Projet Stats

2023-05-05

Introduction

Les données sont prises de https://www.kaggle.com/ . Les données concernent les joueuses de volleyball et contiennent des informations sur leur date de naissance, leur taille, leur poids, leur hauteur de saut (spike), leur hauteur de bloc, leur poste de jeu et leur pays d'origine.

La date de naissance, la taille et le poids sont des caractéristiques physiques des joueuses qui peuvent avoir une incidence sur leurs performances. La hauteur de saut (spike) et de bloc sont également des facteurs importants dans le volleyball, car ils mesurent la capacité d'une joueuse à attaquer et à défendre. Le poste de jeu de la joueuse est également une information importante, car chaque poste a des responsabilités et des rôles différents sur le terrain. Enfin, le pays d'origine peut également avoir une incidence sur le style de jeu et le niveau de compétition des joueuses.

En utilisant ces données, il est possible de mener des analyses statistiques pour étudier la relation entre ces caractéristiques et les performances des joueuses sur le terrain, ainsi que pour étudier les différences entre les joueuses de différents pays et postes de jeu. Ces informations peuvent être utiles pour les entraîneurs et les recruteurs pour sélectionner les joueuses les plus adaptées à leur équipe, ainsi que pour les chercheurs qui s'intéressent aux facteurs qui influencent les performances sportives.

Variables pour l'analyse

Nous allons effectuer une analyse statistique des données fournies, en commençant par comparer les fréquences des nationalités des joueuses. Ensuite, nous étudierons les capacités et aptitudes des joueuses, à savoir la hauteur maximale d'attaque et la hauteur maximale de blocage, en fonction de leur poste. Ces deux variables sont qualitatives.

Pour approfondir cette étude, nous analyserons également les aptitudes des joueuses en fonction de leurs postes. Cette analyse portera sur des variables quantitatives, ce qui nous permettra d'explorer plus en détail les différences et les tendances entre les postes occupés par les joueuses.

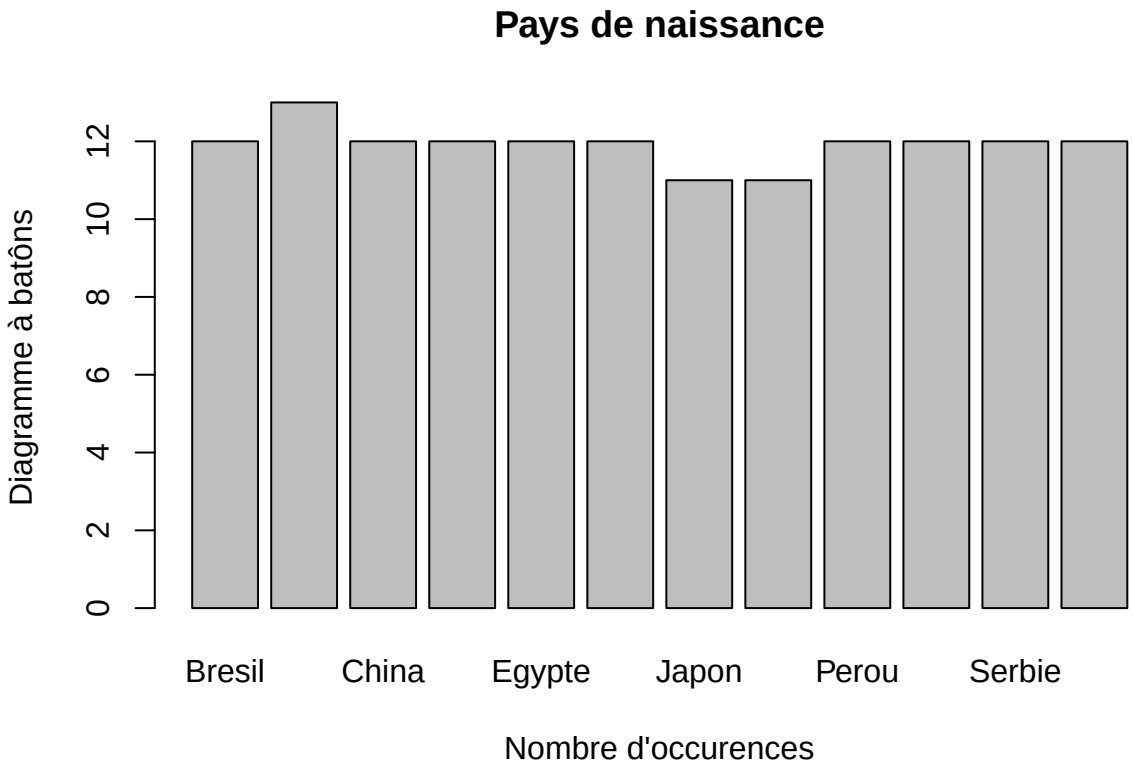
Chargement de données

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

Analyse univariée sur le pays de naissance des joueuses

Nous allons d'abord analyser la fréquence des pays de naissance des joueuses

```
pays <- data['pays']
barplot(table(pays), main = "Pays de naissance", xlab = "Nombre d'occurences", ylab = "Diagramme à batô
```

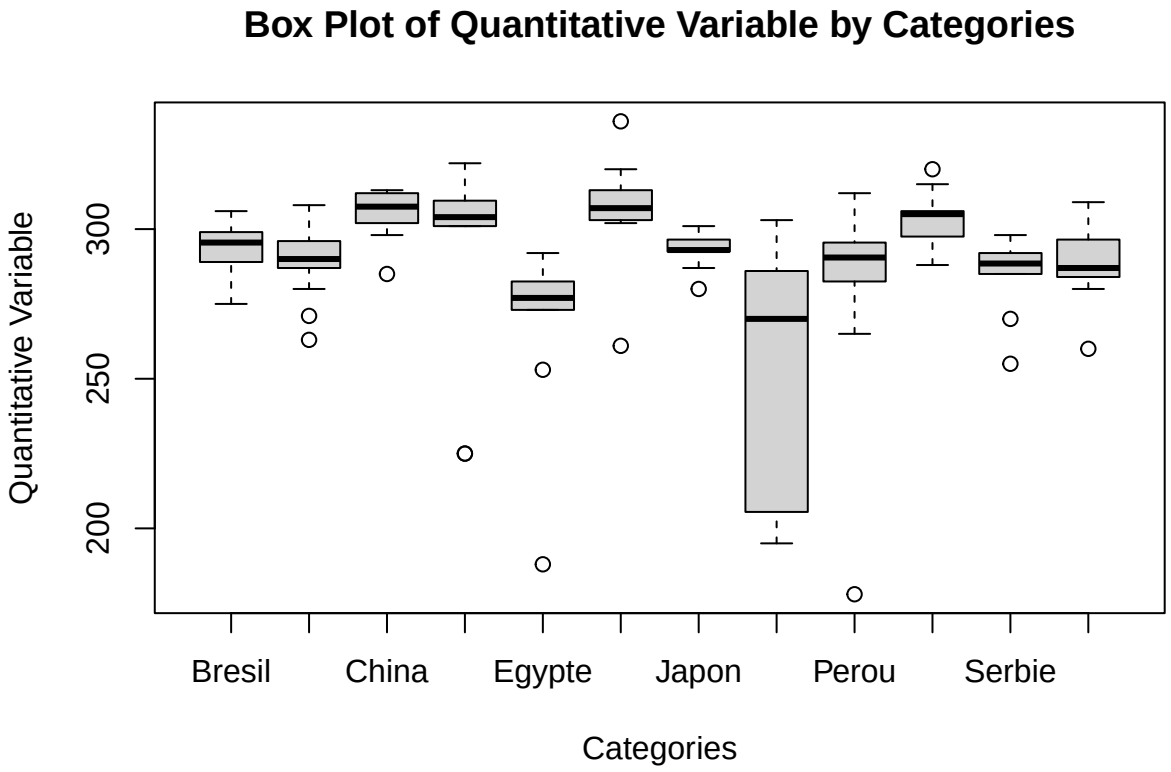


Sur ce diagramme en batôns nous obsevons que la fréquence de la plupart des pays de naissance des joueuses est égale à 12. Il n'y a que la Japon et Mexique qui sont les pays de naissance de 11 joueuse et la Bulgarie avec le maximum de ce diagramme : 13.

Pour approfondir cette analyse nous allons étudier les aptitudes des joueuses en fonction de leur pays de naissance:

```
pays <-data$pays
attaque <- data$attaque

boxplot(attaque ~ pays, main = "Box Plot of Quantitative Variable by Categories", xlab = "Categories", y
```



Sur cette boîte à moustache nous constatons que le pays avec la plus grande variance des aptitudes des joueuse est le

Nous avons remarqué grâce à l'analyse ci-dessous, qu'il y avait une donnée erronée dans notre jeu de données puisqu'il y avait une joueuse qui avait une hauteur de block à 0. Nous avons donc fait le choix de retirer cette joueuse du jeu de données.

Analyse univariée sur la hauteur de block

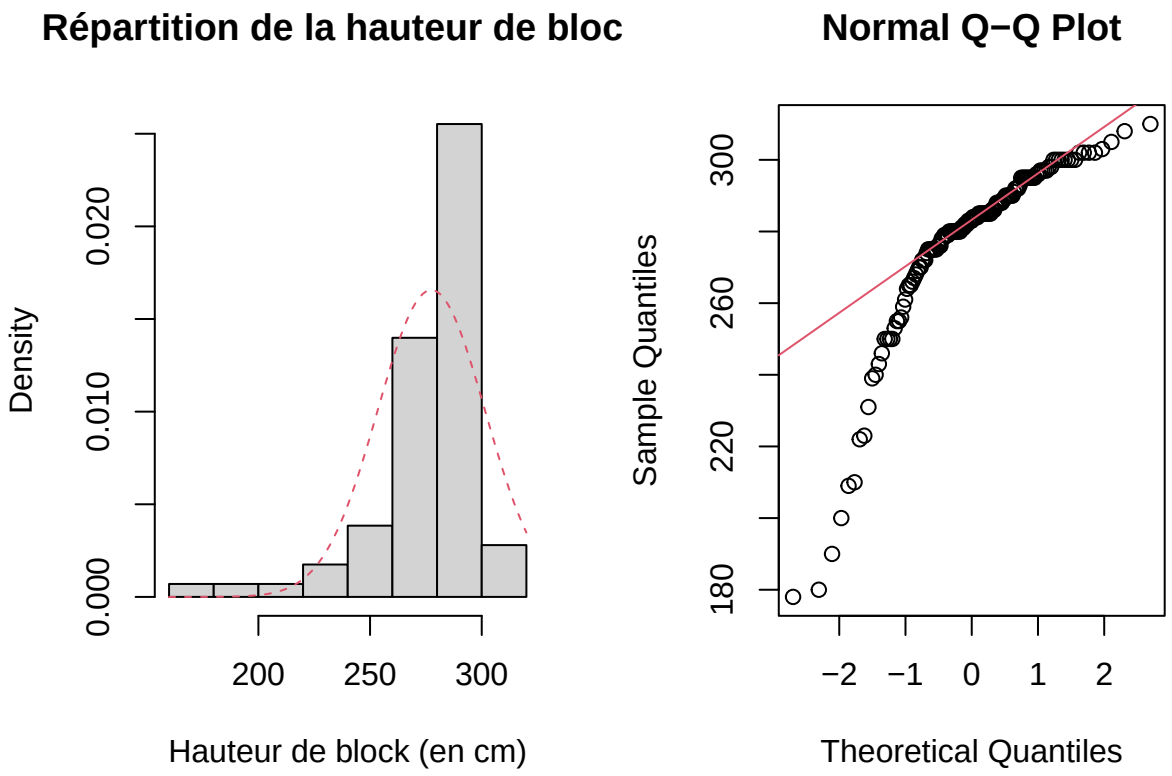
```
block <- data$block
moyenne <- mean(block)
print(paste("La hauteur moyenne de block est de", round(moyenne, 2)))
```

```
## [1] "La hauteur moyenne de block est de 277.43"
```

```
variance <- var(block)
print(paste("La variance de la hauteur de block est de", round(variance, 2)))
```

```
## [1] "La variance de la hauteur de block est de 580.01"
```

```
# Analyse gaussienne
par(mfrow=c(1,2))
hist(block , main="Répartition de la hauteur de block ",xlab="Hauteur de block (en cm)",prob=T)
curve(dnorm(x, mean(block), sd(block)), col=2, add=TRUE, lty=2)
qqnorm(block)
qqline(block, col=2)
```




```
# Calcul de l'intervalle de confiance à 95% pour la moyenne
intervalle_de_confiance <- t.test(block)$conf.int
print(intervalle_de_confiance)
```

```
## [1] 273.4524 281.4148
## attr(,"conf.level")
## [1] 0.95
```

```
# Calcul de la proportion de blocs supérieurs à 280
proportion <- length(block[block > 280]) / length(block)
print(paste("La proportion de block dont la valeur est supérieure à 280 est de", proportion*100, "%"))
```

```
## [1] "La proportion de block dont la valeur est supérieure à 280 est de 56.6433566433566 %"
```

```
# Calcul de l'intervalle de confiance à 95% pour la proportion
int_confiance_prop <- prop.test(sum(block > 280), length(block))$conf.int
print(int_confiance_prop)
```

```
## [1] 0.4810458 0.6481979
## attr(,"conf.level")
## [1] 0.95
```

Dans l'analyse