TRAITEMENT STATISTIQUE DES DONNÉES

&

DATAMINING

Léo Paul BOITEL

Bienfait MAMBOTE KIDIELA
Marc KPATCHA MASSABALO
Hakima ACHAK

2023-2024

Sujet: Analyse des performances des athlètes dans le cadre des Jeux Olympiques
Population cible: Les athlètes olympiques (Homme/Femme)

Entités statistiques étudiées: les données spécifiques des athlètes : performances, médailles, tendances historiques
Phénomène à prédire:

- Le poids et la taille des athlètes ont un impact sur leurs performances dans différents sports
- La participation aux JO a évolué selon les pays, les sports et le genre depuis 1890

I-Définition des objectifs

Introduction

Les Jeux Olympiques est un événement multisports parmi les plus anciens et les plus prestigieux au monde. Ils captivent l'attention de milliards de personnes dans le monde entier. Depuis leur rénovation moderne en 1896, les Jeux se sont développés bien au-delà de leur origine antique, incarnant un mélange unique de compétition sportive, de culture, et de principes universels tels que l'excellence, l'amitié, et le respect. L'attraction des Jeux Olympiques réside dans leur capacité à rassembler les athlètes des quatre coins du monde, chacun portant les espoirs et les rêves de sa nation, dans un esprit de fair-play et de paix.

Les Jeux Olympiques de cette année qui se dérouleront à Paris accueilleront 10 500 athlètes a annoncé le comité internationale Olympiques avec une stricte parité homme et femme. Une première dans l'histoire des événements. En France, 248 athlètes dont 124 femmes et 124 hommes participent aux Jeux Olympiques .

Afin d'être dans l'actualité de cet événement, nous avons entrepris une étude approfondie en explorant un vaste ensemble de données des Jeux Olympiques disponible sur Kaggle.

L'exploration de ces données offre une opportunité d'analyser en profondeur les tendances et les modèles qui se dessinent à travers les âges, les performances physiques, les dynamiques de genre, les changements sociaux-culturels. Cette exploration peut révéler comment les caractéristiques physiques des athlètes influencent leurs performances dans différentes disciplines, comment la participation et les succès des athlètes ont évolué en fonction de leur sexe, ou encore comment les pays ont progressé ou régressé dans le classement des médailles au fil du temps.

Pour ce faire, nous effectuerons des d'analyses descriptives et prédictives spécifiques afin d'observer et découvrir des modèles significatifs qui influencent la dynamique de la compétition olympique.

Description des données

Nous avons utilisé 3 fichiers csv:

Le premier est notre jeu de données principale : ‘dataset_olympics.csv’. Nous avons par la suite intégré deux autres fichiers csv : ‘noc_regions.csv’ et ‘countries by continents.csv’ pour avoir des plus d’informations afin d’appuyer nos analyses.

II-Inventaire des Données Existantes

Présentation du jeu de données

Les données portent sur les jeux olympiques de 1896 à 2016. Il y a 70 000 lignes et 15 colonnes. Voici les colonnes du jeu de données, avec leur description :

Données	Description	Type	Type informatique	Exemple
ID	Identifiant unique de l'événement olympique	Quantitative-discrète	Integer	1, 2
Name	Nom de l'athlète	Qualitative - Nominale	String	Usain St .Leo Bolt
Sexe	Sexe de l'athlète	Qualitative-Nominale	String	Masculin(M) ou Féminin(F)
Âge	Âge de l'athlète lors de la compétition	Quantitative discrète	Float	25.0, 30.0
Height	Taille de l'athlète (en cm)	Quantitative-Continue	Float	180.0 cm, 165.0 cm
Weight	Poids de l'athlète(en kg)	Quantitative-Continue	Float	75.0 kg, 55.0 kg
Team	Nom de l'équipe nationale	Qualitative-Nominale	Float	United State , France
NOC	Code du Comité national olympique	Qualitative-Nominale	String	USA, FRA
Games	Année et saison des Jeux olympiques	Qualitative-Nominale	String	"2008 Summer", "2016 Winter"
Year	Année des Jeux olympiques	Quantitative - Discrète	Integer	2008, 2016
Season	Saison des Jeux olympiques (été ou hiver)	Quantitative-discrète	String	Summer, Winter
City	Ville hôte des Jeux olympiques	Qualitative-Nominale	String	Paris, London
Sport	Sport pratiqué par l'athlète	Qualitative-Nominale	String	Football
Event	Événement spécifique dans le sport	Qualitative-Nominale	String	Sailing Mixed 8 mètres

Medal	Type de médaille remportée (or, argent, bronze, ou NA si aucune médaille)	Qualitative-Ordinale	String	Gold, bronze

III-Prétraitement des données:

Le fichier des données en format csv est bien structuré. Les noms des colonnes sont explicites (les explications concernant les valeurs abrégées de la colonnes Code du comité national olympique “NOC”’ se trouvent dans un autre fichier “noc_region.csv”). Néanmoins, nous avons entrepris plusieurs actions pour optimiser notre jeu de données : nous avons reformaté les types de données pour économiser de l'espace mémoire, examiner attentivement les valeurs manquantes pour garantir la fiabilité de nos données, éliminer les doublons et ajuster les noms de certaines colonnes afin de les rendre plus pertinents pour notre analyse.

1. Nettoyage des données

1.1. Analyse des données manquantes

Le jeu des données contient **96397** cellules vides (9,2% du jeu des données).

Pourcentage des données manquantes par colonnes:

ID 0.00 %
Name 0.00 %
Sex 0.00 %
Age 3.90 %
Height 23.22 %
Weight 24.43 %
Team 0.00 %
NOC 0.00 %
Games 0.00 %
Year 0.00 %
Season 0.00 %
City 0.00 %
Sport 0.00 %
Event 0.00 %
Medal 86.16 %

On peut émettre l’hypothèse que les données n’étaient pas bien conservées avant l’expansion de l’informatique ou que les méthodes de recensement des athlètes n'étaient pas perfomant.

Nous avons décidé d'analyser en profondeur les colonnes contenant les valeurs manquantes.

a) Données manquantes pour les colonnes "Height" et/ou "Weight"

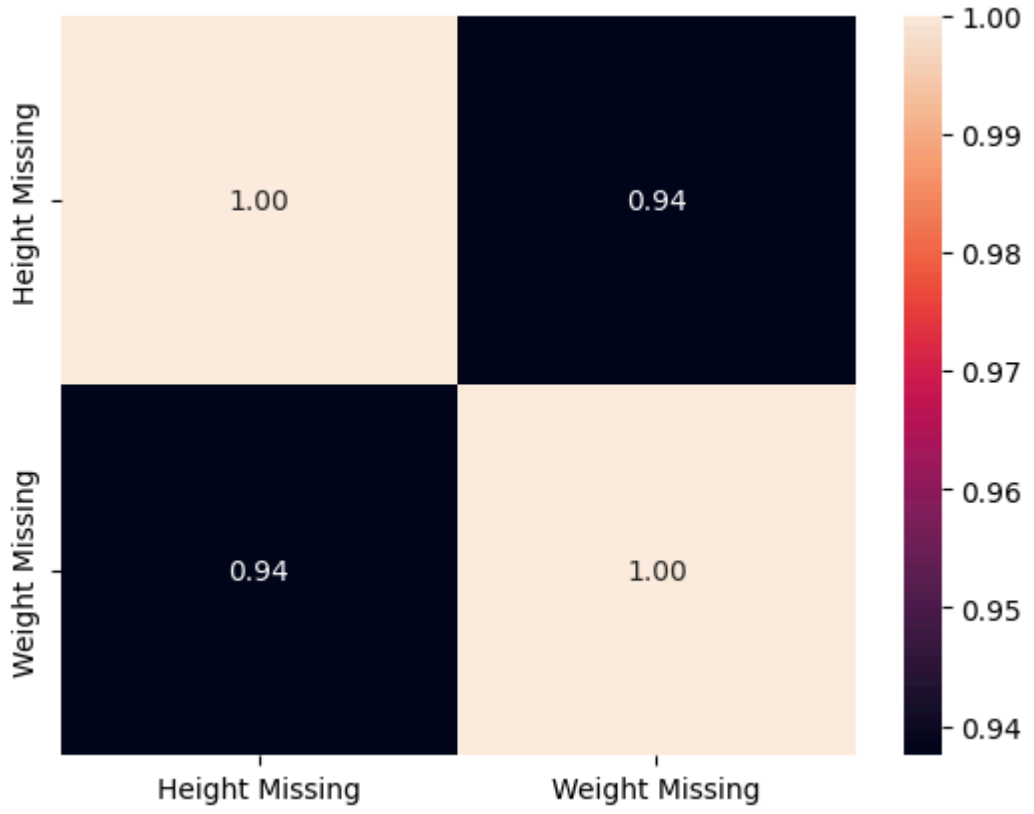
L'absence de valeurs de taille pour 23,22 % et de poids pour 24,43 % des athlètes peut poser des problèmes lors de l'analyse des performances des sportifs.

<u>Vérification de la corrélation entre les des données manquantes des colonnes Height and Weight</u>

Le pourcentage de lignes avec des valeurs manquantes à la fois pour le poids et la taille est de 22,68 %.

Le pourcentage de lignes avec uniquement des valeurs manquantes pour la taille ('Height') est de 0,54 %, tandis que celui avec uniquement des valeurs manquantes pour le poids ('Weight') est de 1,75 %. Ces deux pourcentages sont proches mais ils ne sont pas exactement égaux.

Nous avons utilisé un diagramme en mosaïque pour visualiser la corrélation entre les données manquantes les deux variables.



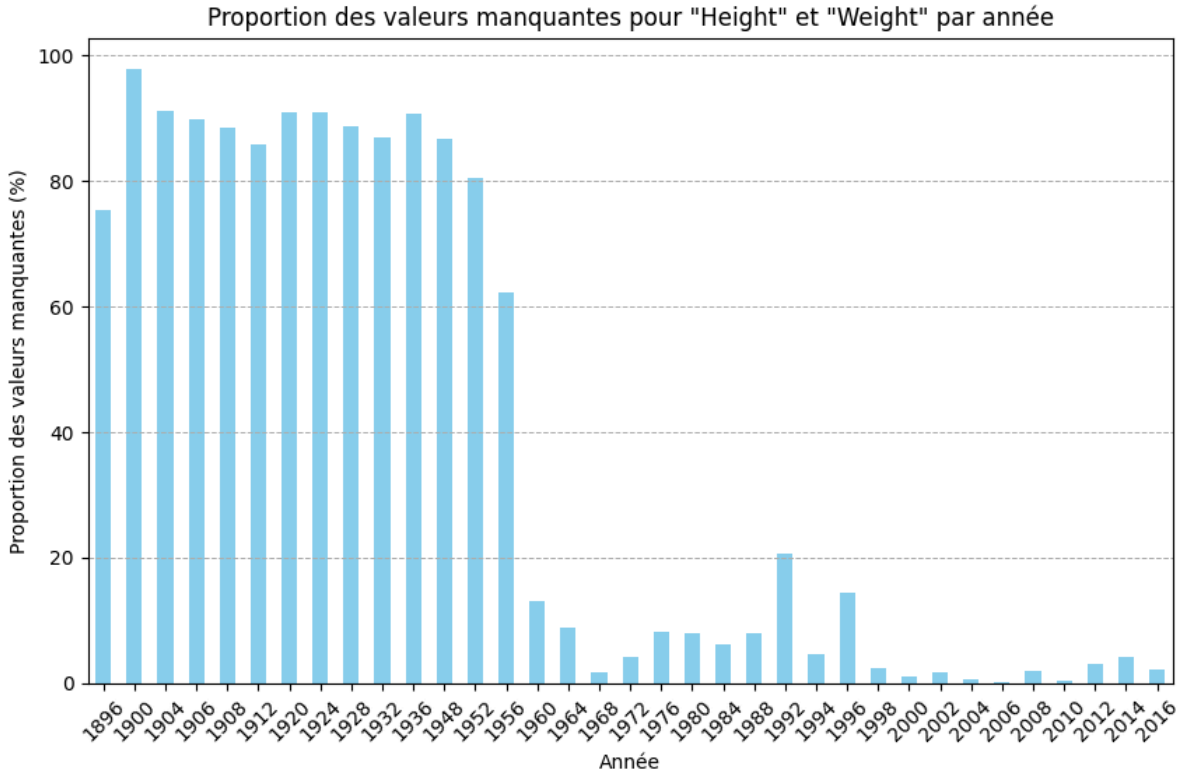
La corrélation entre "Height Missing" et "Weight Missing" est de 0.94. C'est un coefficient de corrélation très élevé, suggérant une forte relation positive entre les valeurs manquantes dans les deux colonnes. En d'autres termes, lorsque la taille est manquante, il est très probable que le poids soit également manquant, et vice versa.

Répartition des données manquantes par année

Afin de vérifier si l'année a un lien avec la qualité des données, nous avons observé les valeurs manquantes du poids et de la taille ensemble et comment elles sont réparties par année. Nous avons déterminé que :

- L'année minimale avec des données manquantes est : 1896
- L'année maximale avec des données manquantes est : 2016

L'année 1900 compte le plus des valeurs manquantes avec : 97,94 %, suivi de l'année 1904 avec 91.23 % et l'année 1920 avec 90.97 %. (Voir ci-dessous un histogramme des valeurs manquantes par année)



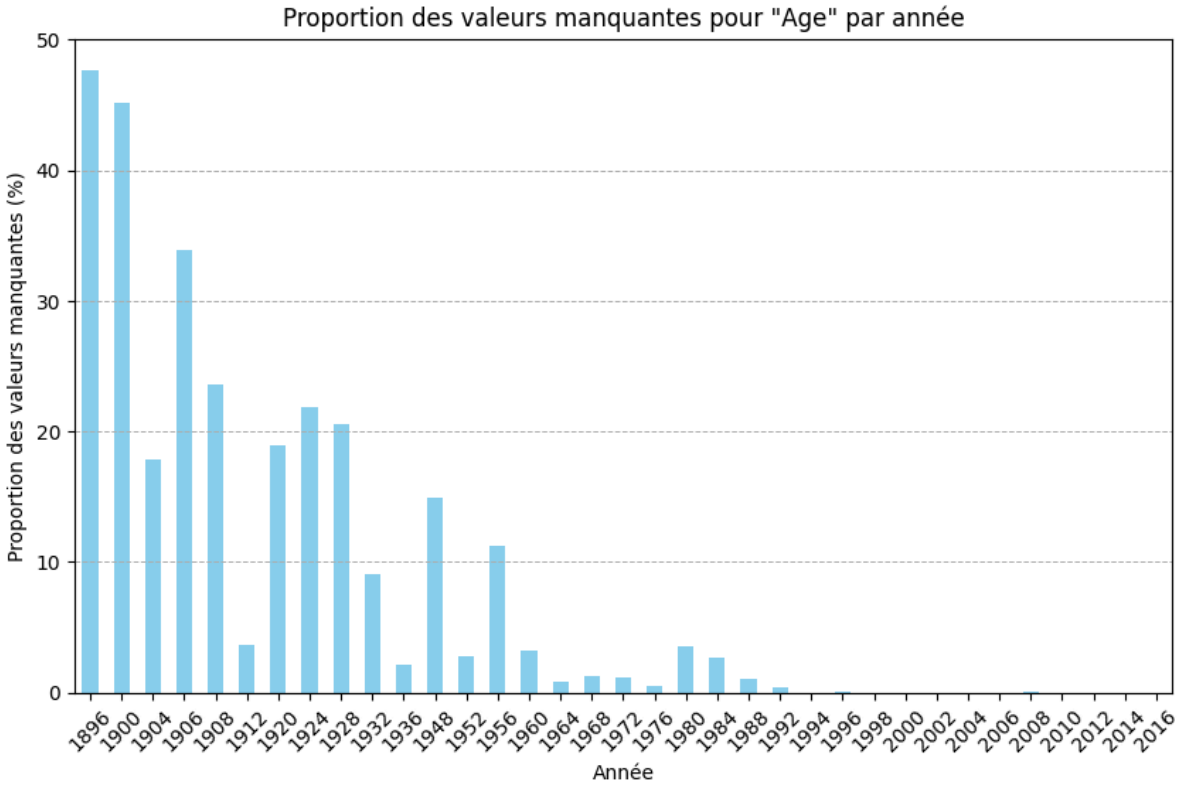
En résumé, les valeurs manquantes du poids ou de la taille sont **surreprésentées** entre l'année 1896 et 1956.
Pour nos analyses des performances des athlètes en fonction de leurs attributs physiques, on se focalise sur les jeux Olympiques organisés entre 1996 et 2016, cette intervalle n'ayant pas beaucoup de valeurs manquantes de poids et/ou de taille. Les années précédentes serviront à apporter du support aux conclusions que nous tirerons des modèles.

b) Données manquantes pour la colonne "Age"

La colonne “Age” contient environ 3,90% de valeurs manquantes dans notre jeu de données.

En analysant la répartition de ces valeurs manquantes par année, nous avons identifié des tendances significatives.

Plus précisément, nous avons constaté que l'année 1896 affiche le plus grand nombre de valeurs manquantes, avec 47,69 %, suivi de près par l'année 1900 avec 45,17 %, et l'année 1906 avec 33,87 %.



Il y a plus de valeurs manquantes avant les années 1990 qui pourrait nous amener à penser que les outils de stockage et d'historisation des données moins performantes dans le passé en seraient la cause mais l'année seule ne suffit pas à expliquer les valeurs manquantes. Par exemple, il y a moins de données manquantes en 1948 qu'en 1980, malgré le progrès technologique entre ces deux périodes.Plusieurs variables pourraient expliquer la mauvaise qualité des données mais ce n’est pas le but de notre travail.

Nos analyses sur l'évolution et l'influence de l'âge sur les athlètes se concentreront principalement sur les données postérieures aux années 90. Les données antérieures seront utilisées pour étayer nos conclusions dans la mesure du possible.

c) Données manquantes pour la colonne “Medal”

La colonne 'Medal' présente un taux élevé de valeurs manquantes, soit 86,16 %. Cette proportion élevée s'explique par le fait qu'une minorité seulement des athlètes parviennent à décrocher des médailles lors de chaque compétition.

1.2. Suppression des doublons

Le dataset comportait **383 lignes dupliquées**, lesquelles ont été éliminées pour ne conserver que des occurrences uniques.

1.3. Valeurs aberrantes

Age

Real number (ℝ)

MISSING

Distinct	68	Minimum	11
Distinct (%)	0.1%	Maximum	88
Missing	2732	Zeros	0
Missing (%)	3.9%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	25.6446453	Memory size	547.0 KiB

Nous avons constaté que la colonne “Age” contenait des valeurs extrêmes

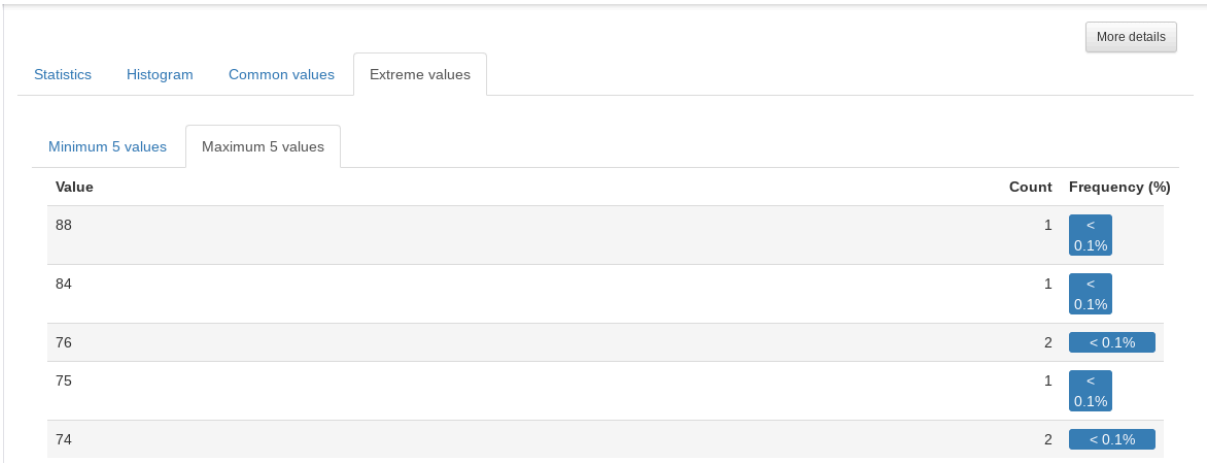
Une valeur minimale de 11 et une valeur maximale de 88

Selon les règlements des Jeux Olympiques, la limite minimale de l’âge varie en fonction des règles des fédérations internationales de chaque sport. La commission exécutive du Comité Internationale Olympique se contente ensuite de valider les décisions. Concernant l'âge maximale, il n'existe aucune limitation pour participer aux Jeux olympiques

En analysant de près les jeux des données on observe :

- La valeur minimale de l’âge est celle de la patineuse artistique britannique Magdalena Cecilia Colledge qui n’avait que 11 ans lors de sa première participation aux jeux olympiques en 1932. Cette valeur est donc une valeur correcte que nous gardons dans notre jeux des données

- La valeur maximale de 88 est celle de Thomas Cowperthwait Eakins qui était un peintre dont les œuvres sur les différents sports olympiens ont contribué à la culture des jeux



Nous avons observé que ces valeurs extrêmes maximales ne concernent que les compétitions artistiques qui ont eu lieu entre 1912 et 1948. Les compétitions d'arts n'étant plus dans les jeux olympiques, nous avons décidé de ne pas les considérer pour les analyses impliquant l'âge et les performances des athlètes à cause du risque des bruits que ces valeurs peuvent causer sur nos analyses

2. Intégration

Le fichier csv 'noc_region.csv' contenant les codes des comités national olympique ainsi que le nom complet des pays a été intégré afin de transformer la colonne 'NOC' de notre jeux des données en 'Code Pays' et associer facilement les différents code avec le nom complet de leur pays.

Le fichier "Countries by Continents.csv" a aussi été intégré pour associer chaque pays à son continent

3. Transformation

a-Transformation des types des données

Voici ci-dessous les types des données avant la transformation.

```
 #  Colonnes   Types
---  ------  --------------  -----
 0   ID       int64
 1   Name     object
 2   Sex      object
 3   Age      float64
```

```
 4   Height  float64
 5   Weight   float64
 6   Team     object
 7   NOC      object
 8   Games    object
 9   Year     int64
 10  Season   object
 11  City     object
 12  Sport    object
 13  Event  object
 14  Medal    object
```

Nous avons modifié les types de données par colonne pour correspondre aux variables quantitatives (discrètes ou continues) et qualitatives :

- Les valeurs de l'âge ont été transformées en quantitative discrète (nombre entier).
- Les données des colonnes contenant des variables qualitatives ont été converties en type 'category', l'équivalent des variables qualitatives dans la bibliothèque Pandas.

```
 #  Colonnes  Types
---  ------  --------------
 0   ID       int64
 1   Name     category
 2   Sex      category
 3   Age      Int64
 4   Height   float64
 5   Weight   float64
 6   Team     category
 7   NOC      category
 8   Games    category
 9   Year     int64
 10  Season   category
 11  City     category
 12  Sport    category
 13  Event    category
 14  Medal    category
```

b-Changement de noms des colonnes et des modalités

On peut connaître le pays de l'athlète à partir de son code du Comité national olympique.
Le code du comité national olympique n'est pas pertinent pour nos analyses mais il correspond à la nomenclature internationale des codes pays. Nous avons donc transformé la colonne 'NOC' en 'Code pays' et inclus une colonne 'Pays' contenant le nom complet des pays.
Les modalités de la colonne 'Sex' ont été modifiées pour une meilleure lisibilité : 'F' est désormais représenté par 'Femme' et 'H' par 'Homme'.

IIII. Analyse descriptive:

1. Analyse des participations aux Jeux Olympiques

Nous trouvons pertinent d'observer en premier lieu l'évolution des participations des hommes et des femmes et leurs performances.

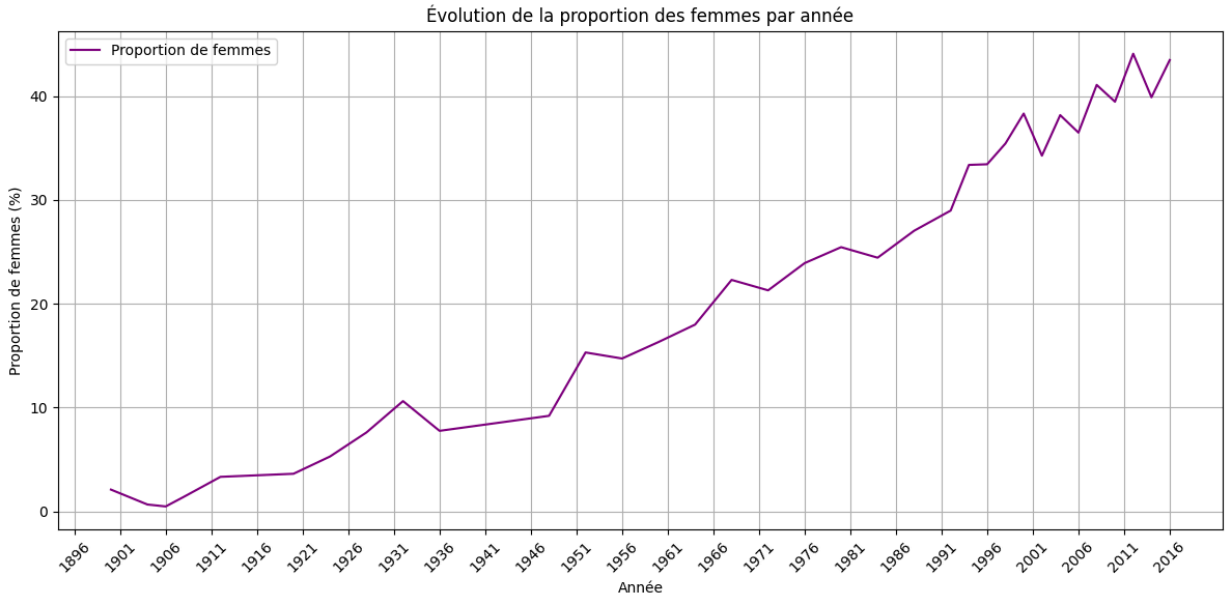
Il est nécessaire de noter que ce ne sont pas les nombres d'athlètes masculins ou féminins mais bien les 'participations'. Un athlète peut participer à plusieurs événements des éditions des jeux Olympiques.

Pour commencer, voici quelques chiffres :

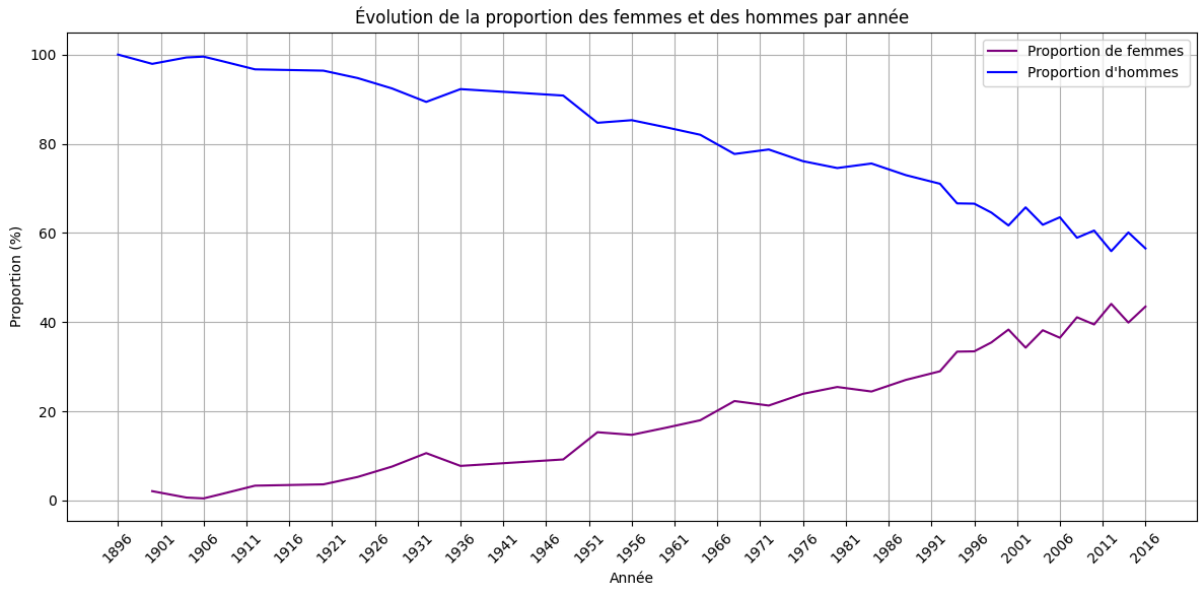
Nombre des participations des hommes dans le jeux des données : 51 531
Nombre des participations des femmes dans le jeux des données : 18 086

Les hommes représentent environ 74 % et les femmes 26 % dans les données que nous avons récoltées.

Ces chiffres nous indiquent que la participation des femmes dans les jeux olympiques est largement inférieure aux hommes. Malgré cela, on peut observer une augmentation à travers les années (voir le graphique ci-dessous. L'axe des Y contient les proportions en pourcentages des femmes et l'axe des Y les années)



La participation des hommes est supérieure à celle des femmes mais elle tend vers la baisse depuis les années 1896. Le graphique ci-dessous illustre l'évolution de la participation des femmes et des hommes (en rouge la courbe des femmes et en bleu la courbe des hommes)

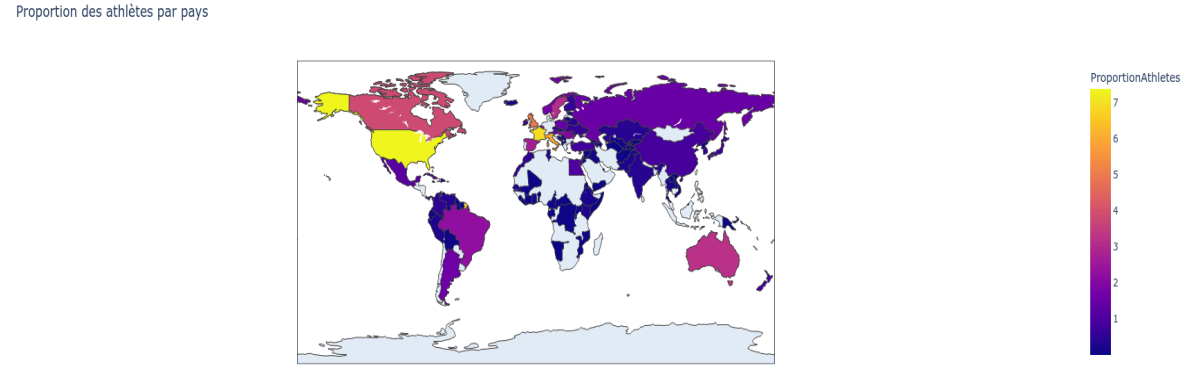


a-Répartition des athlètes par pays

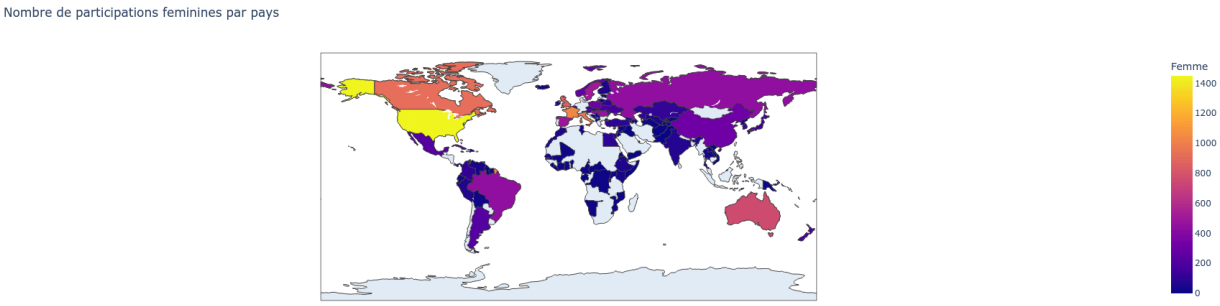
On pourrait éventuellement analyser la proportion des athlètes féminins et masculins par pays mais on aurait un résultat biaisé parce que certains pays n'ont qu'un seul athlète aux Jeux Olympiques dans notre dataset (c'est le cas du Kiribati qui n'a qu'une athlète. La proportion des athlètes féminins sur l'ensemble des athlètes de ce pays serait de 100% et celle des hommes 0 %).

Nous avons trouvé pertinent d'examiner comment les 26% des athlètes féminins et les 74% des athlètes masculins sont répartis dans le monde.

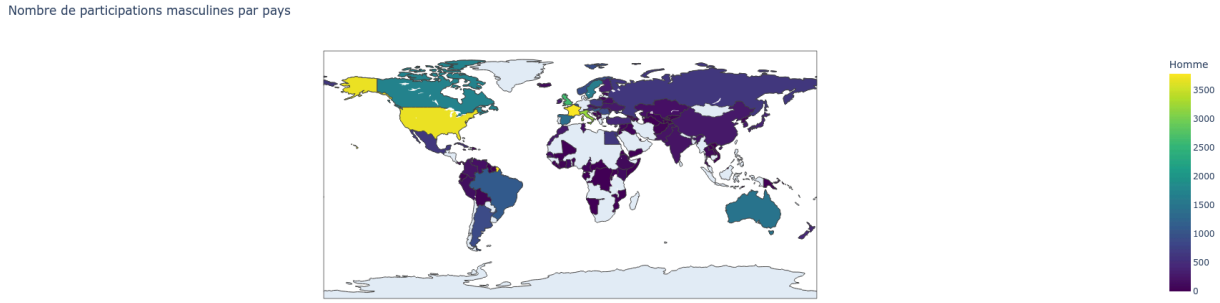
Le graphique ci-dessous montre la répartition des athlètes par pays. Les USA ont le plus grand nombre d'athlètes avec environ 7.37 % aux Jeux Olympiques suivis de la France avec 6.93 %.



Si on distingue les genres dans le graphique ci-dessous. On observe aussi que les USA a le plus grand nombre de participations d'athlètes féminins avec 1 450 participantes.



Concernant la répartition des hommes, c'est la France qui a le plus grand nombre de participations d'hommes avec 3 788 participants masculins.

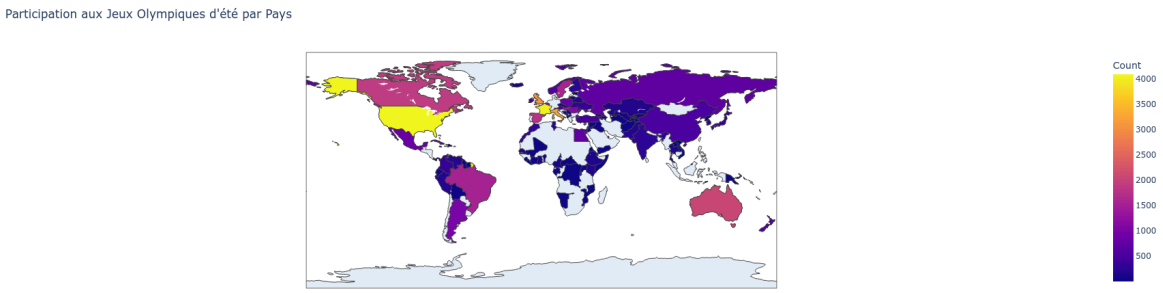


Nous avons étudié la participation des athlètes en général. Les jeux olympiques se déroulent en deux saisons (Hiver et été)

On pourrait affiner nos analyses pour voir comment les chiffres évoluent

b- Analyse par saison

Répartition par pays des participations aux jeux olympiques d'été



La carte ci-dessus illustre la participation des athlètes aux Jeux olympiques d'été par pays. Les différentes couleurs indiquent l'intensité de la participation — plus la couleur est foncée, plus le nombre d'athlètes participant aux Jeux est élevé pour un pays donné.

Voici quelques observations basées sur les couleurs et les légendes de la carte :

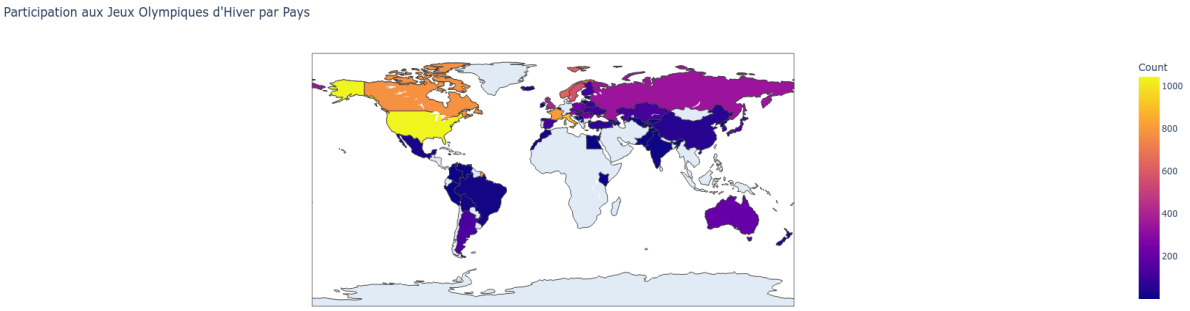
- Les pays colorés en violet foncé semblent avoir la plus forte participation, avec le nombre d'athlètes atteignant ou dépassant 4000.
- Des teintes plus claires de violet et de bleu représentent des nombres décroissants de participants. Les pays en bleu clair, par exemple, ont des représentations plus modestes aux Jeux Olympiques, avec des chiffres se situant autour de 500 participants.

- Les pays sans couleur ou colorés en blanc ou gris représentent ceux sans participation ou avec des données manquantes pour les Jeux d'été.

- La diversité de la participation est visible à travers le monde, certains continents comme l'Amérique du Nord, l'Europe et certaines parties de l'Asie présentant des couleurs plus foncées, indiquent une participation plus élevée.

- L'Australie se distingue également par une couleur significativement plus foncée par rapport à ses voisins de l'Océanie, ce qui suggère un nombre plus élevé d'athlètes olympiques d'été.

c- Répartition par pays des participations aux jeux olympiques d'Hiver



La carte ci-dessus présente la participation aux Jeux Olympiques d'hiver par pays. Tout comme la précédente carte des Jeux d'été, les couleurs ici varient en fonction du nombre de participants de chaque pays, avec une échelle de couleur allant du violet (moins de participants) au jaune (plus de participants).

Observations basées sur les informations visuelles de la carte :

- Les pays avec la participation la plus élevée aux Jeux Olympiques d'hiver sont colorés en jaune, indiquant un nombre de participants proche de 1000 selon la légende des couleurs.

- Il y a clairement moins de pays avec une participation élevée comparé aux Jeux d'été, ce qui peut s'expliquer par la nature plus spécialisée des sports d'hiver et une accessibilité géographique et climatique limitée à ces activités.

- Plusieurs pays, en particulier dans les régions où les sports d'hiver sont populaires comme l'Europe et l'Amérique du Nord, montrent des nombres de participants relativement élevés (en orange et rouge).

- Les pays aux climats plus chauds ou sans tradition de sports d'hiver montrent une participation beaucoup plus faible ou nulle, ce qui est indiqué par la couleur bleue foncée ou l'absence de couleur.

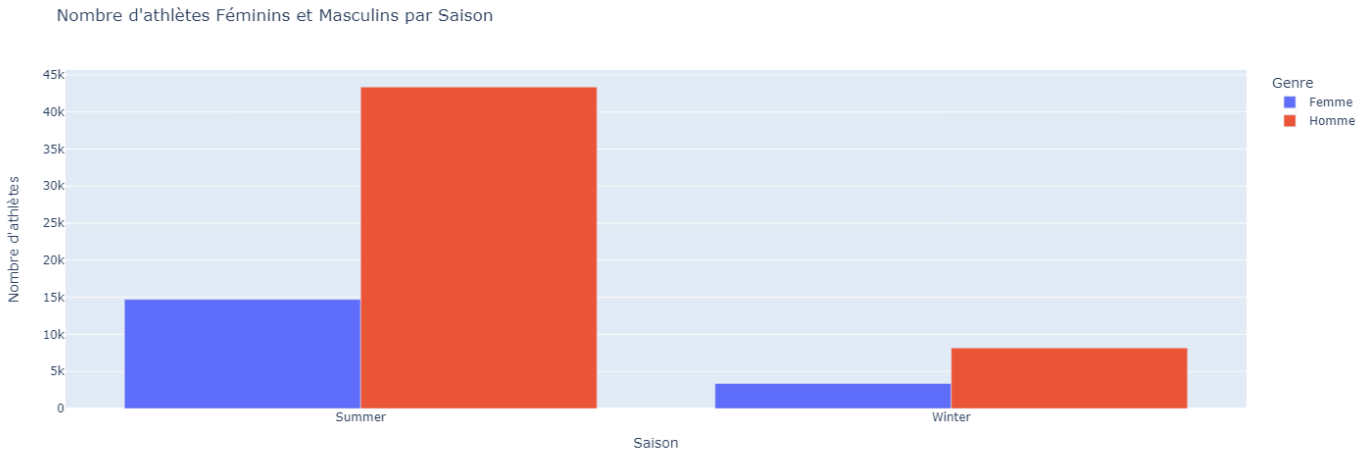
- Les régions de l'Afrique, du Moyen-Orient et d'une grande partie de l'Asie du Sud-Est affichent peu ou pas de participation, ce qui reflète probablement à la fois des facteurs climatiques et des différences dans l'intérêt ou les investissements pour les sports d'hiver.

d- Répartition des participations aux jeux olympiques d’hiver / été par sexe

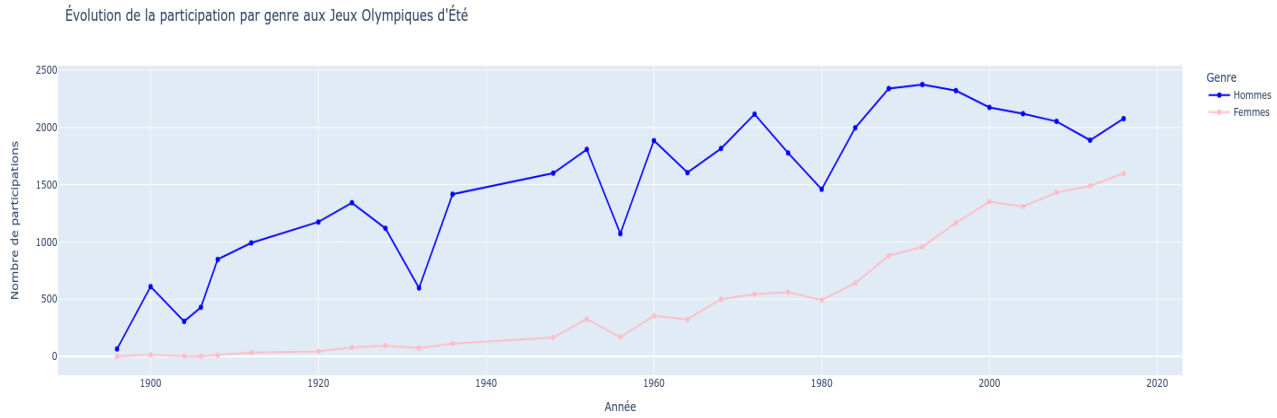
En observant la participation des hommes et des femmes aux jeux Olympiques d’hiver et d’été, on remarque qu’il y a plus de participations de manière générale aux Jeux Olympiques d’été:

- Il y a 43 369 participations d’hommes dans les jeux d’été contre seulement 8 162 aux jeux d’hiver

- Et il y a 14 715 participations féminines dans les jeux d’été contre seulement 3 371 aux jeux d’hiver



d.1 Evolution de la participation des hommes et des femmes dans les jeux Olympiques d'été



Ce graphique représente l'évolution du nombre de participations par genre aux Jeux olympiques d'été au fil du temps. Voici quelques observations que l'on peut tirer de ce graphique :

- Augmentation constante des participants masculins : On peut voir que le nombre d'hommes participants a augmenté de manière assez constante au fil du temps, avec quelques fluctuations notables.

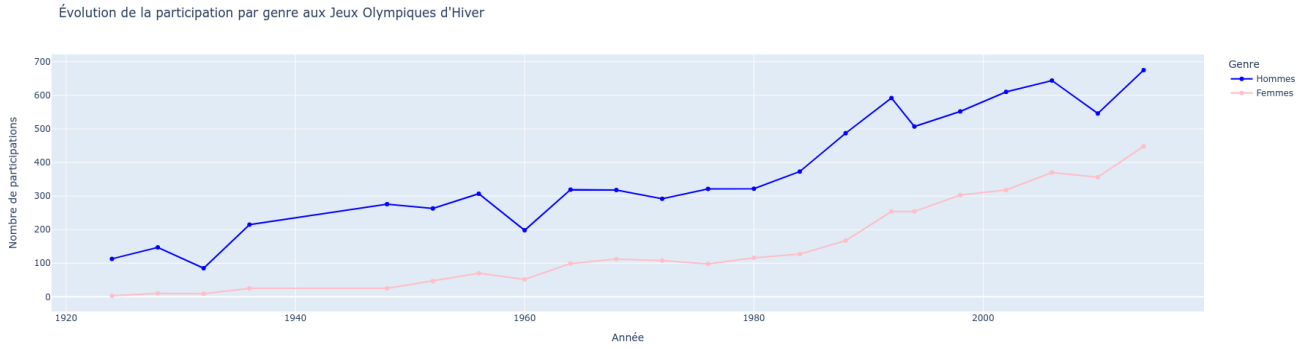
- Faible représentation féminine initiale, mais croissance constante : Le nombre de femmes participantes était initialement très faible comparé à celui des hommes, mais il montre une tendance à l'augmentation constante depuis leur introduction dans les jeux.

- Dips correspondant probablement aux années de guerre : Il y a des chutes prononcées dans le nombre de participants à certains moments du graphique, notamment autour des années 1910 et 1940, qui correspondent probablement aux périodes de la Première et de la Seconde Guerre mondiale, lorsque les Jeux Olympiques n'ont pas eu lieu ou ont été fortement affectés par la guerre.

- Réduction de l'écart entre les genres : Bien que les femmes aient commencé avec beaucoup moins de participantes, l'écart entre les hommes et les femmes semble se réduire avec le temps, ce qui indique une amélioration de la parité de genre dans les Jeux Olympiques d'été.

- Dans les années les plus récentes sur le graphique, on voit que la participation féminine continue de croître, tandis que celle des hommes semble se stabiliser.

d.2 Evolution de la participation des hommes et des femmes dans les jeux Olympiques d'hiver



Le graphique ci-dessus montre l'évolution du nombre de participants par genre aux Jeux Olympiques d'hiver. Voici les points clés :

- Participation masculine plus élevée : Tout comme pour les Jeux olympiques d'été, il y a plus d'hommes que de femmes qui participent aux Jeux olympiques d'hiver, mais la différence n'est pas aussi marquée.

- Croissance régulière : Le nombre de participants, tant masculins que féminins, augmente régulièrement sur la période affichée.

- Comme avec les Jeux d'été, il y a des creux dans le graphique qui peuvent correspondre à des événements mondiaux comme la Seconde Guerre mondiale

- Accélération de la participation féminine : Le taux d'augmentation de la participation féminine semble s'accélérer, particulièrement dans les dernières décennies, ce qui peut indiquer une plus grande inclusion et des opportunités pour les femmes dans les sports d'hiver.

- Dips et stabilisation : Il y a des dips notables dans la participation masculine à certains points, suivis d'une stabilisation ou d'une reprise de la croissance

- Rapprochement des courbes : La courbe de participation féminine semble se rapprocher progressivement de la courbe masculine, ce qui peut refléter une tendance vers l'égalité des sexes en termes de participation aux Jeux.

- Stabilité récente : Dans les dernières années du graphique, la participation masculine semble plus stable avec une légère tendance à la hausse, tandis que la participation féminine continue de croître.

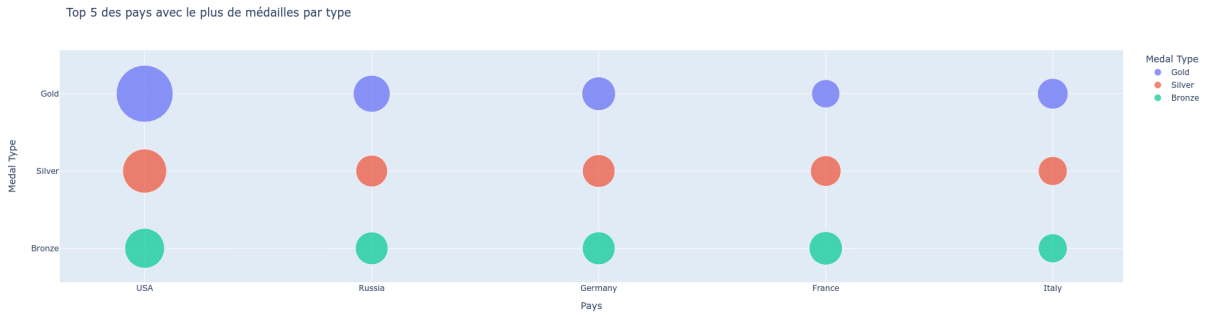
2. Analyse des performances des athlètes par pays

Les performances des athlètes dans les jeux olympiques se mesurent avec les nombres des médailles classées dans selon bronze, silver, gold.

Pour commencer, analysons les pays avec les plus de médailles. Le tableau ci-dessous montre les cinq pays les plus médaillés. Les USA sont les plus médaillés, suivis de la Russie et L'Allemagne

Medal Pays	Gold	Silver	Bronze	Total
USA	747	448	366	1561
Russia	315	232	247	794
Germany	261	246	249	756
France	184	216	257	657
Italy	217	193	194	604

Le graphique offre une représentation visuelle des pays les plus médaillés



Les pays les moins médaillés sont représentées dans le tableau suivant

Medal / Pays	Gold	Silver	Bronze	Total
Ivory Coast	1	0	0	1
Philippines	0	1	0	1
Dominican Republic	1	0	0	1
Jordan	1	0	0	1
Moldova	0	0	1	1

3. Analyse des sports

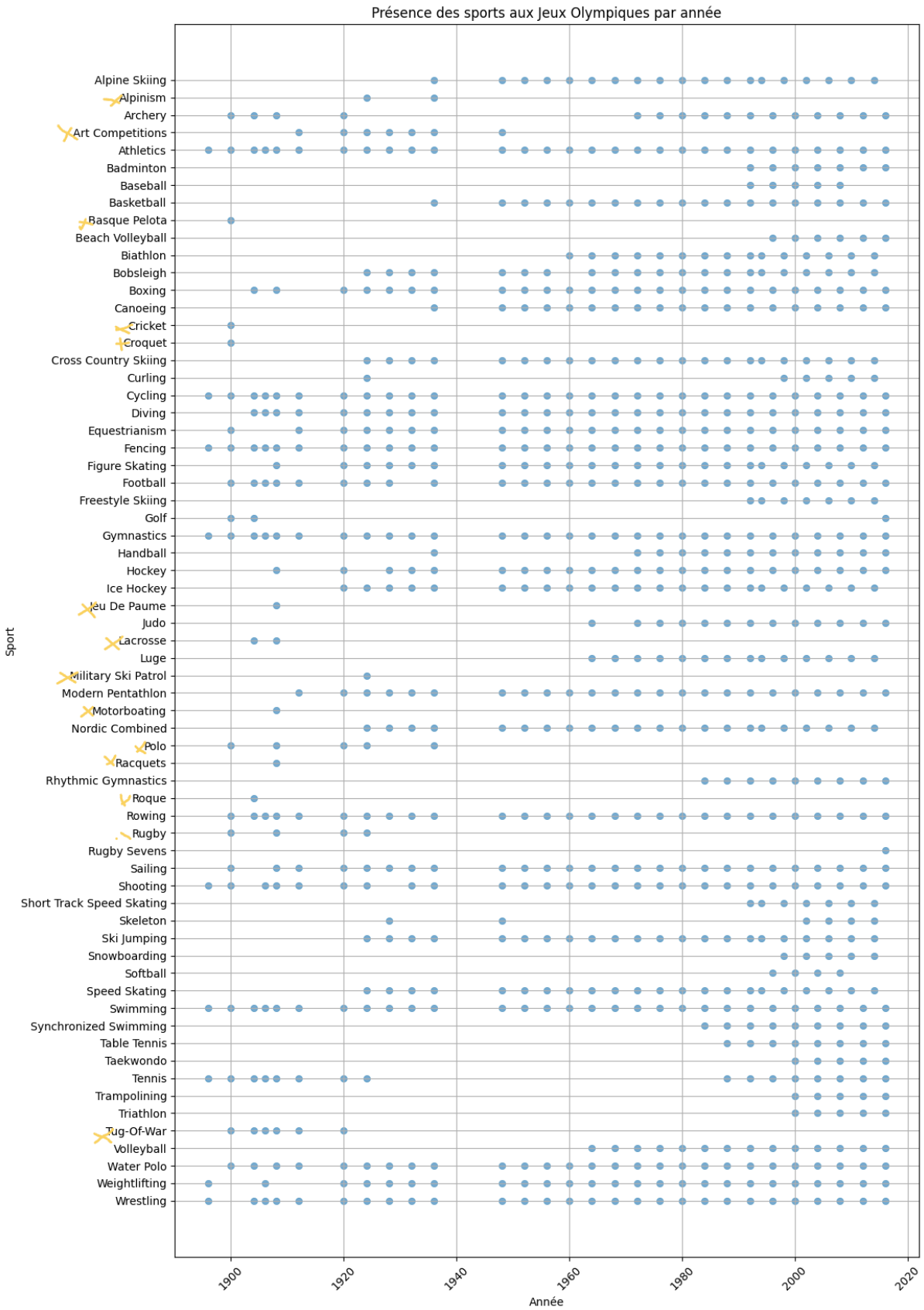
Le jeu des données contient des sports qui ne font plus parties des jeux Olympiques . Voici la liste des ces sports :

Art Competitions : Autrefois considérées comme une discipline olympique de 1912 à 1948, les compétitions d'art ne font plus partie des Jeux Olympiques.
Polo : Dernière apparition aux Jeux Olympiques en 1936.
Tug-Of-War (Tir à la corde) : Inclus dans les Jeux Olympiques jusqu'en 1920.
Lacrosse : Présent aux Jeux Olympiques de 1904 et 1908.
Cricket : Apparu uniquement aux Jeux Olympiques de 1900.
Military Ski Patrol (Patrouille militaire à ski) : Un événement de démonstration aux Jeux Olympiques d'hiver de 1924.
Croquet : Apparu uniquement aux Jeux Olympiques de 1900.
Alpinism (Alpinisme) : Reconnu avec des médailles olympiques au début du 20e siècle pour des réalisations alpines notables pendant la période d'une Olympiade.
Racquets : Présent aux Jeux Olympiques de 1908.

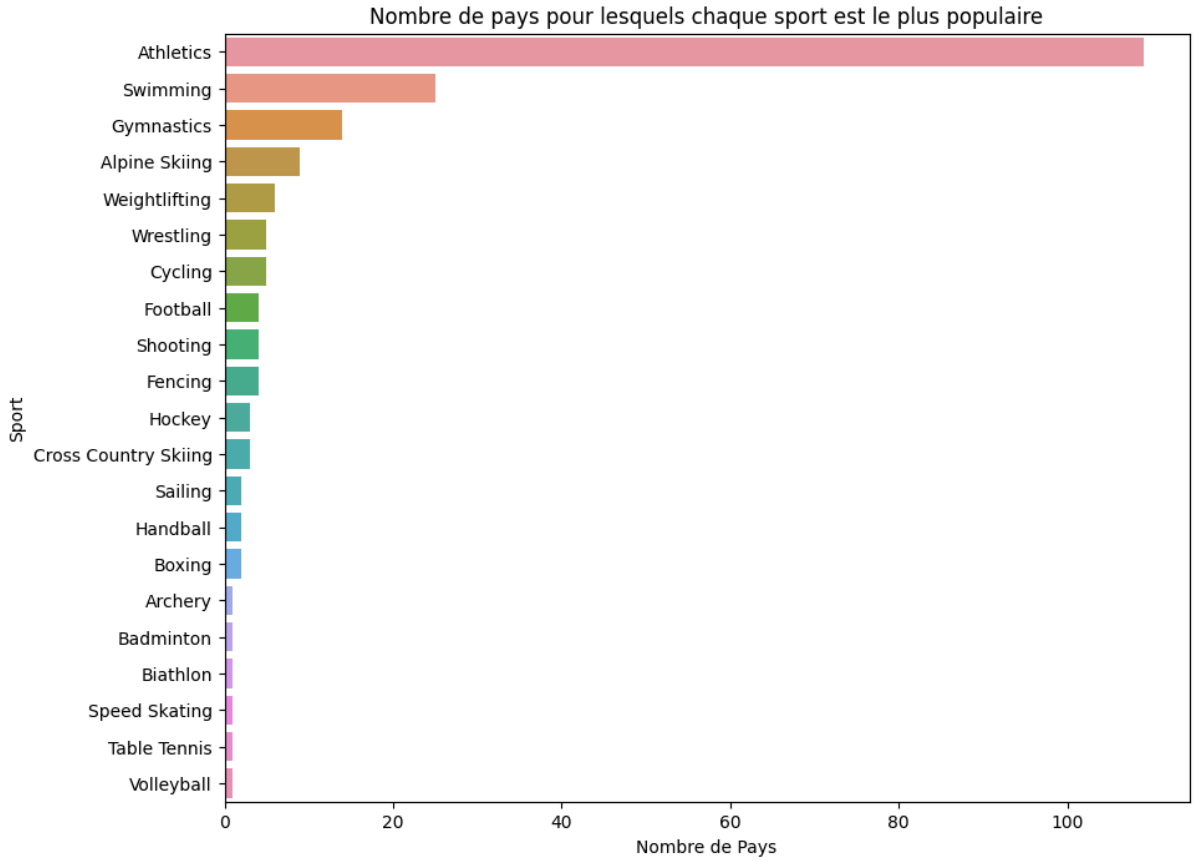
Motorboating (Motonautisme) : Inclus dans les Jeux Olympiques de 1908.
Roque : Une version américaine du croquet, uniquement incluse dans les Jeux de Saint-Louis en 1904.
Jeu De Paume : Présent aux Jeux Olympiques de 1908.
Basque Pelota (Pelote basque) : Apparue aux Jeux Olympiques de 1900 et plus tard comme sport de démonstration.

a-Evolution de la présence des sports dans les Jeux Olympiques

Le nuage des points ci-dessous montre l'évolution des sports à travers les années incluant les sports qui ne font plus partis des jeux Olympiques jusqu'en 2016(le Rugby reviendra aux prochains jeux)


Présence des sports aux Jeux Olympiques par année
Sport
Année
Alpine Skiing
Alpinism
Archery
Art Competitions
Athletics
Badminton
Baseball
Basketball
Basque Pelota
Beach Volleyball
Biathlon
Bobsleigh
Boxing
Canoeing
Cricket
Croquet
Cross Country Skiing
Curling
Cycling
Diving
Equestrianism
Fencing
Figure Skating
Football
Freestyle Skiing
Golf
Gymnastics
Handball
Hockey
Ice Hockey
Jeu De Paume
Judo
Lacrosse
Luge
Military Ski Patrol
Modern Pentathlon
Motorboating
Nordic Combined
Polo
Racquets
Rhythmic Gymnastics
Roque
Rowing
Rugby
Rugby Sevens
Sailing
Shooting
Short Track Speed Skating
Skeleton
Ski Jumping
Snowboarding
Softball
Speed Skating
Swimming
Synchronized Swimming
Table Tennis
Taekwondo
Tennis
Trampolining
Triathlon
Tug-Of-War
Volleyball
Water Polo
Weightlifting
Wrestling
1900
1920
1940
1960
1980
2000
2020

b-Popularité des sports



A travers le graphique en barres ci-dessus on peut observer :

- Athlétisme (Athletics) : C'est de loin le sport le plus populaire, avec le plus grand nombre de pays pour lesquels c'est le sport le plus pratiqué. Cela peut refléter le fait que l'athlétisme est un ensemble de disciplines que beaucoup de pays pratiquent.

- Natation (Swimming) : Le deuxième sport le plus populaire selon le nombre de pays, ce qui suggère que la natation est également largement pratiquée et valorisée à travers le monde.

- Gymnastique (Gymnastics) : Vient ensuite, ce qui pourrait indiquer que la gymnastique est un sport de premier plan dans un nombre significatif de pays

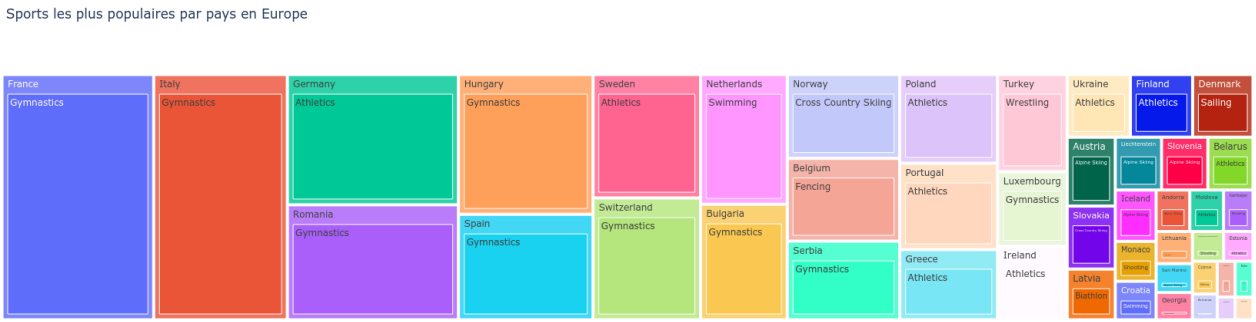
- Ski Alpin (Alpine Skiing) et Haltérophilie (Weightlifting) : Ces sports sont également populaires, bien que dans un nombre plus restreint de pays. Ceci pourrait être dû aux préférences régionales, au climat (pour le ski alpin), ou à la disponibilité des installations et de l'équipement nécessaire pour s'entraîner.

- Les autres sports : Ils ont une popularité variable, avec certains sports comme la lutte (Wrestling), le cyclisme (Cycling), et le football (Football) étant les plus populaires dans au moins quelques pays.

- Sports avec moins de représentation : Les sports en bas du graphique comme le tennis de table (Table Tennis), le patinage de vitesse (Speed Skating), et le volley-ball (Volleyball) sont moins populaires en termes de nombre de pays pour lesquels ils sont le sport le plus pratiqué. Cela ne signifie pas nécessairement que ces sports ne sont pas pratiqués, mais plutôt qu'ils ne dominent pas en termes de popularité par rapport à d'autres sports dans de nombreux pays.

Le graphique donne un aperçu de la répartition mondiale des sports prédominants. Pour les sports qui sont les plus populaires dans de nombreux pays, cela pourrait refléter l'accessibilité universelle et l'intérêt général pour ces sports, tandis que pour les autres, cela pourrait indiquer des spécialités régionales ou culturelles.

c- Sports populaires par pays d'Europe



Dans ce treemap, nous pouvons observer les points suivants :

- Gymnastique (Gymnastics) semble être le sport le plus populaire dans plusieurs pays européens comme la France, l'Italie, l'Espagne, la Suisse, et la Bulgarie.

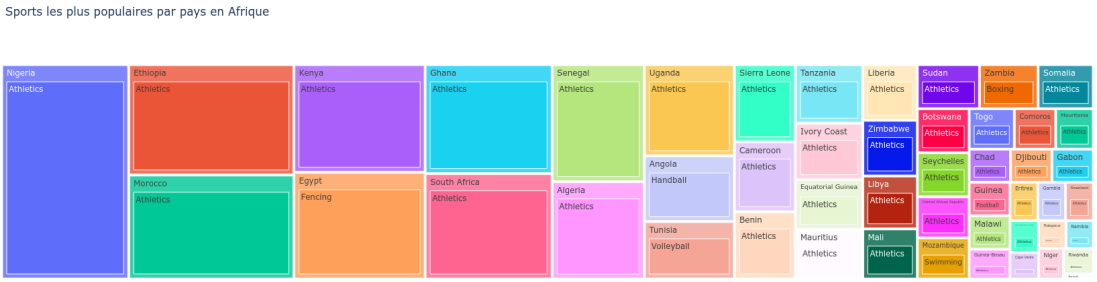
- Athlétisme (Athletics) est également un sport populaire, apparaissant comme le plus pratiqué dans des pays tels que l'Allemagne, la Suède, la Pologne, et l'Irlande.

- Certains sports sont spécifiques à certains pays, comme la Lutte (Wrestling) en Turquie, ce qui peut refléter une tradition ou une expertise particulière dans ce sport.

- Des sports d'hiver tels que le Ski Alpin (Alpine Skiing) et le Biathlon sont les plus populaires dans des pays avec des conditions climatiques appropriées, tels que l'Autriche et la Lettonie respectivement.

- La voile (Sailing) apparaît comme le sport le plus populaire au Danemark, ce qui pourrait être dû à la géographie du pays qui possède de longues côtes.

d- Sport le plus populaire par pays africains



D'après le treemap :

- Athlétisme (Athletics) est le sport dominant en Afrique, étant le plus populaire dans la plupart des pays représentés. Cela reflète probablement la forte tradition de l'athlétisme sur le continent, surtout dans les courses de distance où des pays comme l'Éthiopie, le Kenya et le Maroc sont réputés pour leurs coureurs d'élite.

- Boxe (Boxing) est signalée comme le sport le plus populaire en Zambie, ce qui pourrait indiquer une tradition ou un investissement particulier dans ce sport dans ce pays.

- D'autres sports comme l'Escrime (Fencing) en Égypte, le Handball en Angola, et le Volleyball en Tunisie montrent la diversité des sports populaires sur le continent, chaque pays ayant ses propres préférences sportives qui peuvent être influencées par la culture, l'histoire, ou les investissements en infrastructure et en formation.

- Des pays comme la Libye, le Mali, et le Mozambique montrent également une seule et unique couleur, indiquant probablement qu'ils ont un sport qui se distingue particulièrement par rapport aux autres.

IV. Analyse prédictive:

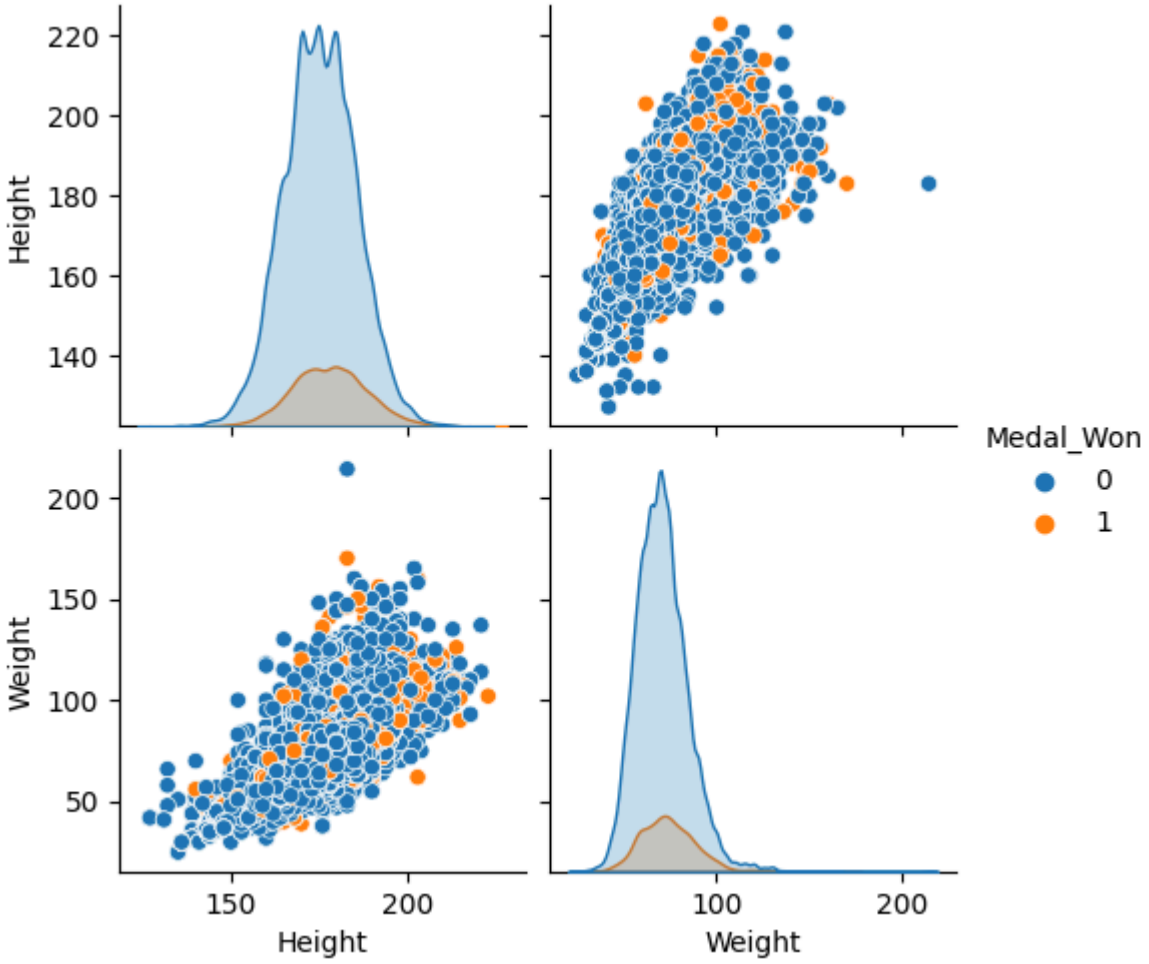
1. Hypothèses

Les hypothèses que nous tenterons de vérifier sont les suivantes:

1. Le poids et la taille des athlètes ont un impact sur les performances des athlètes dans différents sports
2. Les performances passées des athlètes peuvent impacter les performances futures

2. Description des analyses

2.1. Corrélation entre le poids, la taille et les performances d'un athlète



Le graphique ci-dessus est un pairplot avec des distributions marginales, montrant la relation entre la taille et le poids des athlètes par rapport à leurs performances (gagner une médaille ou non, indiqué par Medal_Won où **1** signifie une médaille gagnée et **0** pas de médaille).

Les résultats de ce graphique nous montrent :

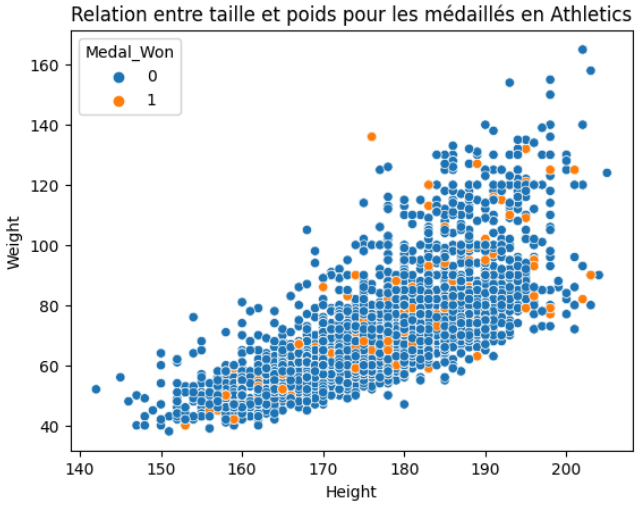
- Répartition de la taille et du poids : Les graphiques en haut à gauche et en bas à droite montrent la distribution de la taille et du poids des athlètes, respectivement. Les pics de ces distributions indiquent les valeurs les plus communes de taille et de poids.

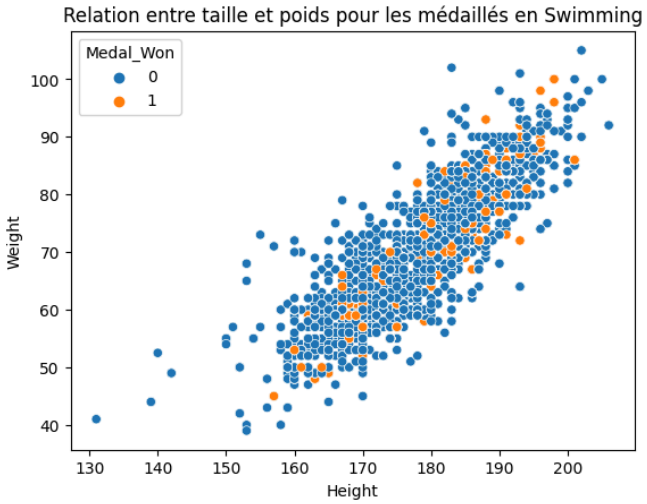
- Corrélation entre la taille et le poids : Le graphique en bas à gauche montre une corrélation positive entre la taille et le poids, ce qui est attendu car les personnes plus grandes ont tendance à peser plus.

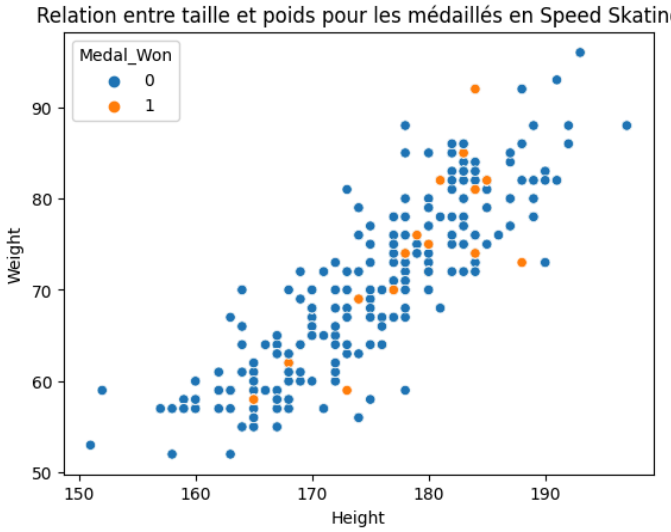
- Influence sur le gain de médailles : Les graphiques qui montrent des points colorés (orange pour Medal_Won égal à 1 et bleu pour 0) donnent une idée de la répartition de la taille et du poids des athlètes médaillés par rapport à ceux qui n'ont pas gagné de médaille.

Le fait de ne pas voir de séparation nette entre les points orange et bleu dans les graphiques de dispersion suggère qu'il n'y a pas de différence distincte dans la taille et le poids entre ceux qui ont gagné des médailles et ceux qui n'en ont pas gagné. Cela pourrait signifier que d'autres facteurs que la taille et le poids sont de meilleurs prédicteurs du gain de médailles, ou que l'influence de la taille et du poids sur le gain de médailles dépend du sport spécifique et ne peut être généralisée pour tous les sports.

2.2 Corrélation poids,taille et médailles par sport





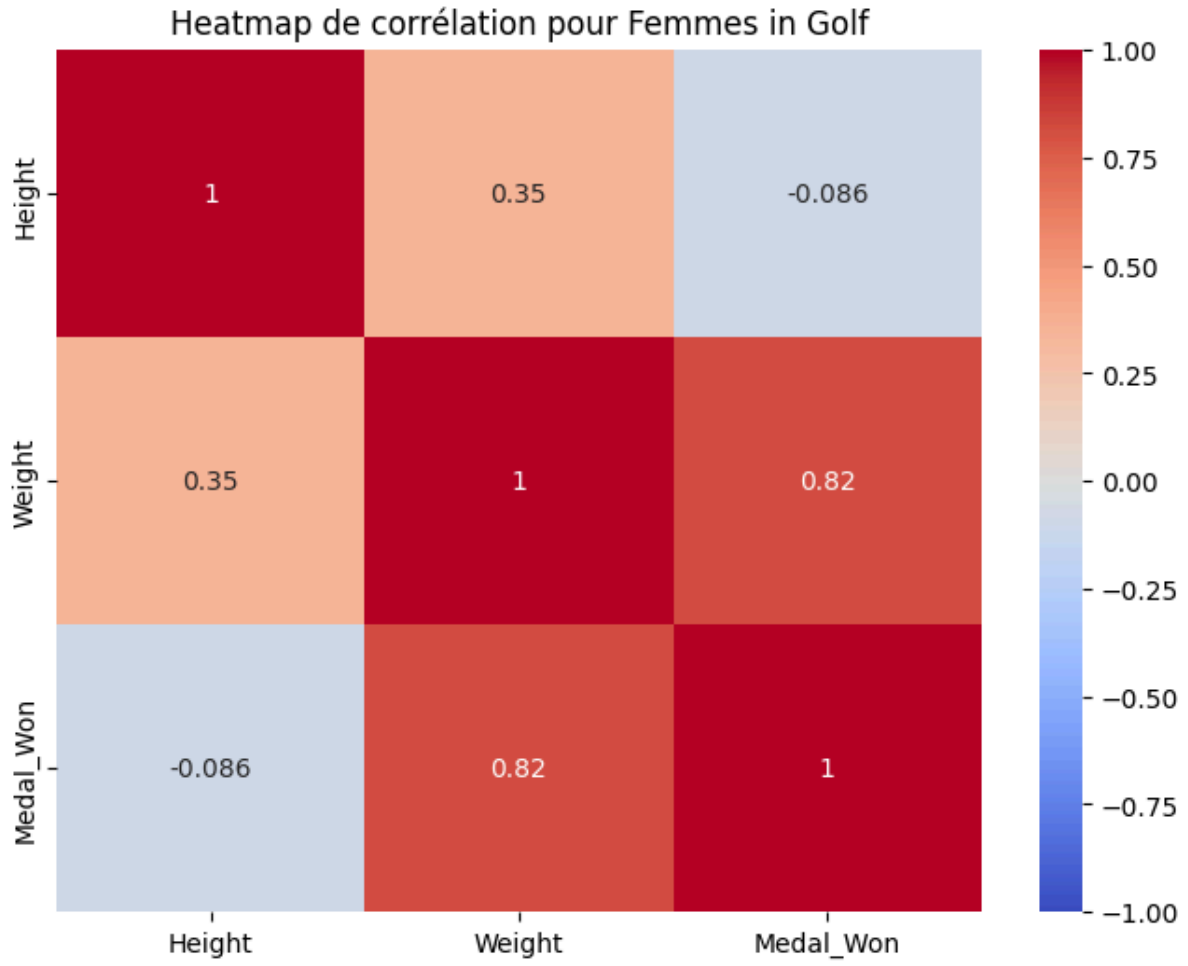


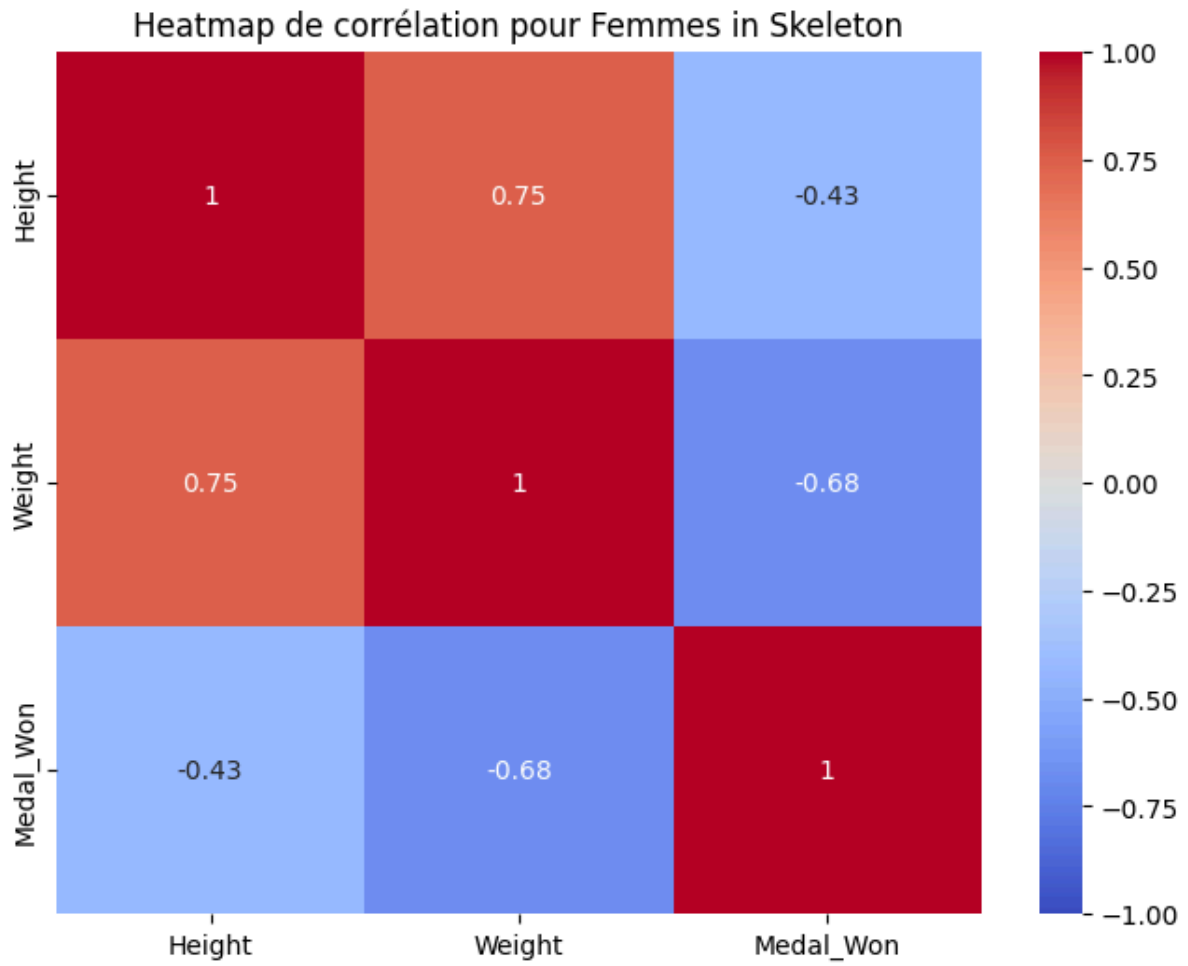
Les graphiques ci-dessous ne sont que des exemples des relations entre le poids et la taille pour les différents sports.

Pour chaque sport étudié, aucune corrélation dépensant le seuil de 50% n'ont été trouvé pour entre la taille, le poids et les médailles remportées. Les caractéristiques physiques du poids et de la taille ne sont pas des prédicteurs dominants de succès dans les différents sports en ce qui concerne le gain de médailles. Cela met en évidence la complexité de la performance sportive et le besoin de prendre en compte plusieurs variables lors de l'analyse des facteurs de succès.

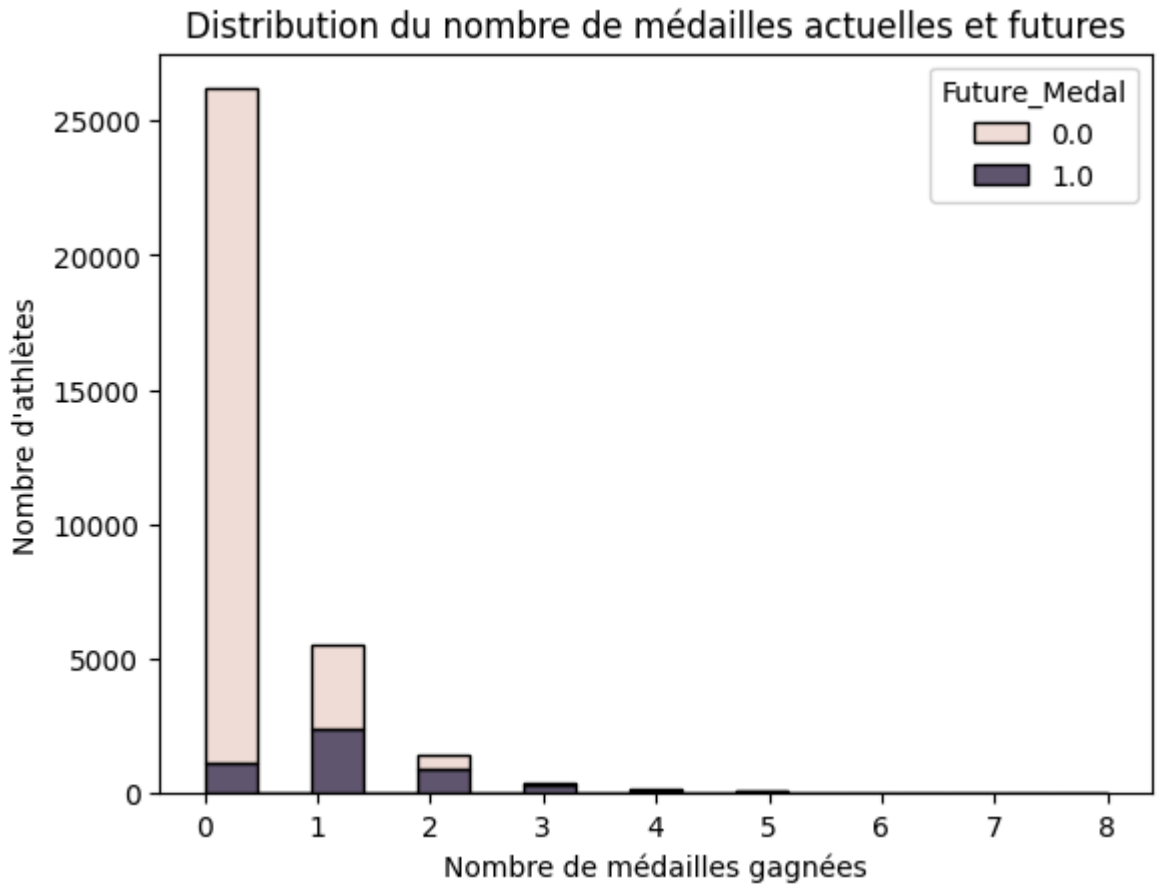
Séparation du jeux des données par sexe

Si on le sépare le jeux des données par sexe on constate que pour les femmes il y a une corrélation élevée entre le poids, la taille pour le sport comme le golf et le skeleton .Cependant ce résultat est fiable car il n'y a pas assez des données sur ces sports pour évaluer la corrélation(il n'y a qu'une médaille gagnée dans le sport de Golf chez les femmes). Le graphe ci- dessous montre cette corrélation:





2.3. Influence des performances passées des athlètes sur les performances futures



Le graphique ci-dessus est est un histogramme empilé, deux catégories sont affichées : les athlètes qui n'ont pas gagné de médailles lors de futurs jeux (Future_Medal = 0.0) et ceux qui en ont gagné (Future_Medal = 1.0), répartis en fonction du nombre de médailles qu'ils ont gagnées précédemment.

Si l'on analyse le graphique :

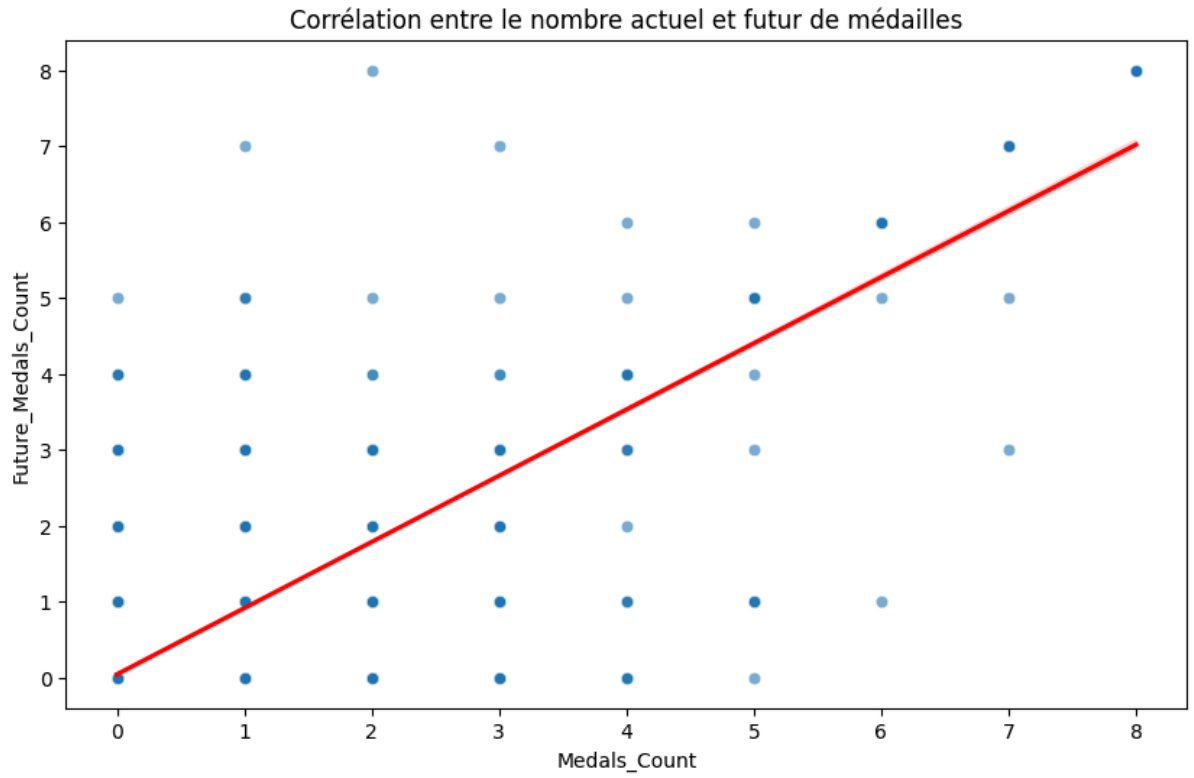
Une grande majorité des athlètes n'ont gagné aucune médaille lors de la compétition actuelle (barre pour 0 médailles gagnées).
Parmi ces athlètes, un petit pourcentage a gagné des médailles lors des jeux suivants (partie supérieure plus foncée de la barre pour 0 médailles gagnées).
À mesure que le nombre de médailles gagnées lors de la compétition actuelle augmente, le nombre total d'athlètes diminue. Cela est attendu car il est plus difficile de gagner de multiples médailles.

Il semble que les athlètes qui ont gagné des médailles lors des jeux actuels ont une plus grande section de barres plus foncées dans les colonnes suivantes (1, 2, etc.), ce qui indique une plus grande proportion de ces athlètes gagnant des médailles lors des futurs jeux par rapport à ceux qui n'ont pas gagné de médailles précédemment.

Cependant, il faut être prudent avec l'interprétation, surtout si les nombres absolus d'athlètes gagnant des médailles lors des jeux actuels sont faibles pour les nombres plus élevés de

médailles gagnées (comme 3, 4, etc.), ce qui pourrait mener à des conclusions imprécises en raison de la petite taille de l'échantillon.

Pour résumer, ce graphique suggère qu'il pourrait y avoir une tendance selon laquelle gagner des médailles lors d'une olympiade actuelle peut être lié à la probabilité de gagner des médailles lors des jeux futurs, et cette probabilité semble augmenter pour ceux qui ont déjà remporté des médailles. Mais pour des conclusions définitives, une analyse statistique plus approfondie serait nécessaire, y compris la prise en compte de la taille de l'échantillon pour chaque catégorie de nombre de médailles gagnées.



Le nuage des points ci-dessus confirme ce qui précède

Axe horizontal (x) - Medals_Count : Le nombre de médailles gagnées par les athlètes dans les jeux actuels.
Axe vertical (y) - Future_Medals_Count : Le nombre de médailles que ces athlètes ont gagnées dans les jeux suivants.

Points bleus : Chaque point représente le nombre d'athlètes (pas nécessairement la somme exacte) qui ont gagné un nombre spécifique de médailles lors des jeux actuels et futurs.
Ligne de tendance rouge : Elle montre qu'il y a une tendance générale où les athlètes qui gagnent plus de médailles dans les jeux actuels ont tendance à gagner plus de médailles dans les futurs jeux.

En analysant le graphique :

- Distribution des Points : La plupart des points semblent être concentrés vers la partie inférieure gauche du graphique, indiquant que la majorité des athlètes gagne peu ou pas de médailles, à la fois actuellement et à l'avenir.

- Tendance : La ligne de tendance ascendante indique qu'il existe une relation positive entre le nombre de médailles gagnées actuellement et à l'avenir - c'est-à-dire que, en général, gagner des médailles maintenant augmente les chances d'en gagner plus tard.

3. Présentation des résultats

Partie 1: Analyse de la Corrélation entre Taille, Poids et Gains de Médailles

Objectif : Examiner l'influence de la taille et du poids des athlètes sur leur capacité à gagner des médailles aux Jeux Olympiques.

Méthodologie : Analyse de corrélation et modélisation prédictive à l'aide de régression logistique pour déterminer la force de la relation entre les variables physiques (taille et poids) et le gain de médailles.

Résultats

- Analyse de Corrélation

Nous avons trouvé que, en général, il n'y a pas de corrélation très élevée entre la taille/le poids et le gain de médailles pour l'ensemble des athlètes.

Des analyses par sexe et par sport ont révélé que, pour certains sports spécifiques, il peut exister des relations plus significatives.

- Modélisation Prédictive

Pour les groupes où une corrélation significative a été observée, des modèles de régression logistique ont été construits (Veuillez voir la construction en détails des modèles construits dans le NoteBook).

Les modèles ont révélé que, même là où des corrélations ont été identifiées, la taille et le poids ne sont pas des prédicteurs extrêmement forts du gain de médailles. Ceci suggère que, bien que ces facteurs physiques puissent jouer un rôle, ils sont loin d'être les seuls déterminants du succès olympique. La taille et le poids peuvent influencer les performances dans certains sports, mais ils ne devraient pas être utilisés comme uniques indicateurs de potentiel olympique.

Partie 2: Influence des Performances Passées sur les Performances Futures

Objectif : Analyser si le gain de médailles lors des Jeux précédents peut prédire le succès dans les éditions futures des Jeux Olympiques.

Méthodologie : Évaluation statistique et visuelle de la continuité du succès olympique en comparant le nombre de médailles gagnées lors des jeux actuels et futurs.

Résultats

- Analyse Statistique

Une corrélation positive a été observée entre le nombre de médailles gagnées lors des Jeux actuels et futurs, suggérant un effet de "momentum".

Cet effet semble augmenter avec le nombre de médailles gagnées; cependant, l'effet diminue après un certain seuil.

- Visualisation

Les nuages de points avec lignes de régression montrent une tendance ascendante, indiquant que les athlètes ayant remporté des médailles ont une probabilité plus élevée de succès futur.

L'effet est plus marqué chez les athlètes ayant gagné plusieurs médailles.Les succès passés sont un indicateur potentiel des succès futurs, mais le lien n'est pas absolu et peut être affecté par de nombreux facteurs dynamiques.

Conclusion générale

L'analyse globale des données olympiques a révélé des insights précieux sur la dynamique de la performance athlétique. Notre exploration a d'abord mis en lumière des modèles de participation et de succès variés selon le sexe et la discipline, indiquant des tendances spécifiques à chaque sport. Ensuite, en examinant les attributs physiques des athlètes, nous avons constaté des corrélations avec le succès sportif, bien que ces dernières ne soient pas suffisamment robustes pour prédire les médailles sur la base unique de la taille et du poids. Par ailleurs, notre investigation prédictive a montré que les athlètes ayant précédemment remporté des médailles ont plus de chances de réussir dans les futurs jeux, un effet renforcé avec l'augmentation du nombre de médailles acquises. Ce phénomène souligne le rôle de l'expérience et du succès antérieur comme des indices de performances futures, dépassant la simple analyse des caractéristiques physiques pour comprendre le succès dans le sport de haut niveau.