

Université Alioune Diop de Bambey
UFR : Sciences Appliquées et Technologies de l'Information et
de la Communication



Département : Mathématiques

Master Statistique et Informatique

Décisionnelle

Projet Intégration de Données

Applications aux entrepôts de données

Présenté par Mr **Amadou BA**

Supervisé par le Professeur
Gaoussou CAMARA

Table des matières

Introduction	2
I. Description d'une page de l'activité que nous souhaitons analyser ainsi que l'apport de l'entrepôt de données en termes d'analyse décisionnelle pour cette activité	3
II. Analyse du jeu de données	4
III. Conception d'un modèle en étoile pour notre entrepôt de données avec Power AMC	5
IV. Utilisation de l'ETL Talend Open Studio pour alimentation de l'entrepôt de données	6
V. Visualisation des données sur Power BI.....	9
VI. Tableau de bord.....	13
Conclusion.....	15

Introduction

Depuis leur réinvention en 1896 par le baron Pierre de Coubertin, les Jeux olympiques ont transcendé les frontières nationales pour devenir l'un des événements les plus universellement célébrés, symbolisant l'excellence athlétique, la camaraderie internationale et l'esprit de compétition. Tous les quatre ans, des athlètes du monde entier se rassemblent pour participer à cette vitrine mondiale du sport, où le talent, la détermination et la passion se rencontrent pour inspirer des millions de personnes à travers le monde.

Dans ce contexte l'objectif du projet est d'approfondir l'approche d'intégration de données, d'entrepôt de données et l'usage de Talend Open Studio et de Tableau Desktop et pour cela, le jeu de données ***Summer_Olympic_medallists_1896-2008*** est mis à notre disposition. Il fournit les médaillés des Jeux olympiques de chaque été entre 1896 et 2008.

En combinant la puissance de l'informatique et de l'analyse statistique, nous cherchons à découvrir les histoires cachées derrière les médailles, à identifier les facteurs clés de succès et à éclairer les décisions futures dans le domaine du sport de haut niveau.

À travers ce projet, nous espérons apporter une contribution significative à la compréhension et à l'appréciation des Jeux olympiques en tant que phénomène mondial, en mettant en lumière les performances exceptionnelles des athlètes, les dynamiques géopolitiques et les tendances émergentes dans le monde du sport. Nous sommes convaincus que cette exploration approfondie des données olympiques nous permettra de découvrir de nouveaux insights et de nourrir une passion renouvelée pour cet événement emblématique qui unit le monde entier dans un esprit de compétition, de camaraderie et d'excellence.

I. Description d'une page de l'activité que nous souhaitons analyser ainsi que l'apport de l'entrepôt de données en termes d'analyse décisionnelle pour cette activité

Les Jeux olympiques, symbole d'excellence sportive et de compétition mondiale, ont une longue histoire qui remonte à 1896. Dans ce projet, nous explorerons le jeu de données "Summer_Olympic_medallists_1896-2008", qui répertorie les médaillés de chaque été entre 1896 et 2012.

Notre objectif est d'approfondir notre compréhension de l'intégration de données, des entrepôts de données et de l'utilisation d'outils tels que Talend Open Studio et Tableau Desktop.

L'activité que nous souhaitons analyser dans ce projet est la performance des pays et des athlètes aux Jeux olympiques d'été de 1896 à 2008 en termes de médailles. Cette analyse se concentrera sur plusieurs dimensions clés, notamment le nombre de médailles remportées par chaque pays, continent, la distribution des médailles par sexe, les tendances des performances par année, et les disciplines sportives les plus prolifiques en termes de médailles.

Le jeu de données "Summer_Olympic_medallists_1896-2008" fournira les bases nécessaires pour cette analyse, incluant des informations détaillées sur les médaillés, les pays d'origine, les événements, les médailles et le sexe des athlètes.

1. **Nombre total de médailles par type de médailles :** Nous étudierons la répartition des médailles d'or, d'argent et de bronze ;
2. **Nombre total de médailles par sexe :** Comparaison des médailles remportées par les athlètes masculins et féminins ;
3. **Nombre total de médailles par pays d'origine et par sexe :** Nous examinerons les performances des 10 des pays les plus médaillés en termes de médailles remportées et aussi en fonction du sexe des athlètes ;
4. **Nombre total de médailles par sexe et par continent :** Comparaison des médailles remportées par les athlètes masculins et féminins et aussi de leur continent d'origine ;
5. **Évolution au fil des années :** Comment le nombre de médailles a-t-il évolué au cours des décennies ?
6. **Types d'épreuves les plus médaillées :** Top 10 des événements les plus médaillés pour identifier les domaines les plus performants ;

Ce projet est une opportunité d'explorer des concepts clés de l'intégration de données et de créer des rapports visuels informatifs à l'aide de Power BI.

Suivons ensemble ce voyage passionnant !

II. Analyse du jeu de données

Notre jeu de données est un fichier Excel renseignant les informations sur les médaillés des jeux Olympiques de 1896 à 2008.

Le jeu de données est constitué de deux feuilles.

La première feuille renseigne sur les médaillés et contient vingt-neuf mille deux cents seize (29216) observations et dix (10) champs qui sont :

- 🚩 City : c'est la ville dans laquelle l'édition a été organisée ;
- 🚩 Edition : c'est l'année d'organisation ;
- 🚩 Sport : c'est le type de sport ;
- 🚩 Discipline : c'est la discipline dans le sport ;
- 🚩 Athlète : c'est le nom le prénom de l'athlète ;
- 🚩 NOC : c'est le code du pays d'origine de l'athlète ;
- 🚩 Gender : c'est le genre de l'athlète ;
- 🚩 Event : c'est le type de l'événement ;
- 🚩 Event_gender : c'est le genre de l'événement ;
- 🚩 Medal : c'est le type de médailles

La deuxième feuille contient deux cent un (201) observations sur les pays des athlètes et les codes des pays et quatre (4) champs qui sont :

- 🚩 Country : c'est le pays d'origine de l'athlète ;
- 🚩 Int Olympic Committee code : c'est le code du pays d'origine de l'athlète ;
- 🚩 ISO code : c'est le code ISO.

III. Conception d'un modèle en étoile pour notre entrepôt de données avec Power AMC

Pour la conception de l'entrepôt de données, nous allons utiliser le logiciel PowerAMC.

PowerAMC (Analysis and Modeling Corporation) est un outil de modélisation utilisé pour concevoir et documenter les systèmes d'information. Développé par Sybase, qui a été ensuite acquis par SAP, PowerAMC est principalement utilisé par les analystes de systèmes, les concepteurs de bases de données et les développeurs pour plusieurs types de modélisations.

On l'utilisera pour la conception de notre entrepôt de données.

On utilisera cet outil pour la conception de notre modèle conceptuel données et le modèle physique de données.



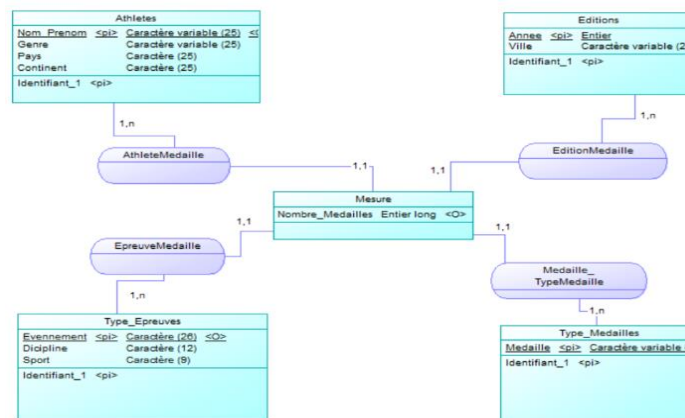
Un modèle en étoile est une approche de conception de base de données utilisée principalement dans les entrepôts de données pour organiser et structurer les données de manière à faciliter l'analyse et la génération de rapports. Il s'agit d'un schéma de base de données relationnelle simplifié qui se compose de deux types de tables : la table de faits et les tables de dimensions.

✚ Model conceptuel de données (MCD)

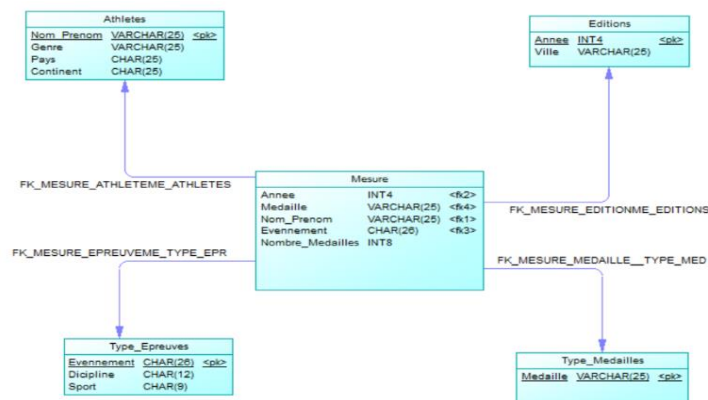
Le modèle conceptuel de données permet de créer des modèles conceptuels de données pour représenter la structure et les relations des données à un niveau abstrait. Il aide à comprendre et à documenter les besoins en données d'une organisation.

Voici notre modèle en étoile pour notre entrepôt de données avec **Power AMC**.

Il est constitué de quatre tables de dimensions qui sont : « Athlètes », « Editions », « Types_Epreuves » et « Type_Médailles », au tour d'une table de fait qui est la table « Mesure ».



✚ Model physique de données (MPD)



IV. Utilisation de l'ETL Talend Open Studio pour alimentation de l'entrepôt de données

Talend Open Studio est un logiciel open source développé par Talend pour l'intégration de données. Il est largement utilisé pour la gestion des données, le nettoyage, la transformation, et l'intégration de données provenant de diverses sources. Son utilisation pour alimenter un entrepôt de données est une pratique courante dans le domaine de la business intelligence et de la gestion des données.



Tout d'abord, on doit configurer des connexions à nos sources de données, telles que des bases de données relationnelles, des fichiers plats, des services web, etc.

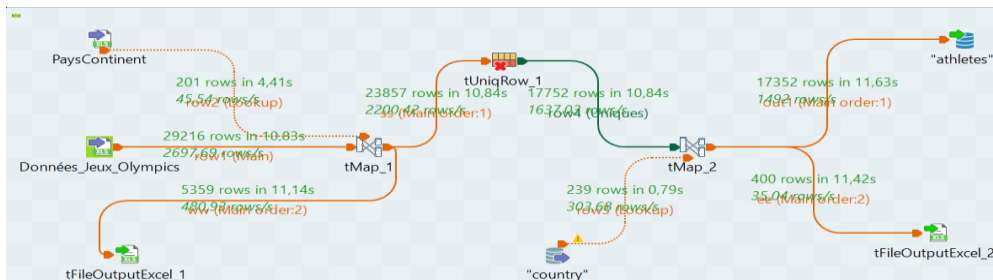
Ensuite, on conçoit un job dans Talend Open Studio pour extraire les données de nos sources, les transformer selon nos besoins et les charger dans notre entrepôt de données. On peut utiliser une variété de composants prédéfinis dans Talend pour accomplir ces tâches, tels que les composants de lecture (pour extraire les données), les composants de transformation (pour nettoyer et transformer les données), et les composants d'écriture (pour charger les données dans l'entrepôt).

Dans notre cas, nos données sont de types Excell.

Alimentation des tables de dimensions ✓ Table Athlètes

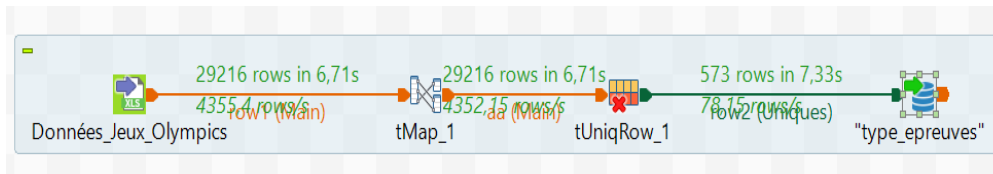
Pour alimenter la table « Athlètes », on prend ces deux fichiers (PaysContinent et Donnees_Jeux_Olympics) en entrée puis les connecter à travers **tMap** et faire une jointure entre le 'NOC' et le 'Int Olympic Committee code', ensuite on crée une sortie : la sortie contient le nom_prenom de l'athlètes, son genre, son pays d'origine, puis on la connecte avec un **tUniqRow** qui nous permettra de sélectionner de façon unique chaque nom_prenom d'un athlète puisque notre clés primaire dans cette table est le nom_prenom de l'athlète. Après cela, on joint cette même sortie à travers un tMap par une jointure entre les deux colonnes de pays qui ressemble avec un autre fichier qui contient les continents correspondant aux pays des athlètes et en fin on plate notre fichier hôte qui est la table à charger.

On a aussi créé deux fichiers de rejet de types Excel pour voir les Athlètes qui n'ont pas de correspondance exacte de pays : le tFileOutputExcel_1 et le tFileOutputExcel_2.



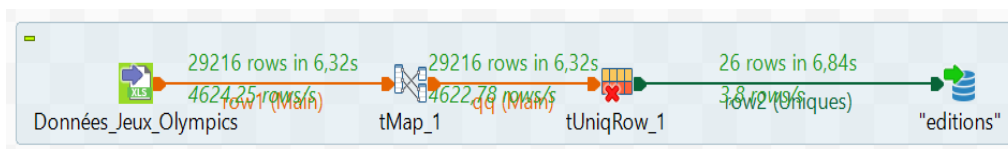
✓ Table Types_Epreuves

Pour alimenter la table « Types_Epreuves », on prend le fichier Données_Jeux_Olympics en entrée puis le connecter à un **tMap** pour sélectionner les colonnes qu'on a besoin pour le chargement, ensuite on place **tUniqRow** qu'on applique sur la colonne « event » qui est notre clé primaire dans cette table puis on la connecte avec la table à charger qui la table « **Types_Epreuves** ».



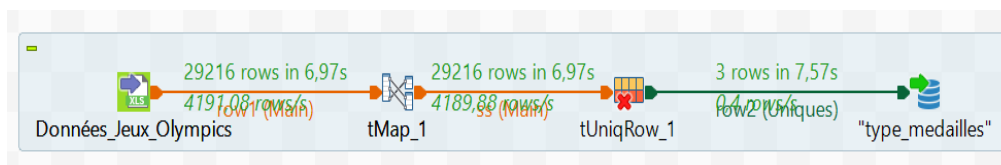
✓ Table Editions

Pour alimenter la table « Editions », on prend le fichier Données_Jeux_Olympics en entrée puis le connecter à un **tMap** pour sélectionner les colonnes qu'on a besoin pour le chargement, ensuite on place **tUniqRow** qu'on applique sur la colonne « edition » qui est notre clé primaire dans cette table puis on la connecte avec la table à charger qui la table « **Editions** ».



✓ Table Types Médailles

Pour alimenter la table « Type_Médailles », on prend le fichier Données_Jeux_Olympics en entrée puis le connecter à un **tMap** pour sélectionner les colonnes qu'on a besoin pour le chargement, ensuite on place **tUniqRow** qu'on applique sur la colonne « medal » qui est notre clé primaire dans cette table puis on la connecte avec la table à charger qui la table « **Type_Médailles** ».



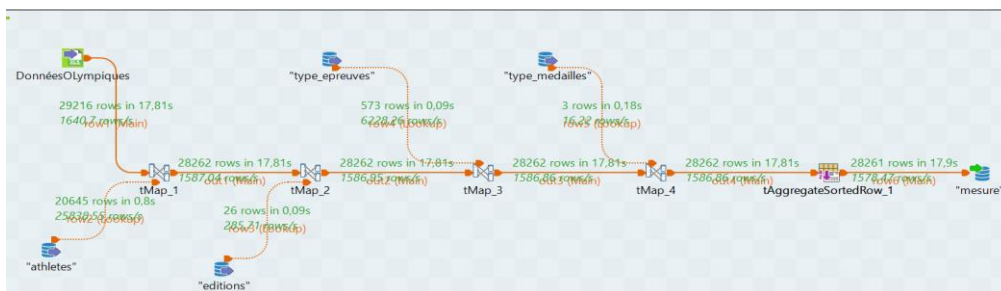
Alimentation de la table de fait

Mesure

Pour l'alimentation de la table de fait « Mesure », il faut impliquer toutes les tables de dimensions car ces clés lui viennent des tables de dimension qui étaient les identifiants primaires vont migrer dans la table de fait et constitueront non seulement des clés étrangères mais aussi des clés primaires dans cette table.

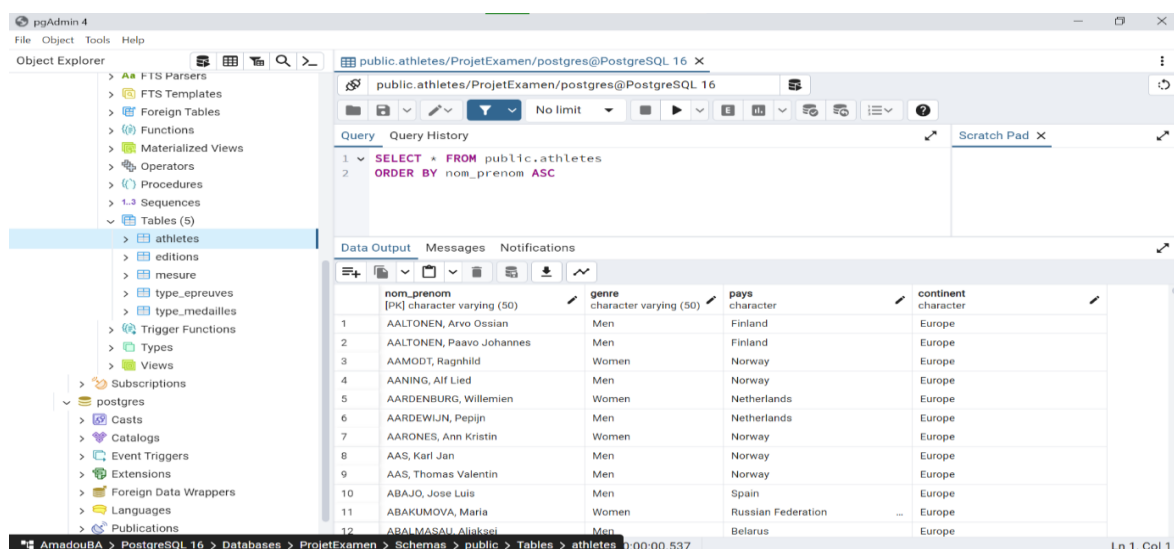
Pour le faire, on prend le fichier `Donnees_Jeux_Olympiques` en entrée puis le connecter avec la table athlètes grâce à un **tMap** et sélectionner les colonnes qu'on a besoin pour le chargement, ensuite on crée une sortie qu'on va ensuite connecter avec la table éditions par un **tMap**, faire la jointure des deux entrées et créer une autre sortie qu'on va ensuite connecter dans un autre **tMap** avec la table `Types_Epreuves` et on crée une autre sortie qu'on va connecter un dernier **tMap** avec la table `Type_Médailles` et créer une sortie qu'on va connecter avec un **tAggregateSortedRow** qui nous permet en même temps de trier les données et de faire des agrégations. Dans notre on compte le nombre de médaille.

Dans cette table, chaque ligne correspond à un athlète qui a joué un événement unique, à une édition unique et à un type de médaille unique.



Exemple de vue d'une table après chargement dans PostgreSQL

Voici un exemple de chargement d'une table après une bonne exécution d'un job dans Talend open Studio.



	nom_prenom	genre	pays	continent
1	AALTONEN, Arvo Ossian	Men	Finland	Europe
2	AALTONEN, Paavo Johannes	Men	Finland	Europe
3	AAMODT, Ragnhild	Women	Norway	Europe
4	AANING, Alf Lied	Men	Norway	Europe
5	AARDENBURG, Willemien	Women	Netherlands	Europe
6	AARDEWIJN, Pepijn	Men	Netherlands	Europe
7	AARONES, Ann Kristin	Women	Norway	Europe
8	AAS, Karl Jan	Men	Norway	Europe
9	AAS, Thomas Valentin	Men	Norway	Europe
10	ABAJO, Jose Luis	Men	Spain	Europe
11	ABAKUMOVA, Maria	Women	Russian Federation	Europe
12	ABALMASAU, Aliaksei	Men	Belarus	Europe

V. Visualisation des données sur Power BI

Power BI est un ensemble de services et d'applications d'analyse de données développés par Microsoft. Il permet de visualiser les données, de partager des informations et d'intégrer des données provenant de diverses sources pour prendre des décisions éclairées.



Visualiser les données sur Power BI est une excellente façon de donner vie à nos données et de les comprendre plus facilement.

Tout d'abord, on doit nous connecter à power BI a Talend Open Studio.

Une fois connecté à nos données, on peut commencer à créer des visualisations. Power BI offre une interface intuitive de type glisser-déposer qui nous permet de créer des tableaux de bord interactifs, des graphiques, des cartes, des tableaux croisés dynamiques et bien plus encore.

Power BI offre également des fonctionnalités avancées de personnalisation pour adapter nos visualisations à vos besoins spécifiques. On peut modifier les couleurs, les polices, les légendes et bien plus encore.

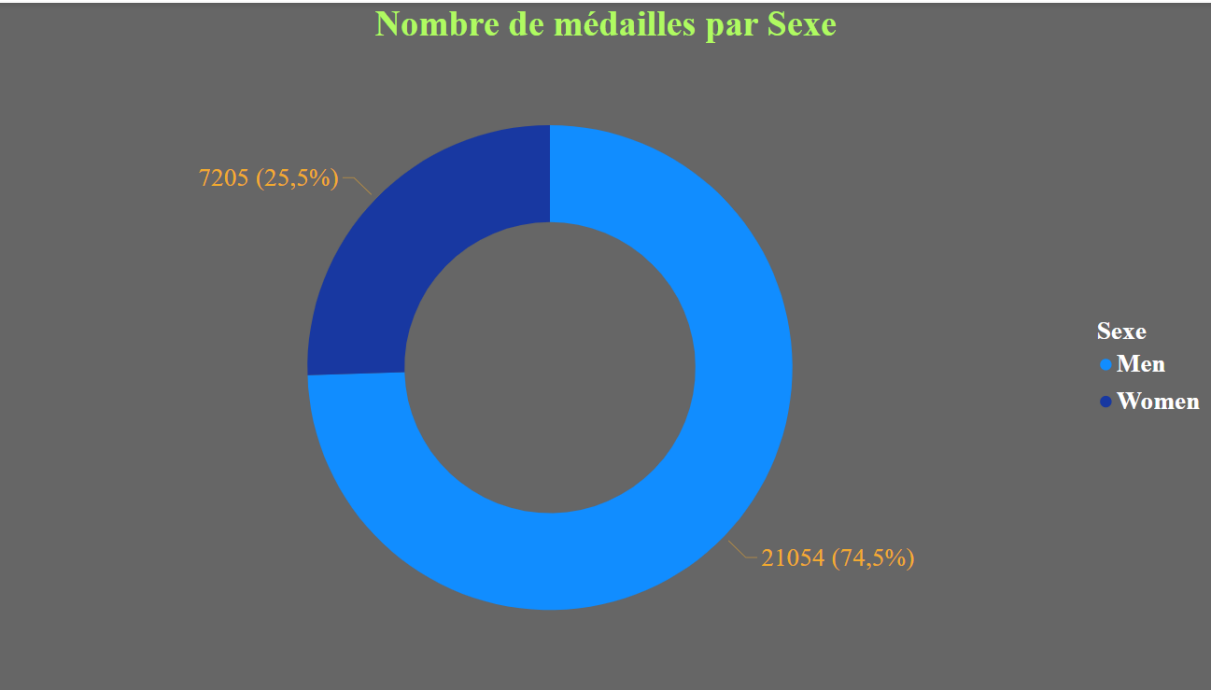
A/ Nombre de médailles par types de médailles

Nombre de medailles par types de médailles	
Silver	9292
Bronze	9408
Gold	9559
Total	28259

On constate que le nombre total de médaille est distribué de façon un peu équilibrée dans les trois différentes types de médailles.

Les variations dans la répartition des médailles par types de médailles peuvent être influencées par des facteurs tels que les stratégies de coaching, les investissements dans les infrastructures sportives et les politiques de financement des programmes olympiques nationaux.

B/ Nombre de médailles par sexe



L'analyse du nombre de médailles par sexe révèle des tendances significatives dans la participation et la performance des athlètes hommes et femmes aux Jeux olympiques, offrant un aperçu de l'égalité des sexes dans le sport de haut niveau. Nous constatons que les hommes ont remporté un total de 21054 médailles, tandis que les femmes ont gagné un total de 7205 médailles.

La répartition des médailles par sexe peut être influencée par des facteurs tels que la disponibilité des programmes sportifs pour les femmes, les différences de financement et de soutien pour les athlètes hommes et femmes, et les normes socioculturelles entourant le sport.

C/ Top 10 des pays les plus médaillés par sexe

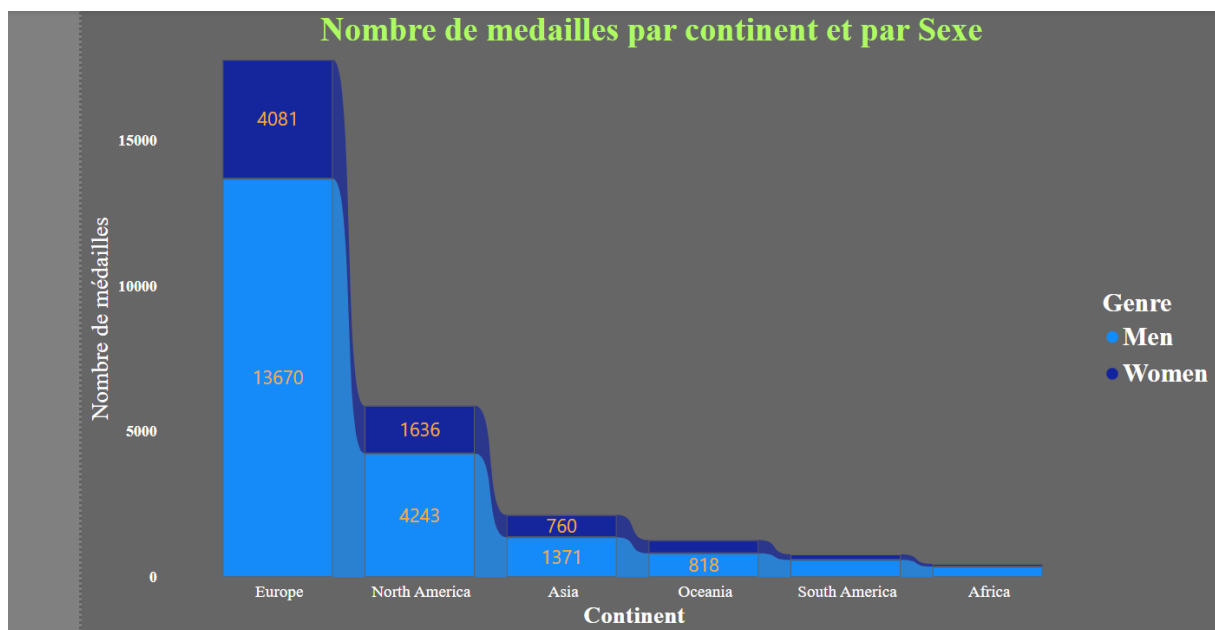


Les dix pays les plus médaillés parmi les hommes aux Jeux olympiques sont United States, Russian Federation, Germany, United Kingdom, France, Italy, Australi, Hungry, Sweden et le Nerthland.

Les performances des États-Unis sont remarquables dans les deux catégories, avec une position de leader chez les hommes et chez les femmes, démontrant leur excellence dans une variété de sports olympiques tandis que la Russie occupe la deuxième place chez les hommes ainsi que chez les femmes.

Les pays les plus médaillés dans cette catégorie sont souvent ceux qui ont une diversité de sports bien développés, couvrant à la fois les sports d'équipe et les sports individuels.

D/ Nombre de médaille par continent et par sexe

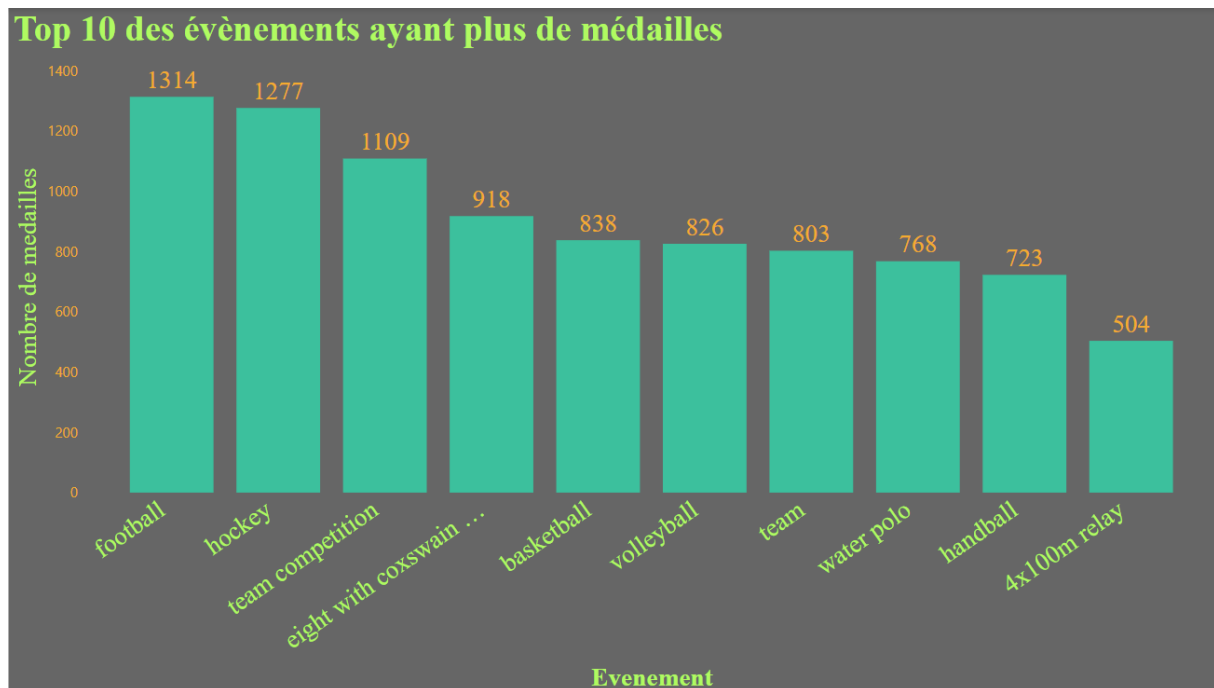


L'analyse du nombre de médailles par sexe et par continent révèle des disparités significatives, mettant en lumière les tendances et les différences de participation entre hommes et femmes dans diverses régions du monde.

Les continents avec les plus grands écarts entre les sexes en termes de médailles sont l'Europe et l'Amérique du Nord, où les hommes dominent largement les femmes en nombre de médailles. En revanche, en Asie et en Océanie, bien que les hommes aient toujours remporté plus de médailles, l'écart est moins prononcé.

Le continent africain enregistre un nombre très faible de médailles qui peut être expliqué par une intégration très tardive à ces jeux olympiques.

E/ Top 10 des événements ayant plus de médailles



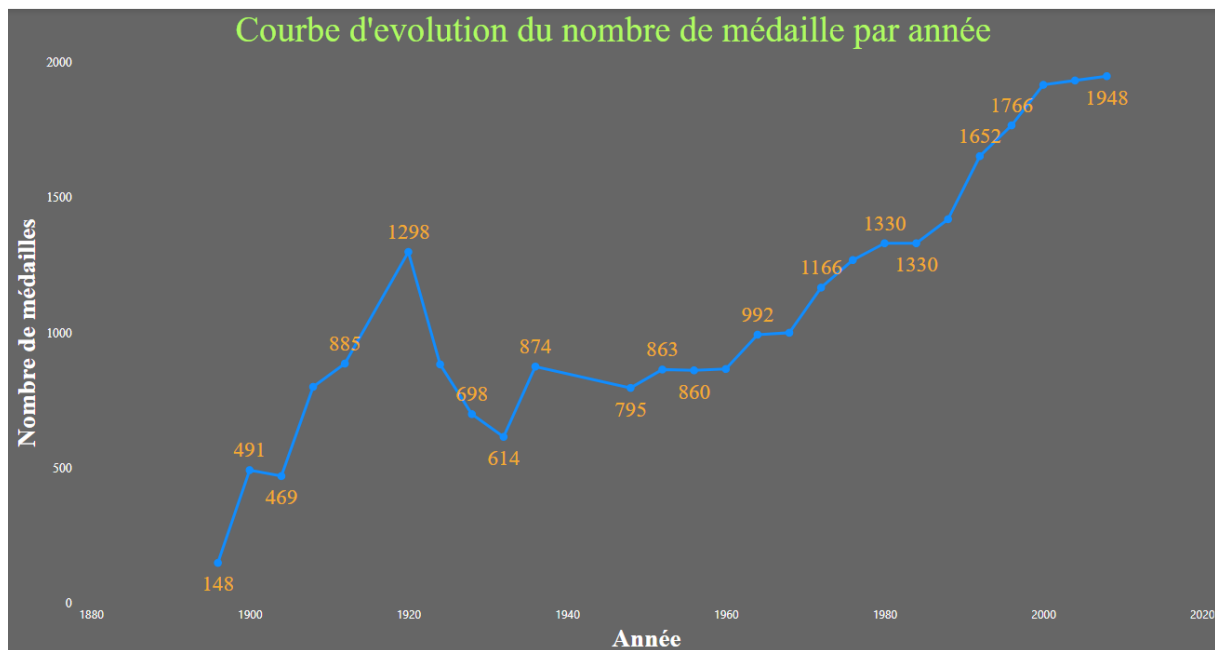
L'analyse des événements olympiques avec le plus grand nombre de médailles révèle les disciplines les plus compétitives et populaires au fil des années.

Il est probable que le football soit en tête, car il s'agit d'un sport d'équipe où de nombreuses médailles sont attribuées à la fois aux hommes et aux femmes et suivi du hockey qui arrive en deuxième position, soulignant sa popularité et le grand nombre d'équipes et de tournois organisés pour ce sport.

Certains événements ont une longue histoire aux Jeux olympiques et sont profondément enracinés dans la tradition sportive, ce qui contribue à leur statut et à leur attractivité pour les athlètes et les spectateurs mais aussi ce sont souvent des sports populaires et largement accessibles, qui attirent un grand nombre d'athlètes.

Globalement le top 10 montre les événements les plus populaires dans les jeux olympiques avec beaucoup de médailles décernées aux athlètes.

B/ Courbe d'évolution du nombre de médailles par année



Commentaire : Cette graphe nous montre une évolution très rapide du nombre de médaille entre les années 1896 et 1920, probablement due à l'introduction de nouveaux sports dans les Jeux olympiques suivi d'une chute entre les 1920 à 1932. A partir des années 1945 le nombre de médaillés a connu une évolution fulgurante jusqu'en 2008.

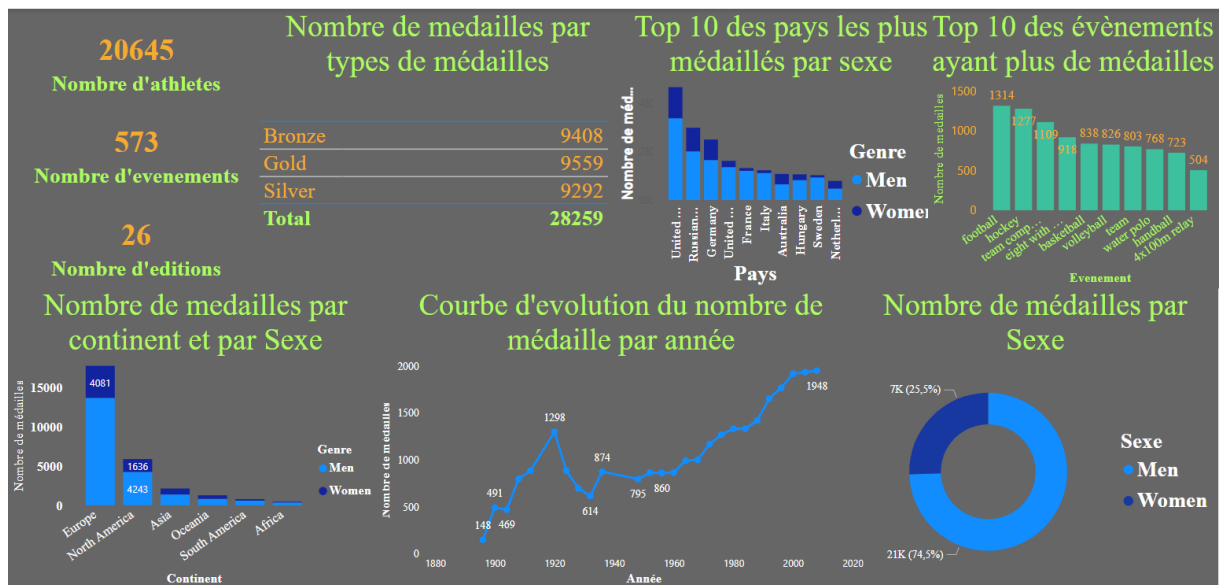
On note aussi l'absence des Jeux olympiques en 1916, 1940, et 1944, causée par les Première et Seconde Guerres mondiales.

Les années de croissance dans le nombre de médailles peuvent être attribuées à des facteurs tels que l'augmentation de la participation des pays, l'amélioration des infrastructures sportives et l'innovation dans l'entraînement et la préparation des athlètes tandis que les années de stagnation ou de baisse dans le nombre de médailles peuvent être le résultat de facteurs externes tels que les conflits politiques, les crises économiques ou les changements dans les politiques sportives nationales.

Globalement, la courbe d'évolution du nombre de médailles olympiques montre comment les Jeux olympiques ont évolué au fil du temps, en réponse à des facteurs historiques, politiques et sportifs divers.

VI. Tableau de bord

Le tableau de bord nous permet de présenter de manière visuelle les résultats de l'analyse des données.



La visualisation des données dans le tableau de bord confirme la prédominance des grandes puissances sportives dans les Jeux olympiques, tout en soulignant les progrès réalisés par les pays émergents et les régions en développement dans le domaine du sport de haut niveau.

L'analyse des tendances temporelles révèle une augmentation constante du nombre de médailles au fil des ans, avec des variations notables en fonction des événements mondiaux, des avancées technologiques et des changements dans les politiques sportives.

Lien de partage de notre tableau de bord :

<https://app.powerbi.com/groups/me/reports/80c69731-4e76-4786-b48c-79930f2d0f08/bf9942303a001590d803?experience=power-bi>

Conclusion

Ce projet d'intégration de données sur les Jeux olympiques a été une opportunité précieuse d'explorer et d'analyser en profondeur les performances sportives mondiales à travers les années.

Nous avons observé des évolutions temporelles dans le nombre de médailles, des disparités régionales dans les performances des pays, et des tendances émergentes dans la participation et la compétitivité dans différentes disciplines sportives. Ces compréhensions ont été essentiels pour éclairer les décideurs, les entraîneurs et les athlètes sur les opportunités et les défis dans le monde du sport de haut niveau.

En outre, ce projet a mis en lumière l'importance de l'intégration de données et de l'analyse avancée pour prendre des décisions éclairées dans le domaine du sport. En utilisant des techniques et des outils modernes d'analyse des données, nous avons pu transformer des volumes massifs de données brutes en informations exploitables et pertinentes pour l'amélioration continue des performances sportives et de l'expérience des Jeux olympiques.