FIVB (Fédération Internationale de VolleyBall)

Begimay KONUSHBAEVA & Camille CABROL

2023-05-05

Contents

Introduction

Notre jeu de données est récupéré du site Kaggle. Elles concernent des joueuses de volleyball de la FIVB (Fédération Internationale de VolleyBall) et contiennent des informations sur leur date de naissance, leur taille, leur poids, leur hauteur de saut, leur hauteur de bloc, leur poste de jeu et leur pays d'origine.

Chargement du jeu de données

```
url <- "women_vb.csv"
data <- read.csv(url,sep=';', header=TRUE)

attach(data)
```

Les données décrivent chaque joueuse de la manière suivante

- **nom** Le nom et le prénom de la joueuse
- **annee_naissance** L date de naissance de la joueuse
- **taille** La taille en cm de la joueuse
- **poids** Le poids de la joueuse en kg
- **attaque** La hauteur maximale d'attaque de la joueuse
- **block** La hauteur maximale du block de la joueuse

- **poste** Le poste occupé par la joueuse : Réceptionneuse/attaquante, centrale, pointue, libero, passeuse,
- **pays** Le pays de naissance de la joueuse

Nous pouvons répartir nos données en 2 catégories :

Quantitatives	Qualitatives
Taille (en cm)	Nom
Poids (en kg)	Année de naissance
Attaque (en cm)	Poste
Block (en cm)	Pays de naissance

Après consultation du contenu de notre jeu de données, nous constatons que notre fichier contient 432 lignes. Or, il y a énormément de doublons. En effet, il y a une répétition du jeu de données qui apparaît en 3 exemplaires. Nous avons donc supprimé les doublons, ce qui nous a amené à 144 lignes. Enfin nous sommes arrivées à 143 après avoir constaté qu'une joueuse avait une hauteur de block à 0 cm et avons donc fait le choix de retirer cette joueuse du jeu de données.

En résumé :

- Taille de l'échantillon = 143
- Unité statistique : individu (joueuses de volleyball)

L'analyse effectuée portera sur les questions suivantes :

- Est-ca que les performances sportives (dans notre cas la hauteur d'attaque) des joueuses changent selon leur pays?
- Est-ce que le poste d'une joueuse dépend de son pays de naissance?
- Est-ce que la hauteur de l'attaque dépend de la taille d'une joueuse?

La hauteur d'attaque selon le pays de naissance

Le pays de naissance des joueuses

Nous allons commencer par effectuer une analyse du pays de naissance des joueuses, qui est une variable *qualitative discrète*.

Pour résumer l'information contenue dans la variable nous commençons par un *tri à plat*.

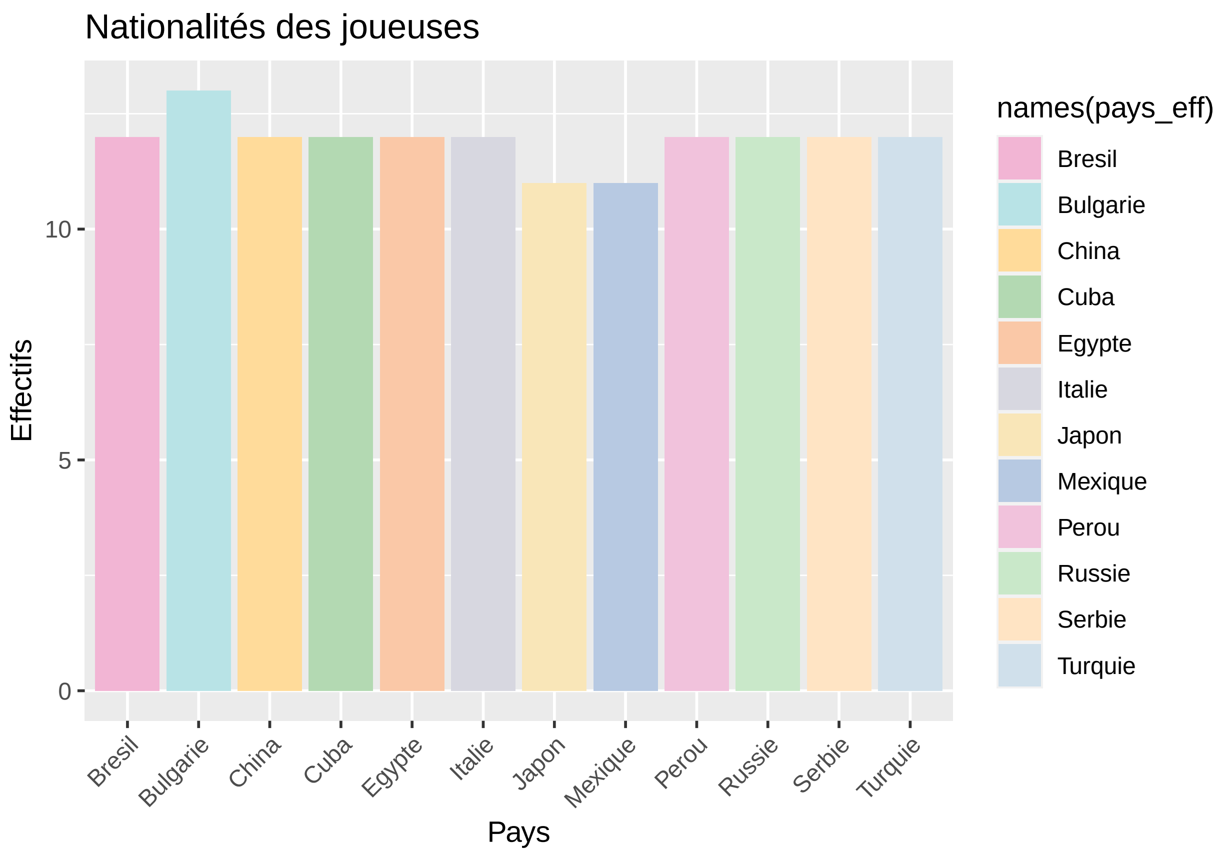
```
# Calculer les effectifs de chaque pays
pays_eff <- table(pays)
```

Répartition des données

```
analyse_pays <- data.frame(names(pays_eff), pays_eff)


pale_colors <- c("#F2B5D4", "#B8E3E6", "#FFDB9A", "#B3D9B2", "#FAC8A7", "#D7D7E0",
                 "#F9E6B8", "#B7C9E2", "#F1C2DC", "#C9E8C9", "#FFE4C4", "#D0E0EB")

# Create the bar plot
ggplot(analyse_pays, aes(x = names(pays_eff),
  y = pays_eff, fill = names(pays_eff))) +
  geom_bar(stat = "identity") +
  scale_fill_manual(values = pale_colors) +
  labs(title = "Nationalités des joueuses",
```

```
  x = "Pays", y = "Effectifs") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  coord_cartesian(clip = "off")
```

```
## Don't know how to automatically pick scale for object of type <table>.
## Defaulting to continuous.
```

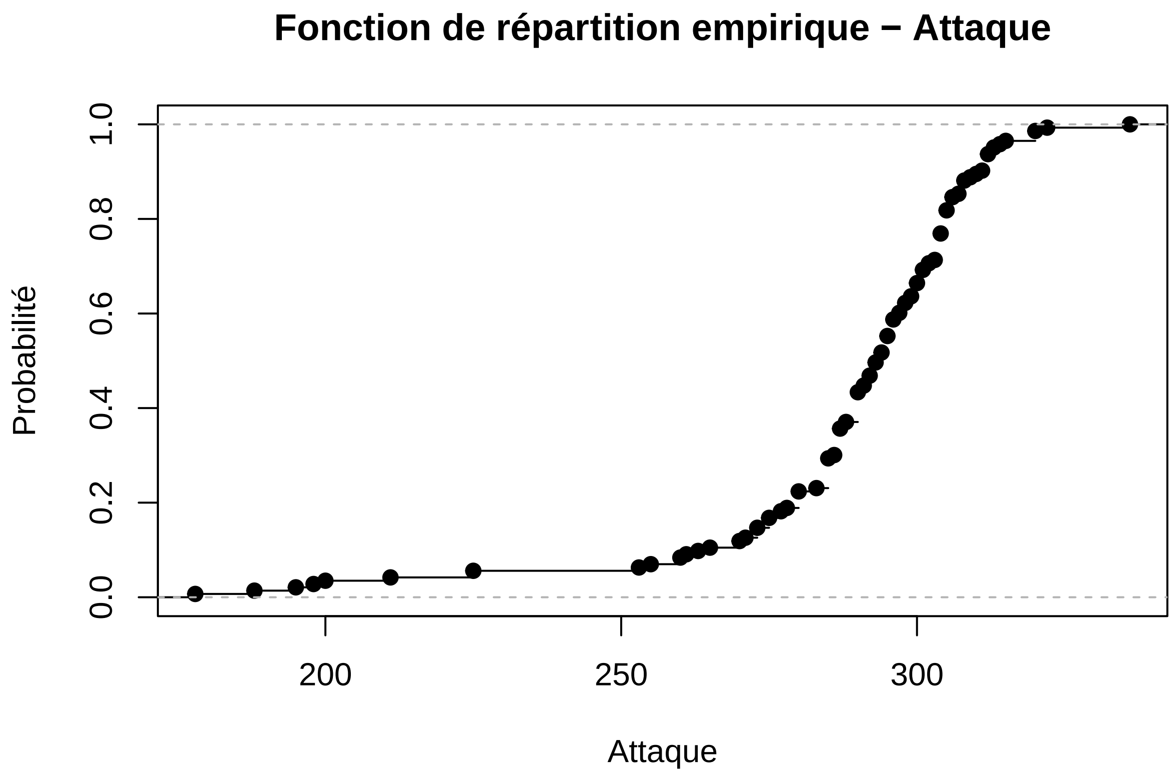


En analysant ce diagramme, il est clair que la plupart des pays de naissance des joueuses ont une fréquence de 12, ce qui signifie qu'il y a un nombre égal de joueuses nées dans ces pays. Cependant, il y a deux exceptions à cette tendance : le Japon et le Mexique, qui sont les pays de naissance de 11 joueuses chacun.

Il est intéressant de noter que la Bulgarie a l'effectif maximal dans ce diagramme, avec 13 joueuses nées dans ce pays. Cela indique que la Bulgarie est le pays d'origine le plus représenté.

La hauteur d'attaque des joueuses

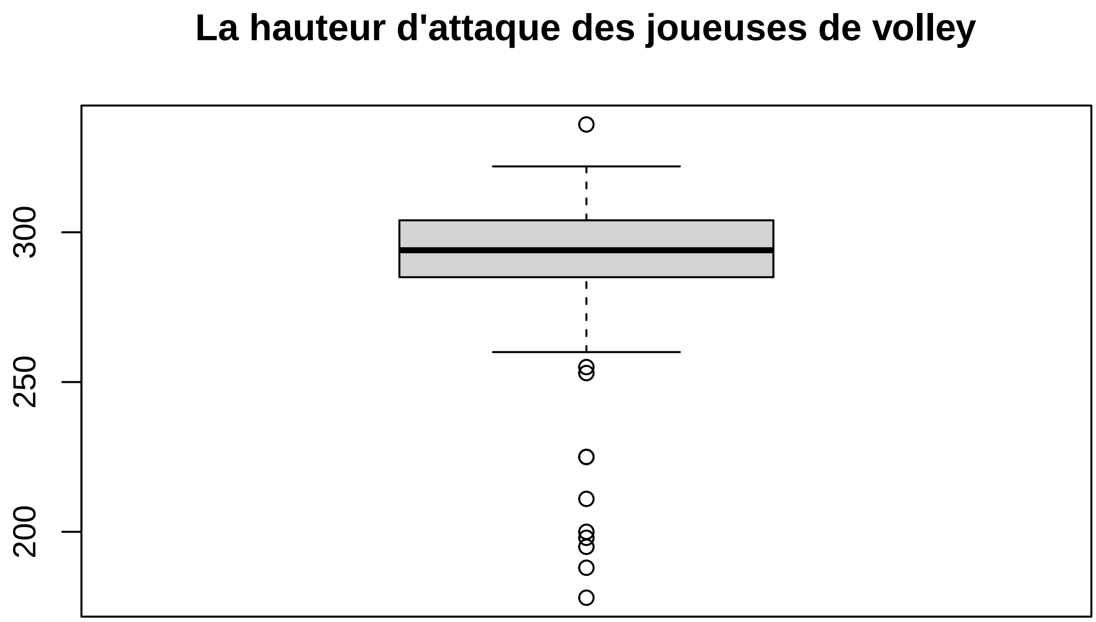
```
# Calcul de la fonction de répartition empirique
ecdf_attaque <- ecdf(attaque)

# Tracé de la fonction de répartition empirique
plot(ecdf_attaque, main = "Fonction de répartition empirique - Attaque", xlab = "Attaque", ylab = "Proba
```



Les nombreux sauts de petite taille observés dans la variable suggèrent fortement que cette variable *quantitative* est *continue.*

Mesures de position

```r
boxplot(attaque, main="La hauteur d'attaque des joueuses de volley")
```



En observant ce diagramme, on peut clairement distinguer la valeur minimale, les quartiles, la médiane et la valeur maximale.

Voici ces valeurs

```
summary(attaque)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   178.0   285.0   294.0   288.8   304.0   336.0 
```

Probabilité et estimations

```
interval=t.test(attaque)
interval$conf.int
```

Estimation de moyenne

```
## [1] 284.5521 293.0283
## attr(,"conf.level")
## [1] 0.95
```

```
mean(attaque)
```

```
## [1] 288.7902
```

Nous pouvons estimer que la vraie moyenne de l'attaque se situe dans l'intervalle [284.5521, 293.0283] avec un niveau de confiance de 95%.

L'estimateur de la moyenne est égal à 288.79

```r
# Calcul de l'intervalle de confiance pour la variance
n <- length(attaque)  # Taille de l'échantillon
alpha <- 0.05  # Niveau de confiance (ici 95%)

lower_bound <- (n - 1) * var(attaque) / qchisq(1 - alpha/2, df = n - 1)
upper_bound <- (n - 1) * var(attaque) / qchisq(alpha/2, df = n - 1)

# Affichage de l'intervalle de confiance
confidence_interval <- c(lower_bound, upper_bound)
confidence_interval
```

Estimation de variance

```
## [1] 527.6508 841.5602
```

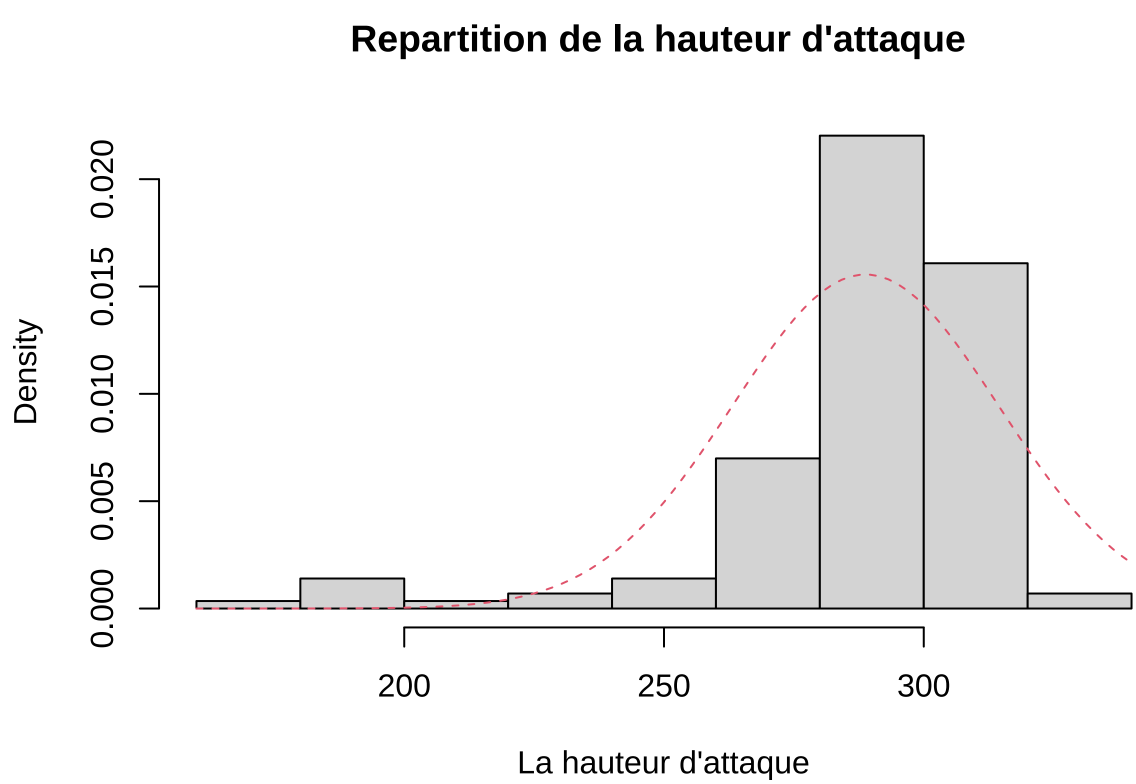
```r
var(attaque)
```

```
## [1] 657.2655
```

De même nous estimons que la variance de la hauteur d'attaque des joueuses se situe dans l'intervalle [527.6508 841.5602] avec un niveau de confiance de 95%.

L'éstimateur de la variance est égale à 657.27

```r
hist(attaque , main="Repartition de la hauteur d'attaque ",xlab="La hauteur d'attaque",prob=T)
curve(dnorm(x, mean(attaque), sd(attaque)), col=2, add=TRUE, lty=2)
```



Caractère gaussien

```
qqnorm(attaque, main = "QQ-Plot", xlab = "Théorique", ylab = "Observé")
abline(mean(attaque),sd(attaque),col=2)
```

Sur ce graphique, nous pouvons observer que l'histogramme de la distriburion de la hauteur d'attaque n'est pas symétrique. Nous voyons que la variable visiblement ne s'approche pas à la loi normale. Nous allons vérifier ces propos à l'aide de tests statistiques.

Pour valider notre estimation, nous procéderons aux tests statistiques.

Test de la normalité de distribution Les hypothèses sont les suivantes :

- H0 : La variable suit la loi normale
- H1 : La variable ne suit pas la loi normale

```
shapiro.test(attaque)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  attaque
## W = 0.77567, p-value = 1.619e-13
```

La p valeur du test est extrêmement faible, soit 1.619e-13, ce qui indique que nous avons suffisamment de preuves pour rejeter l'hypothèse nulle H0 selon laquelle les données suivent une distribution normale.

Analyse bivariée

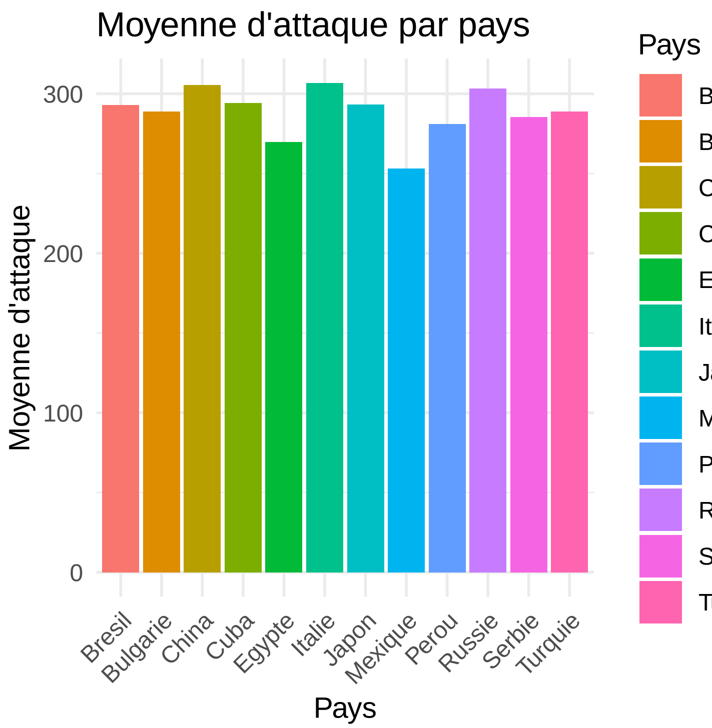
La hauteur d'attaque selon le pays

Est-ca que les performances sportives (dans notre cas la hauteur d'attaque) des joueuses changent selon leur pays?

Après avoir étudié les pays de naissance des joueuses et leurs hauteurs d'attaque, nous pouvons désormais analyser la hauteur d'attaque selon le pays de naissance des joueuses.

La hauteur d'attaque des joueuses est une variable *quantitative continue*, tandis que leur pays de naissance est une variable *qualitative discrète*.

```r
# Calculer la moyenne d'attaque pour chaque pays
moyenne_attaque <- aggregate(attaque, by = list(pays), FUN = mean)

colnames(moyenne_attaque) <- c("Pays", "Moyenne_attaque")
# Créer le diagramme à barres avec ggplot2
ggplot(moyenne_attaque, aes(x = Pays, y = Moyenne_attaque, fill = Pays)) +
  geom_bar(stat = "identity", position = "stack") +
  labs(title = "Moyenne d'attaque par pays", col= c("#F2B5D4", "#B8E3E6", "#FFDB9A", "#B3D9B2", "#FAC8A
 "#F9E6B8", "#B7C9E2", "#F1C2DC", "#C9E8C9", "#FFE4C4", "#D0E0EB"),
 x = "Pays", y = "Moyenne d'attaque") +
  theme_minimal() +
  theme(plot.margin = margin(1, 5, 1, 1, "cm")) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  coord_cartesian(clip = "off")
```



Moyennes de la hauteur d'attaque de chaque pays

Nous pouvons voir ici que les moyennes les plus elevées sont au Brésil, en Chine, en Italie et en Serbie.

```
mean(data$attaque[data$pays=="Bresil"])
```

```
## [1] 292.9167
```

```
mean(data$attaque[data$pays=="China"])
```

```
## [1] 305.5833
```

```
mean(data$attaque[data$pays=="Italie"])
```

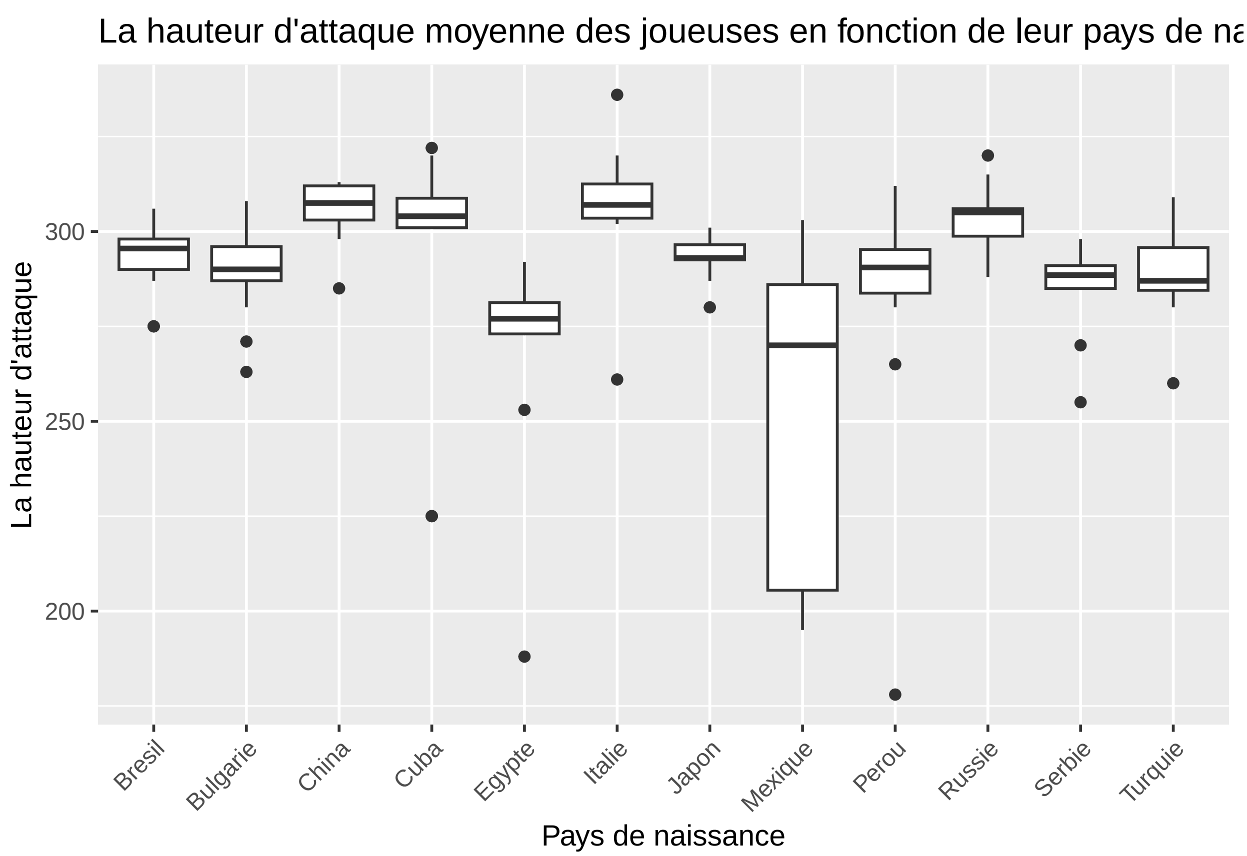
```
## [1] 306.75
```

```
mean(data$attaque[data$pays=="Serbie"])
```

```
## [1] 285.3333
```

La moyenne la plus elevées est en Italie

```
attaque<-data$attaque

ggplot(data, aes(x = pays, y = attaque)) +
  geom_boxplot() +
  labs(title = "La hauteur d'attaque moyenne des joueuses en fonction de leur pays de naissance",
       x = "Pays de naissance",
       y = "La hauteur d'attaque") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Ce graphique en boîte permet d'observer les quartiles et la médiane de la hauteur d'attaque des joueuses des différents sous-groupes, représentés par leur nationalité. Chaque boîte représente la distribution de la hauteur d'attaque pour un pays spécifique.

Nous pouvons voir les pays qui ont les meilleurs aptitudes sportives : la Chine, la Cuba, l'Italie et la Russie (les pays avec une médiane supérieure à 300). Il est interesssant de noter aussi que le pays avec la plus grande variaces est le Mexique ce qui est très différent des autres pays

```
var(data$attaque[pays=="Mexique"])
```

```
## [1] 1858
```

La variance de la hauteur d'attaque de ce sous groupe (pays) est 3 fois plus que la variance du group entier (tous les pays)

Maintenant que nous avons examiné les résultats graphiques, nous pouvons utiliser ces données pour effectuer des tests statistiques. Voici les tests que nous effectuerons en utilisant les résultats de la hauteur d'attaque des différentes nationalités :

- Test de comparaison de moyennes : Sur le graphique, nous remarquons une similitude des moyennes de la hauteur d'attaque entre la Chine et Cuba. Afin de vérifier si cette similitude est statistiquement significative, nous allons effectuer un test d'égalité des moyennes entre ces deux pays.
- Test de comparaison de variances : De même nous observons une similarité entre les variances enre la Cuba et l'Egypte, nous allons donc effectuer un test d'égalité des moyennes pour ces deux pays et ensuite un test d'égalité des variances.

Tests statistiques

- Test d'égalité des moyennes de la hauteur d'attaque entre la Chine et la Cuba. Les hypothèses sont les suivantes
  - H0 Les moyennes sont égales
  - H1 Les moyennes sont différentes

Tests d'égalité des moyennes

```
attaque_cuba <- attaque[pays == "Cuba"]
attaque_china <- attaque[pays == "China"]
t.test(attaque_cuba, attaque_china)
```

```
##
##  Welch Two Sample t-test
##
## data:  attaque_cuba and attaque_china
## t = -1.1724, df = 12.346, p-value = 0.2632
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -32.805071   9.805071
## sample estimates:
## mean of x mean of y
##  294.0833  305.5833
```

La p-valeur est grande (26,32%) par rapport au risque de 5% de se tromper. Nous pouvons donc accepter l'hypothèse nulle HO selon laquelle les deux moyennes sont égales et rejetter l'hypothèse H1 selon laquelle les deux moyennes sont différentes. Nous constatons également que la différence entre ces deux moyennes se trouve dans l'intervalle [-32.8, 9.8] avec un niveau de confiance de 95%.

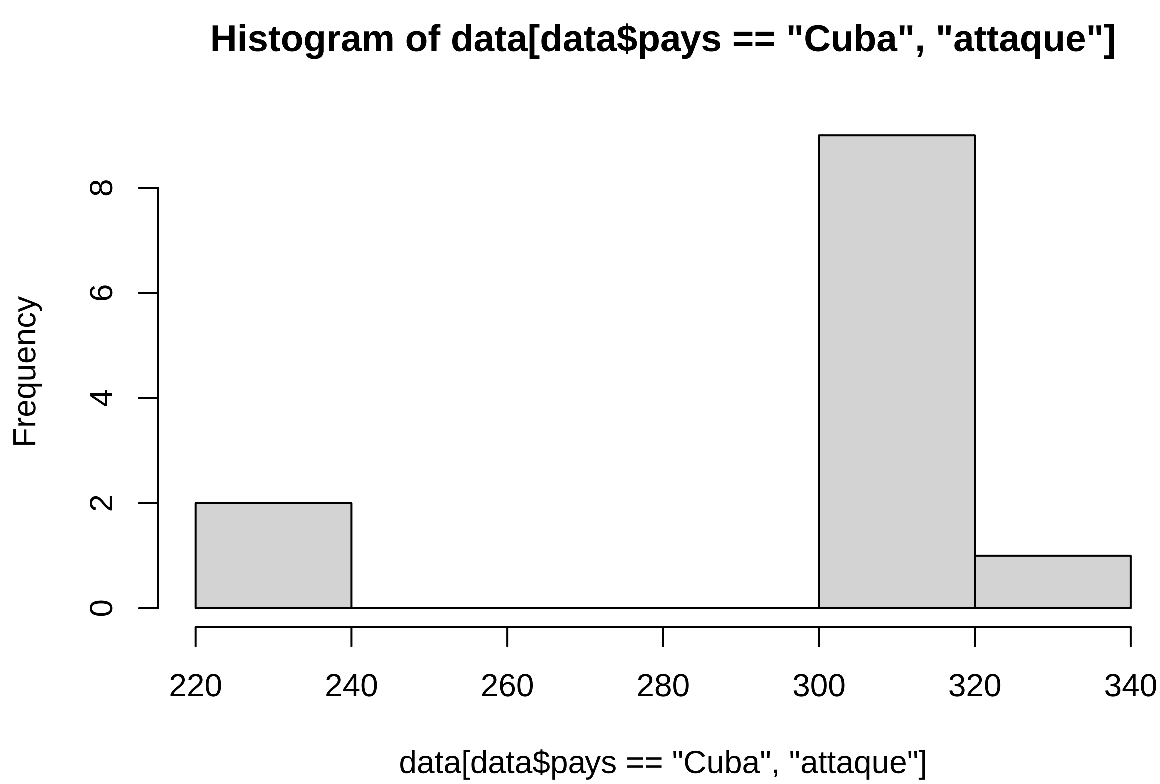
Tests d'égalité des variances

Pour effectuer des tests d'égalité des variances, il est recommandé de vérifier d'abord l'adéquation à la distribution normale des échantillons concernés. Dans notre cas, nous avons deux sous-groupes : Cuba et Égypte. Nous allons donc procéder à des tests de normalité pour ces deux groupes.

Les hypothèses sont les suivantes :

- H0 : La variable suit la loi normale
- H1 : La variable ne suit pas la loi normale

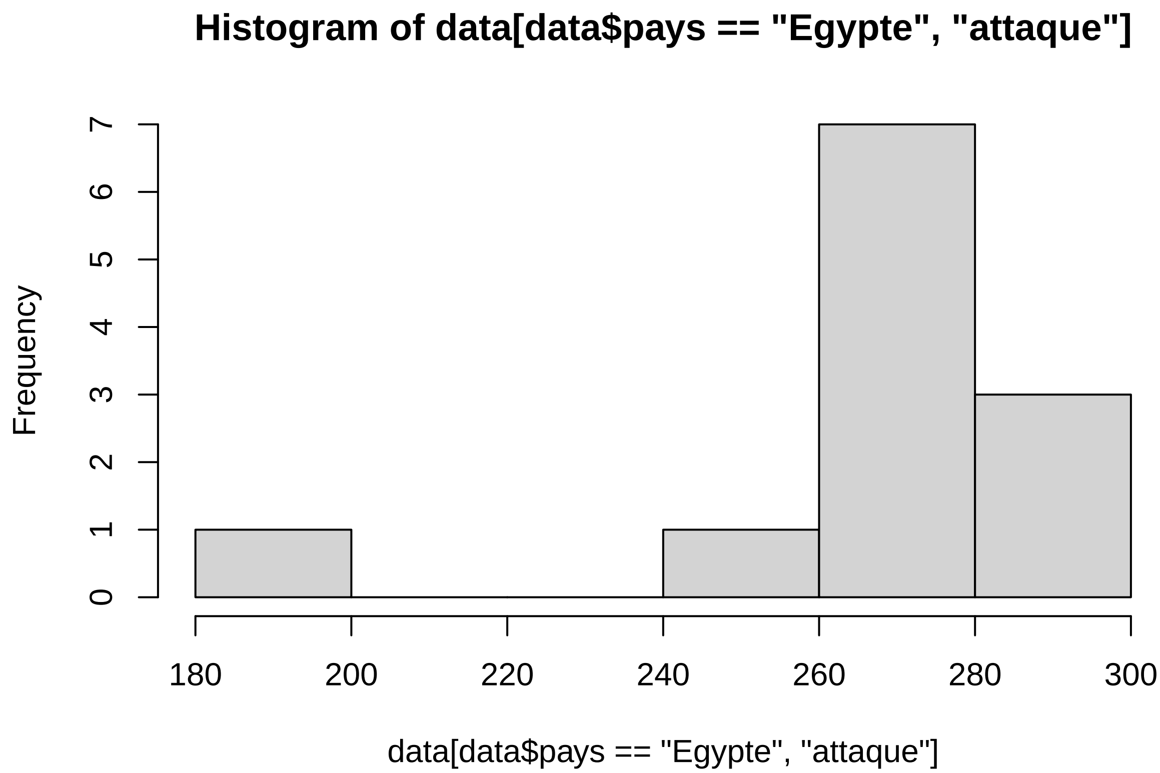
```
hist(data[data$pays == "Cuba", "attaque"])
```




```
shapiro.test(data[data$pays == "Cuba", "attaque"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data[data$pays == "Cuba", "attaque"]
## W = 0.64948, p-value = 0.0002866
```

```
hist(data[data$pays == "Egypte", "attaque"])
```




```
shapiro.test(data[data$pays == "Egypte", "attaque"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data[data$pays == "Egypte", "attaque"]
## W = 0.63874, p-value = 0.0002302
```

Pour les deux tests la p-valeur obtenue est petite, donc nous ne pouvons pas rejjeter l'hypothèse nulle selon laquelle la variable suit la loi normale. Nous supposons donc par la suite que ces deux échantillons ne suivent pas une distribution normale.

Vu que nos échantillons ne suivent pas une distribution normale nous allond vérifier l'égalité des variances de ces deux échantillons avec un test Fisher

Test d'égalité des variance entre la Cuba et l'Egypte :

- H0 Les variances sont égales
- H1 Les variances sont différentes

```
attaque_egypt <- attaque[pays == "Egypte"]
var.test(attaque_cuba, attaque_egypt)
```

```
##
##  F test to compare two variances
##
## data:  attaque_cuba and attaque_egypt
## F = 1.4356, num df = 11, denom df = 11, p-value = 0.5588
## alternative hypothesis: true ratio of variances is not equal to 1
```

```
## 95 percent confidence interval:
##  0.4132863 4.9869538
## sample estimates:
## ratio of variances
##           1.435632
```

La p-valeur est petite mais légérement plus grande que le risque de 5% nous ne pouvons donc pas rejetter l'hypothèse nulle selon laquelle les variances sont égales.

Nous supposons dans la suite que les variances de ces deux échantillons sont égales.

Test d'égalité des moyennes de la hauteur d'attaque entre la Cuba et l'Égypte. Les hypothèses sont les suivantes :

- H0 Les moyennes sont égales
- H1 Les moyenne sont différentes

```
t.test(attaque_cuba, attaque_egypt, var.equal = T)
```

```
##
##  Two Sample t-test
##
## data:  attaque_cuba and attaque_egypt
## t = 1.9622, df = 22, p-value = 0.06251
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.384496 50.051163
## sample estimates:
## mean of x mean of y
##  294.0833  269.7500
```

La p-valeur est plus petite que dans le premier test mais elle reste supérieure au risque de 5%. Nous n'avons donc pas suffisament de preuve pour rejetter l'hypothèse nulle HO selon laquelle les deux moyennes sont égales par consequent nous l'acceptons. Nous observons que la différence entre ces deux moyennes se trouve dans l'intervalle [-1.38, 50.05] avec un niveau de confiance de 95%.

Les postes des joueuses selon leur pays de naisssance

Est-ce que le poste d'une joueuse dépend de son pays de naissance?

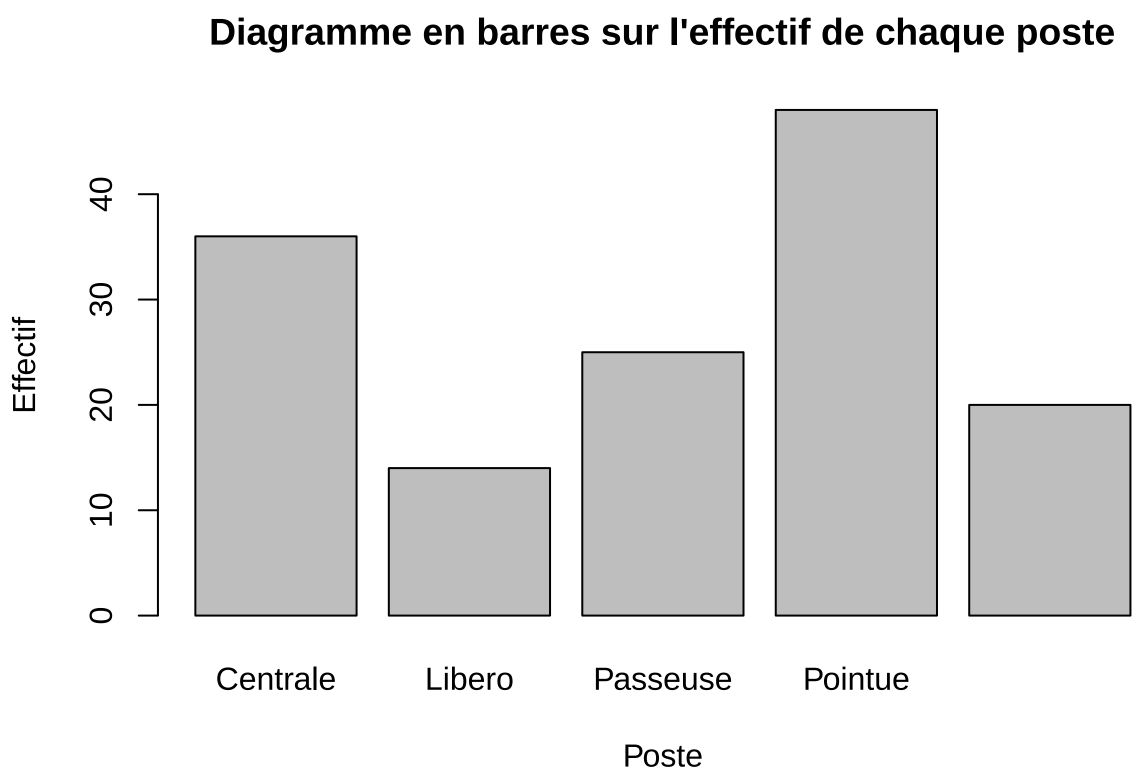
Poste

Pour résumer l'information contenu dans notre variable, nous réalisons un tri à plat.

```
# Utilisation de la fonction table() pour effectuer le tri à plat
tri_a_plat <- table(poste)
tri_a_plat
```

```
## poste
##                  Centrale                    Libero                  Passeuse
##                        36                        14                        25
##                   Pointue Receptionneuse/attaquante
##                        48                        20
```

```
barplot(tri_a_plat, main = "Diagramme en barres sur l'effectif de chaque poste", xlab = "Poste", ylab =
```



33,5% de notre échantillon est représenté par les pointues. Une répartition est quasi équitable entre les passeuses et les réceptionneuses/attaquantes, tandis que l'on retrouve un peu plus du double de centrale que de libero qui est le poste le moins représenté dans notre échantillon.

```r
# Création du tableau de contingence
table_contingence <- table(pays, poste)

# Affichage du tableau de contingence
print(table_contingence)
```

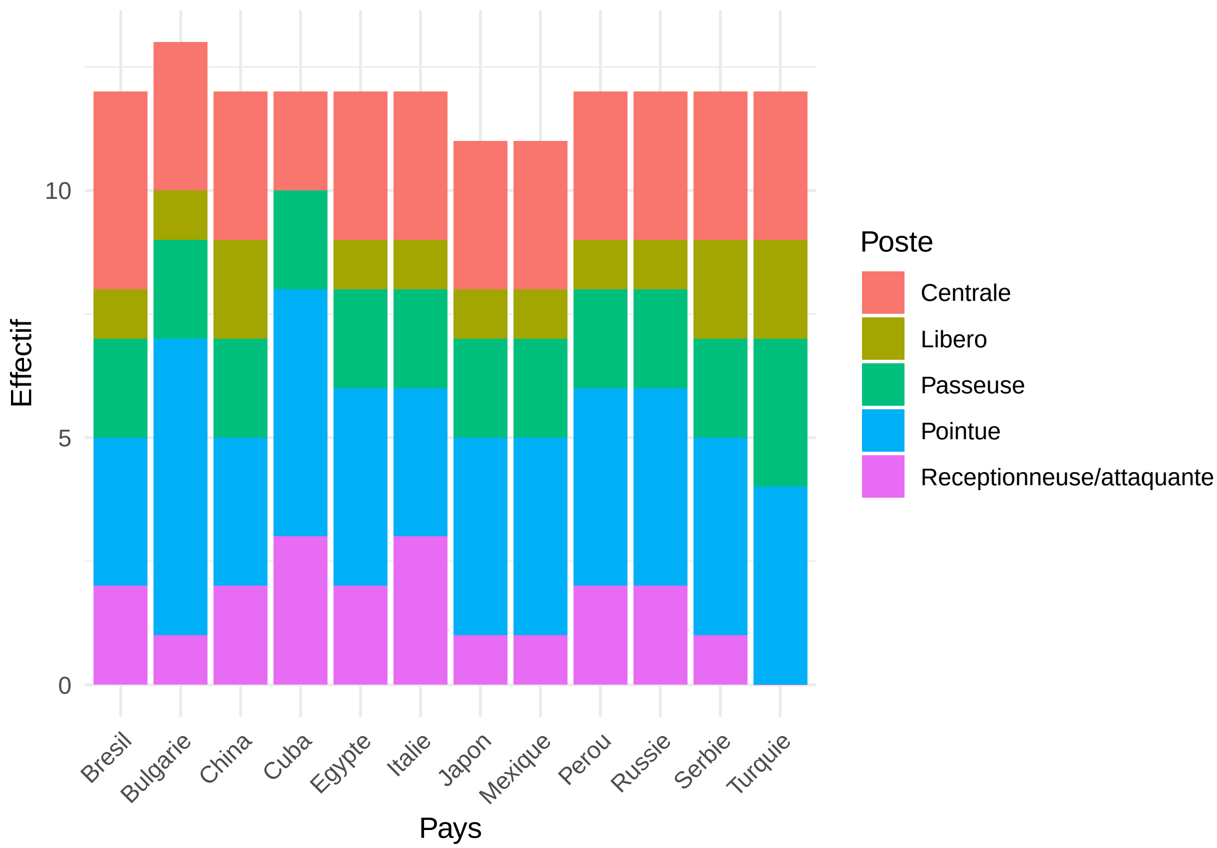
```
##           poste
## pays        Centrale Libero Passeuse Pointue Receptionneuse/attaquante
##   Bresil           4      1        2       3                         2
##   Bulgarie         3      1        2       6                         1
##   China            3      2        2       3                         2
##   Cuba             2      0        2       5                         3
##   Egypte           3      1        2       4                         2
##   Italie           3      1        2       3                         3
##   Japon            3      1        2       4                         1
##   Mexique          3      1        2       4                         1
##   Perou            3      1        2       4                         2
##   Russie           3      1        2       4                         2
##   Serbie           3      2        2       4                         1
##   Turquie          3      2        3       4                         0
```

Nous pouvons observer ici qu'il n'y a pas de receptionneuse/attaquante en Turquie. Nous vouyons que le pays avec le plus grand nombre de joueuses occupant

- le poste de centrale est le Brésil
- le poste libero sont la Serbie, la Turquie et la Chine
- le poste de passeuse est loa Turquie
- le poste de pointue est la Bulgarie
- le poste de receptionneuse/attaquante est la Cuba et l'Italie

Pour illustrer ces données nous proposons un diagramme en barres avec les différents pays et les effectifs de postes pour chaque pays.

```
df <- as.data.frame.table(table_contingence)

# Renommer les colonnes du data frame
colnames(df) <- c("Pays", "Poste", "Effectif")

# Création du graphique en barres
ggplot(df, aes(x = Pays, y = Effectif, fill = Poste)) +
  geom_bar(stat = "identity", position = "stack") +
  labs(x = "Pays", y = "Effectif", fill = "Poste") +
  theme_minimal()+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Test d'indépendence des deux variables

- H0 Les deux variables sont appariées
- H1 Les deux variables sont indépendantes

```
chisq.test(data$pays, data$poste)
```

```
## Warning in chisq.test(data$pays, data$poste): Chi-squared approximation may be
## incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  data$pays and data$poste
## X-squared = 11.076, df = 44, p-value = 1
```

La p-valeur est de 1, donc on a suffisamment de preuves pour rejetter l'hypothèse nulle. Nous concluons que les deux variables sont indépendantes.

Nous pouvons également observer cela sur le diagramme en barres.

La hauteur d'attaque selon la taille

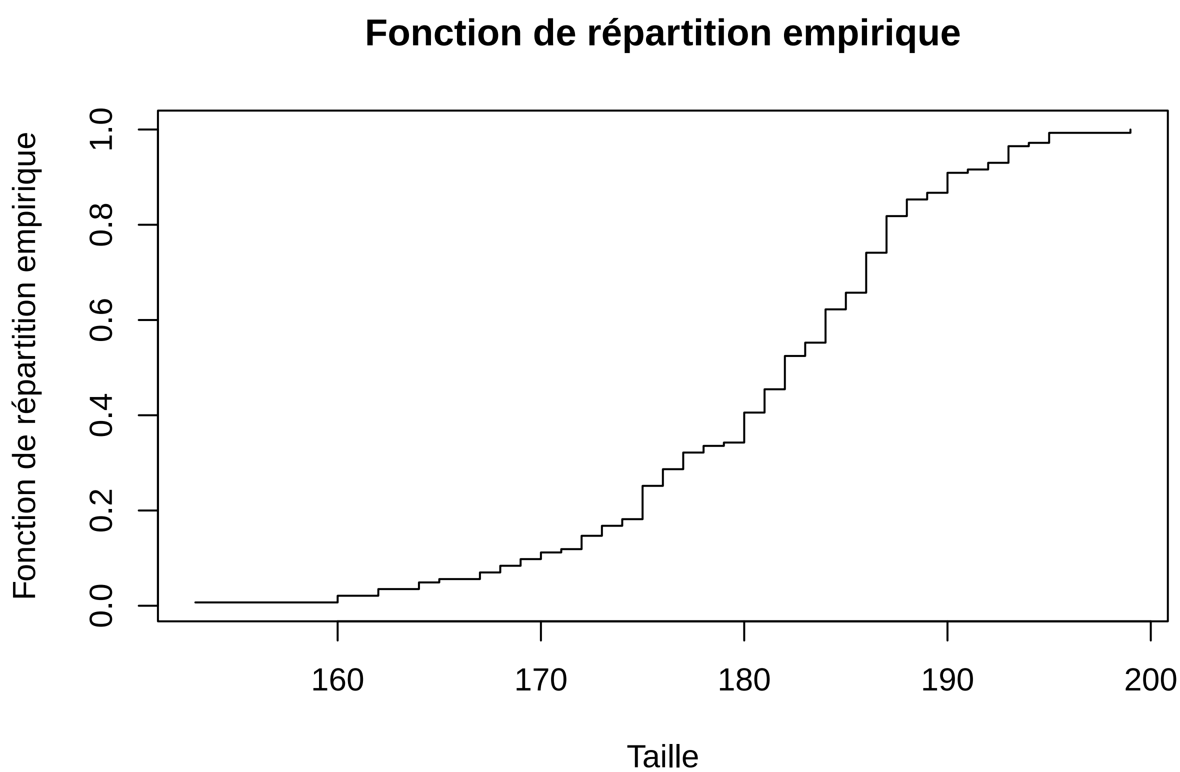
La taille des joueuses

```
#Spécification du nombre de classes souhaitées
nombre_classes <- 3

# Calcul des bornes des classes
bornes_classes <- seq(min(taille), max(taille), length.out = nombre_classes + 1)

# Utilisation de la fonction cut() pour regrouper les données en classes
regroupement_classes <- cut(taille, breaks = bornes_classes, include.lowest = TRUE)

# Extraction des niveaux des classes
niveaux_classes <- levels(regroupement_classes)

# Calcul de l'effectif, de la fréquence et de la fréquence cumulée pour chaque niveau
effectifs <- table(regroupement_classes)
frequences <- prop.table(effectifs)
freq_cumulees <- cumsum(frequences)

# Création d'un tableau avec les résultats
resultats <- data.frame(Niveaux = niveaux_classes, Effectifs = effectifs, Fréquences = paste(round(fr
resultats
```

```
##     Niveaux Effectifs.regroupement_classes Effectifs.Freq Fréquences
## 1 [153,168]                      [153,168]             12      8.39%
## 2 (168,184]                      (168,184]             67     46.85%
## 3 (184,199]                      (184,199]             64     44.76%
##   Fréquences_cumulées
## 1               8.39%
## 2              55.24%
## 3                100%
```

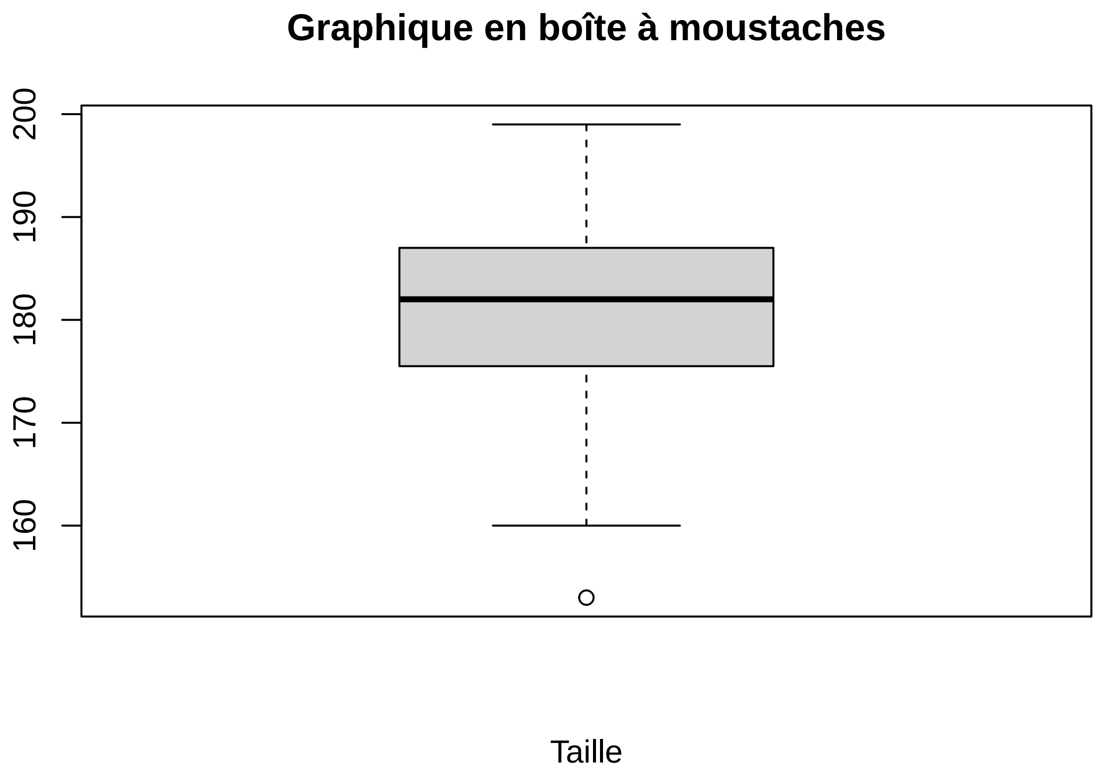
Nous pouvons constater que très peu de joueuses mesurent moins de 168 cm : seulement 12 joueuses, ce qui correspond environ 8% de notre effectif total. Cela nous amène à dire qu'il y a deux classes principales [168, 184] et [184,199] qui comptent respectivement 67 et 64 joueuses. Par rapport à notre effectif total, cela correspond environ à 47% pour la première classe et 45% pour la deuxième classe.

```r
# Calcul de la fonction de répartition empirique
fonction_repartition_empirique <- function(x) {
  sum(taille <= x) / length(taille)
}

# Tri des données brutes
donnees_triees <- sort(taille)

# Création du graphique de la fonction de répartition empirique
plot(donnees_triees, sapply(donnees_triees, fonction_repartition_empirique), type = "s",
     xlab = "Taille", ylab = "Fonction de répartition empirique",
     main = "Fonction de répartition empirique")
```



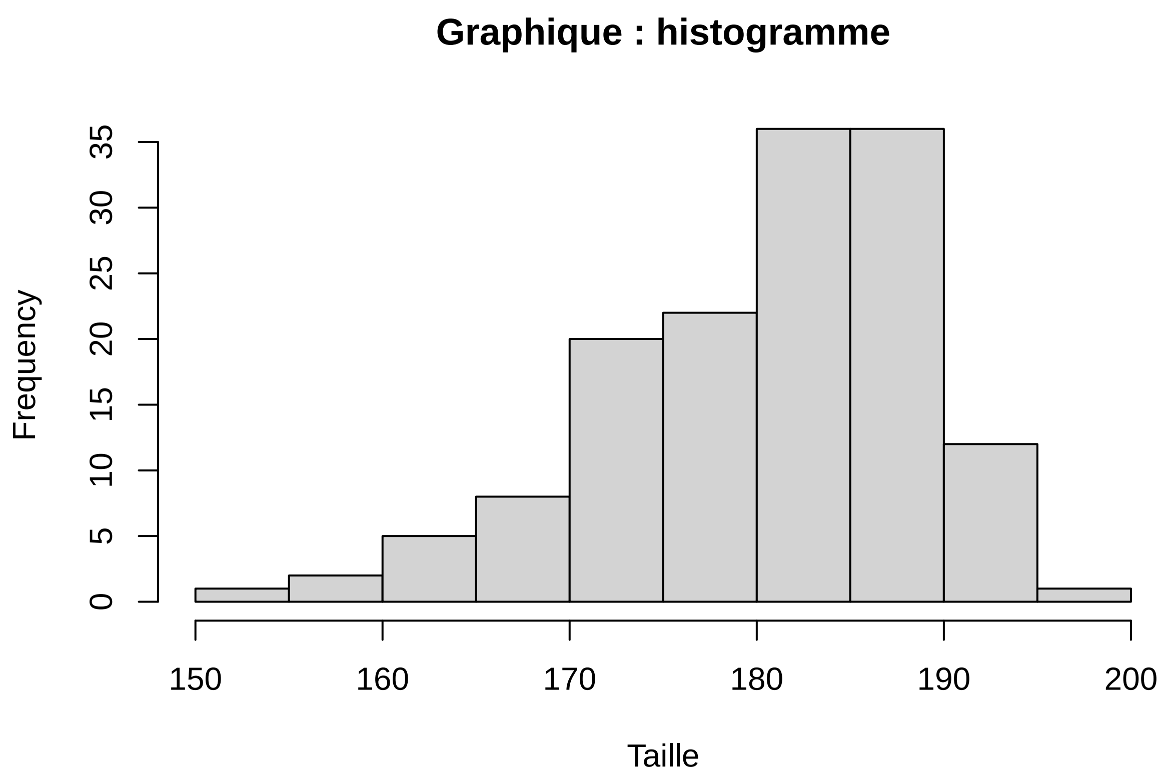
Il découle de ce graphique qu'il y a des sauts de petite taille en quantité. Cela prouve bien que notre variable quantitative est continue. Le fait que la courbe continue soit croissante représente la proportion cumulée des valeurs inférieures à un certain seuil (ici le seuil vaut 1). Pour approfondir cette analyse, regardons un graphique en boîte à moustaches.

```r
# Création du graphique en boîte à moustaches
boxplot(taille,
        main = "Graphique en boîte à moustaches",
        xlab = "Taille")
```



La médiane est d'environ 182 cm. Elle est proche du centre de la boîte ce qui nous suggère que la distribution est symétrique. Le premier quartile Q1 vaut environ 175 cm et le troisième quartile Q3 vaut environ 187 cm. Le point individuel situé en dehors des moustaches signifie qu'il n'y a qu'une seule valeur inhabituelle. Regardons un histogramme afin de voir l'effectif des tailles.

```
# Création du graphique : l'histogramme
hist(taille,
     breaks = "FD",
     main = "Graphique : histogramme",
     xlab = "Taille")
```



Ici, le ‘breaks = “FD” ’ signifie que nous souhaitons utiliser la méthode de calcul automatique des intervalles, appelée “Freedman-Diaconis”. Cette méthode prend en compte la taille de l’échantillon et la variation des données pour déterminer le nombre optimal d’intervalles.

Nous pouvons constater que la taille de nos joueuses va de 150 cm à 200 cm. On retrouve une grande concentration de joueuses dont la taille se situe entre 180 cm et 190 cm (35%).

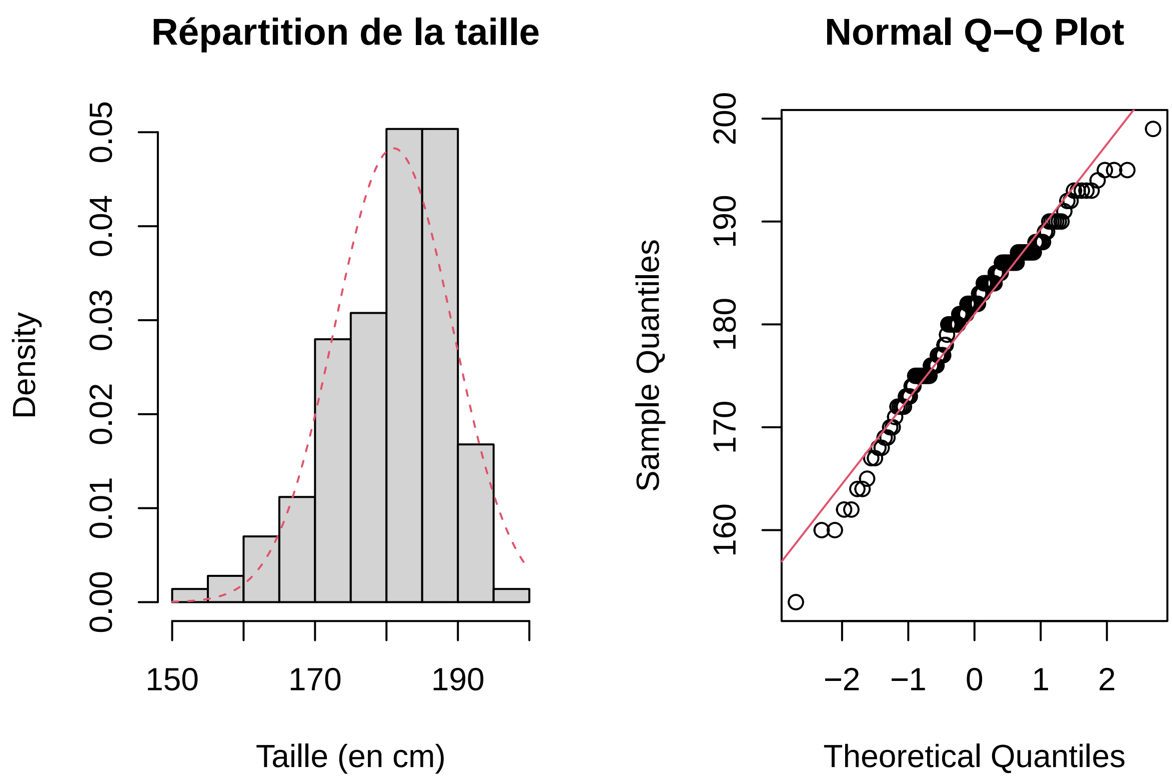
```
summary(taille)
```

Quartiles, minimum, maximum

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   153.0   175.5   182.0   181.0   187.0   199.0
```

Caractère gaussien Nous allons regarder si notre variable peut être approchée par un modèle Gaussien.

```
par(mfrow=c(1,2))
hist(taille , main="Répartition de la taille ",xlab="Taille (en cm)",prob=T)
curve(dnorm(x, mean(taille), sd(taille)), col=2, add=TRUE, lty=2)

qqnorm(taille)
abline(mean(taille), sd(taille), col=2)
```



On observe que l'histogramme de la répartition de la taille des joueuses est plutôt symétrique. La variable s'approche à la distribution normale.

Pour vérifier si la variable suit la loi normale nous allons procéder aux test de normalité.

- H0 : la variable suit la loi normale
- H1 : la variable ne suit pas la loi noramale

```
shapiro.test(taille)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  taille
## W = 0.9681, p-value = 0.002032
```

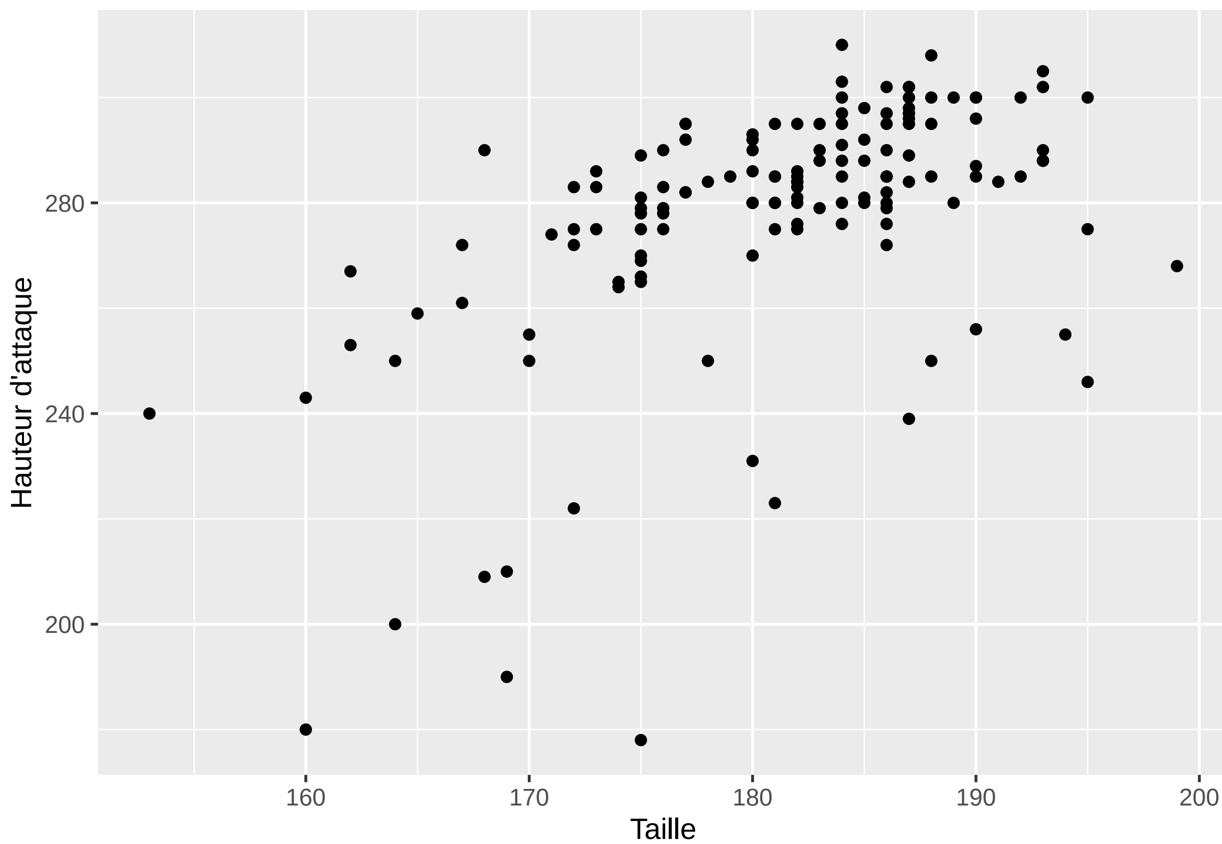
La p-valeur est faible, nous pouvons rejetter l'hypothèse nulle. Donc, contrairement à nos observations la variable ne suit pas la loi normale.

La hauteur de l'attaque selon la taille des joueuses

Est-ce que la hauteur de l'attaque dépend de la taille d'une joueuse?

Nous supposons que la hauteur d'attaque d'une joueuse peut dépendre de sa taille.

```
ggplot(data,aes(x=taille,y=block))+ geom_point()+
xlab("Taille")+
ylab("Hauteur d'attaque")
```



Pour vérifier cette supposition nous allons voir s'il y a une corrélation entre ces deux variables

Corrélation

```
cor(taille, attaque)
```

```
## [1] 0.6355188
```

Le coefficient de corrélation linéaire (de Pearson) de 0,64 est plus proche de 1 que de 0. Cela suggère qu'il y a une tendance pour les valeurs plus grandes de "taille" à être associées à des valeurs plus grandes de "attaque"

Régression linéaire

Vu que les deux variable sont *quantitatives continues* nous pouvons appliquer la régression linéaire.

Définissons nos variables. Nous allons utiliser la taille et la hauteur d'attaque en cm.

```
reg_multi <- lm(attaque~taille)
```

```
summary(reg_multi)
```

```
##
## Call:
## lm(formula = attaque ~ taille)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -88.954  -5.208   3.046  11.533  36.855
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -68.2765    36.5700  -1.867    0.064 .
## taille        1.9727     0.2018   9.774   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.86 on 141 degrees of freedom
## Multiple R-squared:  0.4039, Adjusted R-squared:  0.3997
## F-statistic: 95.53 on 1 and 141 DF,  p-value: < 2.2e-16
```

Nous retrouvons ici Beta 1: –68.28 et Beta 2: 1.97. Donc, on peut tracer notre droite y = Beta 1 + Beta 2 * x

La valeur de R^2 étant 0.4039 est éloignée de 1. Donc nous ne pouvons pas dire qu'il y a forcement une corrélation entre ces deux variables.

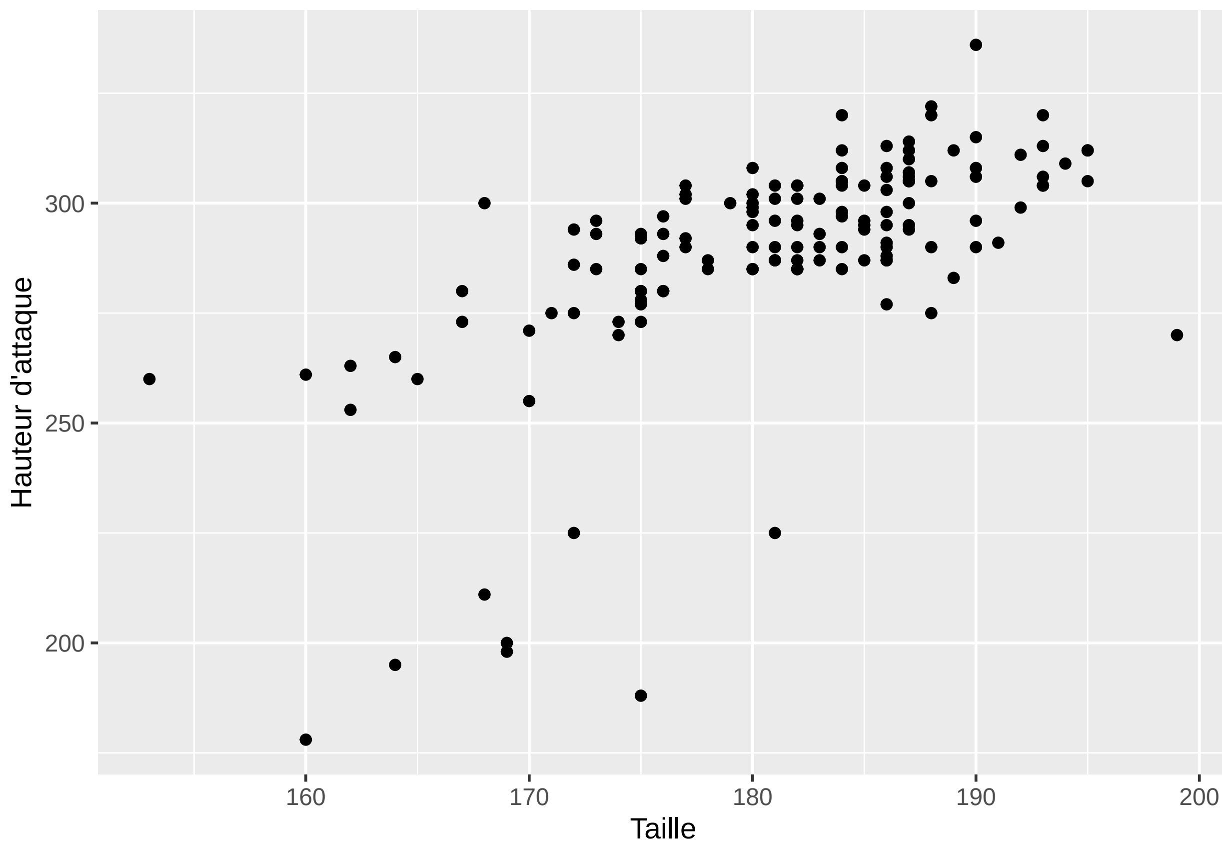
Mais 0,4039 n'est pas extrêmement faible non plus. Nous pouvons dire que ce résultat provient du fait que nous avons un échantillon de taille limitée ne nous permettant pas d'avoir une analyse assez solide.

```
anova(reg_multi)
```

```
## Analysis of Variance Table
##
## Response: attaque
##            Df Sum Sq Mean Sq F value    Pr(>F)
## taille      1  37695   37695  95.531 < 2.2e-16 ***
## Residuals 141  55637     395
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La valeur de SSE est de 55637, tandis que celle de SSM est de 37695. Bien qu'il y ait une différence entre ces deux valeurs, elle n'est pas significative. Cependant, cela nous conduit à la conclusion que nous ne pouvons pas établir de corrélation avec nos données.

```
ggplot(data,aes(x=taille,y=attaque))+ geom_point()+
xlab("Taille")+
ylab("Hauteur d'attaque")
```

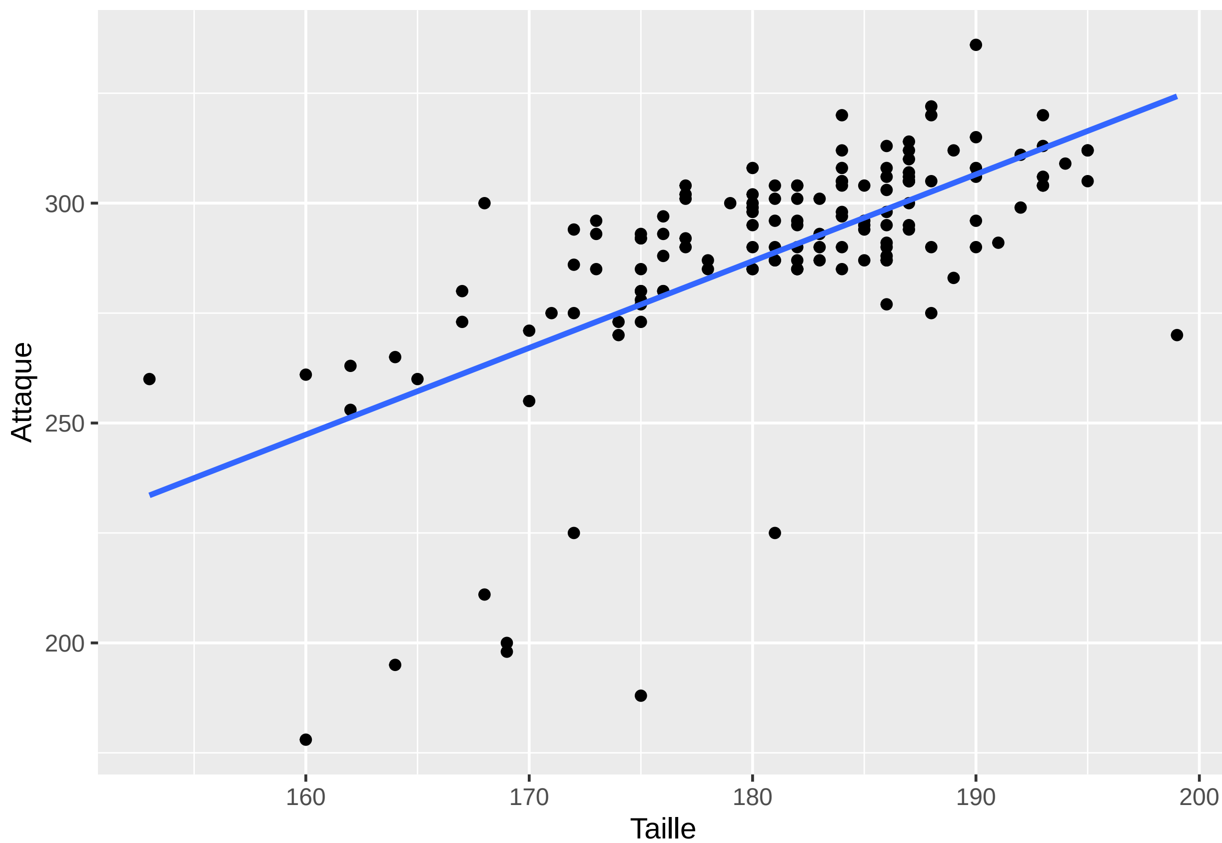


- Abcisse - la taille des joueuses
- Ordonnée - la hauteur d'attaque des joueuses

On observe qu'il y a une grande partie des valeurs où l'on pourrait tracer une droite. Cependant, il y a également suffisamment de valeurs qui s'éloignent nettement des autres vers le bas de l'abscisse.

```
ggplot(data,aes(x=taille,y=attaque))+ geom_point()+
stat_smooth(method="lm",se=FALSE)+ xlab("Taille")+
ylab("Attaque")
```

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



Comme constaté précédement, la droite tracé regroupe effectivement une partie des valeurs entre 250 et 350 (sur l'abcisse). Les valeurs plus éloignées de cette droite se trouvent vers le bas de l'abcisse.

Conclusion

Pour conclure les analyses statistiques effectuées sur cet échantillon nous reprenons les qustions que nous avons posé au début

Est-ca que les performances sportives (dans notre cas la hauteur d'attaque) des joueuses changent selon leur pays?

Nous avons vu que la hauteur d'attaque moyenne change selon le pays de naissance d'une joueuse.Nous avons trouvé à l'aide des tests que les moyennes de la hauteur d'attaque sont égales entre Cuba et Chine, Cuba et l'Égypte. Les variance sont égales entre Cuba et Égyote.

Est-ce que le poste d'une joueuse dépend de son pays de naissance?

Suite à notre analyse nous n'avons pas pu établir de lien entre le poste et le peys de naissance d'une joueuse.

Est-ce que la hauteur de l'attaque dépend de la taille d'une joueuse?

Nous pouvons supposer qu'il existe une corrélation entre la taille et la hauteur d'attaque d'une joueuse. Les joueuse plus grandes ont tendances d'avoir la hauteur d'attaque plus élevée. Mais nous ne pouvons pas faire de conclusion avec les valeurs obtenues. C'est possible qu'il y a une corrélation mais notre échantillon est limité pour en déduire une.