Visualization of Massive Data

Module présentée par Nicolas Lehir

Document récapitulatif réalisé par Nicolas Deviers

Nicolas Deviers
Nathanael Melouki
Antoine La Mache

Choix du dataset :

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

Ce dataset rassemble de nombreuses données sur tous les participants aux Jeux Olympiques depuis l'édition de Rio de 1896.
Il contient les champs suivants :

1. **ID** - Unique number for each athlete
2. **Name** - Athlete's name
3. **Sex** - M or F
4. **Age** - Integer
5. **Height** - In centimeters
6. **Weight** - In kilograms
7. **Team** - Team name
8. **NOC** - National Olympic Committee 3-letter code
9. **Games** - Year and season
10. **Year** - Integer
11. **Season** - Summer or Winter
12. **City** - Host city
13. **Sport** - Sport
14. **Event** - Event
15. **Medal** - Gold, Silver, Bronze, or NA

Il est intéressant car très fourni, et offre un large potentiel d'informations que l'on pourrait en extraire en recoupant les différents champs.

Le dataset est situé dans l'archive archipe.zip, dans le dossier dataset. Il faut l'extraire afin d'exécuter les scripts.

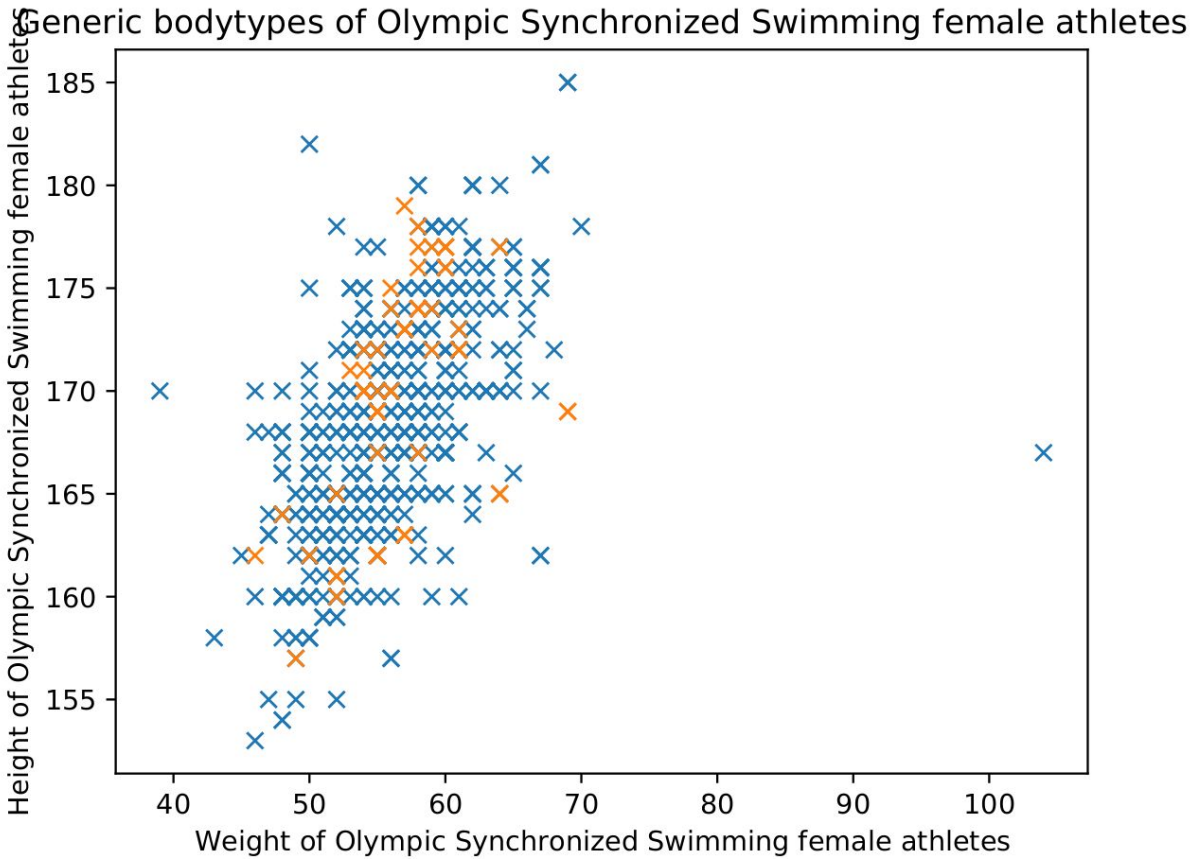
Visualization :

1 - Scatter (Nicolas Deviers)

Deux versions de scatter ont été réalisées, qui sont les suivantes :

La première version compare la taille et le poids des athlètes pour chaque discipline olympique où ces données sont fournies. Un fichier est généré par sexe aussi, les mensurations étant différentes. Sur le graphique, les athlètes ayant obtenu une médaille d'or ont leur point coloré d'une couleur distincte. Cela permet de voir s'il existe un gabarit optimal pour un sport. Puisque la professionnalisation du sport est assez récente (pouvoir vivre de son sport sans travailler à côté et donc s'entraîner à fond sans contrainte), je n'utilise que les données datant d'après 1980.

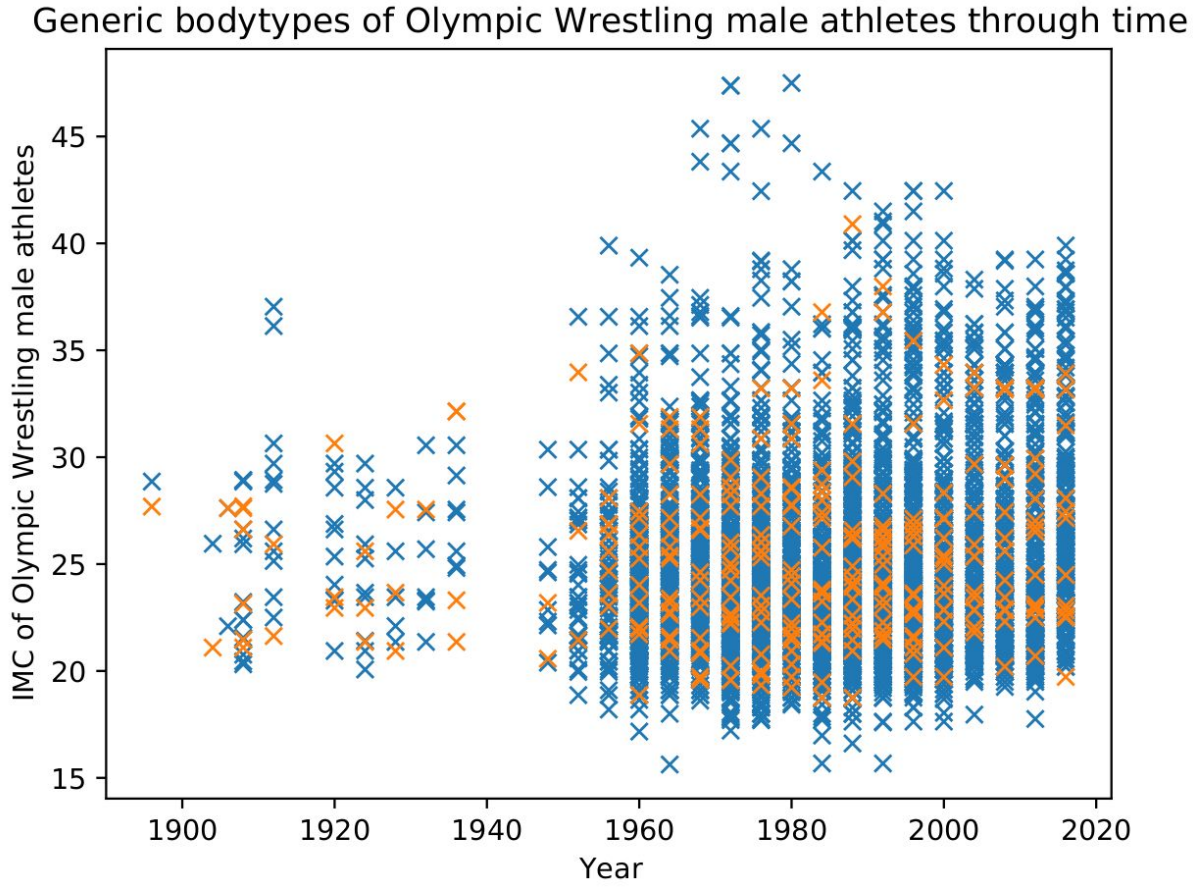
Exemple :



Ce schéma nous montre les mensurations des athlètes de nage synchronisée féminine depuis 1980. On y remarque que tous les points sont relativement concentrés dans un rectangle entre 1m60 et 1m78 de hauteur et entre 48 kg et 65 kg de masse. Cependant, la majorité des médailles d'or sont rassemblées entre 1m68 et et 1m78 de hauteur et entre 51 kg et 61 kg de masse

La seconde version compare l'IMC des mêmes athlètes au cours des années, et ce depuis 1896 pour toutes les disciplines. On pourra y voir à quel point cette même professionnalisation du sport à pu impacter le physique des athlètes au travers d'un siècle d'amélioration constante des techniques d'entraînement.
La formule de l'IMC est le poids divisé par le carré de la taille.

Exemple :



Ce schéma nous démontre l'évolution de l'IMC des lutteurs masculins depuis 1896. On y remarque tout d'abord une augmentation importante du nombre de participants - représenté par le nombre de croix - après la seconde guerre mondiale accompagné un élargissement impressionnant des variations d'IMC. La ou celui-ci était confiné à une fourchette contenue entre 20 et 30 avant la guerre, il se retrouve disséminé sur un véritable rateau oscillant entre 18 jusqu'à 40 ! Cependant la majorité des champions Olympiques restent confinés entre 20 et 30.

Les graphes générés sont stockés dans le dossier scatterImage/ a côté des scripts. Les fichiers des graphes générés par la seconde version sont reconnaissables par leur nom commençant par 'ZZ_'.
Une aide à l'utilisation du script est incorporée dans celui-ci avec l'argument '-h'.

L’exécution du programme étant assez longue à la taille du dataset et au nombre de disciplines, l’utilisateur peut décider de ne l’exécuter que sur une seule discipline en la précisant comme argument au programme à son exécution. La liste des disciplines disponibles est disponible avec l’argument -h.

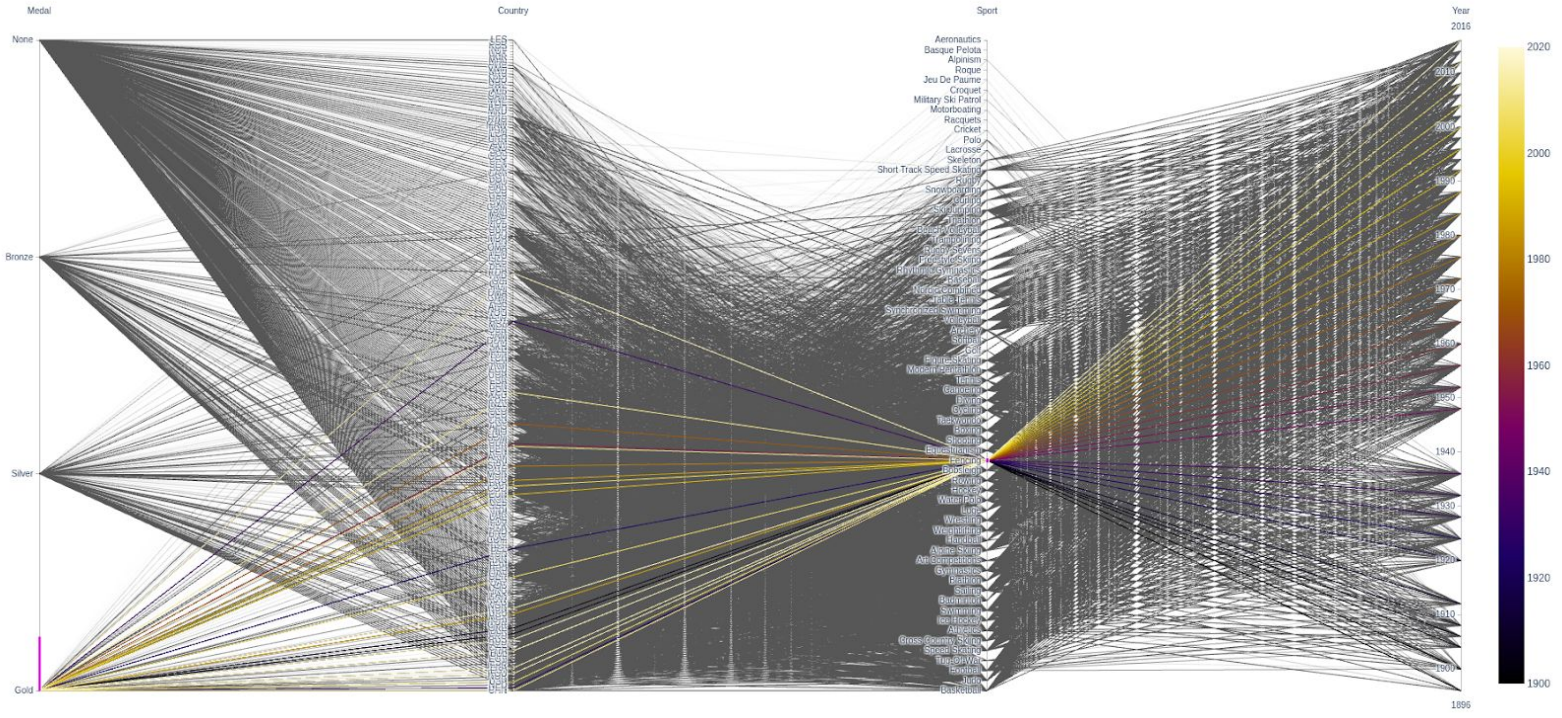
2 - Parallel Coordinates (Nicolas Deviers)

Le schéma de Parallel Coordinates est extrêmement complet. Il recoupe les colonnes des années, des sports, des pays et des médailles. On peut y voir en somme toutes les médailles obtenues par toutes les équipes dans tous les sports et ce pour toutes les éditions des jeux Olympiques depuis 1896.

Le schéma est généré par la librairie Plotly et s’ouvre dans le navigateur. Il est particulièrement lourd car il manœuvre 4 colonnes de 271116 données, plus d’un millions de points sont de ce fait chargés. Un bon ordinateur est conseillé pour son utilisation.

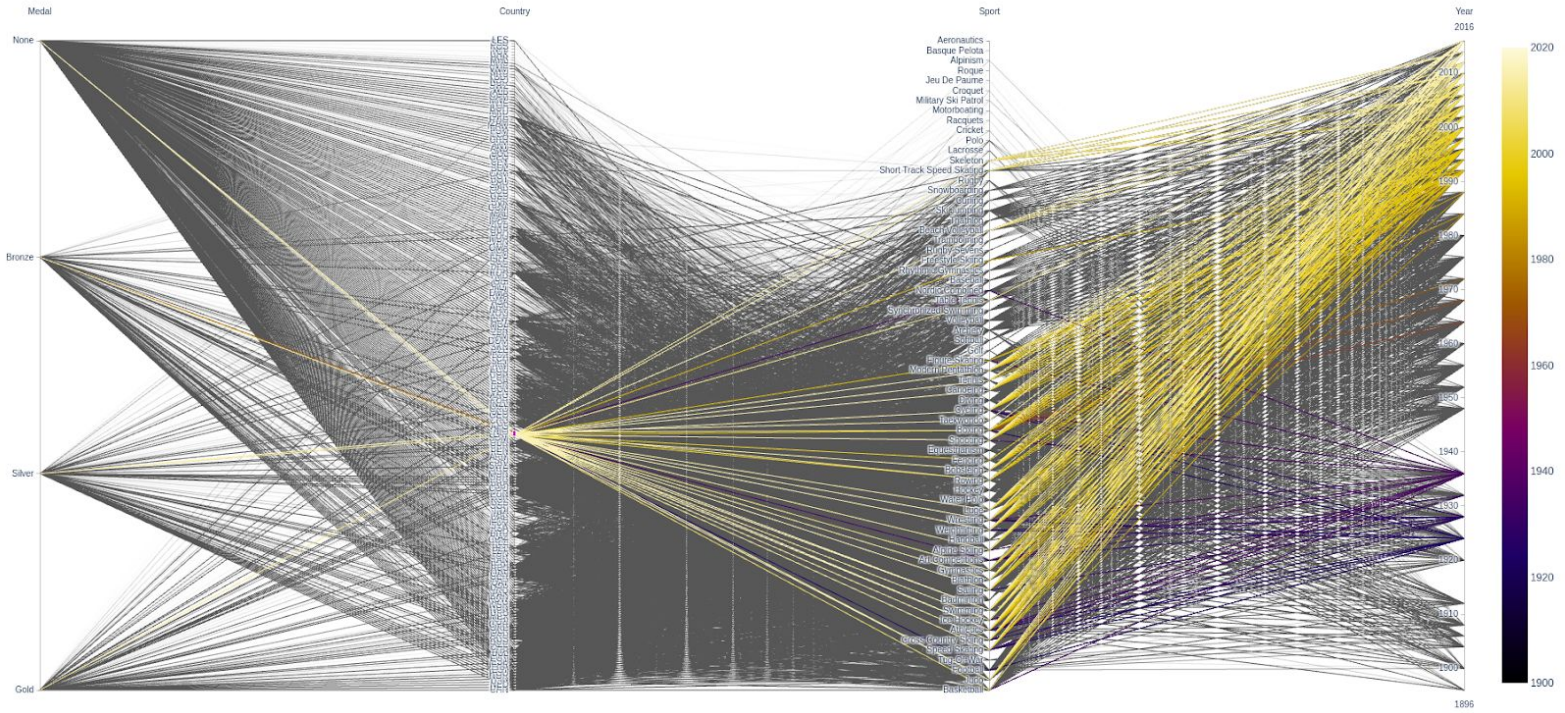
Le format des données du dataset n’étant pas compatible avec le schéma Plotly, celui-la est donc réinterprété afin de n'intégrer que des entiers dans le schéma représentant l’index de la réelle donnée, collée sur les parallèles sous forme de légende de Schéma. Ainsi un écran de grande résolution est recommandé, sinon quoi les indicatifs des noms des pays risquent de se confondre.

Exemple :



Voici un exemple de l'utilisation de ce schéma. En isolant la discipline Escrime et la médaille d'Or, je suis en capacité de retrouver la nationalité de tous les vainqueurs depuis 1896 ainsi que l'année de leur victoire.

Exemple :



Une seconde utilisation du schéma. En ne sélectionnant qu'un pays en particulier, on peut retrouver toutes les disciplines pour lesquelles il participait ou alors les années ou il a pu présenter une délégation.

Quantitative Analysis :

Par manque de temps, je n'ai pas pu réaliser cette troisième partie du projet.

Librairies :

pandas : utilisation de la fonction read_csv pour extraire seulement les colonnes du dataset nécessaires à l'exécution du programme.

plotly : utilisation des fonctions Figure et Parcoords pour la création et mise en forme du graph ParallelCoords

myplotlib : utilisation des fonctions plot, title, xlabel, ylabel, savefig et close pour la création, mise en forme et enregistrement des graphes des scripts Scatter