



# RAPPORT JEUX OLYMPIQUES

**Fatma BEN LTAIEF**  
**Hajar Chakir**  
**Gabriel EL DAHDAH**  
**Nadia ZIDAN**



# SOMMAIRE

01	<b>Introduction</b>
02	<b>Présentation des données</b>
03	<b>La différence de performances des pays</b>
04	<b>Les facteurs qui contribuent à la performance des JO</b>
05	<b>Conclusion</b>

# Introduction

Les Jeux Olympiques représentent un événement sportif de grande envergure qui se déroule tous les quatre ans. À chaque édition, les Jeux sont organisés dans un pays différent, sélectionné par le Comité International Olympique (CIO).

Cette compétition offre aux pays hôtes une opportunité unique de briller sur la scène internationale et de dynamiser leur économie. Des milliers d'athlètes venus du monde entier y participent, représentant leurs nations dans diverses disciplines et épreuves sportives.

Le rapport s'appuie sur un jeu de données provenant du site officiel des Jeux Olympiques, couvrant les informations sur les athlètes olympiques de 1896 à 2016. Ce vaste ensemble de données permet une analyse détaillée des performances des athlètes et des nations à travers les décennies.

L'objectif de ce rapport est de répondre à la question suivante : **“Quels sont les facteurs qui contribuent à la performance des pays aux Jeux Olympiques modernes ?”** Pour ce faire, nous analyserons plusieurs aspects clés. Tout d'abord, nous examinerons la répartition des médailles par pays pour identifier les nations dominantes. Ensuite, nous analyserons l'impact du nombre de participants sur le nombre de médailles remportées. Nous évaluerons également la parité hommes-femmes dans les équipes nationales et identifierons les sports dans lesquels chaque pays excelle.

En outre, nous explorerons les caractéristiques physiques des athlètes, telles que l'âge, la taille et le poids, pour comprendre leur influence sur les performances olympiques. Nous étudierons également l'impact de la localisation géographique des Jeux et l'avantage potentiel pour les pays hôtes.

## 2-Présentation des données

Le jeu de données est divisé en deux feuilles Excel.

La première, appelée "[athlete\\_event](#)", contient des informations sur les athlètes et compte 271 116 lignes.

Elle comprend 15 variables, dont 11 qualitatives et 4 quantitatives.

Les variables qualitatives sont:

- Nominales : -**Id** (identifiant des athlètes)
  - **Name** (nom de l'athlète)
  - **Sex** (sexe de l'athlète)
  - **Team** (équipe de l'athlète)
  - **Games** (année et compétition)
  - **Noc** (code du pays)
  - **Season** (saison de participation)
  - **City** (ville de participation)
  - **Sport** (sport pratiqué)
  - **Event** (nom de l'épreuve)

Ordinale : -**Medal** (médaille obtenue)

Les variables quantitatives sont :

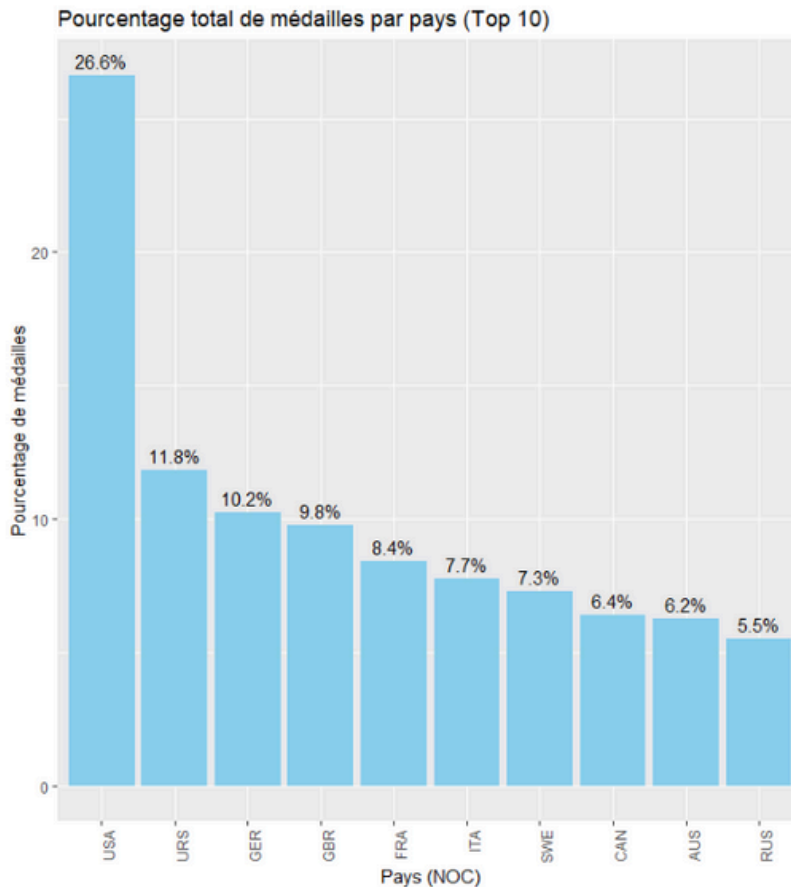
- Continues : -**Age** (âge de l'athlète)
  - **Height** (taille de l'athlète)
  - **Weight** (poids de l'athlète)
- Discrète : -**Year** (année de participation)

La seconde feuille, nommée "[noc\\_regions](#)", contient des informations sur le lieu d'origine des athlètes. Elle comprend 3 variables et 231 lignes. Les variables Noc (code du pays), Region (région d'origine) et Notes (précisions sur la région d'origine), sont toutes qualitatives nominales.

Dans l'ensemble, il y a 135 571 athlètes. L'âge des athlètes varie de 10 à 97 ans. Il y a 1185 équipes différentes venant de 230 pays pour 66 disciplines différentes. Plus de 28 500 médailles ont été attribuées (28530) à plus de 20797 des athlètes.

# 3-La différence de performances des pays

## 3.1 : les pays les plus performants aux Jeux Olympiques



L'analyse univariée permet de comparer les performances des pays en fonction du nombre total de médailles remportées aux Jeux Olympiques. Le graphique en barres montre le pourcentage de médailles obtenues par les dix pays les plus performants, ce qui permet de visualiser les différences de succès olympique entre eux. Les variables analysées sont les pays (NOC) et le nombre de médailles (Medal).

Les États-Unis se distinguent nettement avec 26.6% des médailles, démontrant une domination claire et constante aux Jeux Olympiques. Cette proportion élevée met en évidence l'efficacité des programmes sportifs et l'ampleur des investissements dans le sport de haut niveau aux États-Unis.

L'Union Soviétique, bien qu'elle n'existe plus en tant qu'entité, figure en deuxième position avec 11.8% des médailles, soulignant son passé glorieux dans les compétitions olympiques. Ce résultat historique montre l'impact durable des performances soviétiques avant la dissolution du pays.

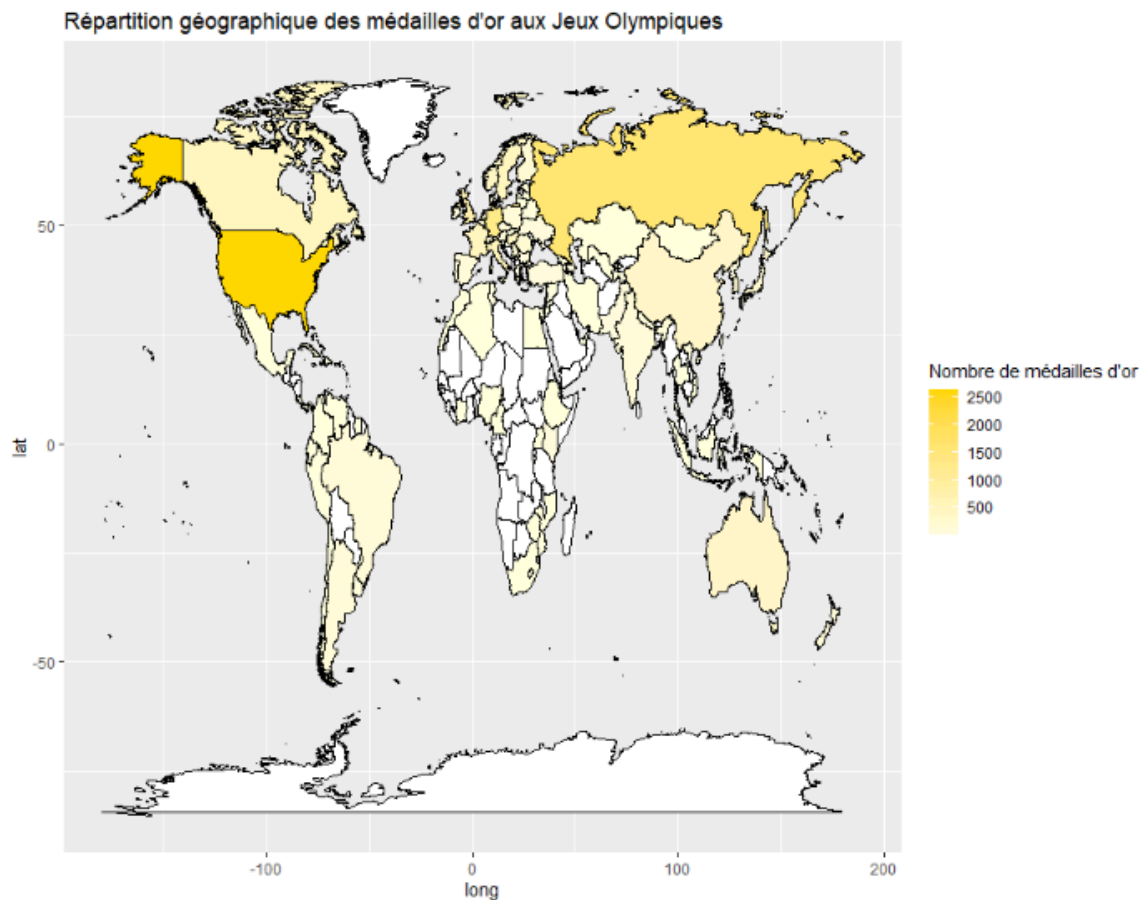
L'Allemagne suit avec 10.2%, reflétant une tradition sportive solide et des performances constantes aux Jeux Olympiques. Le Royaume-Uni avec 9.8% et la France avec 8.4% montrent également leur compétitivité et leur capacité à remporter un nombre significatif de médailles.

L'Italie et la Suède, avec respectivement 7.7% et 7.3% des médailles, montrent leur présence constante et leurs bons résultats dans les compétitions. Le Canada et l'Australie, avec 6.4% et 6.2% des médailles, illustrent la force des pays du Commonwealth dans le sport olympique.

Enfin, la Russie, avec 5.5%, complète le top 10, marquant sa continuité dans les performances olympiques malgré les changements géopolitiques. L'analyse de ce graphique révèle les tendances et les différences dans les performances olympiques des pays, offrant une vue d'ensemble des nations les plus performantes aux Jeux Olympiques.

# 3-La différence de performances des pays

## 3.2 : La répartition géographique des médailles aux Jeux Olympiques



La carte montre que les États-Unis et certains pays d'Europe du Nord se démarquent par un nombre élevé de médailles d'or. Les États-Unis apparaissent en tête avec une concentration significative de médailles, indiquée par une teinte plus sombre.

En Europe, la Russie, l'Allemagne et le Royaume-Uni sont également des pays où le nombre de médailles d'or est élevé. Cette distribution souligne l'importance historique et continue de ces nations dans les compétitions olympiques.

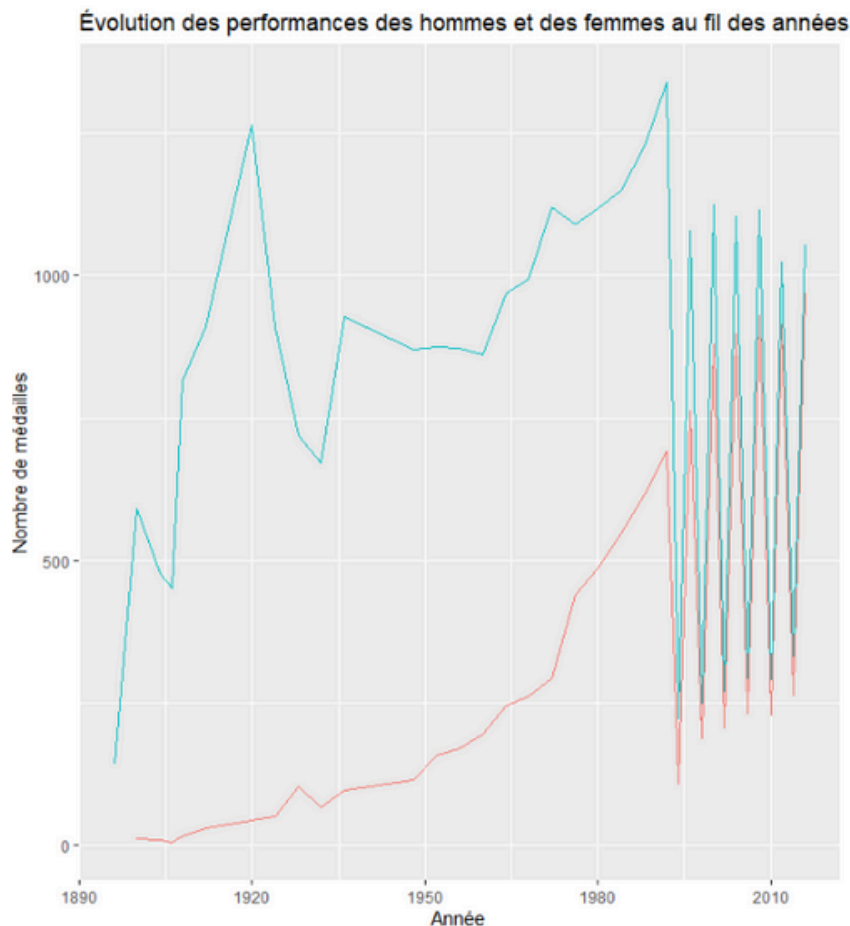
La teinte plus claire sur d'autres parties du globe indique une répartition moindre des médailles d'or, bien que certains pays comme la Chine montrent également des performances significatives.

L'absence de teintes foncées dans plusieurs régions d'Afrique, d'Amérique du Sud et d'Asie du Sud-Est reflète une présence plus limitée de médailles d'or, mettant en lumière les disparités géographiques en termes de succès olympique.

Le test du chi-2 que on a fait, a évalué la relation entre les pays et l'obtention de médailles d'or. Les résultats montrent une valeur de chi-2 de 168940 avec 240 degrés de liberté et une P-valeur inférieure à  $2.2e-16$ . La relation entre les pays et l'obtention de médailles d'or est donc statistiquement significative, indiquant que les différences observées ne sont pas dues au hasard. Il existe une association significative entre le pays d'origine et la probabilité de remporter une médaille d'or, soulignant les disparités géographiques en matière de succès olympique et l'importance des programmes sportifs nationaux.

# 3-La différence de performances des pays

## 3.3 : l'évolution la performance des hommes et des femmes au fil des années



L'analyse bivariée de l'évolution des performances des hommes et des femmes aux Jeux Olympiques permet de comparer le nombre de médailles remportées par chaque sexe au fil du temps. Le graphique présente deux courbes distinctes, une pour les hommes et une pour les femmes. Les variables analysées sont l'année (Year), le sexe (Sex) et le nombre de médailles (Medal).

Le graphique montre une augmentation continue du nombre de médailles remportées par les femmes depuis le début du 20e siècle, atteignant un pic notable à partir des années 1980. Cette tendance reflète non seulement une participation accrue des femmes aux compétitions olympiques, mais aussi une amélioration de leurs performances au fil du temps. Grâce à l'augmentation du nombre d'épreuves disponibles pour les femmes, celles-ci ont désormais presque autant d'opportunités de médailles que les hommes.

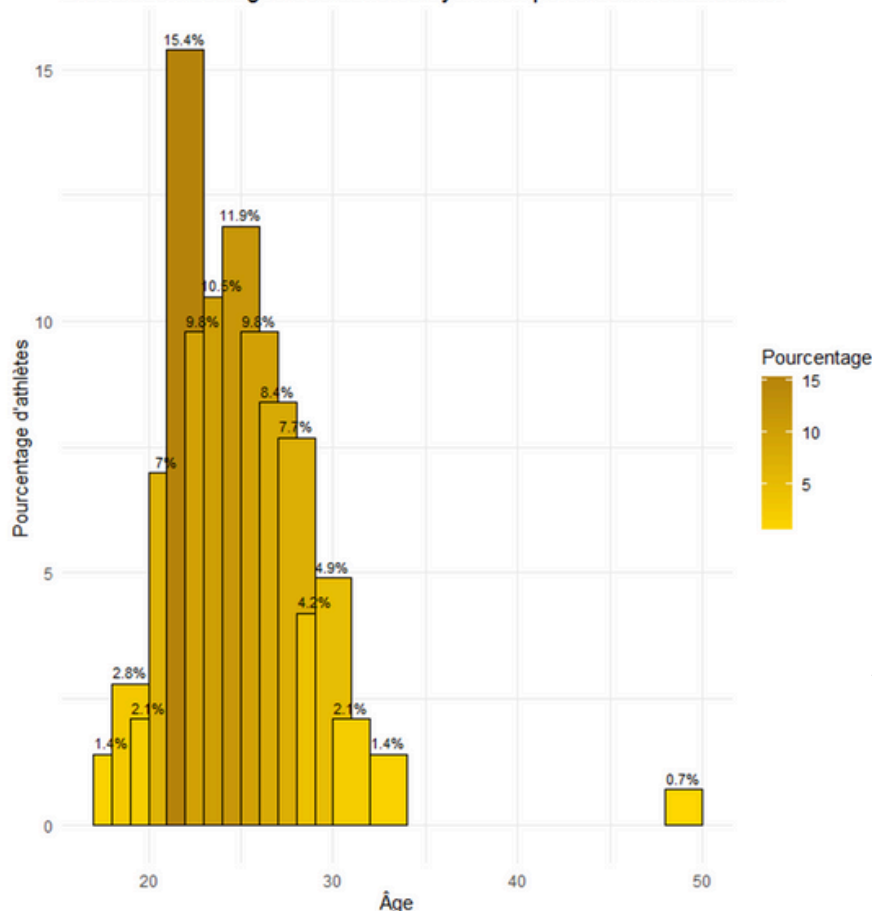
Pour les hommes, le nombre de médailles a fluctué au début du siècle mais montre une augmentation soutenue jusqu'à la fin du 20e siècle. Cependant, le nombre de médailles semble se stabiliser avec des variations moins prononcées dans les dernières décennies.

La comparaison des deux courbes indique que, bien que les hommes aient historiquement remporté plus de médailles, l'écart entre les performances des hommes et des femmes se réduit progressivement. Les années récentes montrent une convergence des performances, avec les femmes rattrapant leur retard par rapport aux hommes. Ce diagramme permet de visualiser l'évolution des performances olympiques en fonction du sexe et de constater que les femmes ont fait des progrès significatifs, réduisant ainsi l'écart avec les hommes au fil des années et participant désormais de manière substantielle aux performances de leur pays.

# 4-Les facteurs qui contribuent à la performance aux Jeux Olympiques

## 4.1 : l'impact de l'âge sur la performance aux Jeux Olympiques

Distribution des âges des athlètes ayant remporté des médailles d'or



Le diagramme en barres ici présente la distribution des âges des athlètes ayant remporté des médailles d'or, offrant un aperçu de l'impact de l'âge sur la performance aux Jeux Olympiques.

On observe que la tranche d'âge la plus représentée est celle des 20-25 ans, avec un pic à 15.4% pour les athlètes de 21 ans. Cette tranche d'âge semble être la période optimale pour atteindre des performances de pointe et remporter des médailles d'or,

probablement en raison d'un équilibre entre la maturité physique, l'expérience et la vigueur athlétique.

Les athlètes de 22 ans et 23 ans suivent de près avec des pourcentages respectifs de 11.9% et 10.6%, confirmant que le début de la vingtaine est une période cruciale pour exceller aux Jeux Olympiques. Les âges de 24 ans et 25 ans maintiennent également une forte présence avec des pourcentages de 9.8% et 9.6%, respectivement.

À partir de 26 ans, la proportion commence à diminuer progressivement, bien que des pourcentages notables soient encore observés. Les athlètes de 27 ans et 28 ans représentent 8.4% et 7.7%, indiquant que les performances de haut niveau peuvent être maintenues jusqu'à la fin de la vingtaine.

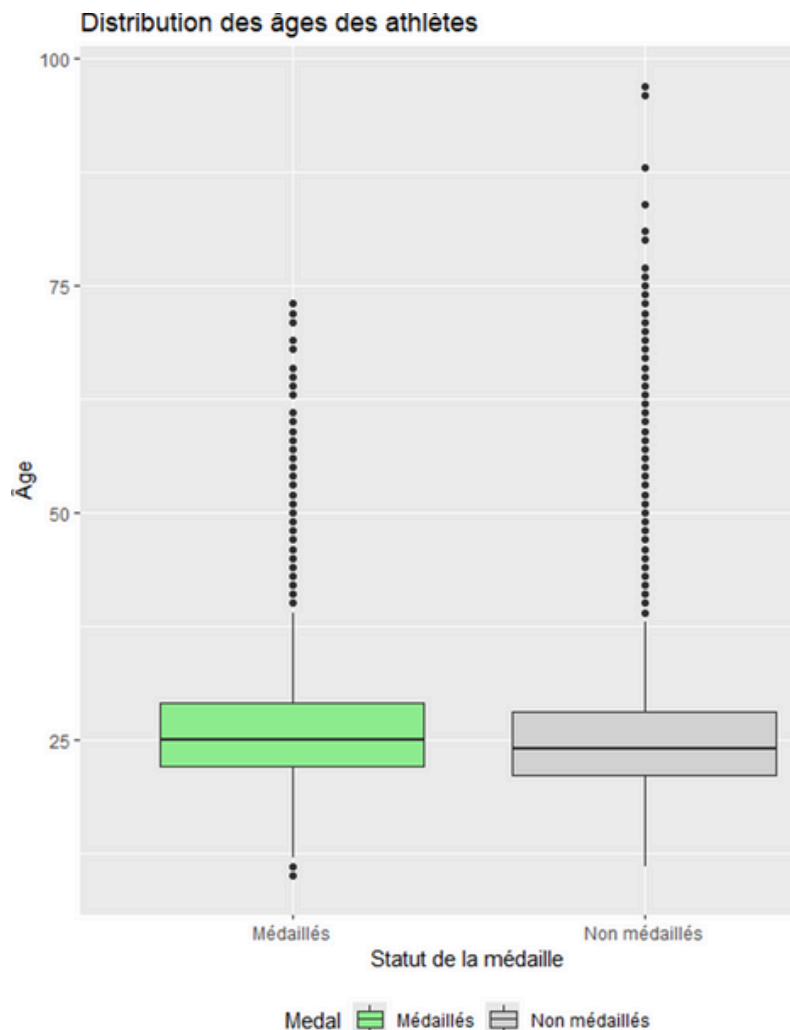
Au-delà de 30 ans, la proportion d'athlètes médaillés diminue de manière plus marquée, avec seulement 4.2% pour les 30 ans, 2.1% pour les 31 ans, et des pourcentages encore plus faibles pour les âges plus avancés. Une exception notable est observée à 50 ans avec 0.7%, ce qui suggère des cas exceptionnels de longévité sportive.

En résumé, ce graphique souligne que la période la plus propice pour remporter des médailles d'or se situe principalement entre 20 et 27 ans, avec un pic de performance autour de 21 à 25 ans. Les athlètes dans cette tranche d'âge bénéficient d'un équilibre optimal entre jeunesse et expérience, leur permettant d'atteindre les sommets de leur discipline.



# 4-Les facteurs qui contribuent à la performance aux Jeux Olympiques

## 4.1: l'impact de l'âge sur la performance aux Jeux Olympiques



La boîte à moustaches ici présente la distribution des âges des athlètes selon leur statut de médaille, différenciant les médaillés des non médaillés.

Pour les athlètes médaillés, représentés en vert, on observe que l'âge médian se situe autour de 25 ans, avec la majorité des âges compris entre 20 et 30 ans. Cela indique que les athlètes qui remportent des médailles sont souvent dans cette tranche d'âge,

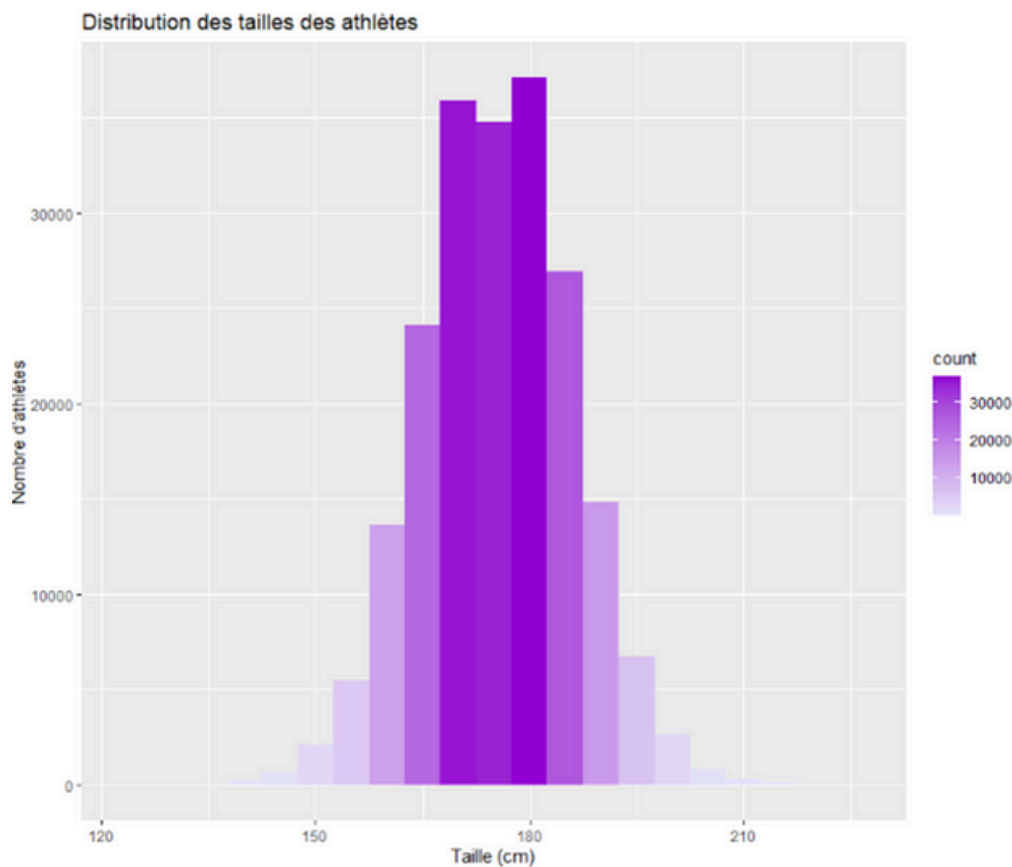
correspondant à une période de leur vie où leurs capacités physiques et leur expérience sont optimales. Les valeurs extrêmes montrent que certains athlètes peuvent encore remporter des médailles bien au-delà de cette tranche d'âge, avec quelques cas rares de médaillés âgés de plus de 75 ans.

En comparaison, pour les athlètes non médaillés, représentés en gris, l'âge médian est légèrement plus bas, autour de 23-24 ans. La distribution des âges est également plus étendue, allant de 15 ans jusqu'à des âges très avancés, voire 100 ans, ce qui suggère une plus grande variabilité dans l'âge des athlètes qui ne parviennent pas à remporter des médailles. Cela pourrait être dû à une variété de facteurs, tels que la maturité physique, l'expérience ou même les opportunités de compétition.

En résumé, ce graphique met en évidence que les athlètes médaillés tendent à être légèrement plus âgés que leurs homologues non médaillés, avec une concentration notable des médailles d'or entre 20 et 30 ans. Les athlètes non médaillés montrent une plus grande dispersion des âges, ce qui pourrait indiquer des différences en termes de préparation, d'expérience ou d'autres facteurs influençant la performance compétitive.

# 4-Les facteurs qui contribuent à la performance aux Jeux Olympiques

## 4.2 : l'impact de la taille et du poids sur la performance aux Jeux Olympiques



L'histogramme ci-dessus présente la distribution des tailles des athlètes, permettant de visualiser la répartition de la stature parmi les participants.

La plupart des athlètes se situent dans une plage de taille comprise entre 160 cm et 200 cm, avec un pic notable autour de 175 cm à 185 cm, où le nombre d'athlètes est le plus élevé, atteignant plus de 30 000 individus dans cette catégorie. Cela indique que la majorité des athlètes ont une taille moyenne à légèrement au-dessus de la moyenne, ce qui peut être optimal pour de nombreuses disciplines sportives.

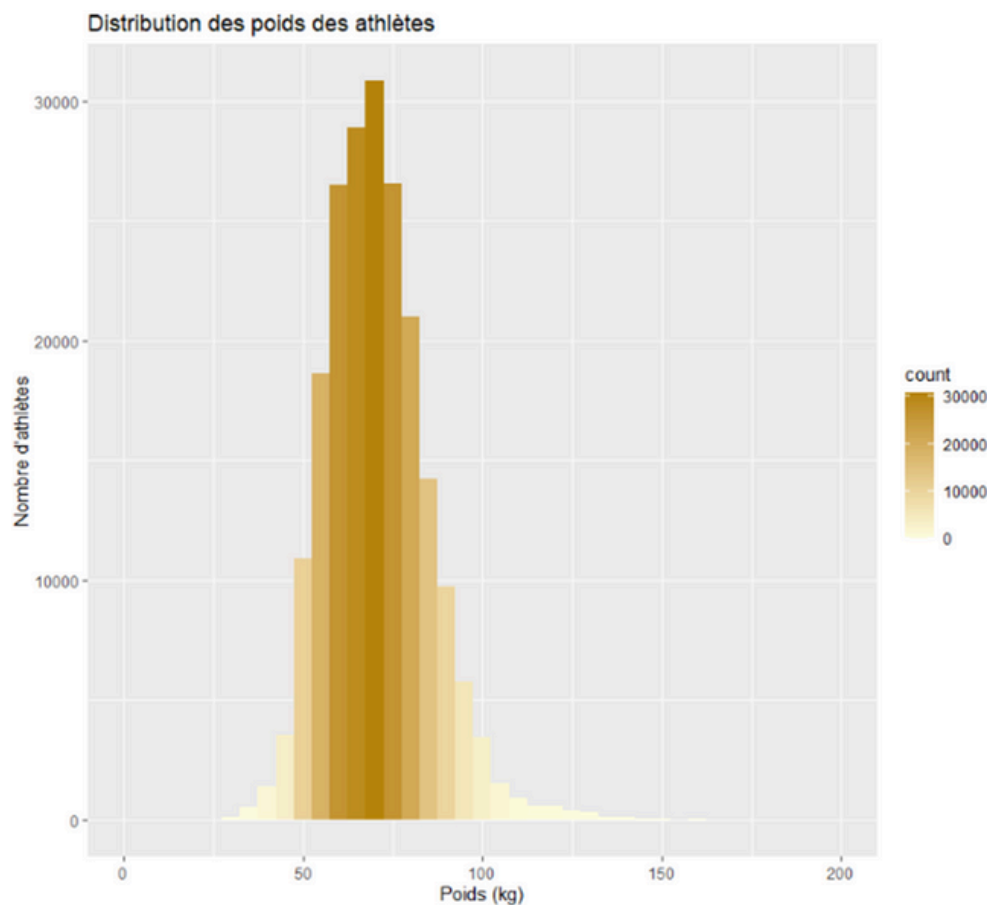
La distribution montre également une diminution progressive du nombre d'athlètes en dessous de 160 cm et au-dessus de 200 cm, suggérant que les tailles extrêmes sont moins communes parmi les athlètes. Cette tendance peut être liée aux exigences spécifiques de certaines disciplines sportives qui favorisent des tailles moyennes, équilibrant la force, l'agilité et l'endurance.

Les valeurs extrêmes, bien que rares, sont présentes et montrent que certains athlètes peuvent réussir et exceller même avec des tailles nettement inférieures ou supérieures à la moyenne, illustrant la diversité morphologique possible dans le sport de haut niveau.

En conclusion, cet histogramme met en évidence que la majorité des athlètes se concentrent autour d'une taille médiane, avec une distribution symétrique et une présence notable dans les catégories de taille moyennes, indiquant une tendance générale vers des proportions corporelles équilibrées favorisées dans le milieu sportif.

# 4-Les facteurs qui contribuent à la performance aux Jeux Olympiques

## 4.2 : l'impact de la taille et du poids sur la performance aux Jeux Olympiques



L'histogramme ci-dessus montre la distribution des poids des athlètes, permettant de visualiser comment les masses corporelles se répartissent parmi les participants.

La majorité des athlètes se situent dans une plage de poids comprise entre 50 kg et 100 kg, avec un pic notable autour de 70 kg, où le nombre d'athlètes est le plus élevé, atteignant plus de 30 000 individus dans cette catégorie. Cela suggère que la plupart des athlètes ont un poids moyen, ce qui peut être idéal pour un large éventail de disciplines sportives.

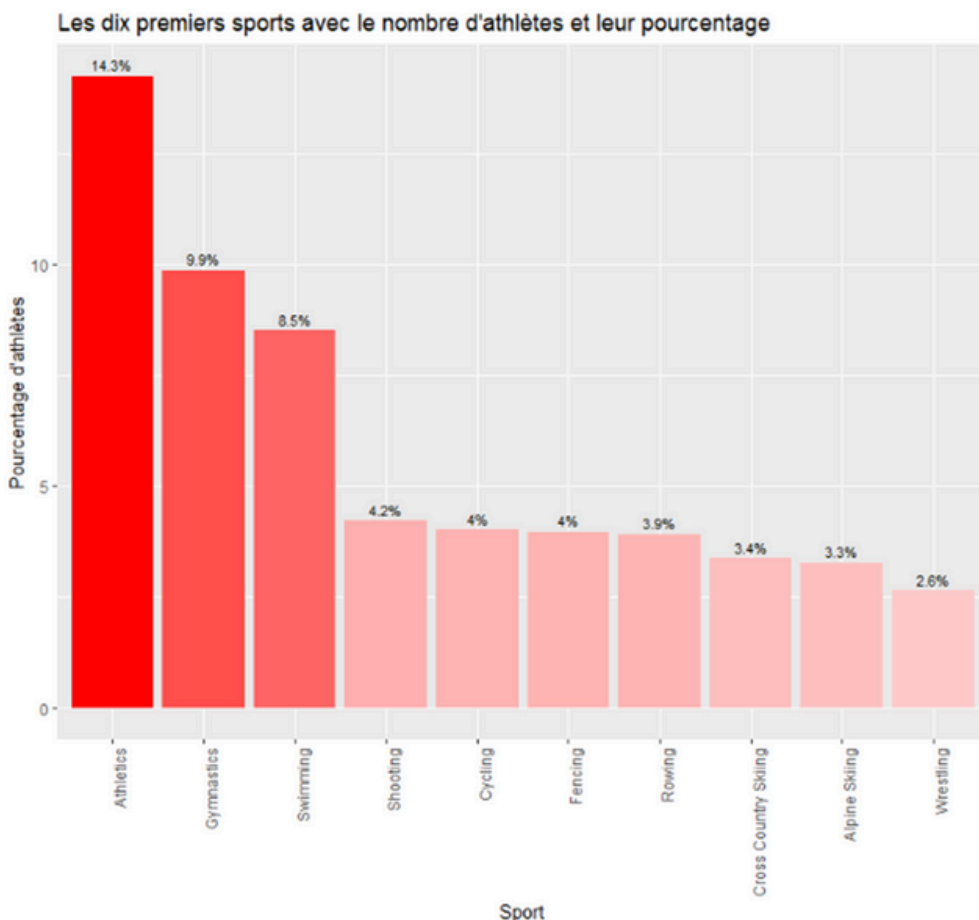
La distribution montre également une diminution progressive du nombre d'athlètes en dessous de 50 kg et au-dessus de 100 kg, indiquant que les poids extrêmes sont moins communs parmi les athlètes. Cette tendance peut être liée aux exigences spécifiques de certaines disciplines sportives qui favorisent des poids moyens, équilibrant la force, l'endurance et la performance globale.

Les valeurs extrêmes, bien que rares, existent et montrent que certains athlètes peuvent réussir et exceller même avec des poids nettement inférieurs ou supérieurs à la moyenne, illustrant la diversité possible des morphologies dans le sport de haut niveau.

En résumé, cet histogramme met en évidence que la majorité des athlètes se concentrent autour d'un poids médian, avec une distribution symétrique et une présence notable dans les catégories de poids moyens, indiquant une tendance générale vers des proportions corporelles équilibrées favorisées dans le milieu sportif.

# 4-Les facteurs qui contribuent à la performance aux Jeux Olympiques

## 4.3 : les sports les plus pratiqués



Le diagramme en barres présenté illustre la distribution des athlètes par sport, mettant en évidence les dix sports les plus pratiqués. Cette distribution permet de comprendre quels sports ont le plus grand nombre de participants, ce qui peut donner des indications sur les disciplines les plus propices à remporter des médailles aux Jeux Olympiques.

L'athlétisme est le sport le plus représenté, avec une large part des athlètes participants, ce qui s'explique par le grand nombre d'épreuves et de disciplines incluses dans l'athlétisme, offrant ainsi de nombreuses opportunités de médailles. La gymnastique suit de près, attirant un grand nombre de participants grâce à ses multiples épreuves. La natation, avec ses nombreuses catégories de courses telles que le freestyle, le dos, la brasse, le papillon et les relais, est également fortement représentée.

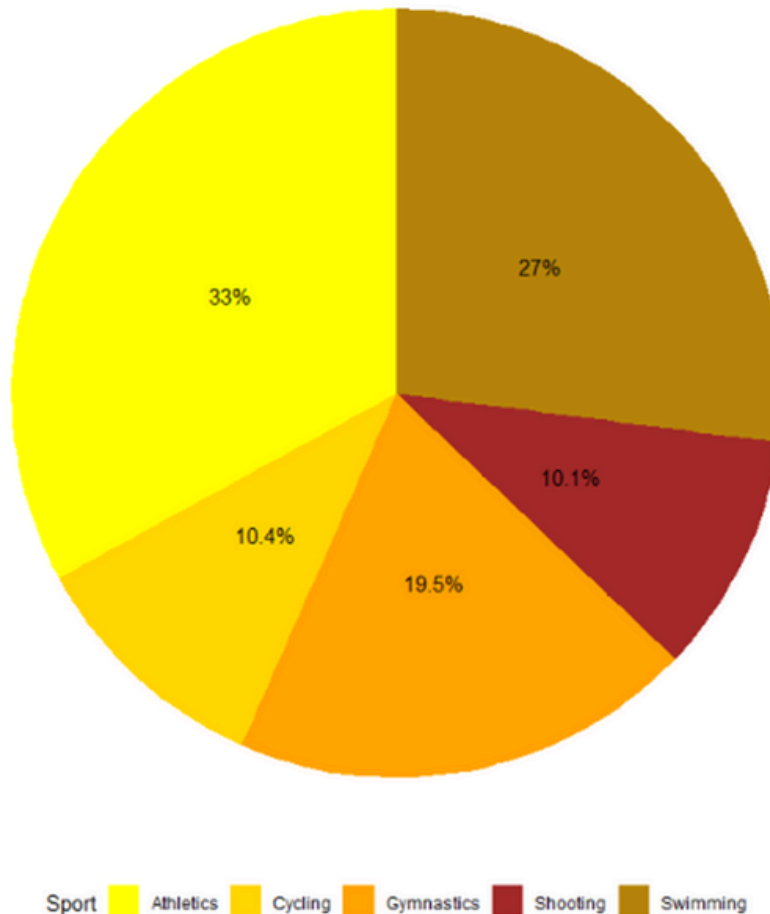
Le tir, bien que moins populaire que les trois premiers, reste un sport notable avec plusieurs épreuves permettant de multiples médailles. Le cyclisme, incluant différentes disciplines comme la route, la piste, le BMX et le VTT, attire aussi un bon nombre de participants, offrant de multiples opportunités de médailles.

L'escrime, avec des épreuves individuelles et par équipes, diversifie les chances de médailles et est bien représentée. L'aviron, avec ses courses variées, attire un nombre significatif d'athlètes. Le ski de fond et le ski alpin, bien que moins populaires, présentent aussi des opportunités pour les athlètes de se distinguer. Enfin, la lutte est également présente, offrant des opportunités dans ses diverses catégories de poids.

# 4-Les facteurs qui contribuent à la performance aux Jeux Olympiques

## 4.3 : les sports les plus pratiqués

Répartition des médailles d'or pour les cinq sports les plus pratiqués



Le diagramme en secteurs (camembert) présente la répartition des médailles d'or parmi les cinq sports les plus performants, offrant une vision claire des disciplines dominantes en termes de médailles d'or aux Jeux Olympiques.

L'athlétisme se distingue comme le sport le plus performant avec 33% des médailles d'or, reflétant non seulement la diversité des épreuves disponibles mais aussi la forte participation des athlètes dans cette discipline, ce qui accroît les chances de succès.

La natation suit de près, représentant 27% des médailles d'or, un chiffre significatif qui souligne l'importance de ce sport aux Jeux Olympiques, avec ses nombreuses courses et catégories offrant de multiples occasions de victoire.

La gymnastique occupe la troisième place avec 19.5% des médailles d'or, démontrant la compétitivité et l'excellence des athlètes dans cette discipline, connue pour ses diverses épreuves techniques et artistiques.

Le tir, bien que moins représenté en termes de participants, assure 10.4% des médailles d'or, ce qui montre l'efficacité des athlètes dans ce sport de précision et de concentration.

Le cyclisme complète ce classement avec 10.1% des médailles d'or, une indication des multiples épreuves cyclistes et de la compétence des athlètes dans cette discipline variée.

Ces données révèlent les sports les plus propices à remporter des médailles d'or aux Jeux Olympiques, mettant en lumière les disciplines où les athlètes excellent et se démarquent par leur performance.

# Conclusion Scientifique

L'analyse des données sur les performances des pays et des athlètes aux Jeux Olympiques de 1896 à 2016 révèle plusieurs tendances significatives. Les États-Unis dominent avec 26.6% des médailles, suivis par l'ex-Union Soviétique (11.8%) et l'Allemagne (10.2%), soulignant l'impact durable des programmes sportifs nationaux robustes. L'analyse géographique montre une forte concentration de médailles en Amérique du Nord et en Europe, avec des disparités significatives dans d'autres régions du monde.

Le test chi-2 confirme une relation statistiquement significative entre le pays d'origine et la probabilité de remporter une médaille d'or, avec une P-valeur inférieure à  $2.2e-16$ . L'âge médian des athlètes médaillés se situe autour de 25 ans, indiquant que la période optimale pour atteindre des performances de pointe est entre 20 et 30 ans. De plus, la majorité des athlètes ont une taille comprise entre 160 cm et 200 cm et un poids entre 50 kg et 100 kg, suggérant que des proportions corporelles équilibrées sont favorisées dans le sport de haut niveau.

## Conclusion Générale

Ce rapport met en évidence les principaux facteurs influençant les performances aux Jeux Olympiques modernes. Les programmes sportifs nationaux, l'âge des athlètes, et les proportions corporelles jouent un rôle crucial dans la réussite olympique. Les États-Unis, avec leur forte infrastructure sportive, montrent une domination continue, tandis que des pays comme l'ex-Union Soviétique et l'Allemagne maintiennent une tradition de succès.

La répartition géographique des médailles indique une prédominance des nations développées, laissant apparaître des opportunités pour une plus grande diversité et inclusion à l'échelle mondiale. En termes de disciplines, l'athlétisme, la natation et la gymnastique se distinguent comme les sports les plus performants, offrant le plus grand nombre d'opportunités de médailles.

Ces conclusions soulignent l'importance d'un soutien structuré et des ressources adéquates pour le développement des athlètes, ainsi que la nécessité d'encourager une participation plus globale pour équilibrer la répartition des succès olympiques.

# ANNEXE

## ##### Importation des données #####

### # Charger les bibliothèques nécessaires

```
library(readxl)
library(writexl)
library(openxlsx)
library(ggplot2)
library(dplyr)
library(sf)
library(ggmap)
```

### # Spécifier le chemin vers les fichiers Excel

```
chemin_fichier<-"C:/Users/fatma/OneDrive/Documents/BUT_SD_2023_2024/DATA/athlete_events.xlsx"
chemin_fichier1<- "C:/Users/fatma/OneDrive/Documents/BUT_SD_2023_2024/DATA/noc_regions.xlsx"
```

### # Importer les fichiers Excel

```
donnees <- read_excel(chemin_fichier)
donnees1 <- read_excel(chemin_fichier1)
```

### # Afficher les premières lignes des données importées

```
head(donnees)
head(donnees1)
```

## ##### Modifier l'âge et le poids #####

### # Modifier l'âge de la personne avec ID 135562

```
person_index <- which(donnees$ID == 135562)
new_age <- 24
donnees$Age[person_index] <- new_age
```

### # Modifier le poids de la personne avec ID 12177

```
index <- which(donnees$ID == 12177)
donnees$Weight[index] <- 218
```

### # Enregistrer les modifications

```
write.xlsx(donnees,
"C:/Users/fatma/OneDrive/Documents/BUT_SD_2023_2024/DATA/athlete_events.xlsx",rowNames= FALSE)
```

### # Vérification des types de données et des valeurs aberrantes

```
donnees$Age <- as.numeric(donnees$Age)
donnees$Height <- as.numeric(donnees$Height)
donnees$Weight <- as.numeric(donnees$Weight)
```

## ##### Jointure des tables #####

### # Effectuer la jointure entre les deux tables

```
resultat_jointure <- merge(donnees, donnees1, by = "NOC")
```

# ANNEXE

## ##### Analyse et visualisation #####

### # Partie 1 : Les différences de performances des pays

#### # Sous-partie 1.1 : les pays les plus performants aux Jeux Olympiques

```
data_1_1 <- resultat_jointure %>% filter(!is.na(Medal)) %>%  
group_by(NOC) %>%  
summarise(TotalMedals = n()) %>%  
arrange(desc(TotalMedals)) %>%  
top_n(10, wt = TotalMedals)
```

#### # Calcul du pourcentage de médailles

```
total_medals <- sum(data_1_1$TotalMedals)  
data_1_1 <- data_1_1 %>%  
mutate(Percentage = (TotalMedals / total_medals) * 100)
```

#### # Diagramme en barres avec pourcentages

```
ggplot(data_1_1, aes(x = reorder(NOC, -Percentage), y = Percentage)) +  
geom_bar(stat = "identity", fill = "skyblue") +  
geom_text(aes(label = sprintf("%.1f%%", Percentage)), vjust = -0.5) +  
labs(title = "Pourcentage total de médailles par pays (Top 10)", x = "Pays (NOC)", y = "Pourcentage de  
médailles") +  
theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

#### # Sous-partie 1.2 : la répartition géographique des médailles d'or aux Jeux Olympiques

```
data_1_2 <- resultat_jointure %>%  
filter(!is.na(Medal) & Medal == "Gold") %>%  
group_by(region) %>%  
summarise(TotalGoldMedals = n())
```

#### # Carte mondiale avec répartition des médailles d'or

```
world <- map_data("world")  
ggplot() +  
geom_map(data = world, map = world, aes(long, lat, map_id = region), fill = "white", color = "black") +  
geom_map(data = data_1_2, map = world, aes(map_id = region, fill = TotalGoldMedals), color = "black") +  
scale_fill_continuous(name = "Nombre de médailles d'or", low = "lightyellow", high = "gold") +  
labs(title = "Répartition géographique des médailles d'or aux Jeux Olympiques")
```

#### # Sous-partie 1.3 : l'évolution de la performance des hommes et des femmes au fil des années

```
data_1_3 <- resultat_jointure %>% filter(!is.na(Medal)) %>%  
group_by(Year, Sex) %>%  
summarise(TotalMedals = n())
```

#### # Courbe de l'évolution des performances

```
ggplot(data_1_3, aes(x = Year, y = TotalMedals, color = Sex)) +  
geom_line() +  
labs(title = "Évolution des performances des hommes et des femmes au fil des années", x = "Année", y =  
"Nombre de médailles")
```



# ANNEXE

## # Partie 2 : Les facteurs qui contribuent à la performance aux Jeux Olympiques

### # Sous-partie 2.1 : l'impact de l'âge sur la performance aux Jeux Olympiques

```
data_2_1 <- resultat_jointure %>% filter(!is.na(Age))
```

#### # Filtrer les données pour ne conserver que les médailles d'or

```
gold_medals <- resultat_jointure %>%  
  filter(Medal == "Gold") %>%  
  na.omit()
```

#### # Calculer les pourcentages

```
total_athletes <- nrow(gold_medals)  
gold_medals <- gold_medals %>%  
  group_by(Age) %>%  
  summarise(Count = n()) %>%  
  mutate(Percentage = (Count / total_athletes) * 100)
```

#### # Histogramme en barres de la distribution des âges des athlètes ayant remporté des médailles d'or

```
ggplot(gold_medals, aes(x = Age, y = Percentage, fill = ..y..)) +  
  geom_bar(stat = "identity", width = 2, color = "black") +  
  scale_fill_gradient(low = "gold", high = "darkgoldenrod", name = "Pourcentage") +  
  labs(  
    title = "Distribution des âges des athlètes ayant remporté des médailles d'or",  
    x = "Âge",  
    y = "Pourcentage d'athlètes",  
    fill = "Pourcentage"  
  ) +  
  geom_text(aes(label = paste0(round(Percentage, 1), "%")), vjust = -0.5, size = 3, color = "black") +  
  theme_minimal()
```

#### # Boîte à moustaches de la distribution des âges des médaillés vs non-médaillés avec couleur verte

#### # Convertir la variable Medal en un facteur avec les niveaux "Médaillés" et "Non médaillés"

```
data_2_1$Medal <- factor(ifelse(!is.na(data_2_1$Medal), "Médaillés", "Non médaillés"),  
  levels = c("Médaillés", "Non médaillés"))
```

#### # Afficher la boîte à moustaches avec les couleurs correctes

```
ggplot(data_2_1, aes(x = Medal, y = Age, fill = Medal)) +  
  geom_boxplot() +  
  scale_fill_manual(values = c("Médaillés" = "lightgreen", "Non médaillés" = "lightgrey")) +  
  labs(title = "Distribution des âges des athlètes", x = "Statut de la médaille", y = "Âge") +  
  theme(legend.position = "bottom")
```

# ANNEXE

## # Sous-partie 2.2 : l'impact de la taille et du poids sur la performance aux Jeux Olympiques

```
data_2_2 <- resultat_jointure %>% filter(!is.na(Height), !is.na(Weight))
```

### # Histogramme de la distribution des tailles avec un dégradé de violet

```
ggplot(data_2_2, aes(x = Height, fill = ..count..)) +  
  geom_histogram(binwidth = 5) +  
  scale_fill_gradient(low = "lavender", high = "darkviolet") +  
  labs(title = "Distribution des tailles des athlètes", x = "Taille (cm)", y = "Nombre d'athlètes")
```

### # Histogramme de la distribution des poids avec un dégradé de jaune et une échelle de 0 à 200 kg

```
ggplot(data_2_2, aes(x = Weight, fill = ..count..)) +  
  geom_histogram(binwidth = 5) +  
  scale_fill_gradient(low = "lightyellow", high = "darkgoldenrod") +  
  labs(title = "Distribution des poids des athlètes", x = "Poids (kg)", y = "Nombre d'athlètes") +  
  xlim(0, 200)
```

### # Boîte à moustaches de la taille des médaillés vs non-médaillés

```
ggplot(data_2_2, aes(x = as.factor(!is.na(Medal)), y = Height)) +  
  geom_boxplot() +  
  labs(title = "Distribution des tailles des athlètes (médaillés vs non-médaillés)", x = "Médaille", y = "Taille (cm)")
```

## # Sous-partie 2.3 : les sports plus propices à remporter des médailles

### # Utiliser la fonction `count()` pour compter le nombre d'athlètes par sport

```
data_2_3_count <- resultat_jointure %>%  
  filter(!is.na(Sport)) %>%  
  count(Sport)
```

### # Calculer le total des athlètes

```
total_athletes <- sum(data_2_3_count$n)
```

### # Sélectionner les dix premiers sports

```
top_10_sports <- data_2_3_count %>%  
  top_n(10, n) %>%  
  arrange(desc(n))
```

### # Calculer les pourcentages

```
top_10_sports <- top_10_sports %>%  
  mutate(Percentage = (n / total_athletes) * 100)
```

### # Créer un dégradé de rouge

```
red_gradient <- c("#FFCCCC", "#FF9999", "#FF6666", "#FF3333", "#FF0000")
```

# ANNEXE

## # Diagramme en barres des dix premiers sports avec le nombre d'athlètes et leur pourcentage

```
ggplot(top_10_sports, aes(x = reorder(Sport, -n), y = Percentage, fill = Percentage)) +  
  geom_bar(stat = "identity") +  
  geom_text(aes(label = paste0(round(Percentage, 1), "%")), vjust = -0.5, size = 3, color = "black") +  
  labs(title = "Les dix premiers sports avec le nombre d'athlètes et leur pourcentage", x = "Sport", y =  
"Pourcentage d'athlètes") +  
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +  
  scale_fill_gradientn(colors = red_gradient, guide = FALSE)
```

## # Diagramme en barres du nombre de médailles par sport

```
data_2_3_medals <- data_2_3 %>% filter(!is.na(Medal)) %>%  
  group_by(Sport) %>%  
  summarise(TotalMedals = n()) %>%  
  arrange(desc(TotalMedals))
```

## # Sélectionner les dix premiers sports

```
top_10_sports <- data_2_3_medals %>%  
  top_n(10, TotalMedals) %>%  
  arrange(desc(TotalMedals))
```

## # Calculer le total des médailles

```
total_medals <- sum(top_10_sports$TotalMedals)
```

## # Calculer les pourcentages

```
top_10_sports <- top_10_sports %>%  
  mutate(Percentage = (TotalMedals / total_medals) * 100)
```

## # Diagramme en barres des dix premiers sports avec le nombre de médailles et leur pourcentage

```
ggplot(top_10_sports, aes(x = reorder(Sport, -TotalMedals), y = Percentage)) +  
  geom_bar(stat = "identity", aes(fill = as.factor(rank(-TotalMedals)))) +  
  labs(title = "Les dix premiers sports avec le nombre de médailles et leur pourcentage", x = "Sport", y =  
"Pourcentage de médailles") +  
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +  
  scale_fill_manual(values = gray.colors(10, start = 0.2, end = 0.8), labels = 1:10) +  
  geom_text(aes(label = paste0(round(Percentage, 1), "%")), vjust = -0.5, size = 3, color = "black")
```

## #diagramme circulaire montrant la répartition des médailles par type (or, argent, bronze) pour un sport spécifique

```
library(ggplot2)
```

## # Sélectionner le sport "Athletics"

```
sport_athletics <- resultat_jointure %>%  
  filter(Sport == "Athletics")
```

## # Compter le nombre de médailles d'or, d'argent et de bronze

```
gold_count <- sum(sport_athletics$Medal == "Gold", na.rm = TRUE)  
silver_count <- sum(sport_athletics$Medal == "Silver", na.rm = TRUE)  
bronze_count <- sum(sport_athletics$Medal == "Bronze", na.rm = TRUE)
```

# ANNEXE

## # Créer un dataframe pour le diagramme circulaire

```
df <- data.frame(  
  Medal = c("Or", "Argent", "Bronze"),  
  Count = c(gold_count, silver_count, bronze_count))
```

## # Définir les couleurs des médailles

```
medal_colors <- c("gold", "grey", "brown")
```

## # Calculer le total des médailles

```
total_medals <- sum(df$Count)
```

## # Calculer les pourcentages

```
df <- df %>%  
  mutate(Percentage = (Count / total_medals) * 100)
```

## # Diagramme en secteurs avec pourcentages

## # Diagramme en secteurs avec pourcentages

```
ggplot(df, aes(x = "", y = Count, fill = Medal)) +  
  geom_bar(stat = "identity", width = 1) +  
  coord_polar("y", start = 0) +  
  geom_text(aes(label = paste0(round(Percentage, 1), "%")), position = position_stack(vjust = 0.5)) +  
  labs()
```

## # Titre du diagramme avec le sport spécifique

```
title = paste("Répartition des médailles pour le sport", "Athletics"),# Légende
```

```
fill = "Type de médaille") +
```

```
scale_fill_manual(values = c("Or" = "gold", "Argent" = "grey", "Bronze" = "brown"),
```

## # Utiliser les couleurs définies

```
breaks = c("Or", "Argent", "Bronze")) + # Définir l'ordre des étiquettes dans la légende  
theme_void() +  
theme(legend.position = "bottom")
```

## #diagramme en secteurs pour les cinq meilleurs sports, en ne considérant que les médailles d'or

## # Sélectionner les cinq meilleurs sports

```
top_5_sports <- resultat_jointure %>%  
  group_by(Sport) %>%  
  summarise(TotalMedals = n()) %>%  
  top_n(5, TotalMedals) %>%  
  arrange(desc(TotalMedals))
```

## # Filtrer les données pour ne conserver que les médailles d'or pour les cinq meilleurs sports

```
gold_medals_top_5 <- resultat_jointure %>%  
  filter(Sport %in% top_5_sports$Sport, Medal == "Gold")
```

## # Compter le nombre de médailles d'or pour chaque sport

```
gold_counts <- gold_medals_top_5 %>%  
  group_by(Sport) %>%  
  summarise(GoldMedals = n()) %>%  
  arrange(desc(GoldMedals))
```

# ANNEXE

## # Calculer le pourcentage des médailles d'or pour chaque sport

```
total_gold_medals <- sum(gold_counts$GoldMedals)
gold_counts <- gold_counts %>%
  mutate(Percentage = (GoldMedals / total_gold_medals) * 100)
```

## # Définir les couleurs pour chaque sport

```
color_palette <- c("Athletics" = "yellow", "Cycling" = "gold", "Gymnastics" = "orange", "Shooting" =
"brown", "Swimming" = "darkgoldenrod")
```

## # Diagramme en secteurs avec les couleurs définies et affichage des pourcentages

```
ggplot(gold_counts, aes(x = "", y = GoldMedals, fill = Sport)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar("y", start = 0) +
  labs(
    title = "Répartition des médailles d'or pour les cinq meilleurs sports",
    fill = "Sport"
  ) +
  scale_fill_manual(values = color_palette) + # Utiliser les couleurs définies
  geom_text(aes(label = paste0(round(Percentage, 1), "%")), position = position_stack(vjust = 0.5)) +
```

## # Afficher les pourcentages

```
theme_void() +
theme(legend.position = "bottom")
```

## #####TEST de KHI-DEUX

### # Création d'un dataframe pour les médailles d'or uniquement

```
df_or <- resultat_jointure[resultat_jointure$Medal == "Gold", ]
```

### # Création d'un tableau de contingence (tableau croisé) pour les médailles d'or par pays

```
table_contingence_or <- table(df_or$Pays, df_or$Medal)
```

### # Test du khi-deux pour les médailles d'or

```
test_khi_deux_or <- chisq.test(table_contingence_or)
```

### # Affichage des résultats du test pour les médailles d'or

```
print(test_khi_deux_or)
```

### # Interprétation des résultats du test pour les médailles d'or

```
if (test_khi_deux_or$p.value < 0.05) {
  cat("Il y a une association significative entre le pays et le nombre de médailles d'or.")
} else {
  cat("Il n'y a pas de preuve d'une association significative entre le pays et le nombre de médailles d'or.")
}
```