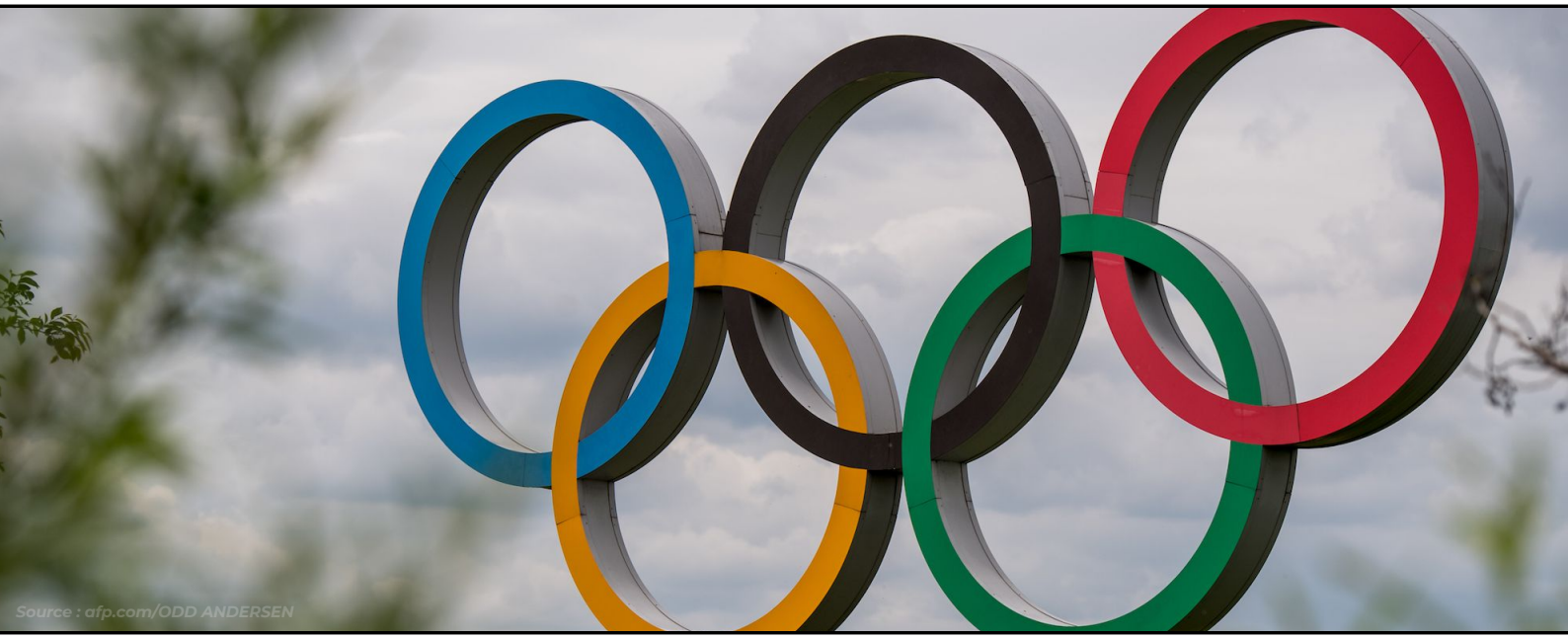
AVEZ-VOUS LE PROFIL D'UN VAINQUEUR OLYMPIQUE ?

LAURA BURLON--ROUX & JONATHAN ANDRIEU | ANALYSE EXPLORATOIRE



Les données d'entrée

Notre étude porte sur un ensemble de données historiques sur les Jeux Olympiques (JO) modernes, depuis les Jeux d'Athènes en 1896 jusqu'aux Jeux de Rio en 2016. Voici les données dont nous disposons :

- **ID** - qui est unique pour chaque athlète
- **Name** - qui est le nom de l'athlète
- **Sex** - qui indique si l'athlète a concurru dans la catégorie homme ou femme
- **Age** - qui précise quel est l'âge de l'athlète au moment de sa participation
- **Height** - qui est la taille de l'athlète (en cm)
- **Weight** - qui est le poids de l'athlète (en kg)
- **Team** - qui est le nom du pays pour lequel l'athlète a concouru
- **NOC** - qui est le code 3 lettres du pays
- **Year** - qui indique quelle est l'année où a eu lieu l'épreuve
- **Season** - qui précise s'il s'agissait des jeux d'hiver ou d'été, seulement 2 valeurs possibles
- **Games** - qui concatène les champs Year/Season
- **City** - qui indique dans quelle ville s'est déroulée l'épreuve
- **Sport** - qui précise de quel sport il s'agit
- **Event** - qui précise l'épreuve sportive
- **Medal** - qui indique si l'athlète a obtenu une médaille d'or, d'argent, de bronze ou rien

A partir de ces 271 116 lignes de données que nous allons analyser, nous tenterons de vous présenter le profil type des personnes qui montent sur les podiums aux Jeux Olympiques. Au fil de votre lecture, nous vous donnerons quelques précieux conseils pour que vous puissiez à votre tour décrocher une médaille !

Pour ce faire, nous n'allons considérer que les dix derniers Jeux Olympiques d'été, depuis ceux de Moscou en 1980, jusqu'à ceux de Rio de Janeiro en 2016. Ainsi, nous ne garderons que les lignes où Season est égal à la valeur "*Summer*" et où la valeur de Year est supérieure ou égale à 1980. Cette sélection de données est rendue possible grâce la fonction *filter()* de la librairie *dplyr*. Nous obtenons maintenant 122 913 lignes sur lesquelles nous allons baser notre analyse.

.

LIEN VERS LE DATASET ET VERS LE CODE EN R

- https://github.com/JohnAndrieu/R_Project

Maximisez vos chances

La première chose à faire est de choisir le sport dans lequel vos chances de remporter une médaille sont les plus importantes. Pour ce faire, nous comparons le nombre d'inscrits à la discipline avec le nombre de médaillés, de sorte à obtenir un ratio. Voici les cinq sports où ce rapport est le plus élevé :

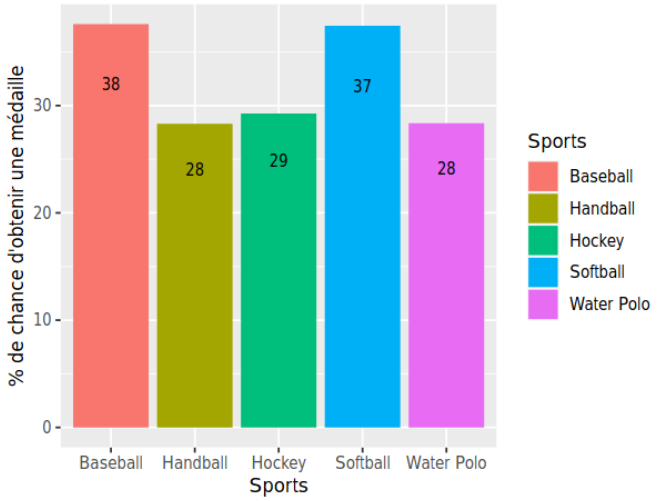


Figure 1 - Top 5 des sports où les chances de médailles sont les plus élevées.

Comme vous pouvez le constater (Figure 1), si vous participez au JO en tant que **baseballeur** vous avez près de 38% de chance d'obtenir une médaille. Nous trouvons ensuite des sports comme le **softball** (37%), le **hockey** (29%), le **waterpolo** (28%) ainsi que le **handball** (28%). Notez que le softball est une discpline se rapprochant du baseball, mais qui est uniquement féminine. Apparue lors des Jeux Olympiques d'été de 1996 à Atlanta et disputée jusqu'à ceux de 2012, cette discipline sera à nouveau présente lors des Jeux de Paris. Le baseball quant à lui est uniquement masculin. Si vous comptiez sur votre revers de tennisman ou votre plus beau justaucorps en gymnastique artisitique, sachez que ces deux sports sont parmi ceux où le ratio médaillés / nombre d'inscrits est le plus faible. Seulement 8% des inscrits obtiennent finalement une médaille... Passez votre chemin !

Nous remarquons que ces 5 sports sont des sports collectifs. Nous pensons que comparativement à des sports individuels, le nombre d'équipes participantes pour des sports collectifs est moindre. Par exemple, en 2016, 12 équipes masculines étaient en lice en handball, alors que 59 hommes concouraient pour le 100 mètres nage-libre en natation.

Nota Bene : toutes les figures contenues dans ce rapport sont disponibles au format .pdf au sein du Git dont le lien est rappelé en page 1.

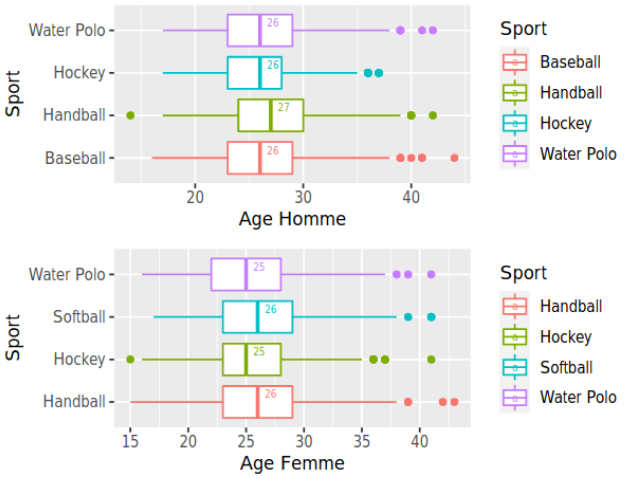


Figure 2 - Distribution de l'âge des médaillés, selon leur sexe

L'âge devrait vous guider

Vous hésitez toujours entre une carrière dans le baseball, le softball, le hockey, le waterpolo ou le handball ? L'âge moyen des médaillés pourrait vous aider à y voir plus clair. Ces boites à moustaches (Figure 2), selon si vous êtes un homme ou une femme, vous permettent d'appréhender la distribution de l'âge des médaillés. Quels que soient les sports ou le sexe de la personne, **l'âge des sportifs médaillés oscille autour de 26 ans**. La répartiton des âges est également similaire d'un cas à l'autre, les lignes horizontales montrent que les 17-37 ans sont représentés, hormis pour le hockey où on retrouve exceptionnellement (les points) des joueurs plus jeunes, âgés de moins de 16 ans, et rarement âgés de plus de 35 ans.

Rien n'est perdu si vous n'êtes pas proche de votre 26ème printemps, même si, tout sport confondu, la moyenne d'âge reste similaire. Certains sports comprennent davantage de jeunes athlètes comme par exemple la gymanstique rythmique où le sportif a environ 18 ans. A l'inverse, en équitation par exemple, il oscille plutôt autour de 35 ans.

LE SAVIEZ-VOUS ?

La plus jeune médaillée aux Jeux Olympiques sur les 40 dernières années été âgée de 13 ans, de nationalité chinoise, et a décrochée la première place en plongeon en 1992.

L'IMC idéal

La section précédente ne vous a pas permis de choisir définitivement un sport ? Vous ne savez pas si votre physique correspond au sport que vous avez pré-sélectionné ? Les prochaines lignes sont faites pour vous ! Il est évident que suivant la discipline, le poids et la taille du sportif varient. Nous avons fait le choix de comparer ces valeurs en nous basant sur l'Indice de Masse Corporelle (IMC), calculé suivant la formule : *Poids* (en kg) / *Taille*^{2} (en cm). Nous distinguons là encore l'affichage de nos calculs, suivant s'il s'agit d'une femme ou d'un homme. Voici les résultats :

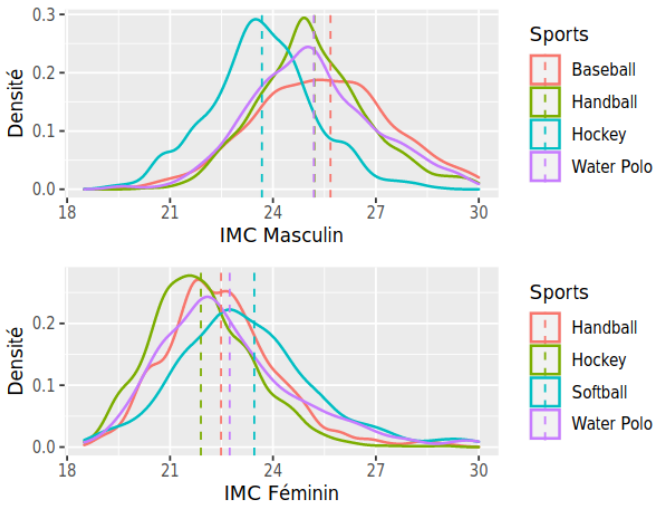


Figure 3 - Distribution de l'IMC des médaillés, selon leur sexe

Comme escompté, les IMC féminins sont en moyenne plus faibles que les masculins : c'est une observation qui est également vraie dans la vie de tous les jours, pas uniquement dans le monde du sport. Les pointillés verticaux représentent la valeur moyenne de l'IMC : comme vous pouvez le constater, les IMC les plus bas correspondent au hockey, avec une valeur de 23,5 pour les hommes, et 22 pour les femmes. Le sport où les IMC sont en moyenne les plus importants est le Baseball / Softball, mais c'est également le sport où la répartition est la plus homogène, ce qui est traduit par la planitude de la courbe : les physiques des sportifs semblent plus hétérogènes que pour le handball par exemple, où on observe un pic franc.

Dans tous les cas, si vous êtes un homme et que votre IMC est compris entre **21 et 28,** ou que vous êtes une femme et que votre IMC est compris entre **19 et 26,** rien n'est perdu pour vous pour ces 5 sports ! En revanche, si cet indice est plus important, il est peut-être préférable pour vous de vous tourner vers des sports comme l'haltérophilie ou le rugby avec un IMC moyen proche de 28 (chez les hommes). Si vous êtes en dessous, tourner vous vers la gymnastique rythmique, avec un IMC moyen autour 17 chez les femmes, ou vers le triathlon pour les hommes avec un IMC moyen proche de 21.

Néanmoins, nous sommes conscients que l'IMC est en réalité un facteur peu judicieux à prendre en compte. En effet, les sportifs ont souvent un IMC relativement élevé car leur masse musculaire représente un poids important. En effet, 2 personnes mesurant la même taille, et pesant le même poids n'ont pas forcément la même silhouette : pour le même poids, le muscle occupe beaucoup moins de place, en termes de volume, que le graisse.

Envisager une naturalisation ?

Las de comparer différentes données pour uniquement 5 sports, nous étendons cette analyse à l'ensemble des 36 sports présents aux Jeux Olympiques depuis 1980.

Dans cette section, nous allons comparer le nombre de médailles gagnées par pays. Pour ce faire nous avons choisi de coupler couleurs et mappemonde : plus le pays est en rouge foncé, plus il a remporté de médailles sur les 40 dernières années. A l'inverse, plus le pays s'approche d'une coloration blanche, moins il en a remportées, La figure 4 est disponible à la prochaine page.

Un coup d'oeil nous permet d'affirmer que le pays en ayant gagnées le plus sont les Etats-Unis avec 2338 médailles sur 40 ans. Loin derrière, nous trouvons l'Australie, la Chine et la Russie avec un nombre de médailles compris entre 800 et 900. La France quant à elle est classée 8ème pays, juste derrière le Royaume-Uni, avec 597 médailles;

LE SAVIEZ-VOUS ?

Le Tir à la Corde (Tug of War) était considéré comme une épreuve olympique de 1900 à 1920. La France a remporté la 2° place en 1900 alors que seulement 2 équipes y participaient.

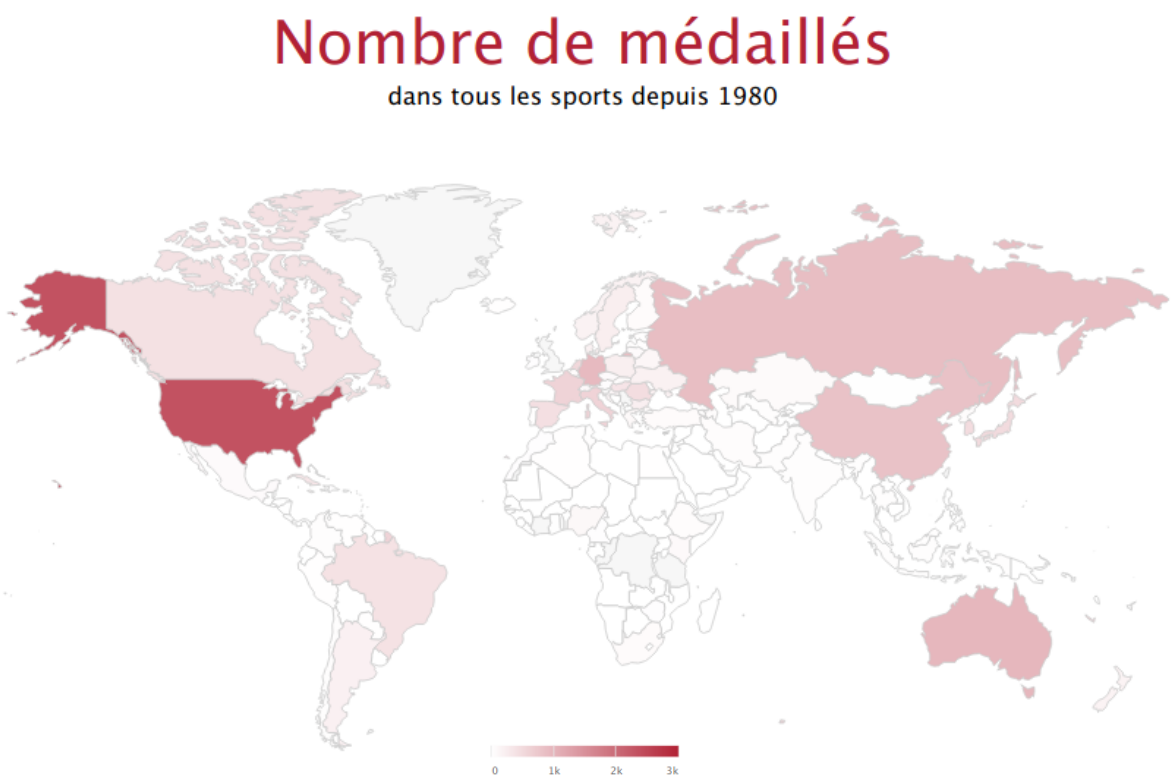


Figure 4 - Nombre de médaillés suivant le pays depuis les Jeux Olympiques de 1980

Dire que concourir pour un certain pays permet d'optimiser ses chances de gagner une médaille, serait certes un raccourci, mais c'est aussi un constat indéniable. **Les pays en cours de développement remportent nettement moins de médailles que les pays développés** d'Occident : par exemple, les Philipines ont remporté 4 médailles, la Côte d'Ivoire 3, ... Le continent africain et l'Asie du sud sont totalement blancs, certains pays n'ayant jamais gagné ! Ces différences sont largement expliquées par les infrastructures dont disposent les pays et le niveau de vie des habitants. Il n'est pas rare de voir des sportifs concourir pour les couleurs de leur pays d'origine mais s'entrainant dans un autre pays, plus développé, pour bénéficier de meilleures conditions de préparation.

Enfin, dans certains pays, le sport féminin en compétition reste confidentiel, notamment en Arabie Saoudite et au Quatar où aucune femme n'a jamais pu prendre part à la compétition

LE SAVIEZ-VOUS ?

L'acteur principal des films Tarzan, Peter John Weissmuller, a été 6 fois médaillé olympique en natation et waterpolo dans les années 1920. Il concourait pour les Etats-Unis.

d'après les données dont nous disposons. Nous avons compté le nombre de '*F*' présents dans la colonne '*Sex*' pour tous les pays.

En 2016, un seul pays au monde n'avait jamais participé aux Jeux Olympiques, c'est le seul pays absent de notre dataset : il s'agit du Vatican. D'après nos recherches, il pourrait néanmoins participer aux Jeux de 2028 puisqu'il vient tout juste de créer sa fédération d'athlétisme.

Ainsi, suivant l'hésmisphère auquel appartient votre pays d'origine, vos chances de gagner une médaille sont clairement différentes.

Envie de tricher ?

Il s'avère que vous n'êtes toujours pas sûr d'obtenir une médaille ? Nous allons regarder ensemble si des cas de fraudes massives ont pu être observés sur les dix derniers jeux olympiques d'été. Nous nous demandons si, lorsqu'un pays est organisateur, il obtient plus de médailles qu'à l'accoutumée. Des tirages au sort truqués pour les poules ? des contrôles anti-dopage plus souples ? ... beaucoup d'éléments pourraient permettrent à un état d'organiser sa réussite.
Nous avons choisi, encore une fois tout sport confondu, de comparer l'évolution du nombre de médaillés par rapport au nombre d'athlètes inscrits pour les dix derniers pays organisateurs.

Si un pic du ratio caLculé est observé pour le pays l'année où il organise, des doutes pourront être avancés.
Pour rappel, voici les dates des Jeux Olympiques et de leurs organisateurs :

- 1980 - Union Soviétique (Moscou)
- 1984 - Etats-Unis (Los Angeles)
- 1988 - Corée du Sud (Séoul)
- 1992 - Espagne (Barcelone)
- 1996 - Etats-Unis (Atlanta)
- 2000 - Australie (Sydney)
- 2004 - Grèce (Athènes)
- 2008 - Chine (Pékin - Beijing)
- 2012 - Royaume-Uni (Londres)
- 2016 - Brésil (Rio de Janeiro)

En ce qui concerne l'Union Soviétique, nous afficherons à partir de 1992 les résultats des athlètes médaillés russes, quelle que soit la date, même si ce choix est discutable. Aussi on observe sur le graphique que certains pays n'ont pas pris part aux Jeux Olympiques de 1980 : en effet, cette année là les Jeux se déroulaient en URSS pour la première fois, et une cinquantaine de pays les ont boycotés pour des raisons géo-politiques dans un contexte de Guerre Froide. En réponse, en 1984, l'URSS ne prendra pas part aux Jeux organisés aux Etats-Unis.

Nous pensions pouvoir observer des pics pour les années où un pays était organisateur mais nous n'avions pas pris en compte le fait que **lorsqu'un pays organise, le nombre de participants et bien plus élevé que d'habitude**, ce qui fait que le ratio médaillés / participants n'augmente pas et peu même parfois baisser. Pour y palier, nous aurions pu classer les pays par nombre de médailles obtenues chaque olympiade. Ainsi, nous aurions pu voir qu'en 1984, la Corée du Sud se classe 4ème au tableau des médailles contre 10ème l'olympiade drécédente et 12ème celle d'après. De même, l'année où l'Espagne organise, elle se classe 6ème, contre 25ème aux Jeux précédents et 12ème aux suivants.

Voici néanmoins la figure que nous avons obtenu. (Figure 5 ci-contre).

Nous n'observons donc aucun pic notable comme nous l'avions espéré. Nous ne dégageons pas non plus de tendances au sein de ces évolutions, nous notons seulement une large baisse du ratio pour la Corée du Sud à partir de 1984. Enfin, la Russie, les USA, et la Chine sont les pays où ce ratio est le plus élevé, ce qui vient confirmer ce que nous avions remarqué dans la section précédente.

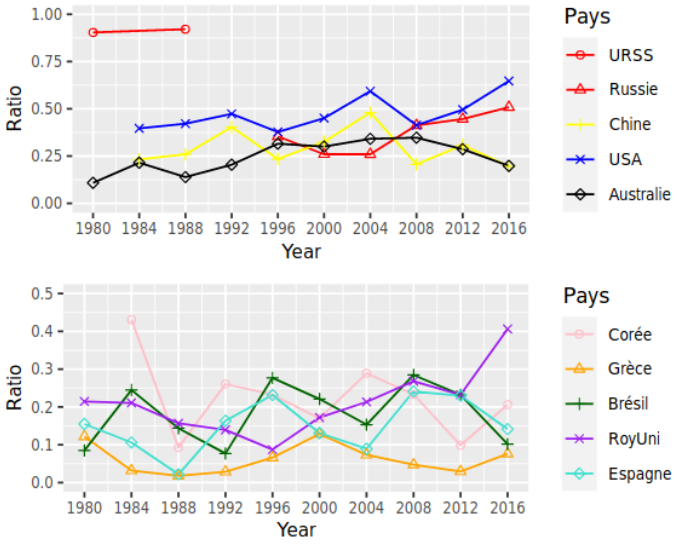


Figure 5 - Evolution du ratio médaillés / participants par pays
⚠ L'axe du ratio n'est pas à la même échelle par soucis de visibilité

Le cas de l'URSS est très intéressant : lorsqu'elle participe en 1980 et 1988, un athlète soviétique avait plus de 9 chances sur 10 de remporter une médaille. C'est un ratio bien plus élevé que pour les autres pays, qui n'a jamais été égalé ensuite. Le sport permettait à l'URSS d'exprimer un soft-power face aux Etats-Unis, apportant ainsi, sans entrer en guerre directe, une connotation physique au sein d'un conflit idéologique.

Conclusion

Ainsi, si vous êtes un homme, baseballeur, âgé 26 ans, ayant un IMC proche de 25 et de nationalité américaine, foncez vous entrainer ! (Nota Bene: en réalité, l'intersection de toutes ces contraintes ne mène pas forcément au profil type idéal, mais nous reprenons la problématique posée au départ)

Nous sommes consicents de n'avoir **exploité qu'une infime partie de ce qui était réalisable** avec ces données. Néanmoins, à travers l'exploitation de ce dataset, ce travail nous a permis d'apprendre à produire des graphes cohérents en fonction des valeurs que nous souhaitions mettre en exergue.

Nous pensons qu'avec ses 271 116 lignes, le dataset est complet. Même les épreuves d'art qui étaient disputées de 1914 à 1948 sont présentes. Nous sommes cependant dubitatifs sur **l'exhaustivité des données**, surtout pour les Jeux Olympiques qui se sont déroulés avant que la numérisation des données soit possible, des erreurs peuvent exister.