



DATA SCIENTEST

MODÈLE DE PRÉDICTION DES RECORDS D'ATHLÉTISME

JEUX OLYMPIQUES

2024



Auteurs :

Armen ASATRYAN
Yannick OREAL
Yves-Marie SALAUN
Xavier GERVAIS

Managers :

Jérémy ROBERT
Raja AARAB

RAPPORT D'EXPLORATION DES DONNÉES

Projet réalisé dans le cadre de la formation
DATA SCIENTIST BOOTCAMP d'avril 2024

ATHLÉTISME DU 01 AU 11 AOÛT 2024



Sommaire

World of Athletics	<u>02</u>
Scrapping de la base	<u>03</u>
Requêtes GraphQL	<u>04</u>
Scripts et Format des données	<u>05</u>
Qualité des données	<u>06</u>
Wikipedia	<u>07</u>
Scrapping	<u>08</u>
Kaggle	<u>09</u>
Qualité Wikipedia & Kaggle	<u>10</u>
Météo	<u>11</u>
Sites choisis	<u>12</u>
Traitement des données Météo	<u>14</u>
Qualité Météo	<u>15</u>

Le GitHub des sources utilisées

[GitHub sources Permalink](#)



World Athletics Présentation du Site

Les JO et plus encore...

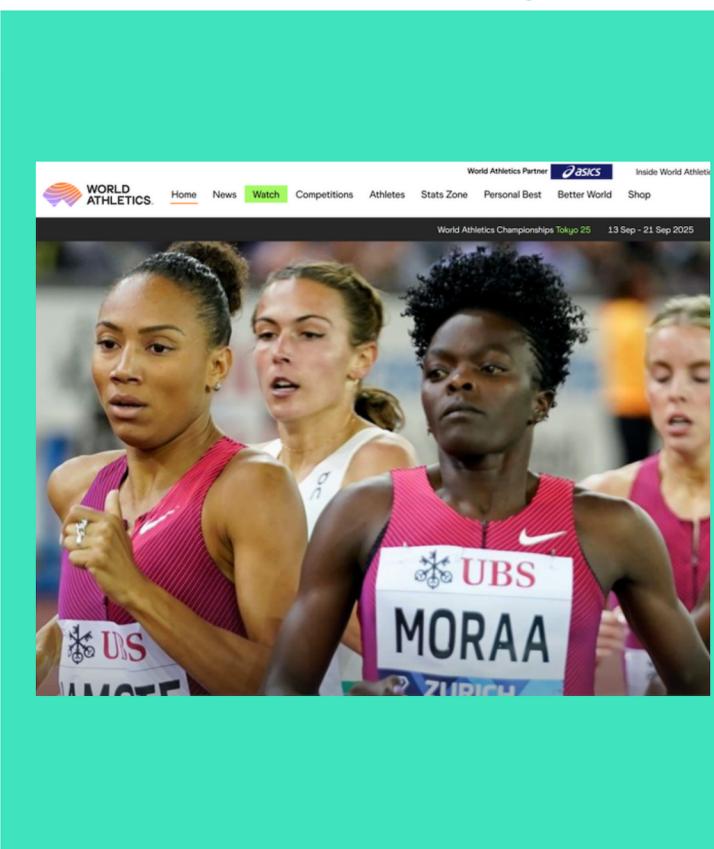


Récupération des données de la Base

Le site worldathletics.org est une base de données officielle exhaustive des épreuves et résultats de tous types de compétitions d'Athlétisme dans le monde. Aussi bien les JO que les mondiaux et les compétitions nationales.

Pour les JO la base couvre tous les résultats, même les phases de qualification, depuis 1996. Pour les Mondiaux, depuis 1983.

Ces données seront essentielles pour l'entraînement de notre modèle.



World Athletics

Scraping de la base

Calendar / Results

SEARCH

FILTER

DD/MM/YYYY DD/MM/YYYY Region Type Discipline Rankings Category Permit

Show only results Yes

This long-term calendar contains all competitions held under World Athletics. Area or National Member Federation Permit as notified to World Athletics through a prior and appropriate application process. The competitions are understood as conducted according to relevant World Athletics Competition and Technical Rules and the published results, unless noted otherwise, are recognised by World Athletics for all its statistical purposes including World Rankings and qualification to major Championships. For the avoidance of doubt, competitions which are not listed, are not recognised by World Athletics. Calendar events are subject to change.

DATE ↑↓	NAME ↑↓	↑↓	VENUE ↑↓	COUNTRY ↑↓	CAT. ↑↓	DISCIPLINE ↑↓	COM.
08 MAY 2024	Frühjahr-Fitness-Check		Bagsværd Stadion, Bagsvaerd (DEN)	DEN	F	Track and Field	
08 MAY 2024	17. Ročník Hvězdného Házení		Městský stadion Střelnice, Domatice (CZE)	CZE	C	Track and Field	World Tour

Disciplines

Les disciplines choisies par nos soins sont celles prévues aux JO 2024.

Nous n'avons retenu que les disciplines mono-compétiteurs.ses (pas de relais...) et mono-épreuves (pas de décathlon...) pour améliorer la fiabilité du modèle.

The screenshot shows the Network tab of a browser's developer tools. It lists numerous XHR (XMLHttpRequest) requests made to the domain `worldathletics.org`. The requests include various endpoints such as `/api/cdn/calical`, `/api/cdn/calical`, `/calendar/apiKey=g0CAbyMg/d4t5o2b8kPl7caNkWEm/00kg.4MGES2Y6515c28path=Fixture/IBCAL`, and `/calendar/apiKey=g0CAbyMg/d4t5o2b8kPl7caNkWEm/00kg.4MGES2Y6515c28path=Fixture/IBCAL`. The requests are primarily GET and POST methods, with some OPTIONS and graphQL requests interspersed. The status codes for most requests are 200 OK, while others are 404 Not Found. The file names listed in the 'Fichier' column include `4e6387e73e4fe6e6e51d9.cs`, `aef97646ddca1a50651.cs`, `339a7ba9e9e58cae432d.cs`, and `lqf74cdt2j7tktozacymhcvbe.appspot..`.

Evènements

15 400

Des compétitions mondiales, JO et Mondiaux, jusqu'aux compétitions nationales.

The screenshot shows the Network tab of a browser's developer tools. It lists multiple XHR (XMLHttpRequest) requests made to the domain `worldathletics.org`. The requests include various endpoints such as `/api/cdn/calical`, `/api/cdn/calical`, `/calendar/apiKey=g0CAbyMg/d4t5o2b8kPl7caNkWEm/00kg.4MGES2Y6515c28path=Fixture/IBCAL`, and `/calendar/apiKey=g0CAbyMg/d4t5o2b8kPl7caNkWEm/00kg.4MGES2Y6515c28path=Fixture/IBCAL`. The requests are primarily GET and POST methods, with some OPTIONS and graphQL requests interspersed. The status codes for most requests are 200 OK, while others are 404 Not Found. The file names listed in the 'Fichier' column include `4e6387e73e4fe6e6e51d9.cs`, `aef97646ddca1a50651.cs`, `339a7ba9e9e58cae432d.cs`, and `lqf74cdt2j7tktozacymhcvbe.appspot..`.

The screenshot shows the Network tab of a browser's developer tools. It lists a single POST request made to the URL `https://lqf74cdt2j7tktozacymhcvbe.appspot-api.eu-west-1.amazonaws.com/graphQL`. The request header includes `X-amzn-user-agent: aws-lambda/3.0.2` and `x-api-key: da2-jgtxe5f2tvcafnp5zh5kmq5oq`. The response status code is 200 OK.

01. Identifier la structure

L'explorateur du navigateur a révélé que la base était organisée via un système de requêtes GRAPHQL.

Onglet Réseau / XHR

02. Les clés de la base

L'examen des headers nous a donné les clés du royaume :

- l'url de requête d'accès POST
- l'x-api-key qui autorise les requêtes

World Athletics

Requêtes GraphQL

Process

L'extension GraphQL playground pour chrome a permis de préciser les schémas des requêtes du site.

Il restait à récupérer les arguments de la requête :

- eventId : pour les évènements
 - competitionId : pour les disciplines

CompetitionId

La liste des competitionId (disciplines qui nous intéressent) a été obtenue par simple examen de la page web des JO de Tokyo 2020.

eventId

La liste des eventIds a été plus compliquée, n'étant pas disponible de cette façon, un script Selenium a été utilisé : scrap_eventid.py

Résultats de requêtes

Il a fallu ensuite extraire les données avec un script python contenant la requête query et les stocker dans des JSON.

Transformation en csv

Dernière étape la transformation des fichiers JSON en CSV pour passer au preprocessing.

The screenshot shows a Java code editor with a search results panel open. The search term is "Results". The results list includes several methods and fields:

- `id: int`
- `urlSlug: String`
- `ids: [Int]`
- `resultsByLimitsDiscipline: String`
- `resultsByLimitsStartDate: String`
- `resultsByLimitsEndDate: String`
- `resultsByLimitsOnlyRegular: Boolean`
- `): resultsByLimit`
- `Query.getCalendarEvents(`
 - `startDate: String`
 - `endDate: String`
 - `query: String`
 - `regionType: String`
 - `regionId: Int`
 - `disciplineId: Int`
 - `rankingCategoryId: Int`
 - `permitLevelId: Int`
 - `currentSeason: Boolean`
 - `competitionGroupId: Int`
 - `competitionGroupSlug: String`
 - `competitionSubgroupId: Int`
- `getCalendarCompetitionResults(`
 - `competitionId: Int`
 - `day: Int`
 - `eventId: Int`

Below the results list, there are two sections: `TYPE DETAILS` and `ARGUMENTS`, each containing a list of items.

Scripts et Format des données

Les scripts utilisés :

01. script Selenium eventId

[Lien vers scrap_eventid.py](#)

03. JSON to CSV

[Lien vers json_to_csv.py](#)

02. script GraphQL

[Lien vers scrap_GraphQL_ids.py](#)

Analyse :

3Go JSON

Données brutes

600k fichiers scrapés.

1 800 000 lignes

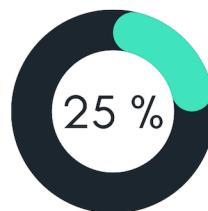
Athlètes et enregistrements

Notre base principale avec leurs temps et les records battus

24 colonnes

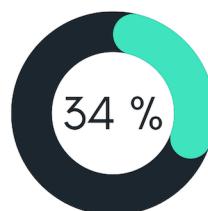
Nom, dates, temps...

12 colonnes seront conservées



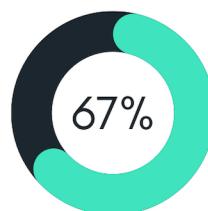
Fichiers utiles

3/4 des compétitions n'ont pas les disciplines que nous cherchions.
150k / 600k



Taux de NA

Les NA sont principalement dans la colonne records ou des colonnes inutiles



Preprocessing

Travail Prévisionnel :

- dichotomisation des Records
- suppressions des compétitions inutiles

Matrices d'exploration :

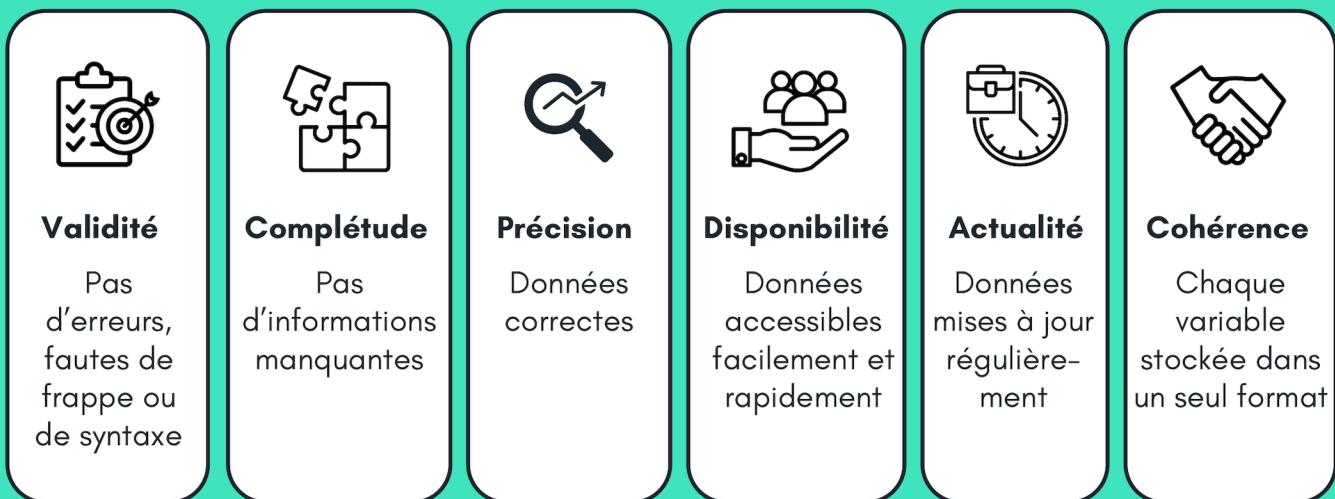
[Lien drive](#)



World Athletics

Qualité des Données

La matrice à respecter



Analyse

- 01. Validité**
- 02. Complétude**
- 03. Precision**
- 04. Disponibilité**
- 05. Actualité**
- 06. Cohérence**

Les données World Athletics sont :

réputées, valides et sans erreurs.

complètes pour celles disponibles (1983-)

celles mesurées sur place par les officiels.

disponibles en consultation libre sur le site.

prises live et reportées intégralement.

présentes en format unique.



Wikipedia

Présentation du Site

Le dictionnaire online

Athletics at the Summer Olympics

文 A 43 languages ▾

Article Talk

Read Edit View history Tools ▾

From Wikipedia, the free encyclopedia

Athletics has been contested at every [Summer Olympics](#) since the birth of the modern Olympic movement at the [1896 Summer Olympics](#). The athletics program traces its earliest roots to events used in the [ancient Greek Olympics](#). The modern program includes [track and field](#) events, [road running](#) events, and [race walking](#) events. [Cross country running](#) was also on the program in earlier editions but it was dropped after the [1924 Summer Olympics](#).

Summary [edit]

Games	Year	Events	Best Nation	Games	Year	Events	Best Nation
1	1896	12	United States	17	1960	34	United States
2	1900	23	United States	18	1964	36	United States
3	1904	25	United States	19	1968	36	United States
4	1908	26	United States	20	1972	38	Soviet Union
5	1912	30	United States	21	1976		
6				22	1980		

Athletics at the Summer Olympics



IOC ATH

Discipline

Code

Governing IAAF

body

Events 48 (men: 24; women: 23; mixed: 1)

Récupération des données

La structure est une page par année, avec plusieurs tableaux indiquant selon les années :

- le récapitulatif des médailles par pays
- le récapitulatif des records battus (mais pas systématiquement)
- les résultats / performances des hommes ou des femmes (à partir de 1928 pour ces dernières) avec les records mais pas systématiquement.

Sur les tableaux de résultats / performances, nous avons un lien "edit" permettant de récupérer les informations principales :

- le nom de l'épreuve
- le nom des médaillés
- les performances (temps ou distance) parfois avec les records battus (OR Olympic Record et WR World Record)

Event	Gold		Silver		Bronze	
	details	Holder	Holder	Holder	Holder	Holder
100 metres	Marcell Jacobs Italy	9.80 AR	Fred Kerley United States	9.84 PB	Andre De Grasse Canada	9.89 PB
200 metres	Andre De Grasse Canada	19.62 NR	Kenny Bednarek United States	19.68 PB	Noah Lyles United States	19.74 SB
400 metres	Steven Gardiner Bahamas	43.85 SB	Anthony Zambrano Colombia	44.08	Kirani James Grenada	44.19
800 metres	Emmanuel Korir Kenya	1:45.06	Ferguson Rotich Kenya	1:45.23	Patrik Dobek Poland	1:45.39
1500 metres	Jakob Ingebrigtsen Norway	3:28.32 OR, AR	Timothy Cheruiyot Kenya	3:29.01	Josh Kerr Great Britain	3:29.05 PB
5000 metres	Joshua Cheptegei Uganda	12:58.15	Mohammed Ahmed Canada	12:58.61	Paul Chelimo United States	12:59.05 SB
10,000 metres	Selemon Barega Ethiopia	27:43.22	Joshua Cheptegei Uganda	27:43.63	Jacob Kiplimo Uganda	27:43.88
110 metres hurdles	Hansle Parchment Jamaica	13.04 SB	Grant Holloway United States	13.09	Ronald Levy Jamaica	13.10
400 metres hurdles	Karsten Warholm Norway	45.94 WR	Rai Benjamin United States	46.17 AR	Alison dos Santos Brazil	46.72 AR
3000 metres steeplechase	Soufiane El Bakkali Morocco	8:08.90	Lamecha Girma Ethiopia	8:10.38	Benjamin Kigen Kenya	8:11.45
4 × 100 metres relay	Italy Lorenzo Patta Marcell Jacobs Fausto Desalu Filippo Tortu	37.50 NR	Canada Aaron Brown Jerome Blake Brendon Rodney Andre De Grasse	37.70 SB	China Tang Xingqiang Xie Zhenye Su Bingtian Wu Zhijiang	37.79 NR



Scrappling et formatage des données

Process

Le format du document est copié manuellement (avec balises html et balises propres au formatage de wikipedia)

Le formatage des pages et des tableaux diffèrent d'année en année (difficile dès lors d'automatiser, cela est dû aux nombreux intervenants qui modifient les pages wikipedia).

Copié collé

Copié/collé manuel du code balisé dans un fichier texte (par année, par genre)

Macro

Utilisation d'une macro sous notepad++ permettant de retirer le plus de caractères de balisage et ajouter le caractère séparateur (;)

Q&A

Traitement manuel des données pour dupliquer le nom de l'épreuve devant chaque nom d'athlète.

Retrait des derniers caractères parasites.

Ajout des informations comme l'année, le pays, la ville, les records battus ou égalés, ces informations se trouvant dans la page "détails" de chaque épreuve (les records battus ou égalés WR / OR / EWR / EOR)

```
<{{FlagIOCmedalist|[[Hero McKenley]]|JPN|1948 Summer}}||46.4
|{{FlagIOCmedalist|[[Mal Whitfield]]|USA|1948 Summer}}||46.8
|
| 800 metres<br/>{{DetailsLink|Athletics at the 1948 Summer Olympics - Men's 800 metres}}
| {{FlagIOCmedalist|[[Mal Whitfield]]|USA|1948 Summer}}||1:49.2
| {{FlagIOCmedalist|[[Arthur Wint]]|JAM|1948 Summer}}||1:49.5
| {{FlagIOCmedalist|[[Marcel Hansenne]]|FRA|1948 Summer}}||1:49.8
|
| 1500 metres<br/>{{DetailsLink|Athletics at the 1948 Summer Olympics - Men's 1500 metres}}
| {{FlagIOCmedalist|[[Henry Eriksson]]|SWE|1948 Summer}}||3:49.8
| {{FlagIOCmedalist|[[Lennart Strand]]|SWE|1948 Summer}}||3:50.4
| {{FlagIOCmedalist|[[Wim Slijkhuis]]|NED|1948 Summer}}||3:50.4
|
| 5000 metres<br/>{{DetailsLink|Athletics at the 1948 Summer Olympics - Men's 5000 metres}}
| {{FlagIOCmedalist|[[Gaston Reiff]]|BEL|1948 Summer}}||14:17.6
| {{FlagIOCmedalist|[[Emil Zátopek]]|TCH|1948 Summer}}||14:17.8
| {{FlagIOCmedalist|[[Wim Slijkhuis]]|NED|1948 Summer}}||14:26.8
|
| 10,000 metres<br/>{{DetailsLink|Athletics at the 1948 Summer Olympics - Men's 10,000 metres}}
| {{FlagIOCmedalist|[[Emil Zátopek]]|TCH|1948 Summer}}||29:59.6
| {{FlagIOCmedalist|[[Alain Mimoun]]|FRA|1948 Summer}}||30:47.4
```

```
<Macro name="clean_scapping" Ctrl="no" Alt="no" Shift="no" Key="0">
<Action type="3" message="1700" wParam="0" lParam="0" sParam="" />
<Action type="3" message="1601" wParam="0" lParam="0" sParam=";" />
<Action type="3" message="1625" wParam="0" lParam="0" sParam=";" />
<Action type="3" message="1602" wParam="0" lParam="0" sParam=";" />
<Action type="3" message="1702" wParam="0" lParam="768" sParam="" />
<Action type="3" message="1701" wParam="0" lParam="1609" sParam="" />
<Action type="3" message="1700" wParam="0" lParam="0" sParam="" />
<Action type="3" message="1601" wParam="0" lParam="0" sParam=";" />
<Action type="3" message="1625" wParam="0" lParam="0" sParam=";" />
<Action type="3" message="1602" wParam="0" lParam="0" sParam="" />
<Action type="3" message="1702" wParam="0" lParam="768" sParam="" />
<Action type="3" message="1701" wParam="0" lParam="1609" sParam="" />
<Action type="3" message="1700" wParam="0" lParam="0" sParam="" />
<Action type="3" message="1601" wParam="0" lParam="0" sParam=";" />
<Action type="3" message="1625" wParam="0" lParam="0" sParam=";" />
<Action type="3" message="1602" wParam="0" lParam="0" sParam="" />
<Action type="3" message="1702" wParam="0" lParam="768" sParam="" />
<Action type="3" message="1701" wParam="0" lParam="1609" sParam="" />
<Action type="3" message="1700" wParam="0" lParam="0" sParam="" />
```

Derniers traitements

Concaténation des fichiers csv obtenus par une commande dos :

copy /b *.csv result.csv

Correction des erreurs (lors de la saisie des informations qui n'étaient pas initialement sur les tableaux scrappés)

Adaptation manuelle de certaines valeurs pour permettre le "perfect matching", notamment le prenom/nom des athlètes car de nombreux diminutifs étaient utilisés, des codes NCO non officiels, et cela afin d'harmoniser avec le dataset kaggle

Découpage du fichier en deux csv distincts séparant les épreuves par type de performances (temps mesuré en h/m/s/ms, distance ou hauteur mesuré en mètres)

120 years of Olympics & Olympics 2020

Les datasets

athlete_events.csv (41.5 MB)						
	Detail	Compact	Column			
1	ID	Name	Sex	Age	Height	
1	134732 unique values	M F Other (227521)	73% 27% 84%	23 24 Other (198453)	8% 8% NA 180 Other (198453)	
1	A Dijiang	M		24	188	
2	A Lamusi	M		23	178	
3	Gunnar Nielsen Aaby	M		24	NA	
4	Edgar Lindenau Aabye	M		34	NA	
5	Christine Jacoba Aaftink	F		21	185	
5	Christine Jacoba Aaftink	F		21	185	
5	Christine Jacoba Aaftink	F		25	185	
5	Christine Jacoba Aaftink	F		25	185	
5	Christine Jacoba Aaftink	F		27	185	
5	Christine Jacoba Aaftink	F		27	185	
6	Per Kout Aloland	M		21	186	

Description des données

Sources :

<https://www.kaggle.com/datasets/vaibhav2025/120-years-of-olympic-history>

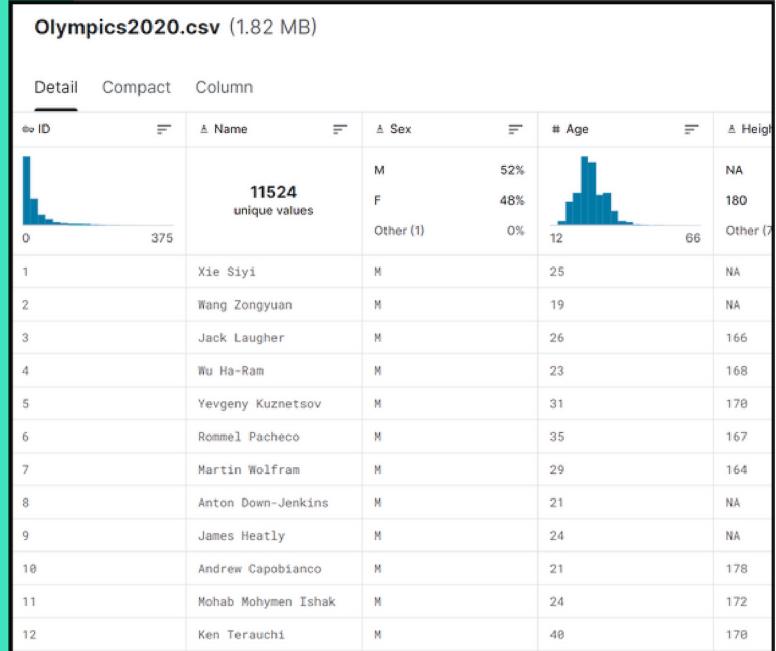
<https://www.kaggle.com/datasets/adamsharifc/olympics-2020-dataset/data>

Résultats des JO de 1896 à 2016 (avec des données qui manquaient sur le dataset wikipedia : l'âge, la taille et le poids de l'athlète)

Résultats des JO de Tokyo 2020 (ayant eu lieu en 2021 suite à la pandémie Covid-19)

Formatage des données

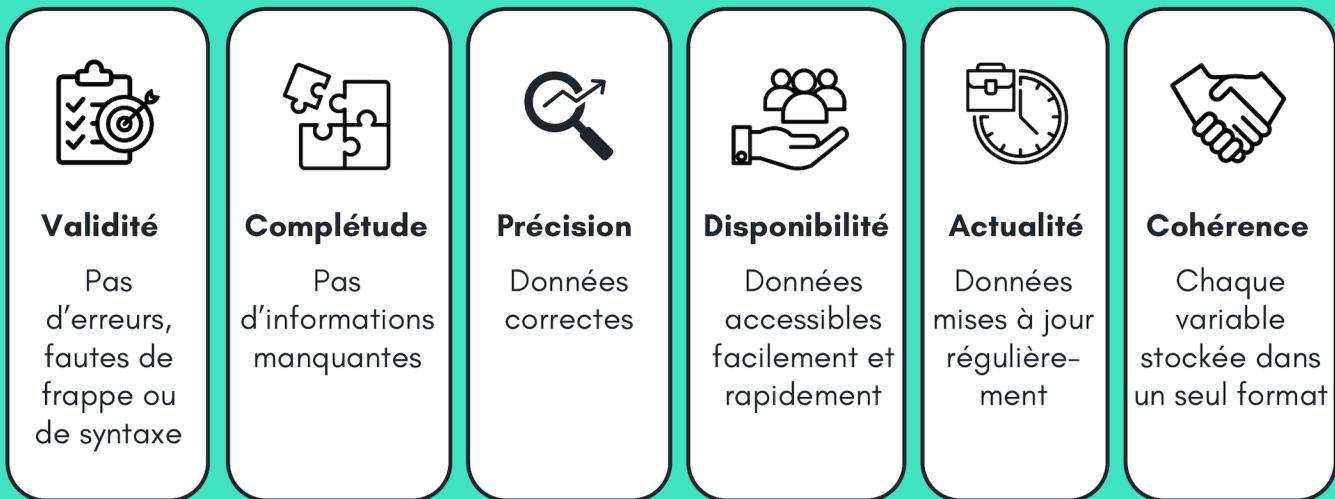
- Harmonisation des noms d'épreuves sur le dataset de 2020 (manuel)
- Suppression des codes spécifiques (DSQ par exemple) sur le dataset de 2020 (manuel)
- Concaténation des deux csv (manuel)
- Filtrage sur seulement la valeur "Athletics" de la variable "Sport" (python) et suppression du champs 'ID'
- Correction des résultats, pour mettre à jour les athlètes disqualifiés (notamment sur les épreuves 2008, 2012 et 2016) afin d'obtenir les résultats à jour (manuel)



Wikipédia & Kaggle

Qualité des Données

La matrice à respecter



Analyse

- 01. Validité**
- 02. Complétude**
- 03. Precision**
- 04. Disponibilité**
- 05. Actualité**
- 06. Cohérence**

Les données Wikipédia & Kaggle sont :

- valides et doublement vérifiées avant publication.
- Débutant en 1896 !
- bonne, pas de changements récents
- disponibles en consultation libre sur le site.
- reportées presque instantanément avant revue
- présentes en format unique.

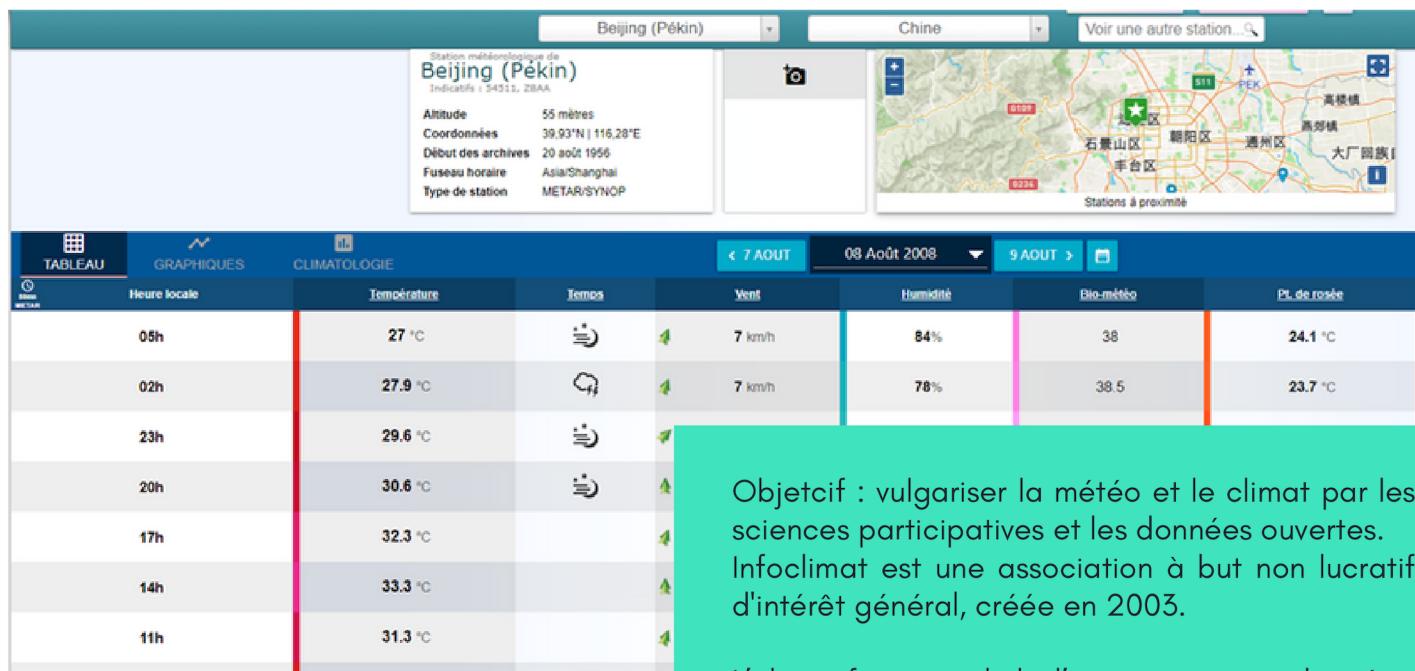


Météo

Les données environnementales

Plusieurs sites mis à contribution

Les principales mesures collectées sont : l'altitude du site accueillant l'évènement, la température du jour, la vitesse du vent et la pluie. Ces données seront analysées pour leur impact sur les performances des athlètes



infoclimat.fr

Plusieurs sources ont été utilisées afin de collecter manuellement les données météo. Les principales mesures collectées sont : l'altitude du site accueillant l'évènement, la température du jour, la vitesse du vent et la pluie.

<https://www.infoclimat.fr/>

Objectif : vulgariser la météo et le climat par les sciences participatives et les données ouvertes. Infoclimat est une association à but non lucratif d'intérêt général, créée en 2003.

L'objectif principal de l'association est de gérer le site internet www.infoclimat.fr, qui propose l'accès aux informations météorologiques mondiales, et notamment aux stations météorologiques semi-professionnelles installées par plus de 700 passionnés, qui mesurent températures, précipitations, vent, parfois depuis plusieurs décennies.

Le site internet propose plus de 6 milliards de données en temps réel et climatologiques (qu'elles soient issues des organismes officiels qui proposent de l'OpenData, de stations installées par l'association, de partenaires, ou de simples citoyens dont la qualité des relevés est contrôlée).



Météo

Les données environnementales

Plusieurs sites mis à contribution

Weatherspark

1.<https://fr.weatherspark.com/>

WeatherSpark.com fournit des rapports détaillés sur les conditions météorologiques typiques de 145 449 sites dans le monde entier.

Heure	Temp.	Alt.	Vent	Vis.	Couverture nuageuse
					Local: 15:00, ven. 9 sept. 1988 UTC: 09:00, ven. 9 sept. 1988 Source: Observation horaire de surface de l'USAF Type de rapport: Rapport météorologique régulier pour l'aviation METAR Pt de rosée: 19,0 °C Autres couches nuageuses: Dégagé dans l'ensemble (900 m)
15:00	28,0 °C	1 015 mbars	13,0 km/h, N	11,2 km	Six oktas - 7/10 - 8/10 (6 900 m)
					Local: 15:00, ven. 9 sept. 1988 UTC: 09:00, ven. 9 sept. 1988 Source: Observation horaire de surface de l'USAF Type de rapport: Rapport météorologique régulier pour l'aviation METAR Pt de rosée: 19,0 °C Autres couches nuageuses: Dégagé dans l'ensemble (900 m)
16:00	28,0 °C	1 014 mbars	13,0 km/h, N	11,2 km	Six oktas - 7/10 - 8/10 (6 900 m)
					Local: 15:00, ven. 9 sept. 1988 UTC: 09:00, ven. 9 sept. 1988 Source: Observation horaire de surface de l'USAF Type de rapport: Rapport météorologique régulier pour l'aviation METAR Pt de rosée: 19,0 °C Autres couches nuageuses: Dégagé dans l'ensemble (900 m)
16:30	28,0 °C	1 014 mbars	13,0 km/h, N	11,2 km	Nuageux dans l'ensemble (6 900 m)
					Local: 15:00, ven. 9 sept. 1988 UTC: 09:00, ven. 9 sept. 1988 Source: Observation horaire de surface de l'USAF Type de rapport: Rapport météorologique régulier pour l'aviation METAR Pt de rosée: 18,0 °C Autres couches nuageuses: One okta - 1/10 or less but not zero (900 m)
17:00	29,0 °C	1 014 mbars	9,36 km/h, NO	11,2 km	Nuageux dans l'ensemble (6 000 m)
					Local: 17:00, ven. 9 sept. 1988 UTC: 07:00, ven. 9 sept. 1988

MSN Météo

1.<https://www.msn.com/fr-fr/meteo>

Plateforme météo du moteur de recherche Microsoft Start qui contient les données météo mondiales jusqu'à l'année 1951.



Informations sur les températures	Comme mensuel	Toutes les années	Résumé (Comme mensuel)	Max	Moyenne	Min
Jour le plus chaud	12 juil.	1 juil.	Haute température (°C)	33	26	20
Jour le plus froid	8 juil.	10 juil.	Basse température (°C)	21	17	15
Jour le plus humide	1 juil.	2 juil.	Precipitation (mm)	2,76	0,51	0
Jour le plus venteux	17 juil.	27 juil.	Vent (km/h)	17,4	8,85	4,4



Météo

Les données environnementales

Plusieurs sites mis à contribution

Données climatiques: Novembre 1956																
Jour	T	TM	Tm	SLP	H	PP	VV	V	VM	VG	RA	SN	TS	FG		
1	16.1	20.6	11.1	-	68	0	19.8	16.7	27.8	-	0					
2	11.6	18.3	7.8	-	60	3.05	26.1	17.4	27.8	-	0					
3	10.8	14.4	6.7	-	68	2.03	24.3	17	22.2	-	0					
4	11.2	13.9	7.8	-	65	2.03	24.3	21.3	33.5	-	0					
5	10.8	15.6	5	-	58	0.51	23.2	9.4	22.2	-						
6	12.7	16.1	4.4	-	69	0	20	10.9	22.2	-						
7	13.7	17.2	9.4	-	67	3.05	23.2	8.5	20.6	-						
8	14.8	19.4	10	-	58	0	23.2	9.1	24.1	-						
9	17.9	25	8.3	-	58	-	18	18.7	31.3	-	0					
10	10.4	16.7	7.8	-	66	19.05	24.3	20.6	37	-						
11	14.1	17.8	6.7	-	73	1.02	26.2	14.3	29.4	-	0					
12	14.9	18.9	11.1	-	71	0.25	27.7	7.4	18.3	-						
13	13.8	17.2	11.1	-	70	0	25.7	12.4	18.3	-	0					
14	13.2	17.8	9.4	-	69	1.02	25.3	10.6	22.2	-						
15	16.1	21.1	8.9	-	85	0	21.7	14.1	37	-	0					

Tutiempo.net

<https://fr.tutiempo.net/>

Plateforme d'information climatique de tous les pays du monde entier avec des données historiques qui remontent à 1929 dans certains cas.

Weatherclimate.eu

<https://www.weatherandclimate.eu/>

Plateforme qui offre de données météorologiques moyennes par mois avec une profondeur historique importante. Ses données sont basées sur les données disponibles en ligne et les sources de la littérature. Cette source a été utilisée surtout pour récupérer les données des événements très anciens.

HISTORIQUE MÉTÉO À HELSINKI

Températures moyennes mensuelles et annuelles de l'air in Helsinki

(selon les données en ligne et les sources de la littérature)

année	jan	fév	mar	avr	may	jun	Jul	aoû	sep	oct	nov	déc	année
1829	-11.3	-14.0	-8.9	-1.7	7.5	14.1	17.7	14.5	12.1	4.5	-3.5	-6.5	2.0
1830	-8.9	-8.5	-2.7	1.1	5.5	12.7	16.3	14.4	10.5	6.2	2.6	-3.1	3.8
1831	-11.2	-5.2	-6.8	2.4	8.8	16.3	18.3	15.6	9.0	5.7	1.0	-3.8	4.2
1832	-4.5	-1.0	-1.7	2.6	6.6	13.7	13.5	14.4	9.0	6.8	0.2	-2.5	4.8
1833	-3.4	-5.0	-5.3	0.4	7.8	14.1	16.4	12.3	11.8	8.1	3.4	-4.4	4.7
1834	10.7	-3.6	-1.6	2.7	8.3	14.6	16.8	18.8	9.7	5.3	-1.1	-3.4	8.4
1835	-3.2	-2.1	-1.6	0.9	6.3	14.7	14.9	13.0	12.1	6.0	-2.8	-10.0	4.0
1836	-7.2	-4.1	0.6	3.2	6.7	11.7	14.4	13.7	9.0	6.7	-0.7	-4.1	4.2
1837	-7.2	-2.5	-5.8	0.6	8.2	14.1	14.7	16.2	10.2	4.8	2.8	-4.0	4.4
2014	-7.4	-0.4	1.6	5.7	10.9	13.6	20.3	17.4	12.3	5.8	2.2	-1.4	8.7
2015	-2.0	-0.1	1.8	5.1	9.6	13.7	16.3	17.3	12.7	5.4	4.6	2.1	7.2
2016	-10.3	-0.6	0.0	4.8	14.3	15.6	17.7	16.0	12.7	4.6	-1.0	-1.0	8.1
2017	-2.9	-3.2	0.6	2.3	10.3	13.8	16.1	16.0	11.2	5.0	2.7	0.7	6.1
2018	-2.4	-8.5	-4.8	5.0	15.3	15.8	21.2	18.3	13.3	6.6	2.7	-2.0	6.7
2019	-6.2	-0.5	-0.3	8.7	10.6	18.0	17.5	16.8	11.2	5.2	1.7	1.4	6.8
2020	1.9	0.5	1.7	4.5	9.5	18.4	16.5	16.8	13.0	8.2	4.3	0.9	8.0
2021	-4.6	-8.2	-0.8	4.5	10.3	19.1	20.9	15.4	9.4	8.0	1.6	-6.8	5.7
2022	-3.3	-2.9	-0.3	3.2	9.6	16.9	18.1	19.0	9.2	7.6	2.3	-3.1	8.4
2023	-1.6	-2.4	-1.7	5.6	10.7	16.3	16.6	17.2	14.9	4.5	-0.4	-4.8	6.2
2024	-8.5	999.9	999.9	999.9	999.9	999.9	999.9	999.9	999.9	999.9	999.9	999.9	999.9



Les données

Répartition des données par source

temperature_C - Temperature
wind_km_h - Vitesse du vent
weather - Pluie constatée (rain / no_rain)

Priorité	Site web	Nb de temperature_C	Nb de wind_km_h	Nb de weather
1	https://www.infoclimat.fr/	422	214	405
2	https://fr.weatherspark.com/	32	32	32
3	https://www.msn.com/	34	3	
4	https://fr.tutiempo.net/	17	17	17
5	https://www.weatherandclimate.eu/	304		
Total général		809	266	454

Température du jour

En fonction de la disponibilité de la donnée, nous avons collecté la température maximale ou moyenne du jour en appliquant un ordre de priorité des ressources mentionnées précédemment. La température est exprimée en degré Celsius.

Vitesse du vent

En fonction de la disponibilité de la donnée, nous avons collecté la vitesse du vent à une valeur précise, la vitesse maximale ou moyenne du jour en appliquant un ordre de priorité des ressources mentionnées précédemment.

La vitesse est exprimée en km/h.

Pluie

Dans le but de simplification de cette donnée elle est transformée en variable binaire qui indique simplement si le jour de l'évènement une pluie est constatée.

Nombre total de jours de JO

32 évènement de 1896 au 2021, 20 pays et 23 villes d'accueil différents sur 5 continents.

Country	City	Continent	Total
Australia	Melbourne	Oceania	17
	Sydney	Oceania	17
Belgium	Antwerp	Europe	24
Brazil	Rio de Janeiro	South America	17
Canada	Montreal	North America	16
China	Beijing	Asia	17
East Germany	Munich	Europe	17
Finland	Helsinki	Europe	32
France	Paris	Europe	191
Germany	Berlin	Europe	17
Greece	Athens	Europe	38
Italy	Rome	Europe	18
Japan	Tokyo	Asia	48
Mexico	Mexico	North America	16
Netherlands	Amsterdam	Europe	88
South Korea	Seoul	Asia	16
Spain	Barcelona	Europe	16
Sweden	Stockholm	Europe	18
United Kingdom	London	Europe	222
United States	Atlanta	North America	17
	Los Angeles	North America	32
	Saint-Louis	North America	29
USSR	Moscow	Europe	16
Total général			939

Qualité des Données

La matrice à respecter

Validité Pas d'erreurs, fautes de frappe ou de syntaxe	Complétude Pas d'informations manquantes	Precision Données correctes	Disponibilité Données accessibles facilement et rapidement	Actualité Données mises à jour régulièrement	Cohérence Chaque variable stockée dans un seul format

Analyse

- 01. Validité**
- 02. Complétude**
- 03. Precision**
- 04. Disponibilité**
- 05. Actualité**
- 06. Cohérence**

Les données météo sont :

- fiables et sans erreurs.
- incomplètes (NaN sur JO très anciens)
- relevées par les stations météorologiques.
- disponibles en consultation libre sur le site.
- collectées à un instant t et sauvegardées intégralement.
- présentes en format unique.