
Cadre Logiciel pour Big Data

Analyse du dataset d'histoire olympique



Master 1 informatique parcours Big Data

Binôme : *BERKANI Yacine & RAMDANI Chaimae*

Enseignante : *JAZIRI Rakia*

6 janvier 2024

Contents

1	Introduction	2
2	Objectifs	2
3	Architecture distribuée	2
3.1	Cluster Hadoop sur GCP avec Dataproc	2
3.2	Configuration du Cluster	3
3.2.1	Description	3
3.2.2	Choix de Configuration	3
3.2.3	Étapes de Création du Cluster	3
4	Analyse de données avec HIVE	3
4.1	Préparation des Données	3
4.2	Création de la Base de Données et de la Table dans Hive	4
4.3	Analyse des Données	4
4.4	La Moyenne du Poids et de la Taille par Sexe	4
4.5	La distribution du poids et de la taille par sexe	4
4.6	Détermination des 10 meilleurs pays en termes de médailles	5
4.7	Analyse du total de médailles par sexe	5
4.8	Classement des pays par performance masculine et féminine	6
4.9	Les 10 sports les plus présents aux Jeux Olympique	6
4.10	Nombre de médailles remportées par sports	6
4.11	Nombre de participants par année	7
4.12	Nombre de médailles remportées par année	7
5	Visualisation des résultats	7
5.1	la distribution des caractéristiques physiques	7
5.2	TOP 10 des pays ayant remporté le plus de médailles	8
5.3	Classement masculin et féminin	8
5.4	Classement par sport	8
5.5	Évolution du nombre de participants	9
5.6	Évolution temporelle du nombre total de médailles	9
6	Conclusion	10

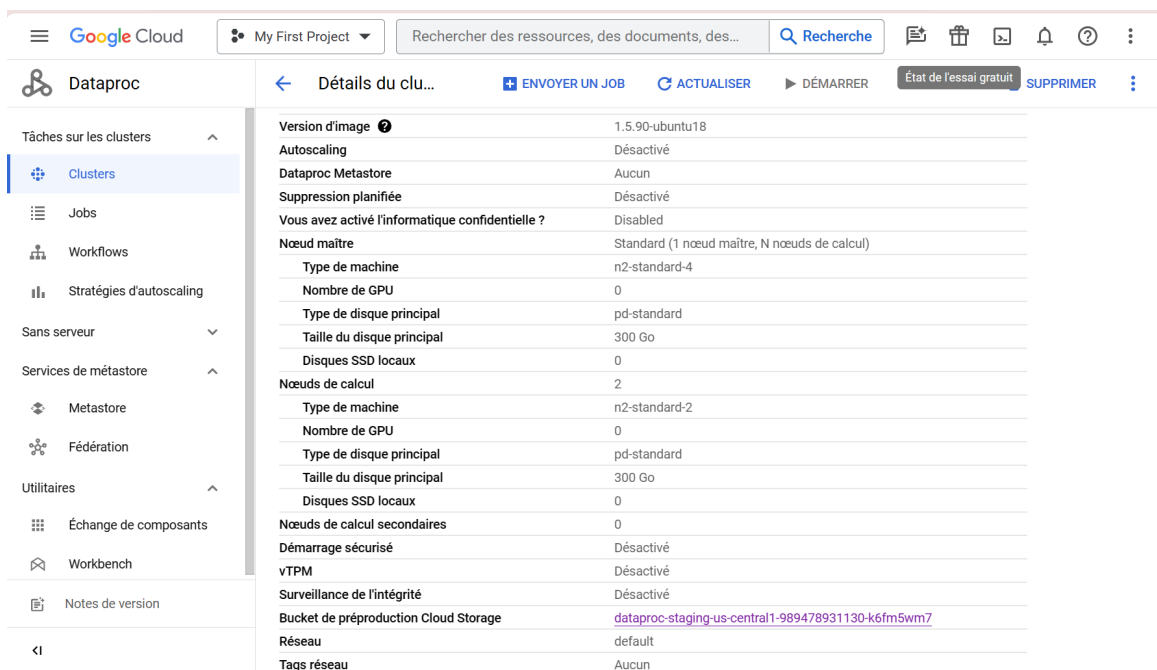
1 Introduction

Notre projet "Analyse du dataset d'histoire olympique" implique l'analyse des données sur 120 ans d'histoire olympique. Les données ont été préparées et stockées dans HDFS, puis analysées pour déterminer les 10 meilleurs pays en termes de médailles (or, argent, bronze). Des analyses spécifiques ont été réalisées sur la moyenne et la distribution du poids par sexe, ainsi que le nombre de médailles remportées par sexe et par année. Des visualisations ont été créées pour chaque analyse, mettant en évidence les tendances et les performances des pays et des athlètes dans les Jeux olympiques.

2 Objectifs

Dans ce projet, nous nous concentrons sur une analyse de données, avec un objectif spécifique : mettre en pratique nos connaissances sur la plateforme Hadoop. En choisissant ce sujet, nous visons à exploiter les concepts et les technologies de Hadoop pour traiter et Analyser du dataset d'histoire olympique.

3 Architecture distribuée



The screenshot shows the Google Cloud Dataproc console interface. The left sidebar contains navigation links for 'Tâches sur les clusters', 'Clusters' (selected), 'Jobs', 'Workflows', 'Stratégies d'autoscaling', 'Sans serveur', 'Services de métastore', 'Utilitaires', and 'Notes de version'. The main panel displays the 'Détails du clu...' for a specific cluster. At the top of the main panel, there are buttons for 'ENVOYER UN JOB', 'ACTUALISER', 'DÉMARRER', 'État de l'essai gratuit', and 'SUPPRIMER'. The cluster details are organized into sections: 'Version d'image' (1.5.90-ubuntu18), 'Autoscaling' (Désactivé), 'Dataproc Metastore' (Aucun), 'Suppression planifiée' (Désactivé), and 'Vous avez activé l'informatique confidentielle ?' (Disabled). The 'Nœud maître' section lists 'Type de machine' (n2-standard-4), 'Nombre de GPU' (0), 'Type de disque principal' (pd-standard), 'Taille du disque principal' (300 Go), and 'Disques SSD locaux' (0). The 'Nœuds de calcul' section lists 'Type de machine' (n2-standard-2), 'Nombre de GPU' (0), 'Type de disque principal' (pd-standard), 'Taille du disque principal' (300 Go), and 'Disques SSD locaux' (0). The 'Nœuds de calcul secondaires' section lists 'Type de machine' (n2-standard-2), 'Nombre de GPU' (0), 'Type de disque principal' (pd-standard), 'Taille du disque principal' (300 Go), and 'Disques SSD locaux' (0). The 'Démarrage sécurisé' section lists 'vTPM' (Désactivé) and 'Surveillance de l'intégrité' (Désactivé). The 'Bucket de préproduction Cloud Storage' is listed as 'dataproc-staging-us-central1-989478931130-k6fm5wm7'. The 'Réseau' section lists 'Réseau' (default) and 'Tags réseau' (Aucun).

Version d'image	1.5.90-ubuntu18
Autoscaling	Désactivé
Dataproc Metastore	Aucun
Suppression planifiée	Désactivé
Vous avez activé l'informatique confidentielle ?	Disabled
Nœud maître	Standard (1 nœud maître, N nœuds de calcul)
Type de machine	n2-standard-4
Nombre de GPU	0
Type de disque principal	pd-standard
Taille du disque principal	300 Go
Disques SSD locaux	0
Nœuds de calcul	2
Type de machine	n2-standard-2
Nombre de GPU	0
Type de disque principal	pd-standard
Taille du disque principal	300 Go
Disques SSD locaux	0
Nœuds de calcul secondaires	0
Démarrage sécurisé	Désactivé
vTPM	Désactivé
Surveillance de l'intégrité	Désactivé
Bucket de préproduction Cloud Storage	dataproc-staging-us-central1-989478931130-k6fm5wm7
Réseau	default
Tags réseau	Aucun

3.1 Cluster Hadoop sur GCP avec Dataproc

Dans le cadre de notre projet, nous avons établi une architecture distribuée en créant un cluster Hadoop sur Google Cloud Platform (GCP) via Dataproc. Cette démarche visait à mettre en place un environnement idéal pour le traitement de données massives. Nous avons bénéficié d'un essai gratuit de 300 \$ sur GCP, couvrant ainsi les coûts liés au projet.

3.2 Configuration du Cluster

3.2.1 Description

Notre cluster Hadoop se compose d'une machine maître dotée de 2 cœurs, 16 Go de RAM, et 300 Go de stockage, et de deux machines esclaves, chacune munie de 1 cœur, 8 Go de RAM, et 300 Go de stockage.

3.2.2 Choix de Configuration

Cette configuration a été choisie pour équilibrer les capacités de calcul et de stockage. La machine maître gère la coordination des tâches, tandis que les machines esclaves se concentrent sur l'exécution parallèle des processus.

3.2.3 Étapes de Création du Cluster

- **Création d'un Compte GCP:** Ouverture d'un compte GCP et activation de l'essai gratuit.
- **Accès à la Console GCP :** Navigation dans la console GCP vers Dataproc.
- **Configuration des Paramètres :** Réglage des paramètres du cluster dans Dataproc.
- **Validation et Lancement :** Vérification des configurations et lancement du cluster.
- **Suivi du Statut :** Surveillance du statut du cluster via la console GCP.

Le cluster Hadoop sur GCP avec Dataproc a fourni une plateforme robuste et adaptable pour le traitement des données distribuées, répondant aux besoins spécifiques de notre projet. La gestion du cluster s'est faite de manière efficace grâce à la console GCP.

4 Analyse de données avec HIVE

Nous allons effectuer une exploration de données statistiques.

Nous utiliserons Hive pour l'analyse de données et la bibliothèque Pandas et Seaborn pour la visualisation de données.

Seaborn utilise la bibliothèque Matplotlib. Sauf que Seaborn configure les graphiques avec des valeurs de style par défaut qui les rendent beaucoup plus beaux visuellement

Pour cette analyse, on utilise le dataset "120 ans d'histoire olympique: athlètes et résultats" que vous pouvez télécharger et lire la description sur le lien ci-dessous

https://www.kaggle.com/datasets/heesoo37/120-years-of-olympic-history-athletes-and-results#athlete_events.csv.

4.1 Préparation des Données

Après avoir téléchargé le jeu de données, nous l'avons déposé dans le Cloud GCP. Ensuite, nous avons créé un répertoire dans HDFS en utilisant la commande **hdfs dfs -mkdir /app**. Puis, nous avons transféré le fichier CSV dans ce répertoire HDFS avec la commande **hdfs dfs -put athlete_events.csv /app/data.csv**.

4.2 Création de la Base de Données et de la Table dans Hive

Nous avons utilisé **PyHive** pour établir une connexion à Hive, puis créé une base de données avec la requête **CREATE DATABASE IF NOT EXISTS bigdata_athlete;** suivie de la création d'une table pour stocker les données olympiques.

```
CREATE TABLE IF NOT EXISTS bigdata_athlete.tab3_athlete (  
    ID BIGINT,  
    Name STRING,  
    Sex CHAR(3),  
    Age INT,  
    Height INT,  
    Weight INT,  
    Team STRING,  
    NOC STRING,  
    Games STRING,  
    Year INT,  
    Season STRING,  
    City STRING,  
    Sport STRING,  
    Event STRING,  
    Medal STRING  
)  
ROW FORMAT DELIMITED  
FIELDS TERMINATED BY ','  
STORED AS TEXTFILE  
LOCATION '/app/master_bigdata/'  
TBLPROPERTIES ("skip.header.line.count"="1");
```

4.3 Analyse des Données

4.4 La Moyenne du Poids et de la Taille par Sexe

Nous avons réalisé une analyse de la moyenne du poids et de la taille par sexe, en incluant également la fréquence d'occurrence de chaque sexe.

<pre># Requête Hive hive_query = f""" SELECT Sex, AVG(Height) AS AverageHeight FROM (SELECT Sex, Height FROM {table_name} WHERE Sex IN ('F', 'M') AND Height IS NOT NULL) subquery GROUP BY Sex ORDER BY Sex """</pre>			<pre># Requête Hive hive_query = f""" SELECT Sex, AVG(Weight) AS AverageWeight, COUNT(*) AS Frequency FROM (SELECT Sex, Weight FROM {table_name} WHERE Sex IN ('F', 'M') AND Weight IS NOT NULL) subquery GROUP BY Sex ORDER BY Sex """</pre>		
sex			sex		
averageheight			averageweight		
0	F	167.845415	0	F	60.040751
1	M	178.835103	1	M	75.712669

4.5 La distribution du poids et de la taille par sexe

Nous avons analysé la distribution du poids par sexe, en prenant en compte la fréquence d'occurrence de chaque sexe.

```

SELECT
    Sex,
    Height,
    COUNT(*) as Frequency
FROM bigdata_athlete.tab3_athlete
WHERE Sex IN ('F', 'M') AND Height IS NOT NULL
GROUP BY Sex, Height
ORDER BY Sex, Height;

```

	sex	height	frequency
0	F	25	6
1	F	28	13
2	F	30	42
3	F	31	23
4	F	32	41

```

SELECT
    Sex,
    Weight,
    COUNT(*) as Frequency
FROM bigdata_athlete.tab3_athlete
WHERE Sex IN ('F', 'M') AND Weight IS NOT NULL
GROUP BY Sex, Weight
ORDER BY Sex, Weight;

```

	sex	height	frequency
0	F	127	6
1	F	131	2
2	F	132	6
3	F	133	5
4	F	135	12

4.6 Détermination des 10 meilleurs pays en termes de médailles

Nous avons identifié les 10 meilleurs pays en termes de médailles, en considérant toutes les catégories - or, argent, et bronze.

```

SELECT
  Team,
  COUNT(*) AS TotalMedals,
  SUM(CASE WHEN Medal = 'Gold' THEN 1 ELSE 0 END) AS Gold,
  SUM(CASE WHEN Medal = 'Silver' THEN 1 ELSE 0 END) AS Silver,
  SUM(CASE WHEN Medal = 'Bronze' THEN 1 ELSE 0 END) AS Bronze
FROM bigdata_athlete.tab3_athlete
WHERE Medal IN ('Gold', 'Silver', 'Bronze')
GROUP BY Team
ORDER BY TotalMedals DESC
LIMIT 10;

```

	team	totalmedals	gold	silver	bronze
0	United States	4190	2028	1221	941
1	Soviet Union	1963	813	584	566
2	Germany	1401	430	455	516
3	Great Britain	1245	417	390	438
4	Australia	1143	286	408	449
5	Canada	1117	401	364	352
6	Italy	966	301	311	354
7	Russia	921	293	297	331
8	Sweden	917	243	319	355
9	France	880	211	316	353

4.7 Analyse du total de médailles par sexe

Nous avons réalisé une analyse le nombre de médailles remportées par sexe en prenant en compte toutes les médailles (or, argent, bronze).

```
SELECT
    Sex,
    COUNT(*) AS TotalMedals,
    SUM(CASE WHEN Medal = 'Gold' THEN 1 ELSE 0 END) AS Gold,
    SUM(CASE WHEN Medal = 'Silver' THEN 1 ELSE 0 END) AS Silver,
    SUM(CASE WHEN Medal = 'Bronze' THEN 1 ELSE 0 END) AS Bronze
FROM bigdata_athlete.tab3_athlete
WHERE Medal IN ('Gold', 'Silver', 'Bronze')
GROUP BY Sex
ORDER BY TotalMedals DESC;
```

	sex	totalmedals	gold	silver	bronze
0	M	20989	6986	6909	7094

4.8 Classement des pays par performance masculine et féminine

Nous avons effectué une analyse dU Classement des 10 meilleurs pays masculine/féminins en termes de victoires médailles

```

SELECT
Team,
Sex,
COUNT(*) AS TotalMedals,
SUM(CASE WHEN Medal = 'Gold' THEN 1 ELSE 0 END) AS Gold,
SUM(CASE WHEN Medal = 'Silver' THEN 1 ELSE 0 END) AS Silver,
SUM(CASE WHEN Medal = 'Bronze' THEN 1 ELSE 0 END) AS Bronze
FROM bigdata_athlete.tab3_athlete
WHERE Medal IN ('Gold', 'Silver', 'Bronze') AND Sex IN ('M')
GROUP BY Team,Sex
ORDER BY TotalMedals DESC
LIMIT 10;

```

	team	sex	totalmedals	gold	silver	bronze
0	United States	M	2626	1267	750	609
1	Soviet Union	M	1386	563	432	391
2	Great Britain	M	939	337	304	298
3	Germany	M	888	283	298	307
4	Italy	M	818	267	253	298
5	Sweden	M	741	206	241	294
6	France	M	708	173	246	289
7	Canada	M	646	256	228	162
8	Australia	M	631	120	239	272
9	Japan	M	519	144	166	209

```

SELECT
Team,
Sex,
COUNT(*) AS TotalMedals,
SUM(CASE WHEN Medal = 'Gold' THEN 1 ELSE 0 END) AS Gold,
SUM(CASE WHEN Medal = 'Silver' THEN 1 ELSE 0 END) AS Silver,
SUM(CASE WHEN Medal = 'Bronze' THEN 1 ELSE 0 END) AS Bronze
FROM bigdata_athlete.tab3_athlete
WHERE Medal IN ('Gold', 'Silver', 'Bronze') AND Sex IN ('F')
GROUP BY Team,Sex
ORDER BY TotalMedals DESC
LIMIT 10;

```

	team	sex	totalmedals	gold	silver	bronze
0	United States	F	1564	761	471	332
1	Soviet Union	F	577	250	152	175
2	China	F	549	167	219	163
3	Germany	F	513	147	157	209
4	Australia	F	512	166	169	177
5	Russia	F	476	179	161	136
6	Canada	F	471	145	136	190
7	Netherlands	F	378	124	124	130
8	East Germany	F	365	156	115	94
9	Great Britain	F	306	80	86	140

4.9 Les 10 sports les plus présents aux Jeux Olympique

Nous avons identifié les 10 sports les plus présents aux Jeux Olympique

```

SELECT
  Sport,
  COUNT(DISTINCT ID) AS NombreParticipants
FROM
  bigdata_athlete.tab3_athlete
GROUP BY Sport
ORDER BY NombreParticipants DESC
LIMIT 10;

```

	sport	nombreparticipants
0	Athletics	21790
1	Swimming	8644
2	Rowing	7558
3	Football	6146
4	Cycling	5800
5	Boxing	5231
6	Wrestling	4956
7	Shooting	4828
8	Sailing	4396
9	Gymnastics	4093

4.10 Nombre de médailles remportées par sports

Nous avons réalisé une analyse statistique des médailles remportées dans chaque sport, incluant le nombre total de médailles ainsi que le décompte des médailles d'or, d'argent et de bronze.

```

SELECT
    Sport,
    COUNT(*) AS TotalMedals,
    SUM(CASE WHEN Medal = 'Gold' THEN 1 ELSE 0 END) AS Gold,
    SUM(CASE WHEN Medal = 'Silver' THEN 1 ELSE 0 END) AS Silver,
    SUM(CASE WHEN Medal = 'Bronze' THEN 1 ELSE 0 END) AS Bronze
FROM bigdata_athlete.tab3_athlete
WHERE Medal IN ('Gold', 'Silver', 'Bronze')
GROUP BY Sport
ORDER BY TotalMedals DESC
LIMIT 10;

```

	sport	totalmedals	gold	silver	bronze
0	Athletics	3493	1149	1178	1166
1	Rowing	2828	923	944	961
2	Swimming	2785	973	918	894
3	Gymnastics	1915	659	629	627
4	Football	1566	513	511	542
5	Hockey	1518	512	498	508
6	Ice Hockey	1504	505	501	498
7	Sailing	1194	430	405	359
8	Basketball	1058	350	359	349
9	Handball	1057	348	357	352

4.11 Nombre de participants par année

On s'intéresse à présent au nombre de participation pour chaque année aux Jeux olympiques

```
SELECT
  Year,
  COUNT(DISTINCT ID) AS NombreParticipants
FROM
  bigdata_athlete.tab3_athlete
WHERE Year IS NOT NULL AND ID IS NOT NULL
GROUP BY Year
ORDER BY Year DESC, NombreParticipants DESC;
```

	year	nombreparticipants
0	2016	11163
1	2014	2741
2	2012	10490
3	2010	2528
4	2008	10869

4.12 Nombre de médailles remportées par année

Nous avons effectué une analyse de nombre de médailles remportées par année aux Jeux olympiques

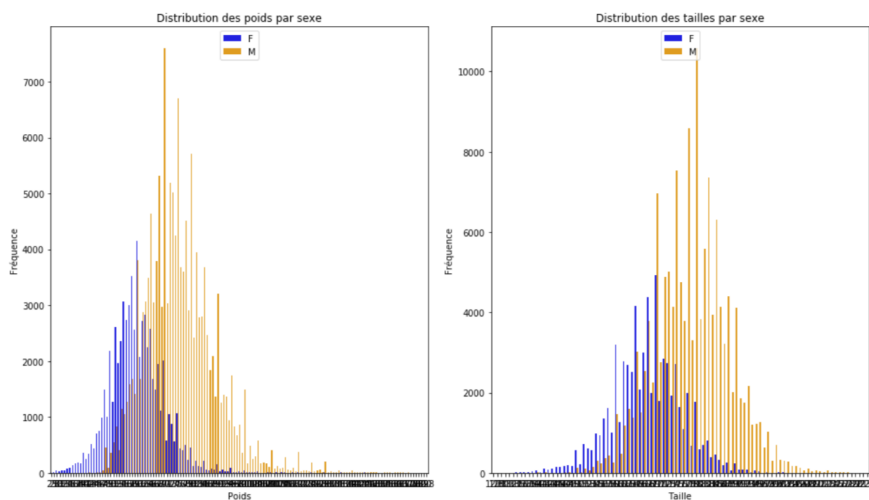
```
SELECT
  Year,
  COUNT(*) AS TotalMedals,
  SUM(CASE WHEN Medal = 'Gold' THEN 1 ELSE 0 END) AS Gold,
  SUM(CASE WHEN Medal = 'Silver' THEN 1 ELSE 0 END) AS Silver,
  SUM(CASE WHEN Medal = 'Bronze' THEN 1 ELSE 0 END) AS Bronze
FROM bigdata_athlete.tab3_athlete
WHERE Medal IN ('Gold', 'Silver', 'Bronze')
GROUP BY Year
ORDER BY Year DESC, TotalMedals DESC;
```

	year	totalmedals	gold	silver	bronze
0	2016	1686	557	549	580
1	2014	556	188	184	184
2	2012	1598	521	521	556
3	2010	476	160	158	158
4	2008	1694	553	555	586

5 Visualisation des résultats

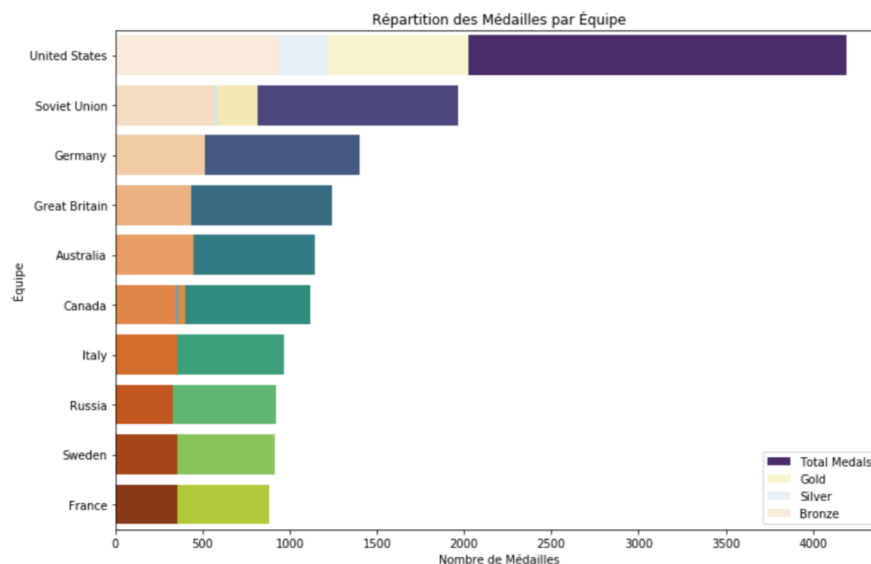
5.1 la distribution des caractéristiques physiques

Pour cette visualisation, il est clairement observable que le poids et la taille des hommes sont généralement plus élevés, suivis de près par ceux des femmes.



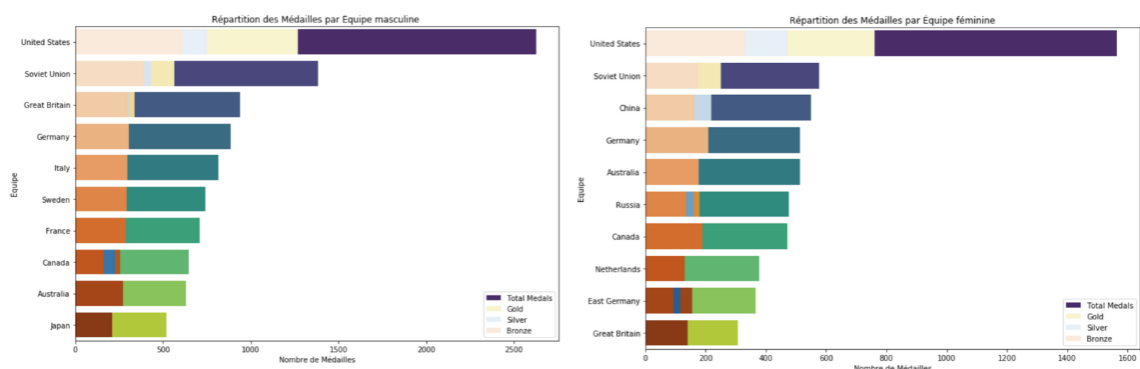
5.2 TOP 10 des pays ayant remporté le plus de médailles

La visualisation des 10 meilleurs pays en termes de vecteur des médailles, intégrant les médailles d'or, d'argent et de bronze, offre un aperçu frappant de la prédominance et de la constance de certaines nations dans l'histoire des Jeux Olympiques, mettant en évidence non seulement leur succès global, mais aussi leur compétitivité à travers différentes catégories de médailles.



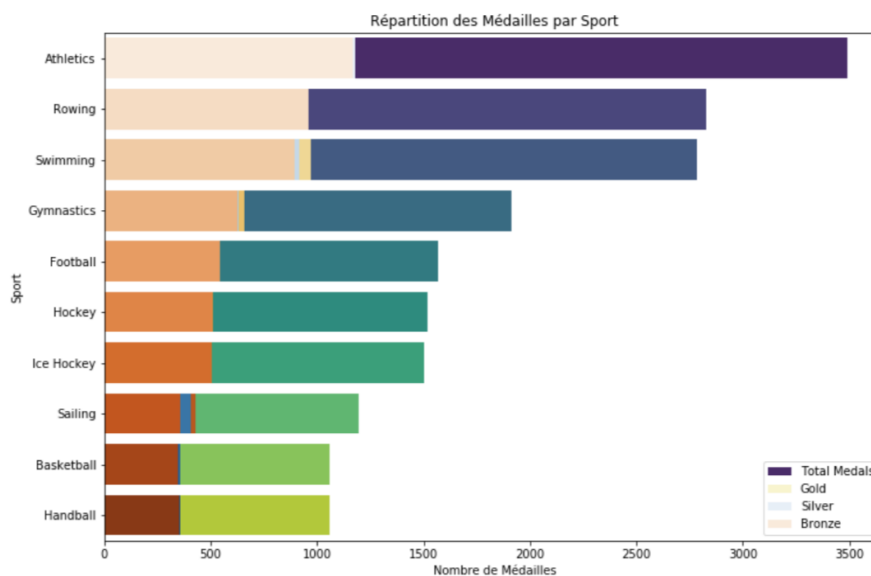
5.3 Classement masculin et féminin

Cette visualisation permet d'apprécier la répartition et l'équilibre des succès sportifs entre les équipes masculines et féminines de différents pays, offrant une perspective unique sur l'évolution et la diversité de la compétition olympique à travers les genres.



5.4 Classement par sport

la visualisation du nombre de médailles par sport révèle que l'athlétisme se distingue nettement comme le sport le plus récompensé aux Jeux Olympiques. Elle montre également une répartition variée des médailles dans les autres disciplines, indiquant une concurrence accrue et une diversité de talents à travers les différentes catégories sportives.



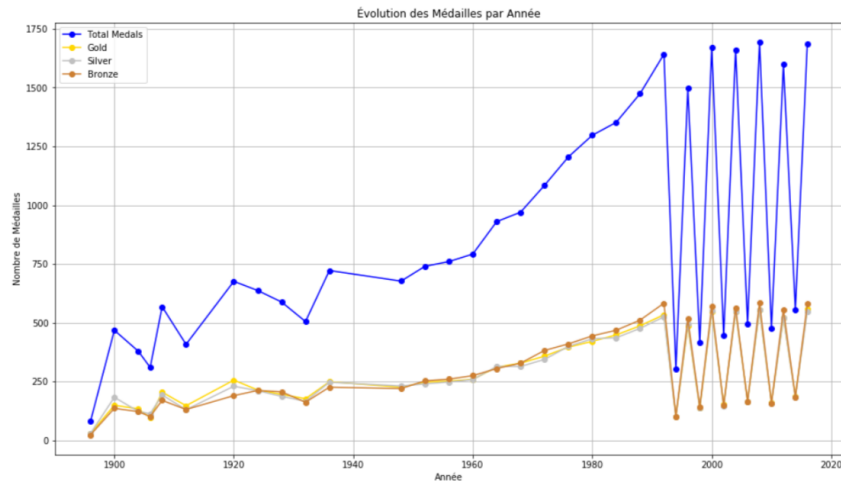
5.5 Évolution du nombre de participants

cette visualisation de l'évolution du nombre de participants aux Jeux Olympiques par année montre une tendance globale à la hausse au fil du temps.



5.6 Évolution temporelle du nombre total de médailles

La visualisation en ligne avec Matplotlib illustre l'évolution du nombre total de médailles, ainsi que des médailles d'or, d'argent et de bronze, au fil des années lors des Jeux olympiques.



6 Conclusion

En conclusion, cette analyse approfondie des données olympiques sur 120 ans a mis en lumière des tendances significatives et des disparités entre les sexes, en termes de caractéristiques physiques des athlètes. Elle a également révélé la dominance de certaines nations dans les médailles olympiques, tout en soulignant la popularité de divers sports. Les visualisations ont permis de tracer l'évolution des participants et des médailles au fil des ans, offrant une perspective enrichissante sur l'histoire et la dynamique des Jeux olympiques.