENTAIRE

Le diagramme en camembert (pie chart) de la répartition des couleurs des yeux chez les hommes aux cheveux bruns offre une représentation visuelle de la distribution de ces caractéristiques. Nous observons une très grande part d'hommes aux cheveux bruns et de couleurs de yeux marron par contre les hommes aux cheveux bruns et aux couleurs de yeux vert sont minoritaires

0.2.15 9°) Utilisons la fonction barplot pour tracer le diagramme en bâton spécifiant la répartition des couleurs des yeux pour les femmes aux cheveux noirs. Ajoutons un titre adéquat

[21]:
```
# Sélectionne les données pour les femmes aux cheveux noirs
data_subset <- x2["Black", , "Female"]

# Crée un diagramme en bâtons
barplot(data_subset, main = "Répartition des Couleurs des Yeux\nFemmes aux␣
↪Cheveux Noirs",
        xlab = "Couleur des Yeux", ylab = "Fréquence")
```

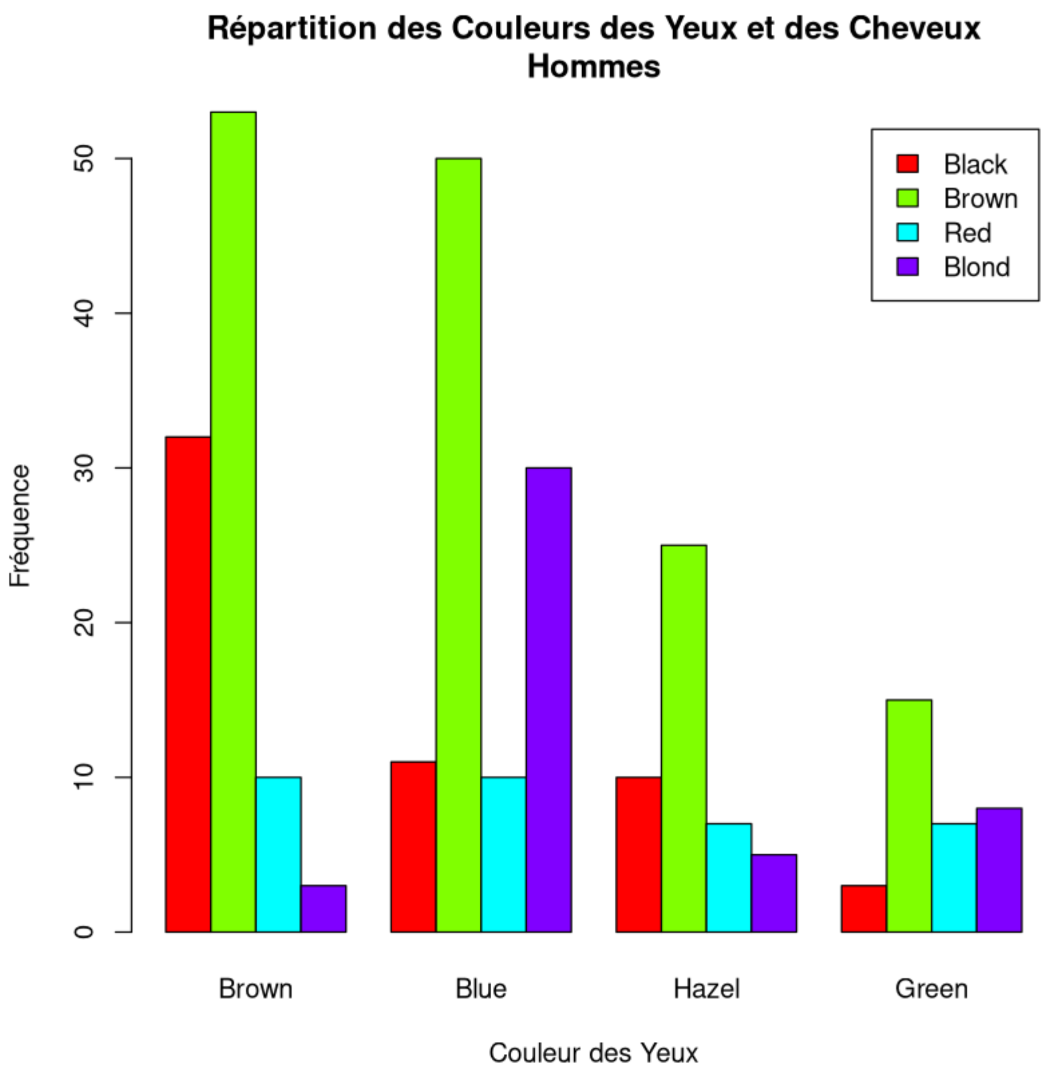


0.2.16 COMMENTAIRE

Le diagramme en bâtons (barplot) de la répartition des couleurs des yeux chez les femmes aux cheveux noirs aide à différencier facilement les catégories sur le graphique.On a plus de femmes aux cheveux noirs et aux couleurs de yeux marron que vert

0.2.17 10°) Représentons le diagramme en bâton spécifiant la répartition en terme de couleurs des yeux et des cheveux pour les hommes, en ajoutant titre et légende.

[22]:
```
data_subset <- x2[, , "Male"]

# Crée un diagramme en bâtons
barplot(data_subset, beside = TRUE,
        main = "Répartition des Couleurs des Yeux et des Cheveux\nHommes",
        xlab = "Couleur des Yeux", ylab = "Fréquence",
        legend.text = rownames(data_subset), col = rainbow(nrow(data_subset)))
```

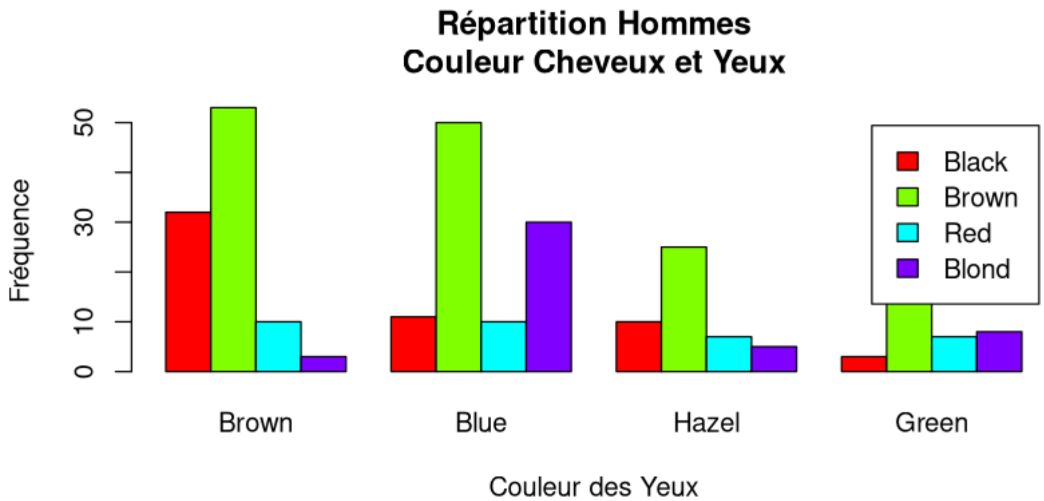


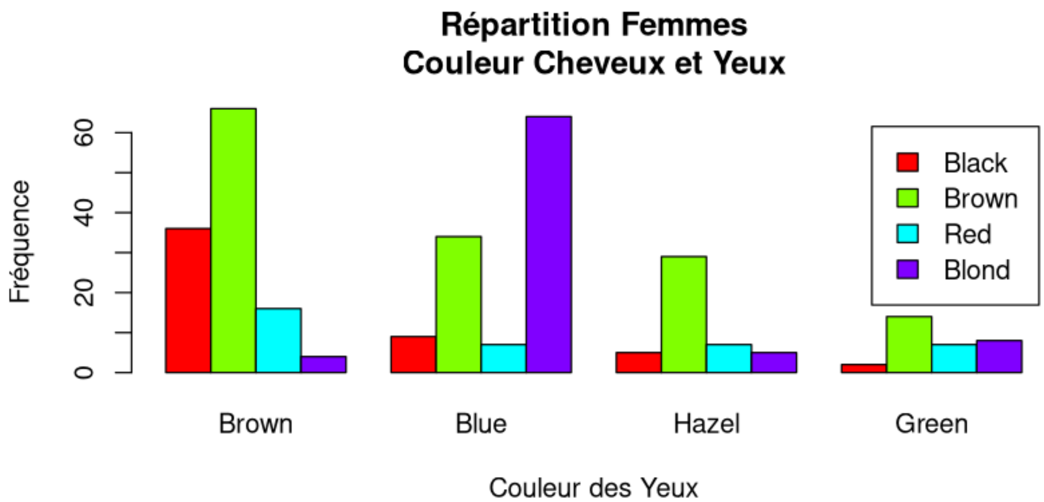
0.2.18 COMMENTAIRE

Le diagramme en bâtons représentant la répartition des couleurs des yeux et des cheveux chez les hommes offre une visualisation claire des combinaisons de ces caractéristiques.Les hommes aux couleurs de yeux marrons et aux cheveux bruns, de couleurs des yeux bleus et aux cheveux rouges, de couleurs des yeux noisettes et aux cheveux bruns, de couleurs de yeux verts et aux cheveux noirs sont rares

0.2.19 11°) Représentons sur une même figure les diagramme de répartition 'couleur de cheveux' et 'couleur de yeux' pour les hommes d'une part, et pour les femmes d'autre part.

[23]:
```
# Crée une fenêtre graphique avec deux rangées et une colonne
par(mfrow = c(2, 1))

# Diagramme en bâtons pour les hommes
barplot(x2[, , "Male"], beside = TRUE,
        main = "Répartition Hommes\nCouleur Cheveux et Yeux",
        xlab = "Couleur des Yeux", ylab = "Fréquence",
        legend.text = rownames(x2[, , "Male"]), col = rainbow(nrow(x2[, ,␣
 ↪"Male"])))

# Diagramme en bâtons pour les femmes
barplot(x2[, , "Female"], beside = TRUE,
        main = "Répartition Femmes\nCouleur Cheveux et Yeux",
        xlab = "Couleur des Yeux", ylab = "Fréquence",
        legend.text = rownames(x2[, , "Female"]), col = rainbow(nrow(x2[, ,␣
 ↪"Female"])))
```





0.2.20 Sauvegarde sous forme fichier pdf

[7]:
```
pdf("cheveux.pdf")

# Crée une fenêtre graphique avec deux rangées et une colonne
par(mfrow = c(2, 1))

# Diagramme en bâtons pour les hommes
barplot(x2[, , "Male"], beside = TRUE,
        main = "Répartition Hommes\nCouleur Cheveux et Yeux",
        xlab = "Couleur des Yeux", ylab = "Fréquence",
        legend.text = rownames(x2[, , "Male"]), col = rainbow(nrow(x2[, ,
 ↪"Male"])))
```

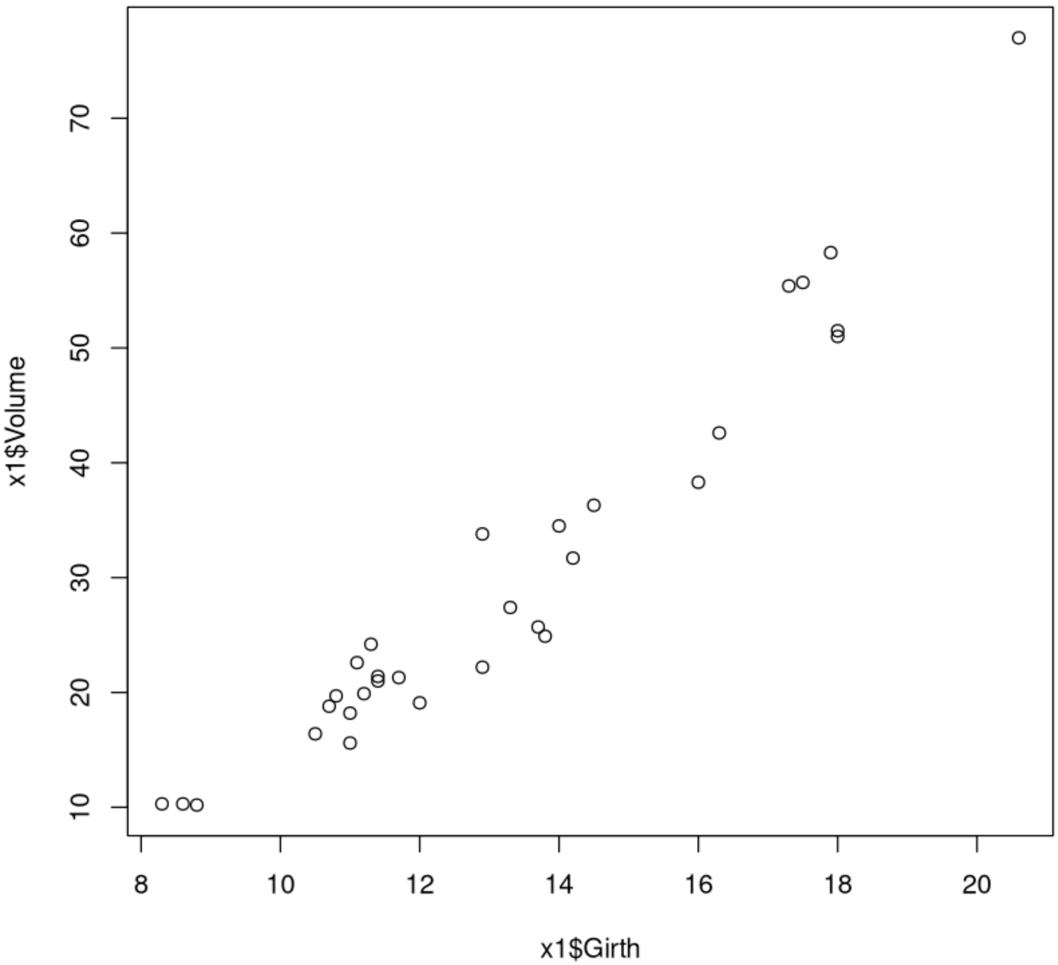
```
# Diagramme en bâtons pour les femmes
barplot(x2[, , "Female"], beside = TRUE,
        main = "Répartition Femmes\nCouleur Cheveux et Yeux",
        xlab = "Couleur des Yeux", ylab = "Fréquence",
        legend.text = rownames(x2[, , "Female"]), col = rainbow(nrow(x2[, ,␣
↪"Female"])))

# Sauvegarde le graphe en PDF
dev.off()  # Ferme la fenêtre graphique actuelle
```

png: 2

0.3 EXERCICE 2

0.3.1 1°) Exécutons le code suivant et expliquons ce qu'il fait:

[10]:
```
x1 = trees # Assignation des données du jeu de données 'trees' à la variable x1.

plot(x1$Girth, x1$Volume) # Création d'un diagramme de dispersion entre les␣
↪colonnes 'Girth' et 'Volume' du jeu de données x1.

pairs(x1) # Création d'une matrice de nuages de points montrant les relations␣
↪bivariées entre toutes les variables de x1.

cor(x1)# Calcul de la matrice de corrélation linéaire entre toutes les␣
↪variables de x1.

cor(x1, method="kendall") # Calcul de la matrice de corrélation de Kendall␣
↪entre toutes les variables de x1 (mesure de corrélation non paramétrique).

cor(x1, method="spearman") # 6. Calcul de la matrice de corrélation de Spearman␣
↪entre toutes les variables de x1 (une autre mesure de corrélation non␣
↪paramétrique).
```



A matrix: 3 × 3 of type dbl

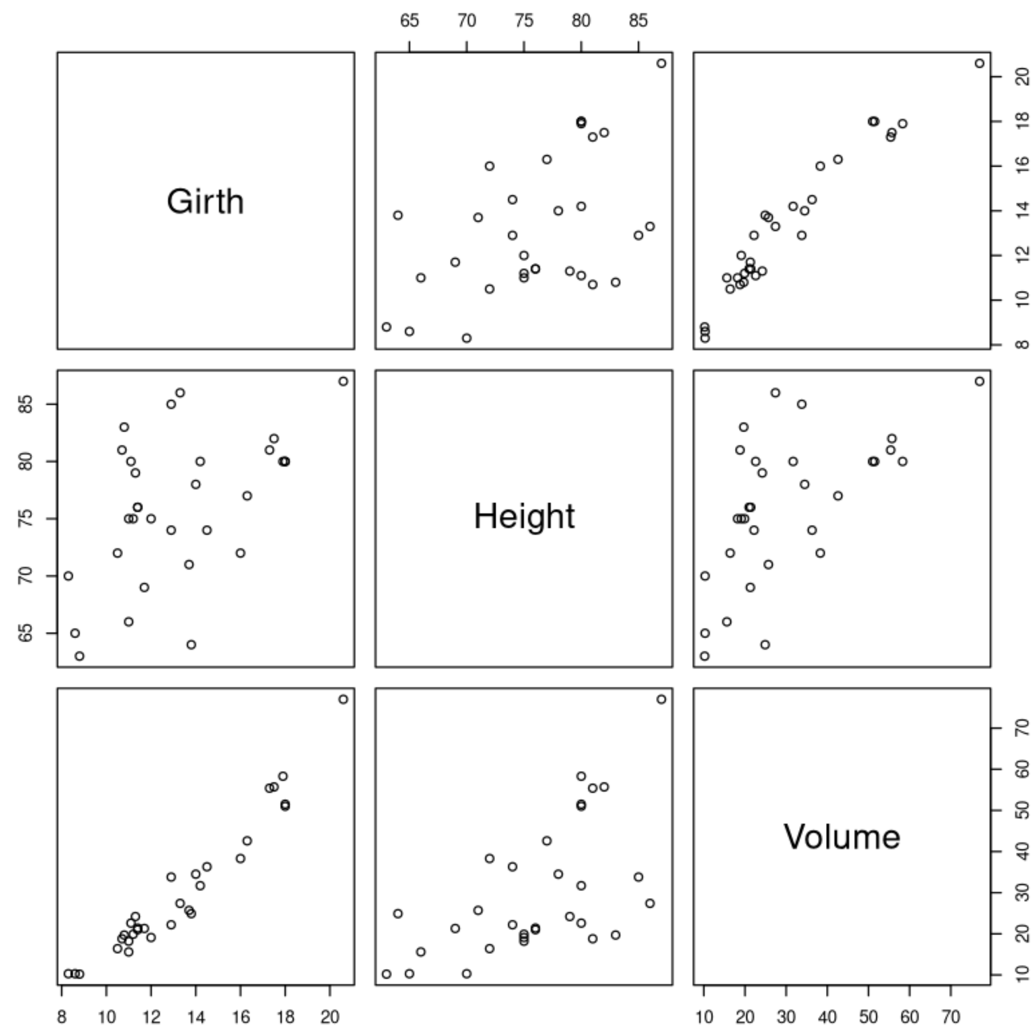
	Girth	Height	Volume
Girth	1.0000000	0.5192801	0.9671194
Height	0.5192801	1.0000000	0.5982497
Volume	0.9671194	0.5982497	1.0000000

A matrix: 3 × 3 of type dbl

	Girth	Height	Volume
Girth	1.0000000	0.3168641	0.8302746
Height	0.3168641	1.0000000	0.4496306
Volume	0.8302746	0.4496306	1.0000000

A matrix: 3 × 3 of type dbl

	Girth	Height	Volume
Girth	1.0000000	0.4408387	0.9547151
Height	0.4408387	1.0000000	0.5787101
Volume	0.9547151	0.5787101	1.0000000



0.3.2 2°) Effectuons des recherches sur internet pour trouver les définitions nécessaires

Voici une explication étape par étape de ce que fait le code :

1. **Assignation du jeu de données à x1** :

   ```
   x1 = trees
   ```

   Cette ligne assigne le jeu de données `trees` à la variable `x1`.

2. **Tracé d'un Nuage de Points (Plot)** :

   ```
   plot(x1$Girth, x1$Volume)
   ```

   Cette ligne trace un nuage de points en utilisant les données de la colonne "Girth" (circonférence) comme valeurs de l'axe x et les données de la colonne "Volume" comme valeurs de l'axe y. Cela permet de visualiser la relation entre la circonférence des arbres et leur volume.

3. **Matrice de Diagrammes de Paires (Pairs Plot)** :

   ```
   pairs(x1)
   ```

   Cette ligne crée une matrice de diagrammes de paires qui montre les relations bivariées entre toutes les variables de `x1`. Chaque cellule de la matrice contient un nuage de points pour la paire respective de variables.

4. **Calcul des Coefficients de Corrélation Pearson, Kendall et Spearman** :

   ```
   cor(x1)
   cor(x1, method="kendall")
   cor(x1, method="spearman")
   ```

   Ces lignes calculent les coefficients de corrélation entre toutes les paires de variables dans `x1` en utilisant différentes méthodes de corrélation :

   - La première ligne utilise la corrélation de Pearson (par défaut).
   - La deuxième ligne utilise la corrélation de Kendall.
   - La troisième ligne utilise la corrélation de Spearman.

   Les coefficients de corrélation quantifient la force et la direction d'une relation linéaire entre deux variables. Les coefficients de Kendall et de Spearman sont des mesures de corrélation non paramétriques qui mesurent les associations monotoniques.

En résumé, le code effectue une exploration visuelle et numérique des relations entre les variables du jeu de données `trees`, en utilisant des graphiques de dispersion, une matrice de diagrammes de paires, et en calculant les coefficients de corrélation pour évaluer la force des relations.

0.4 EXERCICE 3

[11]:
```
x1=iris # Assignation des données du jeu de données 'iris' à la variable x1.
```

[12]:
```
str(x1) # Affichage de la structure des données du jeu de données x1.
```

```
'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 …
```

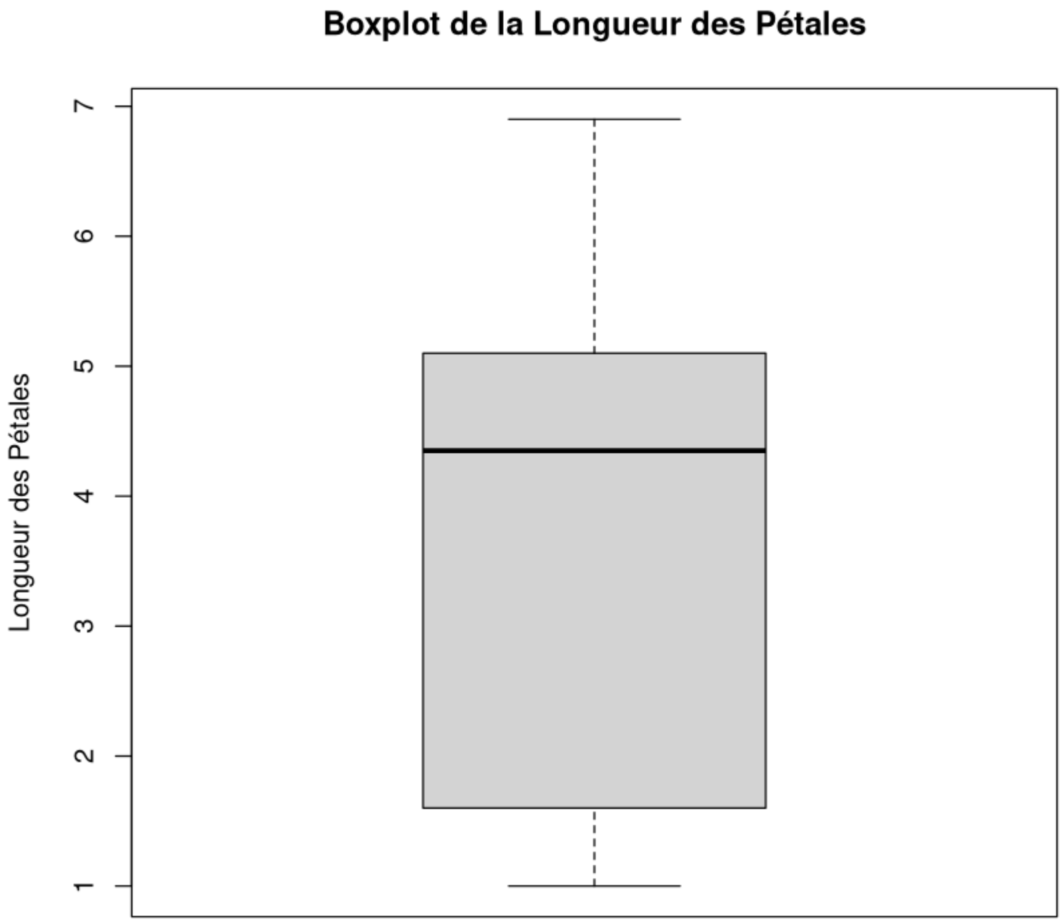
[13]:
```
head(x1) # Affichage des premières lignes du jeu de données x1.
```

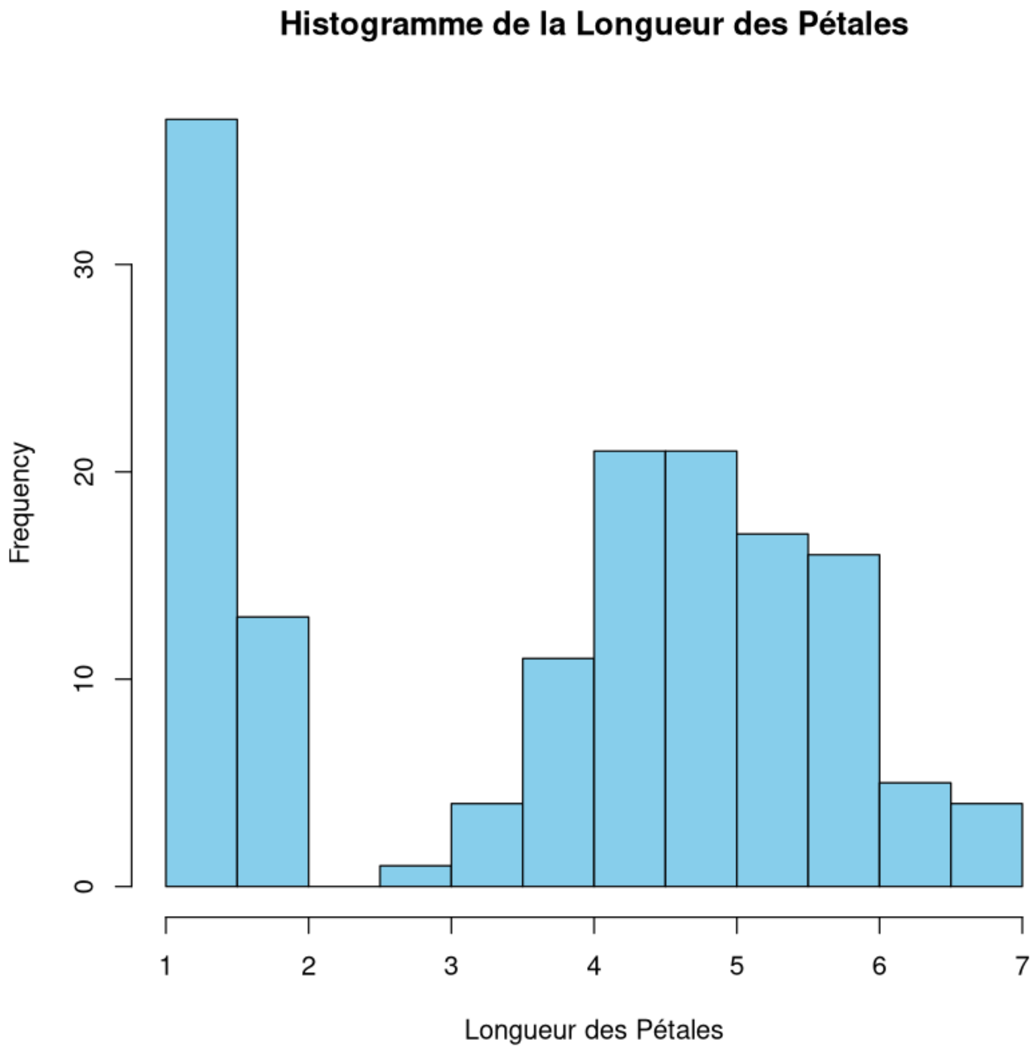
A data.frame: 6 × 5

	Sepal.Length <dbl>	Sepal.Width <dbl>	Petal.Length <dbl>	Petal.Width <dbl>	Species <fct>
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

0.4.1 1°)Traçons le boxplot et l'histogramme des longueurs de pétales des 150 observations

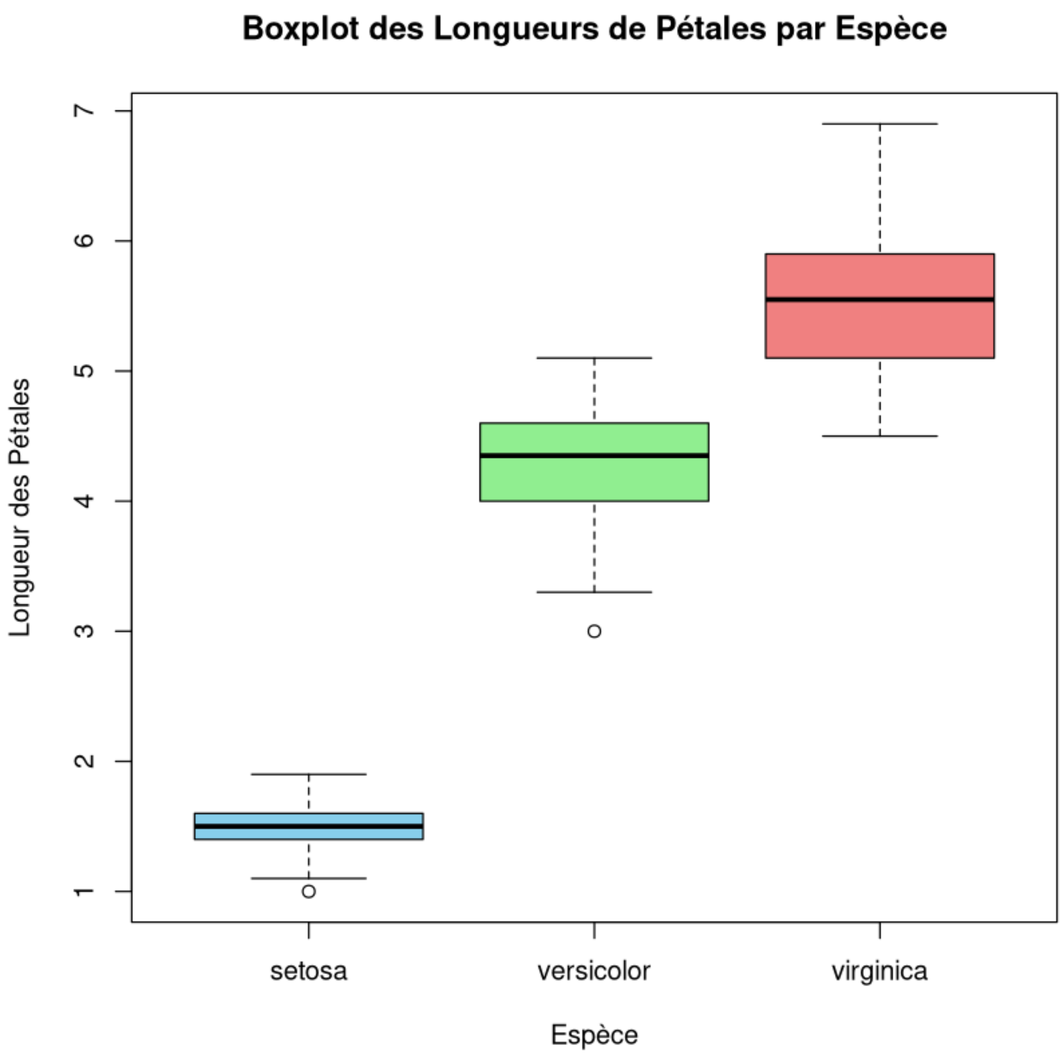
[12]:
```
# Charger la bibliothèque ggplot2 si elle n'est pas déjà installée
# install.packages("ggplot2")
library(ggplot2)

# Charger l'ensemble de données Iris
data(iris)

# Tracer le boxplot des longueurs de pétales
boxplot(iris$Petal.Length, main="Boxplot de la Longueur des Pétales",␣
 ↪ylab="Longueur des Pétales")

# Tracer l'histogramme des longueurs de pétales
hist(iris$Petal.Length, main="Histogramme de la Longueur des Pétales",␣
 ↪xlab="Longueur des Pétales", col="skyblue", border="black")
```


Boxplot de la Longueur des Pétales
Longueur des Pétales
1
2
3
4
5
6
7

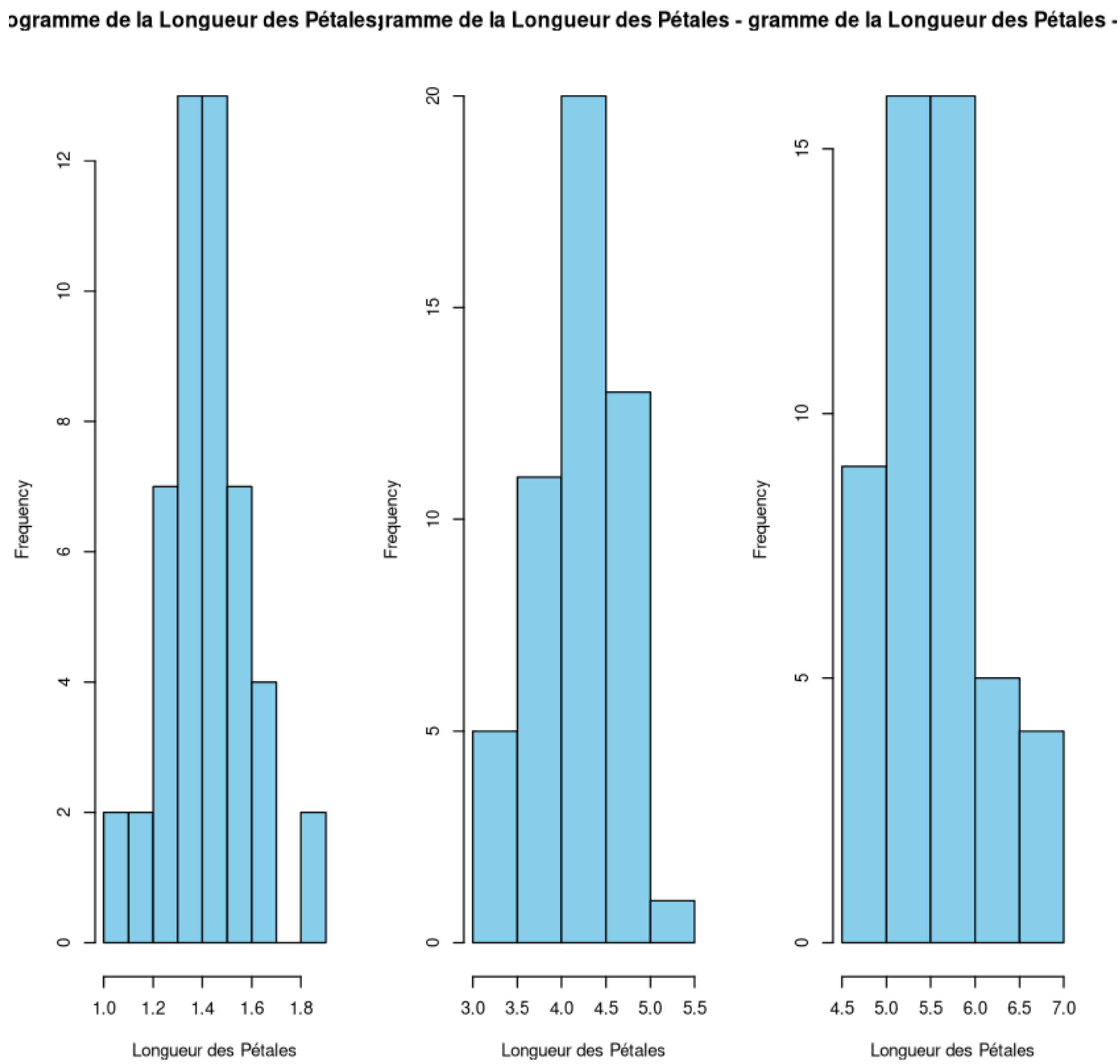


0.4.2 2°)Même question en différenciant les trois échantillons constitués par les trois espèces. On pourra pour cela étudier l'argument formula et la commande boxplot.

[14]:
```
# Trace le boxplot des longueurs de pétales en différenciant les espèces
boxplot(Petal.Length ~ Species, data=iris,
        main="Boxplot des Longueurs de Pétales par Espèce",
        ylab="Longueur des Pétales", xlab="Espèce",
        col=c("skyblue", "lightgreen", "lightcoral"),
        border="black")
```



[15]:
```
# Trace l'histogramme des longueurs de pétales en différenciant les espèces
par(mfrow=c(1,3))  # Divise la fenêtre graphique en 1 ligne et 3 colonnes

for (species in unique(iris$Species)) {
  subset_data <- subset(iris, Species == species)
  hist(subset_data$Petal.Length,
       main=paste("Histogramme de la Longueur des Pétales -", species),
       xlab="Longueur des Pétales", col="skyblue", border="black")
}

par(mfrow=c(1,1))  # Rétablit la configuration par défaut de la fenêtre␣
↪graphique
```



0.4.3 COMMENTAIRE

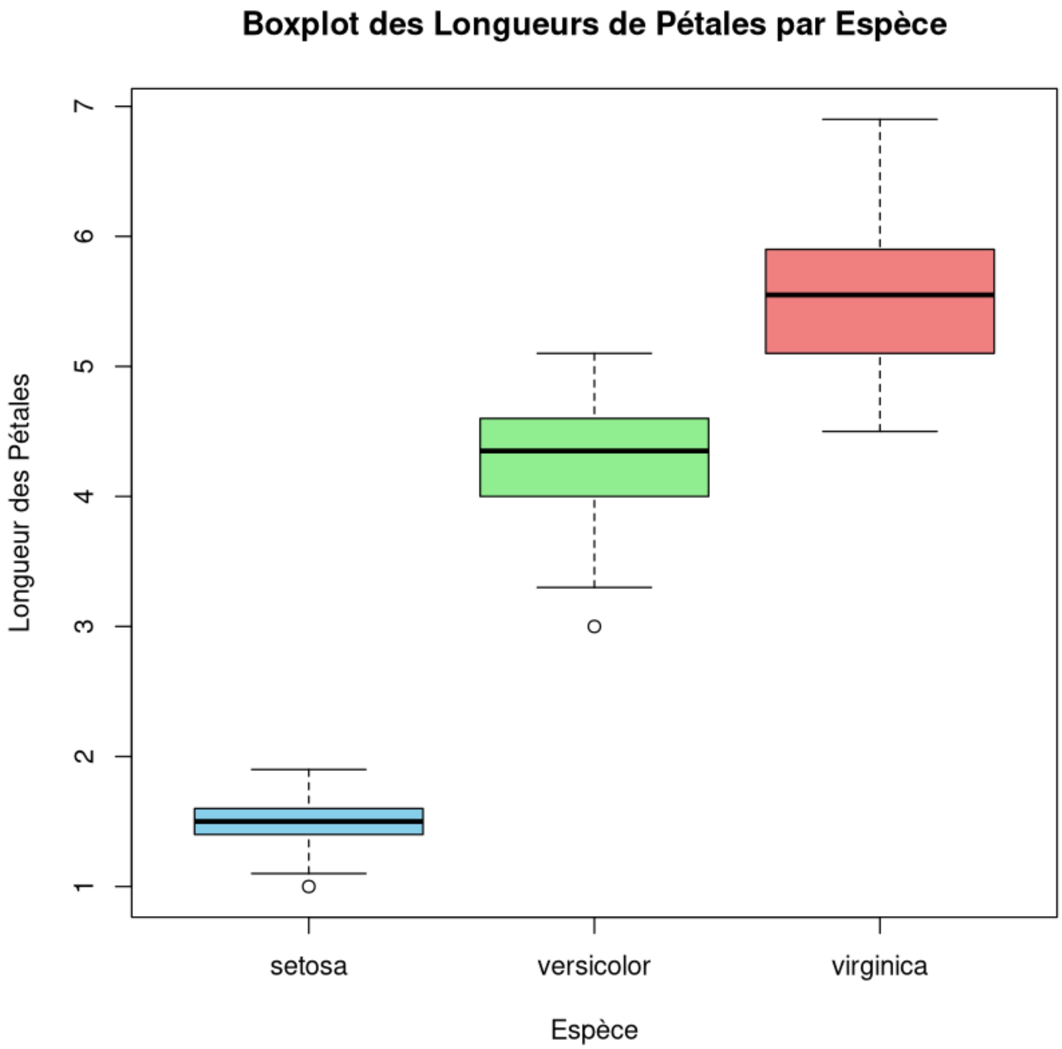
Ces graphiques permettent de visualiser la distribution des longueurs de pétales pour chaque espèce dans l'ensemble de données Iris.

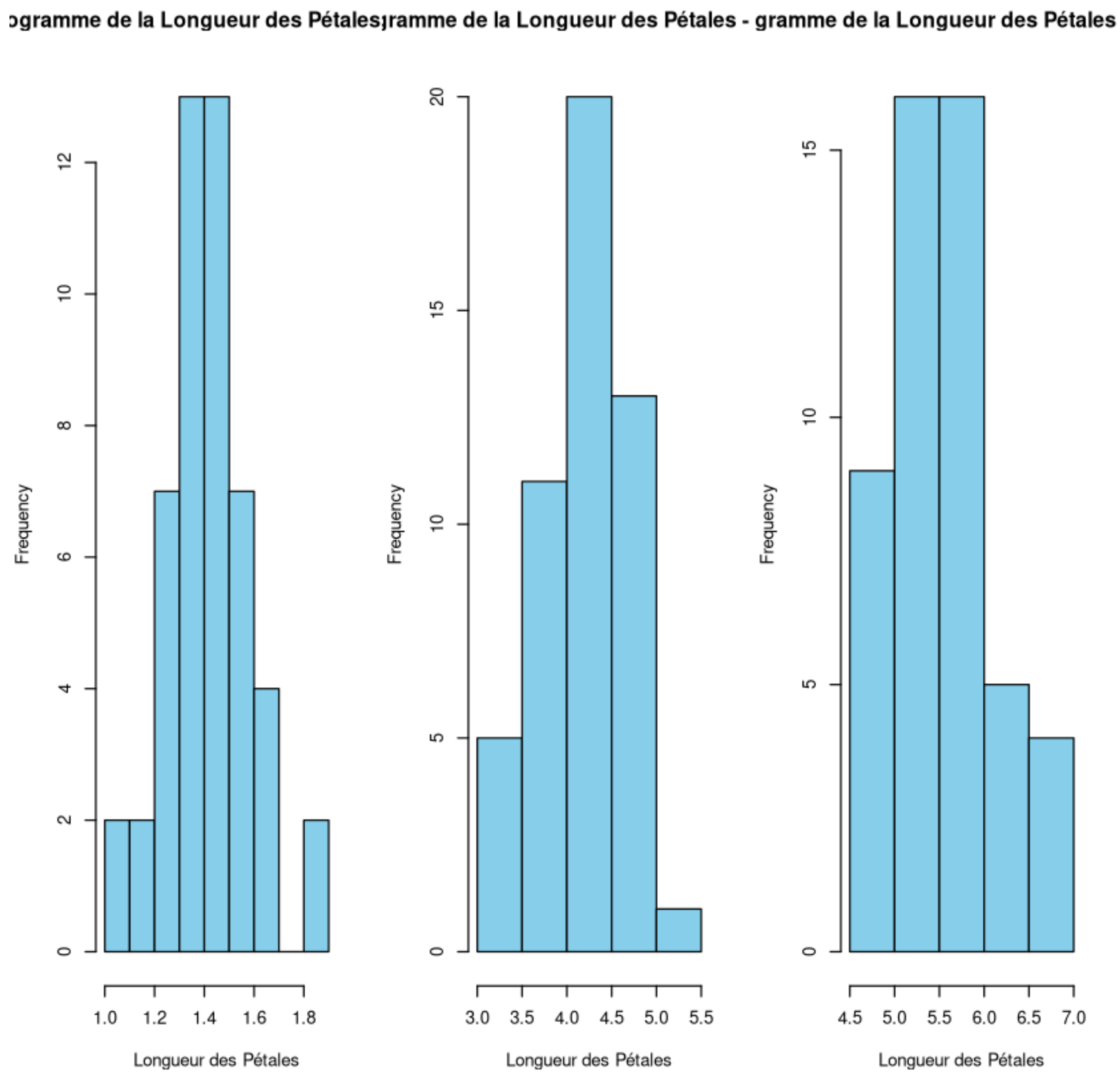
0.4.4 3°) Comment illustrer la problématique soulevée dans la question précédente?

Pour illustrer la problématique soulevée dans la question précédente, qui portait sur la différenciation des échantillons constitués par les trois espèces dans l'ensemble de données Iris, nous pouvons comparer visuellement les distributions des longueurs de pétales pour chaque espèce à l'aide de boxplots et d'histogrammes. Ces graphiques permettent d'observer les tendances et les variations au sein de chaque espèce.

En ajoutant un peu de contexte à l'illustration, supposons que nous cherchons à comprendre comment les longueurs de pétales varient entre les différentes espèces d'Iris (setosa, versicolor, virginica). L'objectif est d'analyser graphiquement ces variations.

[16]:
```
# Charge la bibliothèque ggplot2 si elle n'est pas déjà installée
# install.packages("ggplot2")
library(ggplot2)

# Charge l'ensemble de données Iris
data(iris)

# Boxplot des longueurs de pétales par espèce
boxplot(Petal.Length ~ Species, data=iris,
        main="Boxplot des Longueurs de Pétales par Espèce",
        ylab="Longueur des Pétales", xlab="Espèce",
        col=c("skyblue", "lightgreen", "lightcoral"),
        border="black")

# Histogrammes des longueurs de pétales par espèce
par(mfrow=c(1,3))  # Divise la fenêtre graphique en 1 ligne et 3 colonnes

for (species in unique(iris$Species)) {
  subset_data <- subset(iris, Species == species)
  hist(subset_data$Petal.Length,
       main=paste("Histogramme de la Longueur des Pétales -", species),
       xlab="Longueur des Pétales", col="skyblue", border="black")
}

par(mfrow=c(1,1))  # Rétablit la configuration par défaut de la fenêtre␣
 ↪graphique
```


Boxplot des Longueurs de Pétales par Espèce
Longueur des Pétales
1
2
3
4
5
6
7
setosa
versicolor
virginica
Espèce



L'illustration comprend un boxplot qui montre la répartition des longueurs de pétales pour chaque espèce, ainsi que trois histogrammes, un pour chaque espèce, pour visualiser les distributions de manière plus détaillée. Cela permettra à quiconque observe les graphiques de mieux comprendre comment les longueurs de pétales varient entre les espèces d'Iris.

0.4.5 4°) Discutons ce que ce type de phénomène illustre. et proposons des pistes de prise en compte en terme de modélisation

Le type de phénomène illustré par l'analyse des longueurs de pétales des différentes espèces dans l'ensemble de données Iris montre des variations et des tendances spécifiques à chaque espèce. Cela met en évidence la diversité biologique entre les groupes et souligne l'importance de prendre en compte ces différences lors de la modélisation.

Discussion sur le phénomène :

1. **Diversité biologique :** Les espèces biologiques peuvent présenter des variations significatives dans leurs caractéristiques physiques. Dans le cas des Iris, les longueurs de pétales varient entre les espèces (setosa, versicolor, virginica).
2. **Identification des caractéristiques distinctives :** L'analyse visuelle des boxplots et des histogrammes permet d'identifier les caractéristiques distinctives de chaque espèce. Par exemple, certaines espèces peuvent avoir des longueurs de pétales plus homogènes, tandis que d'autres peuvent montrer une plus grande variabilité.

Prise en compte dans la modélisation :

1. **Modèles spécifiques à chaque espèce :** Si les différences entre les espèces sont importantes, il peut être pertinent de construire des modèles spécifiques à chaque espèce plutôt qu'un modèle unique pour l'ensemble des données. Cela permettrait de mieux capturer les variations inhérentes à chaque groupe.
2. **Variables indicatives :** En plus des caractéristiques communes, l'inclusion de variables spécifiques à chaque espèce pourrait améliorer la précision du modèle. Ces variables indicatives pourraient être des caractéristiques uniques à chaque espèce qui contribuent significativement à la variabilité des données.
3. **Interactions entre variables :** Examiner les interactions entre les caractéristiques et les espèces peut être crucial. Par exemple, une caractéristique spécifique peut avoir un impact différent sur la longueur des pétales en fonction de l'espèce.
4. **Validation croisée spécifique à l'espèce :** Lors de la validation du modèle, il peut être judicieux d'adopter une approche spécifique à l'espèce pour garantir que le modèle est robuste pour chaque groupe.
5. **Diagnostic des résidus par espèce :** Examiner les résidus par espèce peut aider à identifier des modèles inadéquats ou des caractéristiques manquantes spécifiques à une espèce.

En résumé, la prise en compte des différences entre les espèces dans la modélisation est essentielle pour obtenir des résultats précis et généralisables. L'utilisation de modèles spécifiques à chaque espèce et l'exploration des caractéristiques spécifiques à chaque groupe peuvent améliorer la capacité du modèle à capturer la complexité du phénomène étudié.

0.4.6 5°) Paradoxe de Simpson

Oui, le paradoxe de Simpson est un phénomène statistique dans lequel une tendance ou un effet qui apparaît dans différents groupes de données disparaît ou s'inverse lorsque ces groupes sont combinés. Cela peut conduire à des conclusions trompeuses si l'on ne prend pas en compte la structure sous-jacente des données.

Explication du Paradoxe de Simpson :

Le paradoxe de Simpson se produit lorsque les relations entre des variables dans différents sous-groupes sont inversées ou modifiées lorsque ces sous-groupes sont combinés. Cela peut se produire en présence de variables de confusion qui influent sur les relations observées.

Illustration avec un Exemple :

Considérons un exemple classique du paradoxe de Simpson impliquant des taux de succès dans deux groupes de patients (A et B) traités par deux médecins différents (M1 et M2). Les données sont les suivantes :

- Groupe A (M1) : 90 patients, taux de succès = 70%
- Groupe B (M1) : 30 patients, taux de succès = 30%
- Groupe A (M2) : 20 patients, taux de succès = 60%
- Groupe B (M2) : 80 patients, taux de succès = 40%

Si nous regardons chaque groupe individuellement, il semble que le médecin M1 a un taux de succès plus élevé dans chaque groupe par rapport au médecin M2. Cependant, si nous combinons les données pour obtenir les taux de succès globaux pour chaque médecin :

- Médecin M1 : (90 x 0.7 + 30 x 0.3) / (90 + 30)   0.625 (62.5%)
- Médecin M2 : (20 x 0.6 + 80 x 0.4) / (20 + 80)   0.4 (40%)

Maintenant, le médecin M2 a un taux de succès global plus élevé que le médecin M1, même si le taux de succès était initialement plus élevé pour M1 dans chaque groupe.

Le paradoxe de Simpson souligne l'importance de prendre en compte les variables de confusion et de ne pas tirer de conclusions hâtives en agrégeant des données sans tenir compte de la structure sous-jacente des groupes. Il met en évidence la complexité de l'interprétation des données dans des contextes où des variables supplémentaires peuvent influencer les relations observées.

https://www.youtube.com/watch?v=ev8zusJ7BCg&t=5s

Ce lien youtube peut également être proposé pour plus d'exemple et de compréhension.

[ ]:

```
```