Noms/Prénoms :

Amokrane Sofiane
Amrani Amaury
Gislot Mattéo
Marquet Félix

Projet de Statistiques :
Les jeux olympiques de Paris 2024

Objectif du projet :

L'objectif principal de ce projet est de fournir une analyse détaillée des résultats des Jeux Olympiques afin d'identifier des tendances, des corrélations et des particularités dans la répartition des performances par pays et par discipline. L'analyse des données a entièrement été réalisée à l'aide des langages de programmation R et Python ainsi que des logiciels Excel, RStudio et Pycharm.

Présentation du contexte des JO :

« Les Jeux olympiques d'été sont une compétition multisports mondiale supervisés par le Comité international olympique et se déroulent tous les quatre ans. » (Wikipédia). Il s'agit de l'un des évènements sportifs les plus prestigieux et les plus suivis au monde. Le taux de réussite d'une nation est également un très bon indicateur de l'investissement et du développement du sport dans le pays. En 2024, ils se déroulent à Paris, capitale de la France, accueillant 206 délégations internationales et 12 625 athlètes.

Présentation des données :

Nous disposons de deux tableurs Excel, regroupant les données des athlètes et des médailles obtenues.

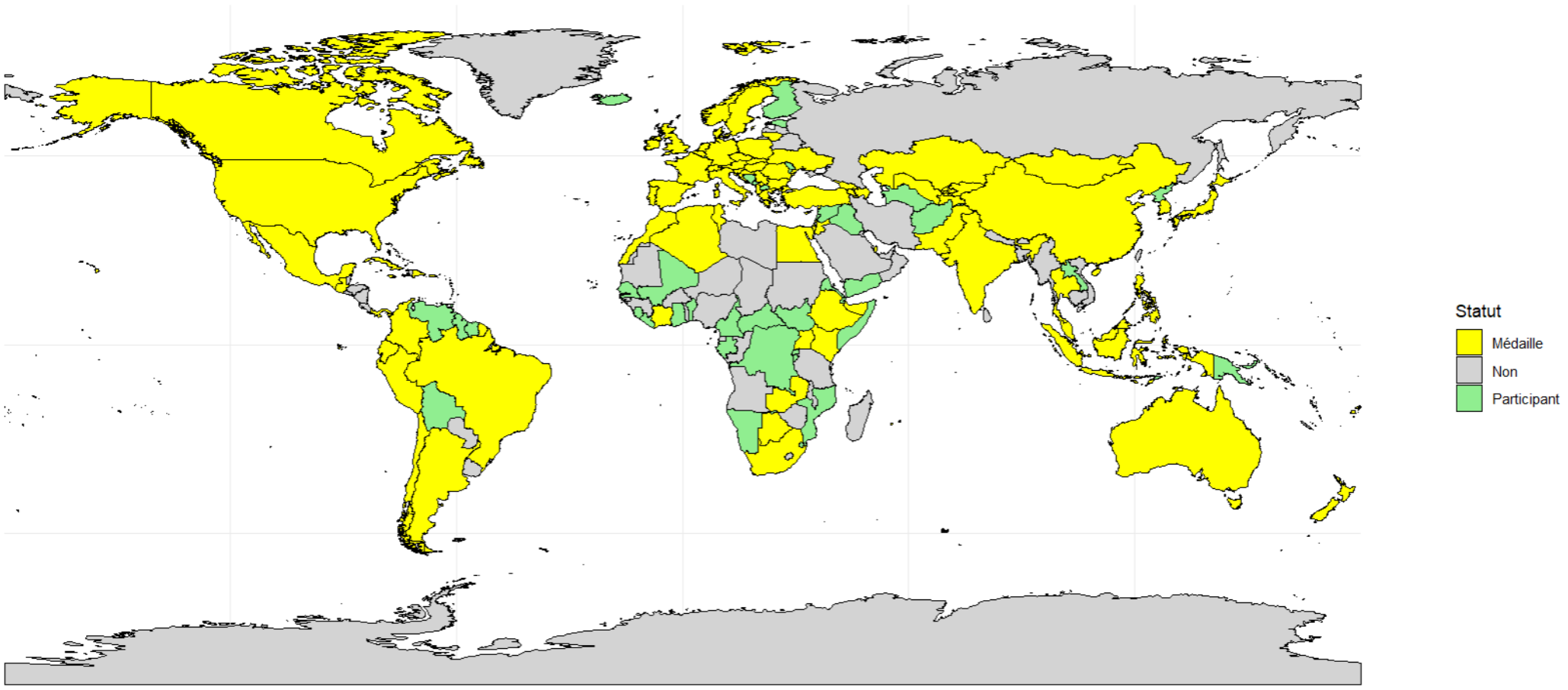
Excel Athlètes		Excel Médailles	
Données	Type de données	Données	Type de données
Nom	Qualitative nominale	Type de médaille	Qualitative ordinale
Prénom	Qualitative nominale	Code de la médaille	Quantitative discrète
Comité olympique	Qualitative nominale	Date d'obtention	Quantitative discrète
Nationalité	Qualitative nominale	Nom	Qualitative nominale
Genre	Qualitative nominale	Genre	Qualitative nominale
Date de naissance	Quantitative discrète	Discipline	Qualitative nominale
Discipline	Qualitative nominale	Epreuve	Qualitative nominale
Epreuve	Qualitative nominale	Pays	Qualitative nominale
Effectif	12625	Effectif	1044

Certaines données, comme l'url de l'épreuve ou les codes des disciplines, jugées inutiles n'ont pas été utilisées.

Problématique :

Au total, sur les 206 délégations participantes aux jeux, seulement 92 nations ont remporté au moins une médaille. La répartition de ces médailles étant inégale en fonction des pays, nous nous poserons la question suivante :

Quels sont les facteurs qui influencent les différences de taux de réussite entre les pays ?



Etudes de cas et discussions :

Le nombre de données par pays ou par discipline étant trop important pour pouvoir être affiché sur un seul graphe, seuls certains cas ont été étudiés.

Top 50 des pays ayant le plus d'athlètes -> à croiser avec le nombre d'hbts par pays comparaison des pourcentages. -> distribution suit une loi normale ?

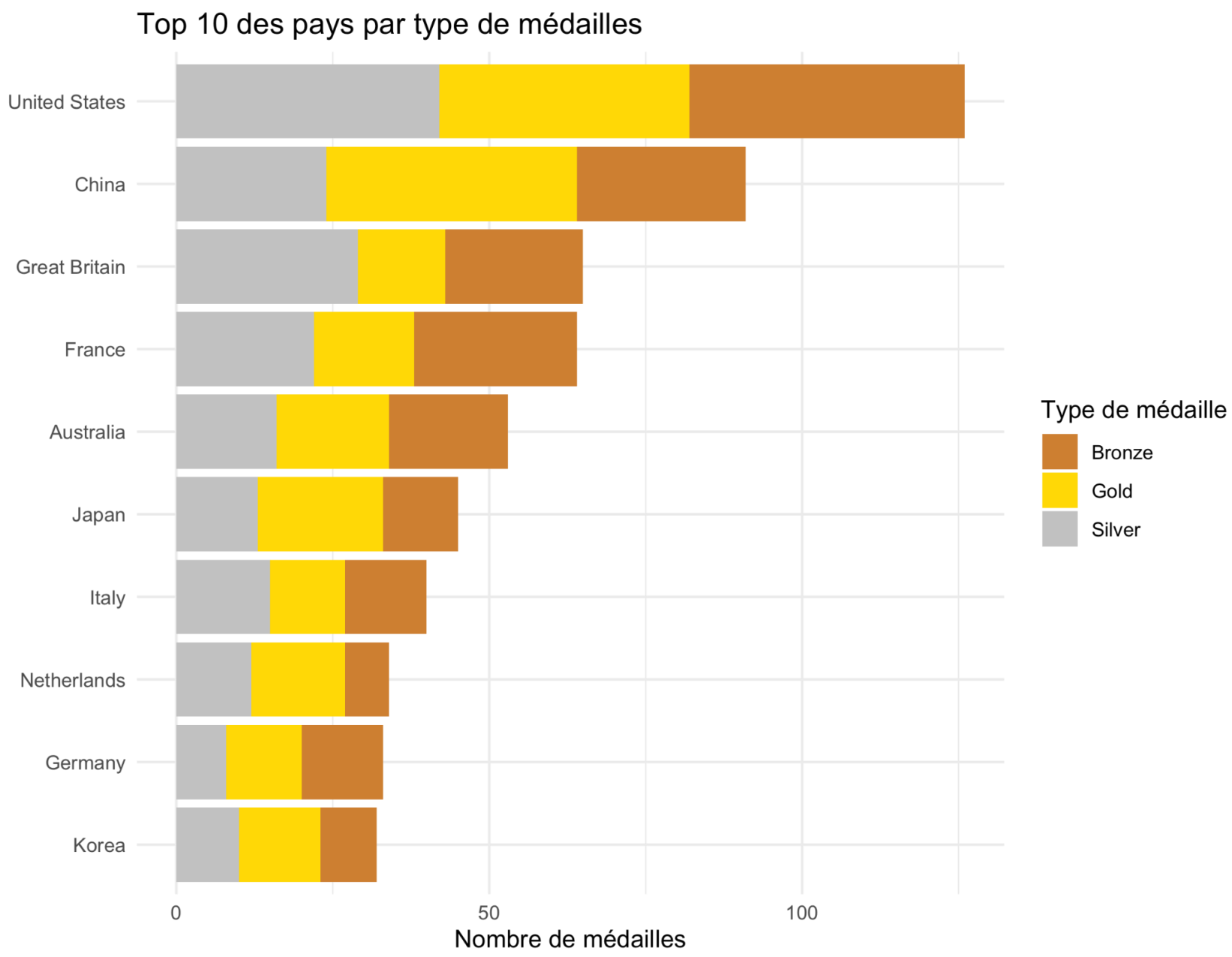
Top 50 des pays ayant le plus d'athlètes -> comparer avec le nb de médailles. Est-ce que l'un explique l'autre -> régression linéaire.

Top 10 des pays avec le plus de discipline -> comparer avec le nb de médailles. Est-ce que l'un explique l'autre -> régression linéaire.

Répartition des âges des athlètes en fonction des pays. -> diviser par le nb d'athlètes du pays -> pourcentage et comparaison au taux de réussite de chaque pays, lui aussi normalisé.

Etude des genre -> nb homme/femme, test de loi normale pour les médailles hommes/femme, répartition des médailles par genre.

Répartition des âges des athlètes en fonction des catégories sportives.

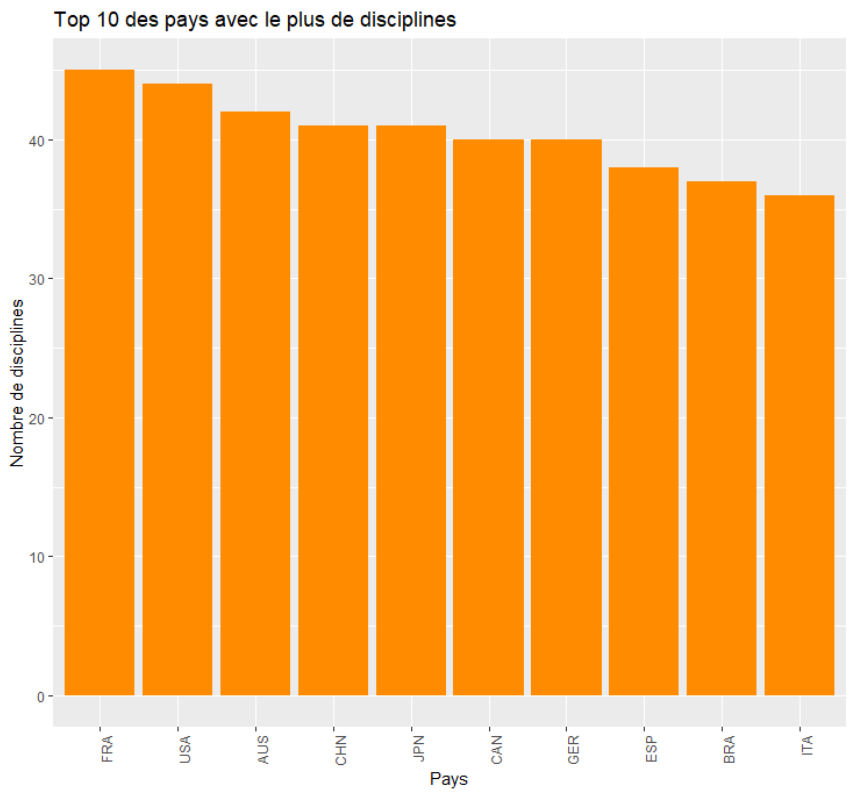


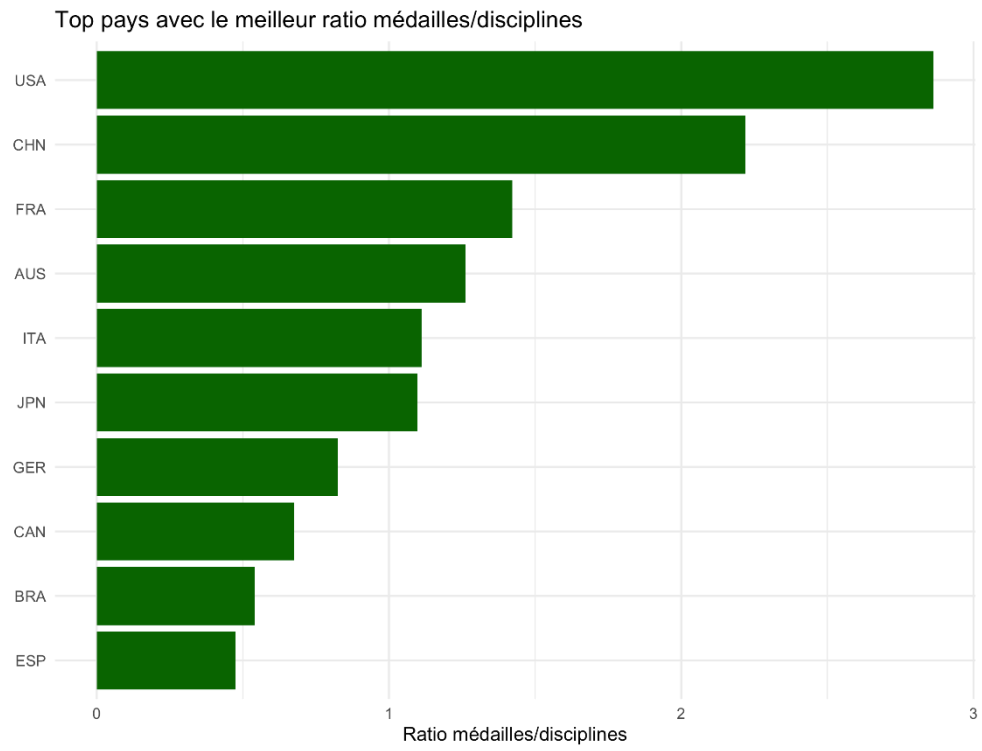
Les États-Unis et la Chine se distinguent comme les leaders incontestés, avec chacun 40 médailles d'or. Les États-Unis dominent en termes de médailles totales grâce à un nombre plus élevé de médailles d'argent et de bronze, totalisant 126 médailles contre 91 pour la Chine.

La moyenne des médailles totales pour les dix premiers pays est d'environ 58,3, avec une médiane de 49, et un écart-type significatif de 30,2, reflétant la disparité entre les pays en tête et ceux en bas du classement.

Les États-Unis, avec 126 médailles, sont bien au-dessus de cette moyenne, tandis que la Corée, avec 32 médailles, est en dessous. En termes de répartition, le Japon se distingue par un ratio élevé de médailles d'or par rapport à ses médailles d'argent et de bronze, indiquant une performance particulièrement forte dans les épreuves où il excelle. En revanche, des pays comme la Grande-Bretagne et la France montrent une distribution plus équilibrée entre les différents types de médailles.

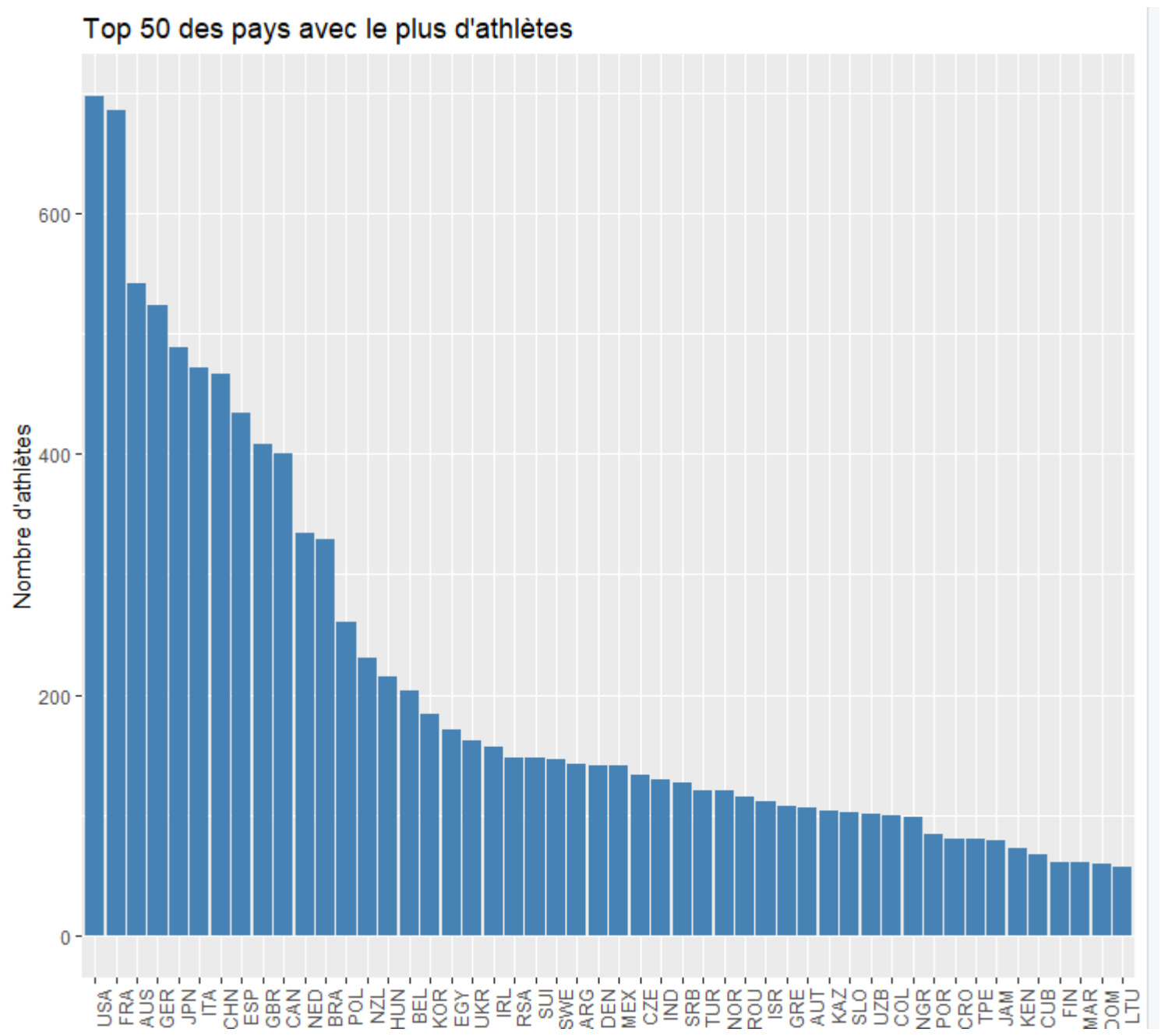
L'analyse de la variance, qui s'élève à 910,2, montre que les différences entre les pays sont statistiquement significatives, soulignant la domination des États-Unis et de la Chine. En conclusion, bien que certains pays aient une performance équilibrée, la compétition est clairement dominée par ceux qui excellent dans l'obtention de médailles d'or.



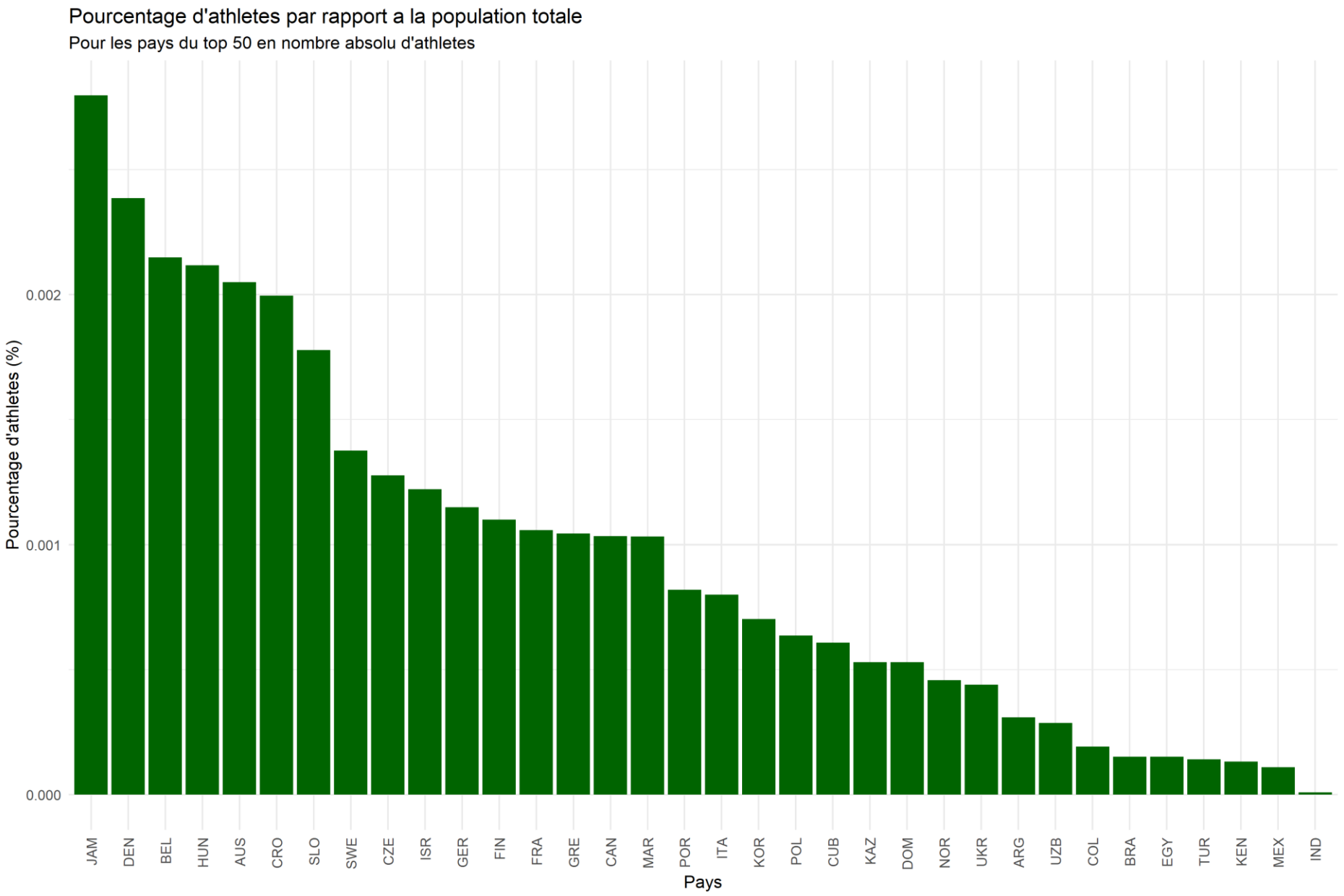


Sur ce graphique affichant les pays avec le plus grand nombre de disciplines différentes, nous pouvons retrouver la plupart des nations situées dans le top 20 du classement global des JO. Nous pouvons donc nous demander si ces deux statistiques sont reliées.

Étant donné que le test du khi2 refuse de s'effectuer car comme l'indique R il est impossible de calculer la valeur de p simulée avec des marges nulles ce qui rend l'approximation du khi2 incorrecte. Nous pouvons donc en déduire que ces deux statistiques ne sont pas reliées.



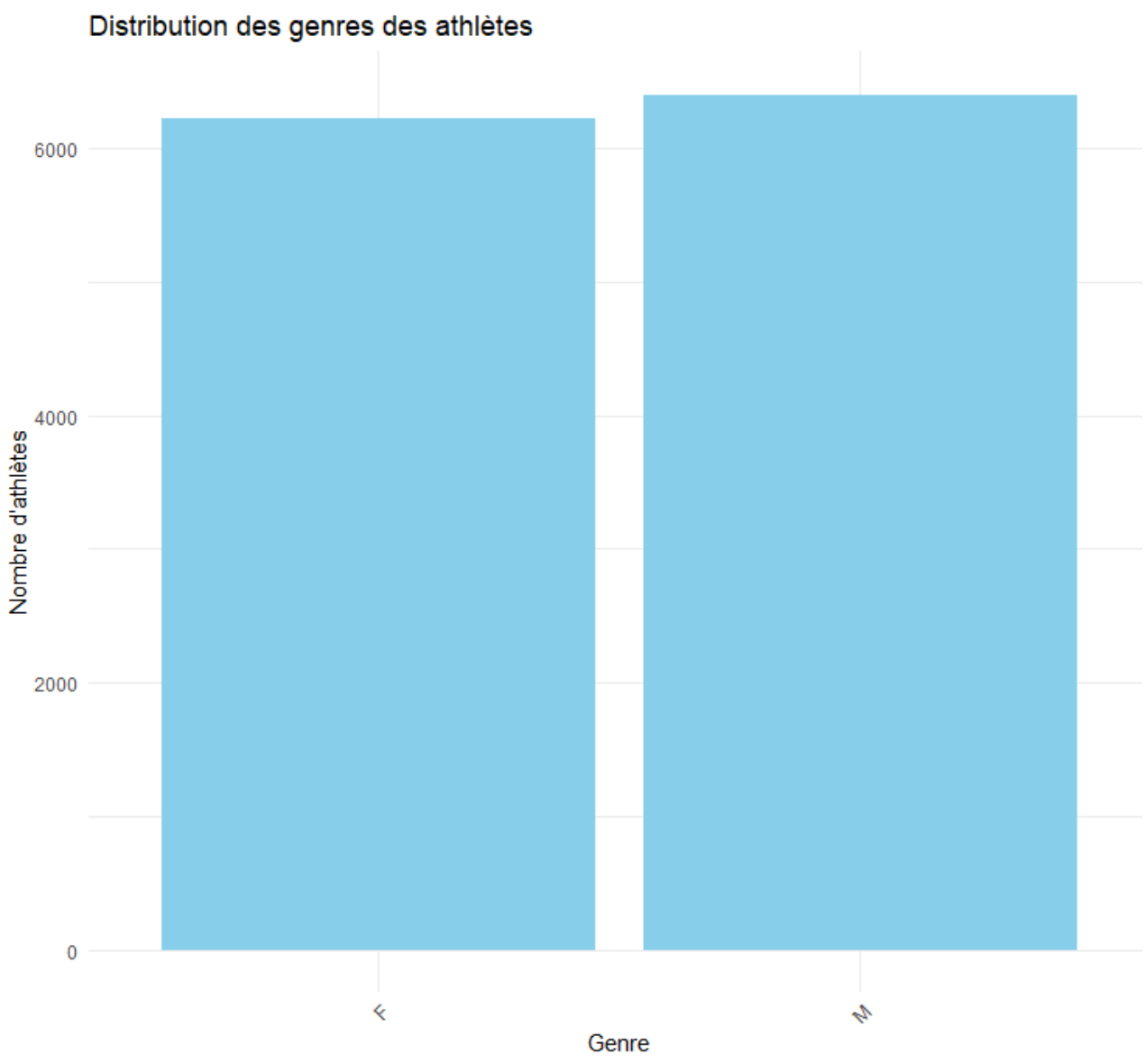
Sur ce graphique nous pouvons observer que le pays avec le plus d'athlètes est les USA avec 696 athlètes suivi de près par la France. Le pays du top 50 avec le moins athlètes est la Lituanie avec 56 athlètes.



Le graphique illustre le pourcentage d'athlètes par rapport à la population totale pour les 50 pays ayant le plus grand nombre absolu d'athlètes, mettant en lumière des disparités intéressantes. La Jamaïque se distingue avec le pourcentage le plus élevé, indiquant une forte proportion d'athlètes par rapport à sa population, suivie par des pays comme le Danemark et la Belgique, qui montrent également des pourcentages relativement élevés. Ces chiffres suggèrent une culture sportive particulièrement développée ou des investissements significatifs dans le sport dans ces nations. En revanche, des pays comme l'Inde et le Mexique, bien qu'ayant un grand nombre d'athlètes en termes absolus, présentent des pourcentages beaucoup plus faibles, reflétant leur population beaucoup plus importante. Cette analyse met en évidence comment la taille de la population influence la représentation proportionnelle des athlètes, soulignant l'importance de considérer les données relatives pour évaluer l'engagement sportif d'un pays. Il est important de noter que les données de population proviennent d'un dataset de 2023 trouvé sur Kaggle (https://www.kaggle.com/datasets/joebeachcapital/world-population-by-country-2023), ce qui peut introduire un léger biais dans l'analyse en raison de la date des données. De plus, le graphique n'affiche que 36 pays au lieu des 50 attendus, car le dataset n'était pas parfait et le code R n'a pas pu identifier tous les pays, expliquant ainsi la présence réduite de pays sur le graphique.

Ne suit pas une loi normale. Pas besoin de test vu le graphique. -> les nations ayant le plus haut pourcentage d’athlètes sont souvent des pays avec de faibles populations.

Étude de Genre :



Description :

On voit bien que la répartition est quasiment égale entre les hommes (male) et les femmes (female), à quelques centaines près.

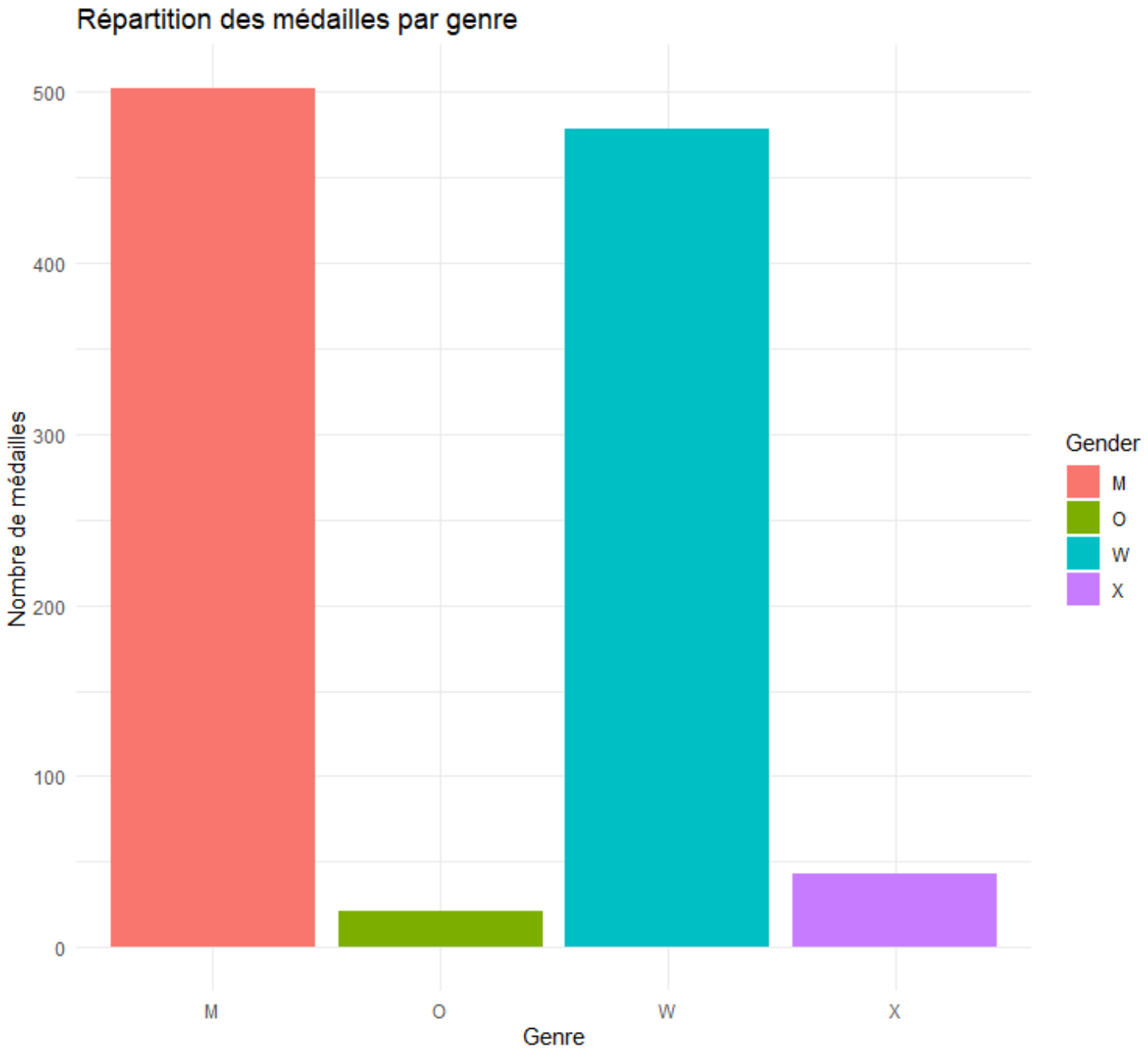
Interprétation et anticipation :
On peut donc s'attendre à une répartition presque égale, avec cependant un léger avantage en nombre de médailles pour les hommes.

Code :

D'abord, on vérifie si la colonne “gender” existe, puis on crée un dataframe pour ggplot. Ensuite, avec ce dataframe, on crée un histogramme. Dans le cas où la colonne "Gender" n'existe pas, un message d'erreur est affiché dans un else if.

Source :

Ancien code R-bloger et docu R



Description :
On constate qu'il y a une répartition assez équilibrée entre les hommes et les femmes, ainsi que deux autres catégories : "Group" et "Non renseigné" (dans le sens où le sexe n'est pas spécifié ou utilisé).

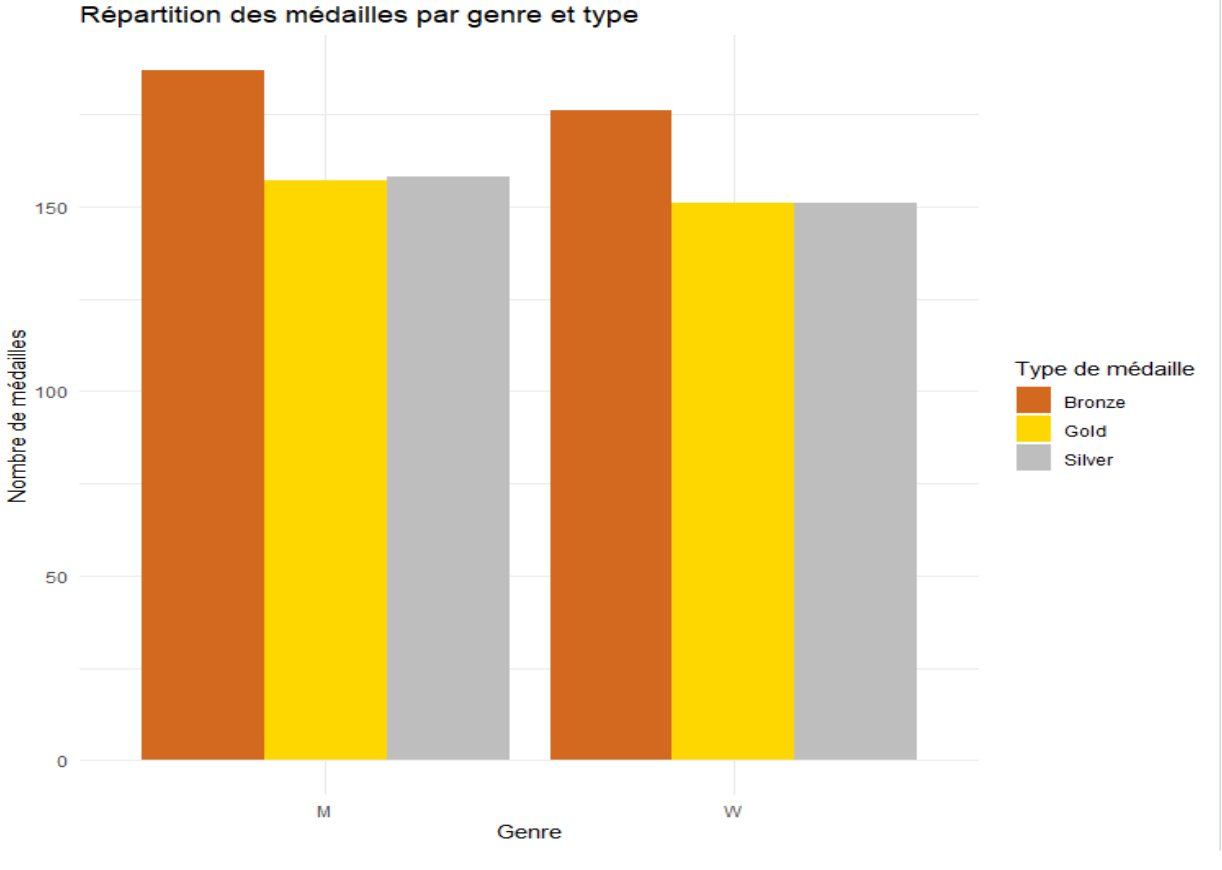
Interprétation et anticipation :
Il est évident que les deux autres genres pourraient fausser nos calculs. Cependant, en raison du faible nombre de données, nous allons les négliger pour les prochains calculs. On observe également que, comme anticipé, la répartition des genres est plutôt équilibrée, bien que l'on note une légère supériorité masculine, même si celle-ci est très infime.

Code :

Comme pour le code précédent, on vérifie d'abord l'existence de la colonne "Gender" dans le dataframe df_medailles. Si elle existe, il regroupe les données par genre et calcule le nombre de médailles par genre à l'aide de group_by et summarise. Ensuite, il crée un dataframe pour ggplot et génère un histogramme montrant la répartition des médailles par genre. Si la colonne "Gender" est absente, un message d'erreur est affiché, comme dans le code précédent.

Source :

Adapté de tutoriels sur R-bloggers et documentation R.



Description :

L'image présente une répartition des médailles par genre (hommes "M" et femmes "W") et par type (bronze, or, argent). On remarque une similarité frappante dans la distribution des médailles entre les genres, avec une légère dominance masculine pour les médailles de bronze.

Interprétation et anticipation :

Il y a un équilibre apparent entre les genres donc une parité dans les performances sportives. La légère prévalence masculine pour le bronze pourrait indique une plus grande participation masculine dans les épreuves où cette médaille est attribuée.

Code :

D'abord prendre la bonne sheet vérifier que la colonne "médal_code" excite affilier les chiffre 1 2 et 3 ah une couleur et ensuite l'utiliser pour comme pour le code au pars avant afficher le nombre de victoire en fonction d'un type de médaille pour chaque sexe puis comparer le différent couleur crée un tableau et afficher le total de tout ça en gg.plot

Code 2.0 :

Avec le tableau on peut donc faire un test de khi^2 et un test de Student le t.test affiche ceci :

```
+   t_test_result <- t.test(Nombre_Medailles ~ Gender, data = df_gold_clean)
+   print(t_test_result)
+ }
Erreur dans t.test.default(x = DATA[[1L]], y = DATA[[2L]], ...) :
  les données sont pratiquement constantes
```

Et le test de khi^2

```
> print(khi2_test)

        Pearson's Chi-squared test

data:  tableau_contingence
X-squared = 0.02105, df = 2, p-value = 0.9895
```

Statistique du khi^2 : La valeur obtenue est 0.02105, ce qui est très faible. Degrés de liberté (df) : Il y a 2 degrés de liberté dans ce test. p-value : La p-value est 0.9895, soit une valeur très élevée. Donc on ne rejette pas l'hypothèse H0 que ce soit vis à vis du test de Student mais aussi le test du khi^2. (Hypothèse H0= Les moyennes des médailles entre les deux groupes sont égales pour le test de Student et pour le khi^2 H0= distribution des médailles est indépendante du genre.)

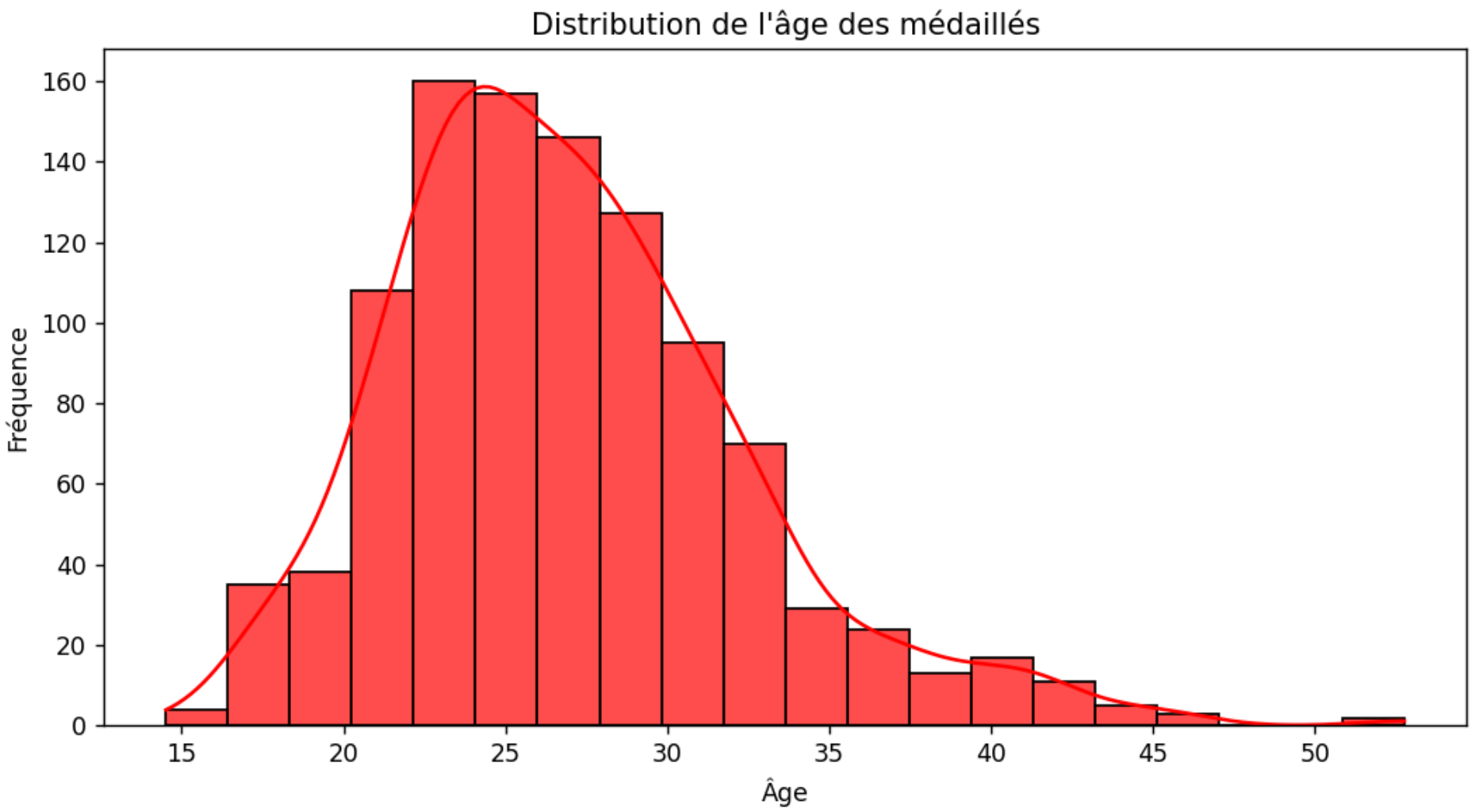
Source :

Cours, Td, tuto R-blogue, documentation R

Conclusions:

Pour conclure, tous ces codes révèlent une répartition globale assez équilibrée entre les genres (hommes et femmes), que ce soit au niveau des athlètes ou des médailles. Il y a une légère supériorité masculine, tant en nombre de participants qu'en répartition des médailles, en particulier pour les médailles de bronze. Les analyses statistiques (test t et khi^2) n'ont pas permis de définir une réelle différence significative entre les genres. En conclusion, il y a une parité dans les performances sportives, sans que cela soit particulièrement notable dans la distribution des médailles entre les hommes et les femmes.

Loi normale Age / Médaille :



Ce graphique représente la distribution de l’âge des médaillés. La fréquence est définie par le nombre d’athlètes. Nous avons effectué trois tests pour analyser en profondeur ces données. Pour commencer,

Test de Student : T-stat= 11.471, p-value=0.00000 Test de Student : Différence significative

Ce test vérifie si l'âge moyen des médaillés est différent de 25 ans, âge de référence dans notre cas. Une **p-value de 0.00000** signifie que la différence des moyennes est **hautement significative**. Cela signifie que l’âge des médaillés est statistiquement différent de celui des autres athlètes. Ensuite,

Test du Khi^2 : Chi^2-stat= 802.241, p-value=0.00000 Test du Khi^2 : Différence significative

Ce test vérifie si la répartition des médaillés selon les classes d’âge est équilibrée. Une **p-value de 0.00000** signifie que la différence entre les classes d’âge des médaillés est **hautement variée**. Cela confirme que la répartition des âges des médaillés n'est pas aléatoire et suit un certain biais. Dans notre cas, il y a plus de médaillés dans la tranche 24 – 30 ans que dans les 50 – 60 ans. Pour finir,

Test de Kolmogorov-Smirnov : KS-stat= 0.058, p-value= 0.00174 Test de KS : Différence significative

Ce test vérifie si la distribution des âges des médaillés suit une distribution normale. Une p-value de 0.00174 (< 0.05) indique qu'on rejette l'hypothèse de normalité, c’est-à-dire, qu’il existe une différence notable avec une loi normale. L’écart mesuré (KS-stat = 0.058) est faible, ce qui signifie que la distribution n’est pas totalement différente d’une loi normale, mais elle présente quand même une déviation significative.

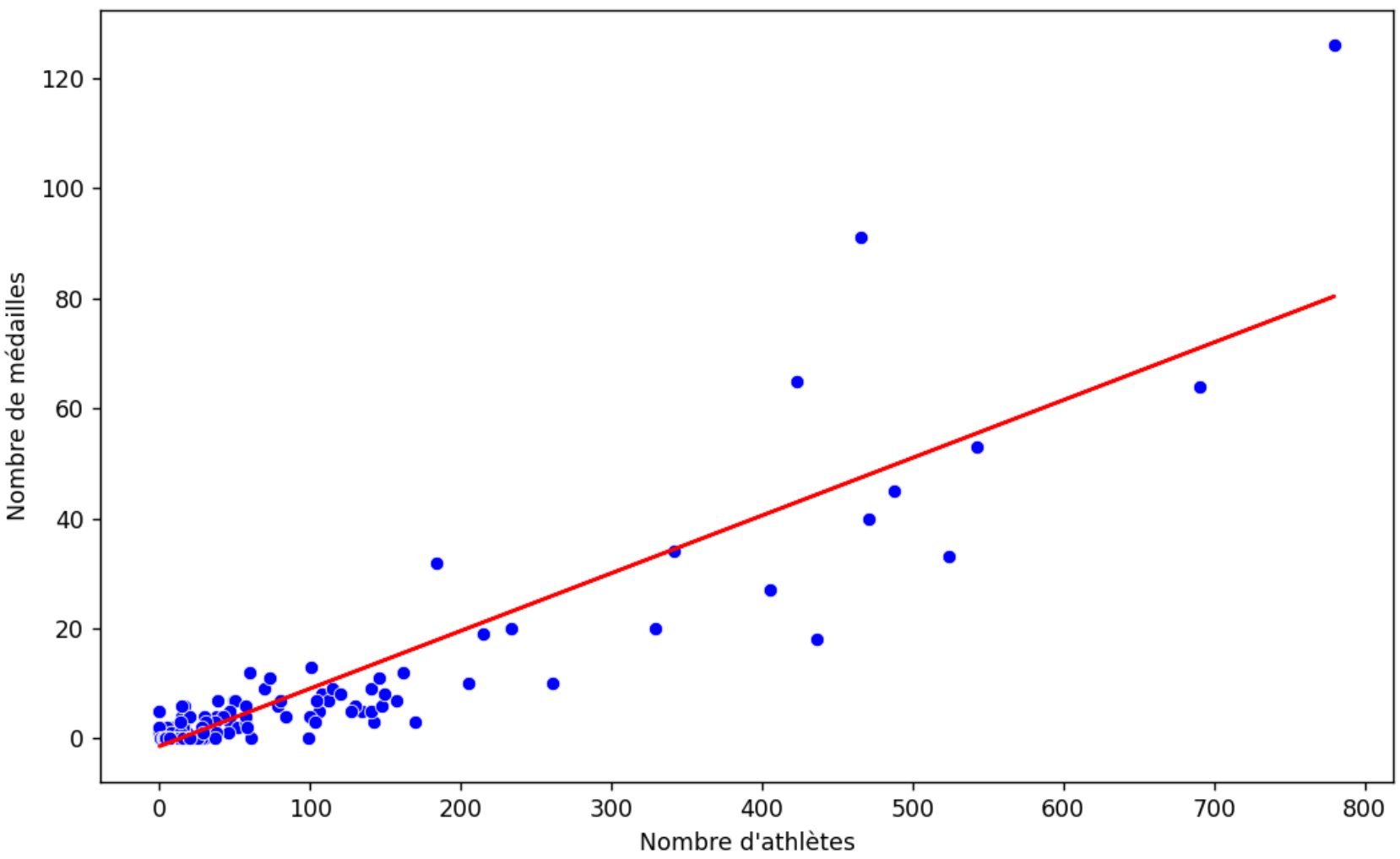
Pour conclure, l’âge des médaillés ne suit pas une loi normale parfaite car tous les tests effectués montrent une différence significative. Mais la distribution reste proche d’une loi normale avec un léger biais, ici une asymétrie : une gaussienne.

Nom code : loiN_age_athlete.py (fichier python car nous avons rencontré des soucis)

Source code :

- pandas → Manipulation et analyse de données tabulaires (excel).
- scipy.stats → Tests statistiques (Student, Khi^2, Kolmogorov-Smirnov, etc.).
- matplotlib.pyplot → Visualisation graphique.
- seaborn → Librairie de visualisation basée sur Matplotlib, plus avancée et esthétique.

Loi normale nombre d’athlète dans un pays / médailles



Ce graphique représente la distribution entre le nombre d'athlètes envoyé par un pays et le nombre de médailles. L'idée est de comprendre si les pays qui envoient un grand nombre d'atlhètes, ont plus de chance d'avoir des médailles. Ici, on remarque que cette affirmation n'est pas respectée parfaitement mais que le concept est là. En effet, outre quelques exceptions, un pays a plus de chances d'avoir beaucoup de médailles, s'il envoie plus d'athlètes. Nous avons effectué trois tests pour analyser en profondeur ces données. Pour commencer,

Test de Student : Stat=6.600, p-value=0.00000 Test de Student : Différence significative

Ce test permet de vérifier si la moyenne du nombre d'athlètes par pays est différente du nombre moyen de médailles remportées par pays. Une **p-value de 0.00000** signifie que la différence des moyennes est **hautement significative**. Cela indique que le nombre d'athlètes envoyés par un pays et le nombre de médailles obtenues **ne sont pas proportionnels**. Ensuite,

Test du Khi^2 : Stat=4867.861, p-value=0.00000 Test du Khi^2 : Différence significative

Ce test permet de vérifier si la distribution des pays selon le nombre d'athlètes et de médailles est équilibrée. Une **p-value de 0.00000** signifie que la différence observée entre les pays en termes de participation et de médailles **n'est pas un hasard**. Cela confirme que certains pays obtiennent beaucoup plus de médailles grâce à leur nombre d'athlètes, tandis que d'autres envoient de nombreux athlètes sans forcément obtenir un grand nombre de médailles.

Test de Kolmogorov-Smirnov : Stat=0.613, p-value=0.00000 Test de KS : Différence significative

Ce test vérifie si la distribution du nombre d'athlètes par pays suit la même loi que la distribution du nombre de médailles par pays. Une p-value de 0.00000 (< 0.05) indique qu'on rejette l'hypothèse d'égalité des distributions. L'écart mesuré (KS-stat = 0.613) est très important, ce qui signifie que le nombre d'athlètes envoyés et le nombre de médailles gagnées suivent deux distributions très différentes. Cela traduit une inégalité entre les pays, où certains surperforment par rapport à leur nombre d'athlètes, tandis que d'autres sous-performent.

Pour conclure, l'âge des médaillés ne suit pas une loi normale parfaite car tous les tests effectués montrent une différence significative. Mais la distribution reste proche d'une loi normale avec une hétérogénéité dans la répartition des médailles.

Nom code : loiN_pays_athlete.py (fichier python car nous avons rencontré des soucis)

Source code :

- pandas → Manipulation et analyse de données tabulaires (excel).
- scipy.stats → Tests statistiques (Student, Khi^2, Kolmogorov-Smirnov, etc.).
- matplotlib.pyplot → Visualisation graphique.
- seaborn → Librairie de visualisation basée sur Matplotlib, plus avancée et esthétique.