Projet Jeux Olympiques - Journal de bord

T. Burnel, N. Grim, M. Griveau, M. Mechentel

Janvier 2024

1 Introduction

Nous sommes une équipe de datajournalistes chargée, à l'approche des Jeux Olympiques de Paris, d'étudier l'existence ou non de liens entre le succès d'un pays aux Jeux Olympiques et sa richesse. Notre rôle sera de comparer le nombre de médailles remportées par pays et par sport à six éditions des Jeux (1996, 2000, 2004, 2008, 2012, 2016) pour cerner la présence ou l'absence d'une logique économique dans la réussite des pays.

Notre hypothèse de départ est la suivante : les pays développés gagnent significativement plus de médailles en raison de leur niveau d'investissement dans le sport. À l'aune de l'examen et du traitement des données, nous avons pour ambition de déterminer si la politique d'investissement dans les infrastructures sportives de ces pays influe – tant positivement que négativement – sur leurs résultats aux différentes épreuves des Jeux Olympiques.

Notre travail consistera, dans un deuxième temps, en une analyse comparative de la politique d'investissement du Royaume-Uni et de la France.

Plusieurs points de convergence distinguent ces deux pays : leur nombre d'habitants, leur économie et leur qualité de pays organisateur des Jeux.

Le résultat du traitement des données constituera une base à la réalisation de datavisualisations et d'une application web pouvant servir de base à l'écriture d'articles journalistiques.

2 Jeux de données

2.1 Jeux du premier *flow*

En amont de la constitution de notre premier *flow*, nos jeux étaient au nombre de deux. Ils présentent les informations suivantes :

— Nombre de médailles remportées par sportif et par pays sur 120 ans [1] ;

— Le PIB des pays et leur investissement dans le domaine du sport [2] ;

Un troisième jeu a dû faire son entrée au cours du traitement pour pallier la maigre qualité des données fournies par le Fonds Monétaire International (*confer infra.* p. 11), singulièrement à propos du PIB. Ce jeu provient du site de la Banque mondiale. [3]

Enfin, un problème récurrent nous a poussé à la recherche d'un quatrième jeu. Lorsque Dataiku identifie des données comme *country*, nous pensions

1. *120 years of Olympic history : athletes and results.* URL : `https://www.kaggle.com/datasets/heesoo37/120-years-of-olympic-history-athletes-and-results` (visité le 17/01/2024).

2. *Download entire World Economic Outlook database, April 2019.* IMF. URL : `https://www.imf.org/en/Publications/WEO/weo-database/2023/October/download-entire-database` (visité le 17/01/2024).

3. *World Bank Open Data.* World Bank Open Data. URL : `https://data.worldbank.org` (visité le 20/01/2024).

pouvoir en extraire les coordonnées et les croiser avec d'autres jeux. L'extension *reverse geocoding plugin* aurait dû nous le permettre – sans succès. Nos suppositions portent sur notre version gratuite de Dataiku et de son nombre réduit de fonctionnalités. Nous avons rapidement trouvé un jeu, mis à disposition sur GitHub,[4] rassemblant les données nécessaires.

Une opération obligatoire dans notre traitement consiste à enrichir ces jeux avec le résultat d'une requête SPARQL, saisie sur le service de Wikidata. Ce premier *flow* a donc a sa base un ensemble de cinq jeux.

2.1.1 Jeu de Kaggle

Le jeu présente 15 colonnes pour 271117 lignes. Il contient des colonnes informatives sur les athlètes (ID ; Name ; Sex ; Age ; Height ; Weight), sur leur pays (Team ; NOC) et sur les Jeux : l'édition, l'année, la saison, la ville d'accueil, le sport et l'épreuve (Games ; Year ; Season ; City ; Sport ; Event) ainsi que sur les médailles remportées (Medal).

La propreté du jeu est notable, seules certaines données sur la taille et le poids des athlètes sont manquantes. La gravité de ces absences est toute relative dans la mesure où ne comptons pas conserver ces données. Un problème plus préoccupant concerne le nom de certains États, aujourd'hui obsolète (Allemagne de l'Est, Yougoslavie, Tchécoslovaquie, Union Soviétique, etc.).

4. Alexandru Paul Copil. *Countries codes and coordinates*. URL : `https://gist.github.com/cpl/3dc2d19137588d9ae202d67233715478` (visité le 25/01/2024).

2.1.2 Jeu du FMI

Le jeu présente 41 colonnes pour 7715 lignes. Il est issu d'un requêtage depuis le site du FMI pour récupérer les données avec lesquelles nous souhaitions travailler. Le potentiel de croisement est limité car un certain nombre de colonnes fournit des codes et des désignations propres au FMI (Country Code ; COFOG Function Name ; COFOG function Code ; Sector Name ; Unit Name ; Unit Code ; Attribute).

29 colonnes correspondent aux années couvertes par le FMI : de 1993 à 2022. Notre intérêt repose singulièrement sur les colonnes de 1996 à 2016 en plus des colonnes *Country Name*, *COFOG Function Name* laquelle correspond aux investissement dans les infrastructures sportives et *Unit Name*, c'est-à-dire l'échelle de mesure, soit en PIB soit en monnaie courante.

Ce jeu est de loin le moins bien structuré et le plus fautif. En guise d'exemple, les trois dernières colonnes (Indicator Code ; Global DSD Time Series Code ; col_41) sont d'une opacité confondante et témoignent du majeur problème du jeu : énormément de données sont vides, nulles ou incompréhensibles (« GERS_G14_GDP_PT » pour la colonne) *Indicator Code*, « A|GB|S1311|W0|S1|G2M|_Z|_Z|GF0801|XDC|_T|_X » pour la colonne *Global DSD Time Series Code*). En somme, ce jeu représente un véritable enjeu de nettoyage et de croisement des données.

2.1.3 Jeu de la Banque mondiale

Le jeu présente 67 colonnes pour 541 lignes – en réalité 271 : il faut soustraire une ligne (l'entête) et diviser le nombre de lignes par deux car chaque ligne pleine est séparée de la suivante par une ligne vide. L'essentiel

des colonnes correspondent à une période chronologique : de 1960 à 2022. Les autres fournissent des informations à propos des pays (Country Name ; Country Code) et leur PIB (Indicator Name ; Indicator Code) – notons que tout est indiqué en dollar courant, il n'y a pas de monnaie domestique.

Les premières décennies souffrent d'un cruel manque de données mais plus nous avançons dans le temps, plus le jeu est complet.

2.1.4 Jeu de GitHub

Le jeu présente 6 colonnes pour 245 lignes. Une colonne *Country* pour le nom des pays, les colonnes *Alpha-2 code*, *Alpha-3 code*, *Numeric code* correspondent à des codes prévus par la norme ISO 3166-1 et *Latitude (average)* et *Longitude (average)* pour les coordonnées géographiques. Le niveau de propreté est remarquable, sans donnée manquante ou fautive.

2.1.5 Enrichissement Wikidata sur la population

Dans l'optique d'enrichir nos données *via* Wikidata, nous avons mis au point une requête SPARQL à même de renvoyer le nombre d'habitants de l'ensemble des pays du monde sur une période de trente ans (1993 - 2003) :

```
SELECT ?paysLabel ?population ?date ?cio
WHERE
{
 ?pays wdt:P31 wd:Q6256.
 ?pays wdt:P984 ?cio.
 ?pays p:P1082 ?populationStatement.
 ?populationStatement ps:P1082 ?population.
```

```
  ?populationStatement pq:P585 ?date.
  FILTER(YEAR(?date) >= (YEAR(NOW()) - 30)).
  SERVICE wikibase:label { bd:serviceParam wikibase:language
      "[AUTO_LANGUAGE],fr". }
}
ORDER BY ?paysLabel ?date
```

Être en mesure de requêter l'intégralité des pays a été la première étape de la construction de notre requête. Le premier triplet permet de spécifier la nature – notée `wdt:P31` – de notre variable inconnue `?pays` en la faisant correspondre à l'objet `country` – noté `wd:Q6256` :

```
?pays wdt:P31 wd:Q6256.
```

Le deuxième triplet constitue un ajout tardif, destiné à faciliter les jointures avec les autres jeux de données. Comme nous travaillons sur les Jeux Olympiques, les pays peuvent être non seulement identifiés par les codes basés sur la norme ISO 3166 mais aussi par les codes du CIO (Comité International Olympique). Nous avons donc récupéré cette donnée grâce à la propriété Wikidata correspondante – notée `wdt:P984` :

```
?pays wdt:P984 ?cio.
```

La deuxième partie de notre requête doit renvoyer le nombre d'habitants par pays en prenant en compte une dimension chronologique. La complexité de cette demande requiert un parcours de graphique en quatre temps.

Pour ce faire, nous avons consulté la liste de préfixes [5] de Wikidata afin de

5. *Wikidata prefixes (E49) - Wikidata.* URL : `https://www.wikidata.org/wiki/`

créer une nouvelle variable, `?populationStatement`. Le troisième triplet a recours à la classe `population` – notée `p:P1082` – et signifie que la variable `?populationStatement` a pour valeur la population des pays :

```
?pays p:P1082 ?populationStatement.
```

L'obtention du nombre d'habitants a nécessité le recours au préfixe `ps`, lequel permet de récupérer la valeur de la propriété relative à la population :

```
?populationStatement ps:P1082 ?population.
```

Enfin, nous avons rédigé le dernier triplet sur la base du préfixe `pq` pour attribuer la valeur chronologique à la variable `?date`. Un filtre y a été appliqué pour exprimer les limites extrêmes de notre période, soit `NOW` pour 2023 et `-30` pour 1993 :

```
?populationStatement pq:P585 ?date.
FILTER(YEAR(?date) >= (YEAR(NOW()) - 30)).
```

La première et dernière ligne permettent de cadrer l'affichage des résultats. Lorsque la requête s'exécute, Wikidata renvoie le nom de chaque pays (`?paysLabel`) par ordre alphabétique, le nombre d'habitants (`?population`) et l'année correspondante (`?date`) par ordre croissant :

```
SELECT ?paysLabel ?population ?date
ORDER BY ?paysLabel ?date
```

Nous nous sommes heurtés à plusieurs obstacles avant de rendre la re-

`EntitySchema:E49` (visité le 20/01/2024).

quête fonctionnelle. Il nous a fallu un certain temps avant de comprendre la nécessité d'un parcours de graphique en quatre temps et du recours à une variable telle que `?paysStatement`. À cet égard, la documentation sur l'utilisation des propriétés `population` [6] et `date` [7] a été d'un grand secours.

Nous avons également pris en exemple la requête *Population in Europe after 1960.* [8] La consulter a été l'occasion d'une meilleure compréhension des propriétés Wikidata ainsi qu'une porte ouverte à la lecture de la documentation et à l'appréhension des requêtes en deux temps (collecte des propriétés d'une classe puis requêtage en fonction de notre besoin).

2.2 Jeux du second *flow*

Notre deuxième *flow* repose sur la comparaison entre les investissements de la France et du Royaume-Uni, en l'occurrence à l'endroit des infrastructures de natation. En amont du traitement, nos jeux de données étaient au nombre de trois et présentent les informations suivantes :

— Investissements dans les infrastructures sportives en France [9] ;

— Investissements dans les infrastructures sportives au Royaume-Uni. [10]

6. *Population*. URL : `https://www.wikidata.org/wiki/Property:P1082` (visité le 20/01/2024).
7. *Point in time*. URL : `https://www.wikidata.org/wiki/Property:P585` (visité le 20/01/2024).
8. *Wikidata Query Service*. URL : `https://w.wiki/8uHH` (visité le 20/01/2024).
9. *Data ES - Base de données*. URL : `https://equipements.sports.gouv.fr/explore/dataset/data-es/information/` (visité le 17/01/2024).
10. *Active Places Power*. URL : `https://www.activeplacespower.com/` (visité le 20/01/2024).

2.2.1 Jeu du ministère des sports et des jeux olympiques et paralympiques

Le jeu présente 114 colonnes pour 143192 lignes. Par souci de concision, nous ne nous attarderons que sur les données qui seront conservées. Les colonnes concernées sont *Type d'équipement sportif* et *Commune Ancienne*, laquelle renseigne les codes postaux.

Sur l'ensemble du jeu, les données sont grossièrement saisies. Les *Type d'équipement sportif* sont par exemple au nombre de 354 sans harmonisation commune dans la nomenclature avec présence de données aberrantes (citons « 3 2 », « PETIT MORNAS » ou encore « Puy »). Nous nous appuierons sur la colonne *Commune Ancienne* non pas sur la colonne *Région Nom* car les données sont également mal saisies et aberrantes (« CC du Pays Châtillonnais », « 4.84405 », etc.).

2.2.2 Jeu de Sport England

À l'endroit du Royaume-Uni : deux jeux nous intéressent. Le premier, nommé *Geographics* présente – sur 25 colonnes et 118655 lignes – des données géographiques à propos d'infrastructures sportives. Chaque infrastructure possède son ID (*FacilityID*) et est rattachée à une région. Les autres colonnes fournissent des informations à propos des régions, selon un principe de colonne jumelles avec d'une part les ID, d'autre part les noms (Parliamentary Constituency ID ; Ward, Local Authority, Country, Active Partnership, Region). Deux colonnes concernent les coordonnées (*Latitude* ; *Longitude*).

Sur la propreté des données, certaines colonnes sont entièrement vides (*Metro Name* ; *Core City Name* ; *LDP Code* ; *LDP Name*) mais nous ne comp-

tons pas les conserver au cours du traitement. Nonobstant, les données des coordonnées correspondent aux infrastructures, non pas aux régions. Cela pourra avoir des conséquences sur les visualisations dans la mesure où la myriade de points ne saura faire ressortir les contours des régions.

Le second jeu se nomme *SwimmingPool* et fournit des données sur la répartition des équipements de natation au Royaume-Uni. L'essentiel des colonnes donne des détails techniques sur les infrastructures (*Length* ; *Maximum Depth* ; *Movable Floor* ; *Seating*, etc.).

L'intégralité des données sont des entiers. Les entrées vides ou équivalentes à zéro sont derechef légion. Cependant, seule la colonne *FacilityID* nous est nécessaire afin de réaliser des jointures avec le jeu *Geographics*.

3 Traitement des données

3.1 Objectif du traitement

3.2 Chaîne de traitement

3.2.1 Semaine du 1er janvier 2024

Nous avons constaté que le jeu de données du FMI est trop avare en données fiables (données manquantes, valeurs aberrantes [11]) à propos des investissements en pourcentage du PIB dans le domaine du sport. Nous avons donc fait le choix d'utiliser les valeurs absolues [12] que nous croiserons avec le PIB par habitant de chaque pays.

11. Donner des exemples.
12. ?

Pour ce faire, nous avions de prime abord songé à rédiger une autre requête SPARQL. Nonobstant, le résultat ne concernait que les années 2021 et 2022 – les années précédentes sont vraisemblablement absentes de Wikidata.

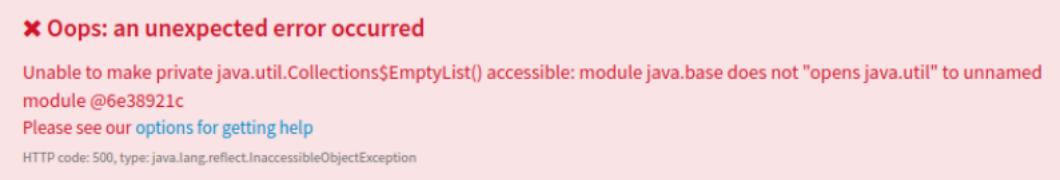
Nous avons donc dû trouver un jeu de données contenant le PIB par habitant. Parmi les données sources de Wikidata se trouvait un jeu hébergé sur le site de la Banque mondiale. Les données correspondent à une profondeur temporelle suffisante pour le cadre de notre projet. Cela nous a permis de résoudre le problème de mauvaise qualité du jeu du FMI.

```
import dataiku
import pandas as pd
from dataiku import pandasutils as pdu

FMI_prepared2 = dataiku.Dataset("FMI_prepared2")
FMI_prepared2_df = FMI_prepared2.get_dataframe()

colonnes_a_traiter = ['1996 - FMI', '2000 - FMI', '2004 - FMI',
    '2008 - FMI', '2012 - FMI', '2016 - FMI']

for colonne in colonnes_a_traiter:
 condition = FMI_prepared2_df[colonne] > 0.1
 FMI_prepared2_df.loc[condition, colonne] /= 10

FMI_prepared2_df.drop_duplicates(subset=['Pays'], keep='first',
    inplace=True)

FMI_Python = dataiku.Dataset("FMI_Python")
```

```
FMI_Python.write_with_schema(FMI_prepared2_df)
```

3.2.2 Semaine du 22 janvier 2024

Mauvaise surprise : alors que nous voulions produire un autre flow afin de remplir la seconde partie de notre objectif et produire des datavisualisations comparant la France et le Royaume-Uni, un bogue manifestement causé par une installation relative à Javascript perturbe l'affichage sur Dataiku. Probablement : problèmes de compatibilité.



ll est probable que cette erreur soit apparue après l'installation d'une version de JDK supérieure à celle prise en charge par Dataiku. Il aurait ainsi fallu rétrograder notre version de JDK, mais nous craignons que cette manipulation entraîne des difficultés dans d'autres projets que nous menons. Nous ne pouvons dès lors nous appuyer que sur un seul ordinateur, qui possède une version inférieure de JDK, ce qui a causé d'importants soucis organisationnels à la fin du mois de janvier.

4 Visualisation des données

4.1 Présentation des visualisations

4.2 Analyse

Qu'apprend-on en regardant les visualisations ? Quels sont les biais ?

5 Sitographie

120 years of Olympic history : athletes and results. URL : https://www.kaggle.com/datasets/heesoo37/120-years-of-olympic-history-athletes-and-results (visité le 17/01/2024).

Active Places Power. URL : https://www.activeplacespower.com/ (visité le 20/01/2024).

Data ES - Base de données. URL : https://equipements.sports.gouv.fr/explore/dataset/data-es/information/ (visité le 17/01/2024).

Download entire World Economic Outlook database, April 2019. IMF. URL : https://www.imf.org/en/Publications/WEO/weo-database/2023/October/download-entire-database (visité le 17/01/2024).

Paul Copil, Alexandru. *Countries codes and coordinates*. URL : https://gist.github.com/cpl/3dc2d19137588d9ae202d67233715478 (visité le 25/01/2024).

Point in time. URL : https://www.wikidata.org/wiki/Property:P585 (visité le 20/01/2024).

Population. URL : https://www.wikidata.org/wiki/Property:P1082 (visité le 20/01/2024).

Wikidata prefixes (E49) - Wikidata. URL : https://www.wikidata.org/wiki/EntitySchema:E49 (visité le 20/01/2024).

Wikidata Query Service. URL : https://w.wiki/8uHH (visité le 20/01/2024).

World Bank Open Data. World Bank Open Data. URL : https://data.worldbank.org (visité le 20/01/2024).

Table des matières