



POLYTECHNIQUE MONTREAL

UNIVERSITÉ D'INGÉNIERIE

INF8808 : Visualisation de données

Plan de projet - Groupe 16 (Plotly)

Canava Lou - 2402245

Gachet Théo - 2402954

Le Manh Ho Mathis - 2055958

Roman Canizalez Roman Alejandro - 2089991

Roy Sébastien - 2146331

Speismann Matthieu - 2409862

Soumis à

Thomas Hurtut

18 mars 2025

Hiver 2025

1. Mise en contexte.....	3
1.1 Contexte.....	3
1.2 Objectifs.....	3
2. Jeux de données.....	4
2.1 Métadonnées.....	4
2.2 Données.....	5
3. Questions cibles.....	8
4. Maquettes.....	10
4.1 Vue générale.....	10
4.2 Visualisation 1.....	11
4.3 Visualisation 2.....	13
4.4 Visualisation 3.....	15
4.5 Visualisation 4.....	18
4.6 Visualisation 5.....	20

1. Mise en contexte

1.1 Contexte

Depuis leur renaissance en 1896, les Jeux Olympiques modernes sont devenus bien plus qu'une simple compétition sportive. Tous les 4 ans (sauf de rares exceptions), ils réunissent les meilleurs athlètes venus du monde entier, qui représentent fièrement leur pays dans de nombreuses disciplines. Ils sont le plus souvent un immense succès, tant populaire que sportif, et font vivre aux fans comme aux athlètes des moments inoubliables d'émotion, comme nous l'a encore rappelé l'édition de Paris 2024.

Les Jeux Olympiques, en plus de 120 ans d'histoire, sont également le reflet de l'évolution du monde et permettent de découvrir ou approfondir, à travers eux, les relations géopolitiques entre pays, ainsi que l'impact des enjeux sociétaux et économiques, dans de nombreux pays du monde.

1.2 Objectifs

L'objectif principal de ce projet est de montrer instinctivement au **grand public**, à des personnes qui, a priori, ne sont pas nécessairement amatrices de sport, les dynamiques principales qui régissent les résultats aux JO. Car en effet, certaines nations dominent historiquement les Jeux, que ce soit dans leur globalité ou bien plus spécifiquement, selon les disciplines.

On cherchera également à proposer des visualisations permettant aux lecteurs d'explorer de possibles explications à ces résultats déséquilibrés. Pour cela, nous nous intéresserons à différentes échelles. Tout d'abord, nous étudierons plusieurs facteurs à l'échelle des pays, tels que leurs économies ou leurs populations. Ensuite nous observerons l'impact des athlètes de manière individuelle, car finalement, ce sont bel et bien eux qui remportent médailles et titres olympiques, récompensant finalement l'ensemble de leurs efforts et de leur talent.

2. Jeux de données

2.1 Métadonnées

Pour bâtir ce projet, nous nous basons sur l'ensemble de données ["Olympics Legacy: 1896-2020"](#), disponible sur la plateforme Kaggle. Ce jeu de données est structuré sous la forme d'un fichier CSV et couvre les Jeux Olympiques modernes depuis 1896 jusqu'à l'édition de 2020, en fournissant de l'information historique sur les athlètes participants, les disciplines sportives, les villes hôtes et les médailles remportées.

Source des données

Concernant la source des données, celles-ci proviennent principalement de la combinaison de plusieurs jeux de données liés aux Jeux Olympiques. La source principale est l'ensemble ["120 years of Olympic history: athletes and results"](#). Mais, celle-ci ne contient pas les Jeux Olympiques de 2020, c'est pourquoi elle a été complétée avec les données de l'édition 2020, disponibles dans la source officielle [Comité International Olympique \(CIO\)](#).

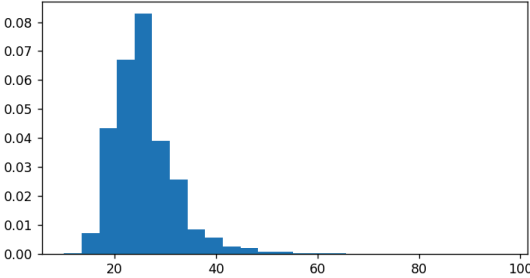
La source principale, le jeu de données "120 years of Olympic history" couvre l'ensemble des Jeux Olympiques modernes, de 1896 (Athènes) à 2016 (Rio de Janeiro). Un auteur indépendant a collecté ces données à partir du site www.sports-reference.com, une base de données historique contenant des informations détaillées sur différents sports, incluant les Jeux Olympiques. Ce site web est le résultat d'un travail de recherche mené par une communauté de statisticiens passionnés par l'histoire olympique et les sports en général.

Considérations à prendre en compte

Dans le cadre de ce projet, il faudra prendre en compte plusieurs facteurs concernant l'ensemble de données :

- **Évolution des disciplines** : Les sports ont évolué au fil du temps. En effet, certaines disciplines ont été ajoutées ou supprimées historiquement.
- **Changements de nom des pays** : Les noms et frontières des pays ont changé, ce qui peut affecter la cohérence des données relatives aux pays représentés.
- **Source créée par la communauté** : sports-reference.com est une source très populaire, utilisée pour un grand nombre d'ensemble de données sur Kaggle. Cependant, ce n'est pas une source officielle comme le Comité International Olympique (CIO). Donc, les données peuvent contenir des erreurs ou omissions, surtout pour les éditions les plus anciennes des Jeux Olympiques.

2.2 Données

Dataset Jeux Olympiques				
Variable	Description	Type de donnée	Intervalle / valeurs des données	Conclusion
ID	Identifiant de l'athlète	int	Valeurs : [1 : 271 115] 286 237 identifiants 271 116 uniques	Un athlète peut avoir plusieurs ID s'il a participé à plusieurs jeux (année différente, discipline différente).
Name	Nom de l'athlète	string	Valeurs : {Sara Bertolasi; David Collins; Morten Jensen; ...} 286 237 enregistrements 146 179 valeurs uniques	Il y a des athlètes différents qui sont des homonymes : ils ont des ID différents mais le même nom et prénom. Le nombre de valeurs uniques est inférieur au nombre d'ID car de nombreux athlètes participent plusieurs fois aux JO.
Gender	Genre de l'athlète	string	Valeurs : {Female (29%); Male (71%)} 286 237 enregistrements (aucune valeur nulle) 2 valeurs uniques	Le genre de tous les athlètes est renseigné.
Age	Âge de l'athlète	int	Valeur : {24; 18; 34; ...} 276 763 valeurs (9 474 valeurs manquantes) 74 valeurs uniques  <p>Minimum : 10 Maximum : 97 Moyenne : 25.6</p>	L'âge de presque tous les athlètes est renseigné. Lorsqu'une case est vide, cela signifie qu'il n'y a aucune information concernant l'âge de l'athlète

Dataset Jeux Olympiques				
Variable	Description	Type de donnée	Intervalle / valeurs des données	Conclusion
			Médiane : 25	
Team	Le nom de l'équipe	string	Valeurs : {China; Denmark; ...} 286 237 enregistrements 1 196 valeurs uniques	Le nom de toutes les équipes présentes. Les équipes peuvent participer à plusieurs JO, c'est pour cela qu'il y a peu de valeurs uniques.
NOC	Le code des pays	string	Valeurs : {CHN; DEN; ...} 286 237 enregistrements 233 valeurs uniques	Les codes correspondent aux noms des pays. La plupart des pays participent à plusieurs éditions des JO. Le nombre de valeurs uniques est élevé puisque plusieurs pays n'existent plus et/ou ont été renommés.
Year	L'année du JO	int	Valeurs : {1896; 2012; 1920; ...} Minimum : 1896 Maximum : 2020 286 237 enregistrements	L'année à laquelle ont eu lieu les JO. Les JO d'été ont lieu tous les 4 ans depuis 1896 et les JO d'hiver ont lieu tous les 4 ans depuis 1924
Season	Les JO d'été ou d'hiver	string	Valeurs : {Summer; Winter} 286 237 enregistrements 2 valeurs uniques	La saison des JO qui correspond à l'été ou à l'hiver
City	La ville d'accueil des JO	string	Valeurs : {Barcelona; London; ...} 286 237 enregistrements 42 valeurs uniques	La ville qui a accueilli les JO. C'est souvent les mêmes villes qui accueillent les JO
Sport	Le type de sport	string	Valeurs : {Basketball; Judo; ...} 286 237 enregistrements 84 valeurs uniques	Le type de sport présent aux JO. La plupart des sports sont régulièrement présents à l'exception de quelques-uns.
Event	La catégorie	string	Valeurs : {Basketball Men's Basketball;	Chaque discipline est

Dataset Jeux Olympiques				
Variable	Description	Type de donnée	Intervalle / valeurs des données	Conclusion
	du sport		Athletics Women's 100 metres; ...} 286 237 enregistrements 1 071 valeurs uniques	scindée en plusieurs catégories (entre femme et homme et entre différents type de pratique (l'Athlétisme contient le sprint 50m, la course 2km, 5km, ...)).
Medal	Les médailles gagnées	string	Valeurs : {Gold; Silver; Bronze} 42 233 enregistrements 3 valeurs uniques	Les trois types de médailles ainsi qu'une case vide lorsqu'aucune médaille n'est discernée. C'est normal qu'il y ait aussi peu d'enregistrements comme il y a seulement 3 médailles par catégorie

3. Questions cibles

**Quels sont les secrets qui façonnent la domination sportive
des grandes nations aux JO ?**

Questions		Priorité
Y'a t-il une hégémonie des grandes nations dans les performances aux JO ?		
1	Les pays avec des populations plus importantes gagnent-ils plus de médailles ?	★★★
2	Les pays les plus riches ont-ils plus de chances de remporter des médailles ?	★★★
3	Y'a-il une corrélation entre le nombre d'athlètes envoyés et le nombre de médailles remportées ?	★★★
4	Certains pays dominent-ils historiquement les mêmes disciplines ?	★★☆
Le pays organisateur a-t-il un avantage lors des Jeux ?		
5	Le nombre d'athlètes représentés évoluent-ils si les JO ont lieu dans leur pays ?	★★★
6	Le nombre de victoires et de médailles est-il plus important à domicile ?	★★★
7	Le pourcentage de médailles ou de victoires est-il plus important à domicile ?	★★★
8	Les pays avec un climat froid/chaud sont-ils plus performants aux JO d'hiver/été ?	★★☆
9	Quelle est l'évolution de ces métriques dans le temps ?	★★☆

10	Quelles sont les différences géographiques de ces métriques?	★★☆
11	Les pays qui organisent les JO choisissent-ils (ajouter/retirer) des disciplines à leur avantage ?	★★☆
Le talent individuel porte-t-il également les nations ?		
12	Quels pays produisent les plus grands athlètes ?	★★★
13	A quel point les meilleurs athlètes impactent-ils le classement de leur pays ?	★★★
14	Quels sont les pays qui "produisent" le plus d'athlètes multi-médailleurs ?	★★☆
15	Existe-t-il des périodes où un pays a bénéficié "d'un âge d'or" porté par quelques talents exceptionnels ?	★★☆
16	A combien d'éditions un athlète doit-il participer avant de remporter sa première médaille selon le pays ?	★★☆
17	Quelle est la longévité des athlètes en fonction des sports et des nations ?	★★☆
18	Quel est le pourcentage d'athlètes d'un certain pays qui participent à plusieurs éditions ?	★★☆

4. Maquettes

4.1 Vue générale

Vue avec un telling

Le but est de réaliser un site web sous la forme d'un article qui reprendrait les codes du storytelling mais sans transitions dynamiques car nous travaillons avec Plotly.

Ainsi, ce “storytelling” est agencé en trois grandes parties qui sont formées par nos trois grandes questions. Pour chaque grande question, nous utiliserons de une à deux visualisations pour répondre aux sous-questions en lien. L'utilisateur du site sera invité à utiliser sa souris pour obtenir des informations supplémentaires en passant au-dessus de chaque visualisation. Il pourra également faire des choix d'affichage et/ou de filtrage de données sur certaines visualisations.

Le tout permettra au lecteur d'en apprendre davantage sur les dynamiques des pays participants aux JO, dans le temps et en fonction des disciplines.

4.2 Visualisation 1

La première visualisation répond aux questions cibles suivantes :

4	Certains pays dominant-ils historiquement les mêmes disciplines ?
11	Les pays qui organisent les JO choisissent-ils (ajouter/retirer) des disciplines à leur avantage ?

Description

La visualisation est un *small multiple* de graphiques “*heatmap*”. Chaque heatmap représente les pays en ordonnées et les éditions en abscisses. Les valeurs représentées sont le nombre de médailles gagnées par le pays en question pendant l’édition en question. Ainsi, plus la case est intensément colorée, et plus le nombre de médailles remportées par ce pays, dans cette discipline lors de cette édition, est important.

Une légende différente par heatmap sera affichée pour prendre en compte que le nombre maximum de médailles qui peuvent être remportées est différent d’une discipline à l’autre

Chacun des multiples du small multiple est une heatmap qui représente une discipline sportive différente. L’utilisateur pourra sélectionner soit les jeux d’hiver soit les jeux d’été. Cela aura un impact sur les disciplines et les éditions affichées. Pour ne pas afficher trop de *small multiple*, nous effectuerons un choix avisé pour les disciplines affichées.

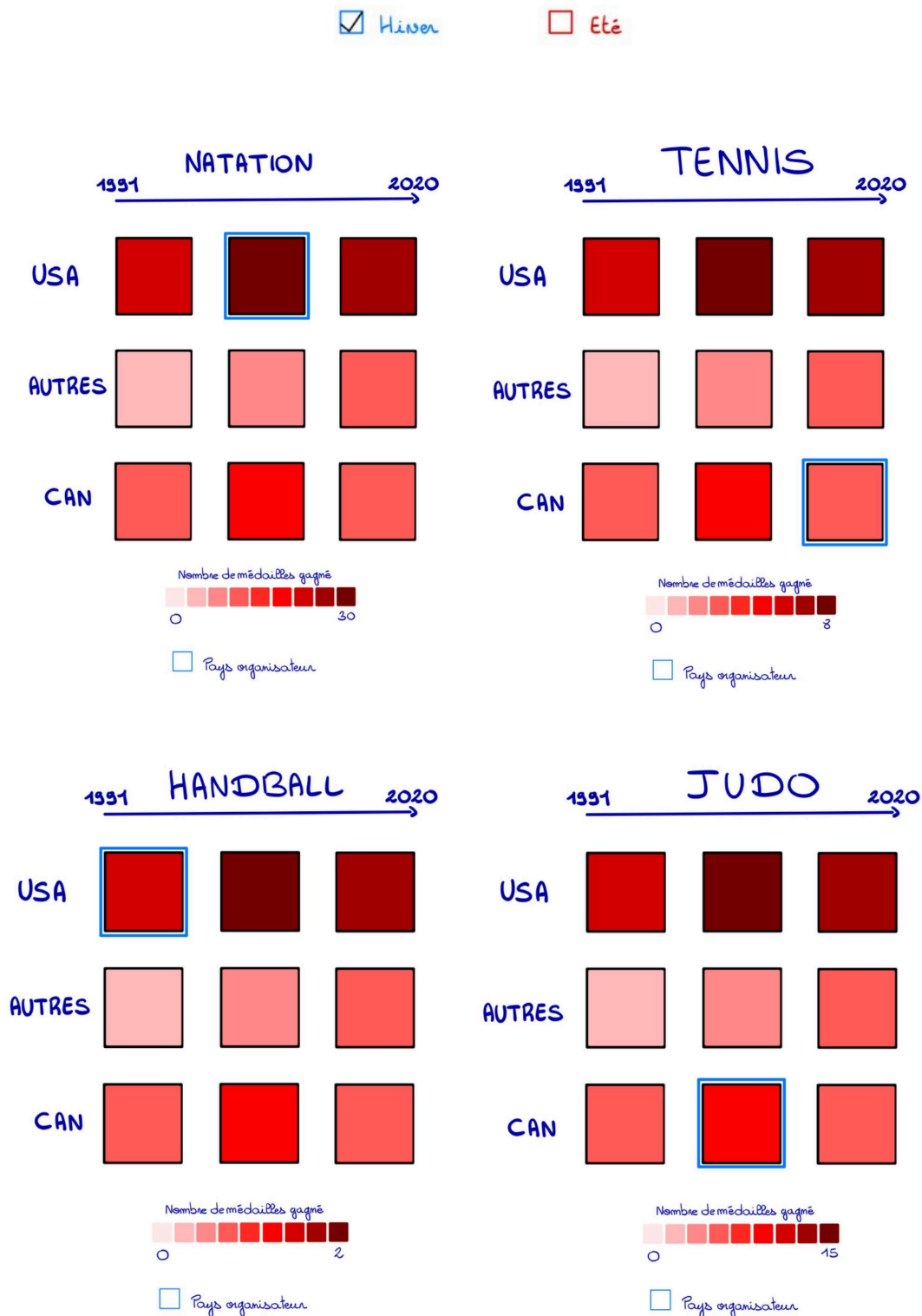
Enfin, un liseré bleu entoure les cases correspondant aux éditions organisées par le pays concerné afin d’étudier si les sports où l’organisateur est historiquement bon (ou non) sont au programme.

Interaction

En passant la souris sur une des cases de la heatmap, l’utilisateur pourra voir s’afficher un *hover* qui mentionne le nombre exact de médailles gagnées, le pays en question, l’année en question et la discipline en question.

Enfin, l’utilisateur peut filtrer en cochant soit les jeux d’hiver soit les jeux d’été. Cela influence les pays affichés car ne seront affichés que les pays ayant gagné le plus de points cumulés sur la période 1992-2020 pour le type d’édition des JO sélectionné par l’utilisateur.

Prévisualisation



Cette visualisation permet ainsi de voir quels pays dominent quelles disciplines au cours des 30 dernières années. De plus, on peut remarquer si le choix des disciplines par un pays organisateur est en sa faveur (ou non) en soulignant en quelle année il a organisé les Jeux.

4.3 Visualisation 2

Cette visualisation répond aux questions cibles suivantes :

1	Les pays avec des populations plus importantes gagnent-ils plus de médailles ?
2	Les pays les plus riches ont-ils plus de chances de remporter des médailles ?
8	Les pays avec un climat froid/chaud sont-ils plus performants aux JO d'hiver/été ?
9	Quelle est l'évolution de ces métriques dans le temps ?
10	Quelles sont les différences géographiques de ces métriques?

Description

Pour répondre à ces questions, nous avons choisi de réaliser deux *“bubble chart”*.

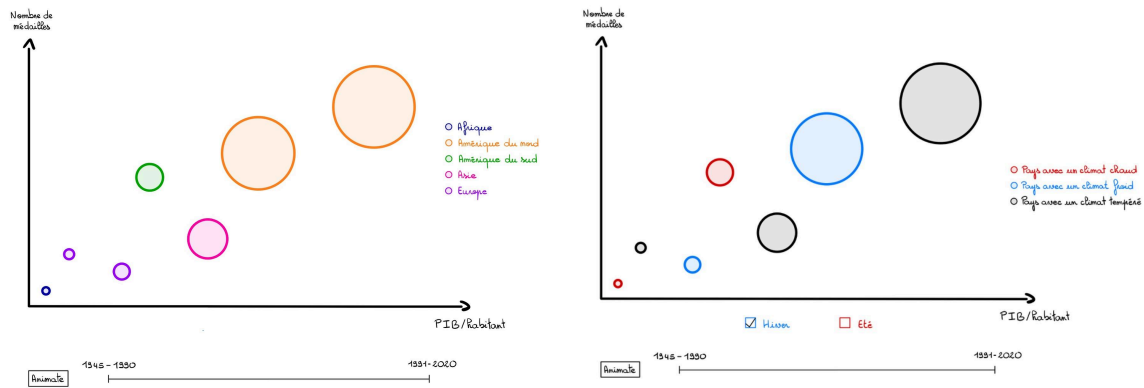
Tout d'abord, chaque bulle représente un pays avec un encodage de couleur par continent. L'axe des abscisses représente le PIB/habitant de chaque pays et l'axe des ordonnées le nombre de médailles moyen remporté par édition par pays. Ensuite, la surface de chaque bulle est proportionnelle à la population du pays. Enfin, il sera possible de sélectionner deux périodes historiques (1945-1990 ou 1991-2020) à l'aide d'un *slider*, pour remarquer une évolution dans le temps, avec une courte animation en passant de l'un à l'autre à l'aide d'un bouton.

Interactions

Dans le premier *bubble chart*, la légende de couleur permet de colorier chaque pays en fonction de son continent d'appartenance pour remarquer d'éventuelles tendances géographiques et ainsi répondre à la question 10.

Dans le second *bubble chart*, il sera possible de sélectionner uniquement les médailles gagnées lors des éditions d'hiver ou d'été, ou les deux afin d'étudier de potentielles différences à l'aide de deux boutons de *“filtre”* différents. De plus, la légende de couleur permet de colorier chaque pays par climat (froid, chaud ou tempéré) afin de répondre à la question 8.

Ainsi, les deux visualisations auront des légendes avec des couleurs différentes pour les bulles. Cela permettra de guider le lecteur dans le récit global de notre site pour répondre à l'ensemble des questions dans un ordre logique.



Ces deux visualisations nous semblent pertinentes car nous utilisons adéquatement le bubble chart pour montrer des corrélations entre deux variables continues placées sur les deux différents axes. De plus, la surface et les axes représentent des variables numériques et les teintes de couleurs représentent des variables catégorielles. Enfin, L'animation permet de montrer une évolution temporelle.

4.4 Visualisation 3

Cette visualisation répond aux questions cibles suivantes :

5	Le nombre d'athlètes représentés évoluent-ils si les JO ont lieu dans leur pays ?
6	Le nombre de victoires et de médailles est-il plus important à domicile ?
7	Le pourcentage de médailles ou de victoires est-il plus important à domicile ?
9	Quelle est l'évolution de ces métriques dans le temps ?
10	Quelles sont les différences géographiques de ces métriques?

Description

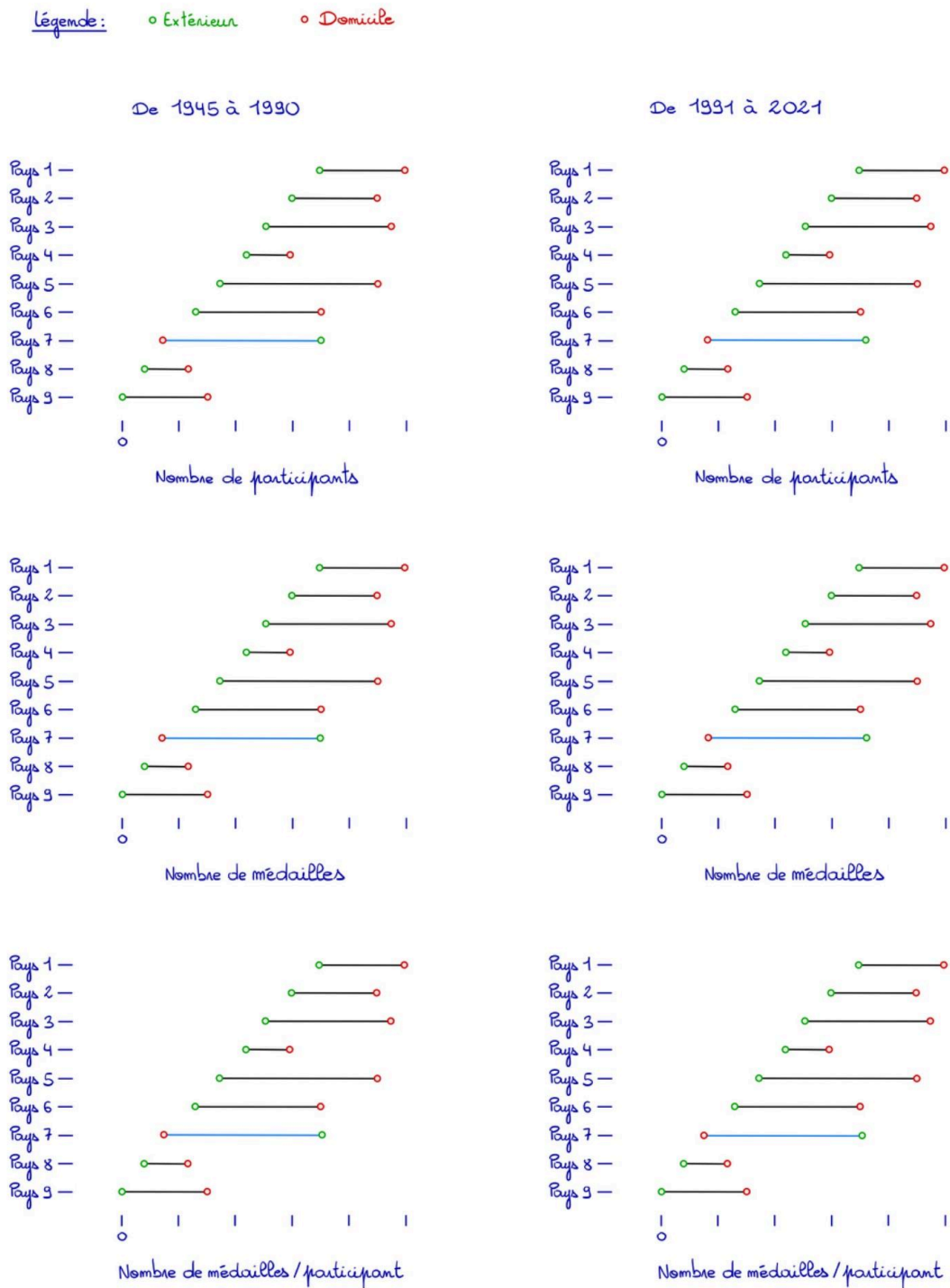
Pour répondre à ces questions, nous avons choisis de réaliser un *small multiple* de graphiques dits "*lollipop charts*" tels que:

- Il y aura 6 lollipop charts rangés en 2 colonnes de trois. Chaque colonne représente les graphiques sur une période de temps donnée : colonne 1 sur 1945-1991 et colonne 2 sur 1991-2020.
- Tous les pays qui ont déjà organisé les JO (au nombre de 27 sur la période de temps considérée de 1945 à 2020) sont placés en ordonnées.
- Les variables en abscisses sont:
 - Pour le premier graphique, le nombre d'athlètes par pays (répond à la question 5)
 - Pour le deuxième graphique, le nombre de médailles remportées par pays (répond à la question 6)
 - Pour le dernier graphique, le nombre de médailles remportées par athlète de ce pays (répond à la question 7)
- Pour chaque graphique, on place alors deux points par pays:
 - Un premier, coloré en vert qui représente les moyennes des variables étudiées lorsque ce pays participe aux JO "à l'extérieur"
 - Un second point, coloré en rouge, avec les moyennes des variables lorsque le pays participe aux JO chez lui (lorsqu'il est organisateur).
- On relie ensuite les deux points par un segment noir pour obtenir le lollipop chart.
- On s'attend à ce que les valeurs soient supérieures lorsque le pays est à domicile. Pour cette raison, dans le cas inverse, on colorie le segment en bleu pour faciliter la lecture.

Interaction

L'utilisateur sera invité à passer sa souris sur une ligne d'un pays pour accéder au détail du graphique.

Prévisualisation



Ainsi, la longueur des traits mesure l'importance de "l'avantage" dont bénéficie le pays organisateur.

Aussi, le fait d'afficher tous les noms des pays permet d'apprécier toutes les informations en un coup d'œil, en particulier si l'on s'intéresse à un pays en particulier, à la différence d'un slope chart où l'on doit survoler le trait pour identifier le nom du pays.

En constatant que le nombre de participants chez soi est toujours supérieur (ce qui est permis par le règlement des JO), plusieurs questions se posent: Est-ce que les pays organisateurs remportent plus de médailles? Si oui, est-ce parce qu'à domicile, un pays compte plus de participants ou bien parce que ceux-ci sont plus performants? Ces trois graphiques permettent au lecteur d'y répondre en suivant ce même raisonnement.

4.5 Visualisation 4

Cette visualisation répond aux questions cibles suivantes :

12	Quels pays produisent les plus grands athlètes ? (athlètes qui ont gagné plus de 5 médailles au cours de leur carrière)
14	Quels sont les pays qui "produisent" le plus de grands athlètes ?

Description

La visualisation est un *small multiple* de graphiques "*packed circle chart*" où:

- Chaque bulle est un athlète qui a gagné 5 médailles au moins tout au long de sa carrière, sur la période 1992-2020. Précision : les athlètes ayant commencé leur carrière avant 1992 mais fini après 1992 seront compris dans le graphique.
- La surface de la bulle est proportionnelle au nombre de médailles gagnées par cet athlète.
- On regroupe les bulles avec un algorithme de force par pays (en ne conservant que 10 pays et une catégorie "Autres"). Les 10 pays visibles seront les 10 pays ayant gagné le plus de points cumulés aux JO (soit d'été soit d'hiver selon le type de JO sélectionné par l'utilisateur) sur la période 1992-2020. Le nombre de points est calculé d'après les règles suivantes : 1 médaille d'or = 3 points, 1 médaille d'argent = 2 points, 1 médaille de bronze = 1 point.

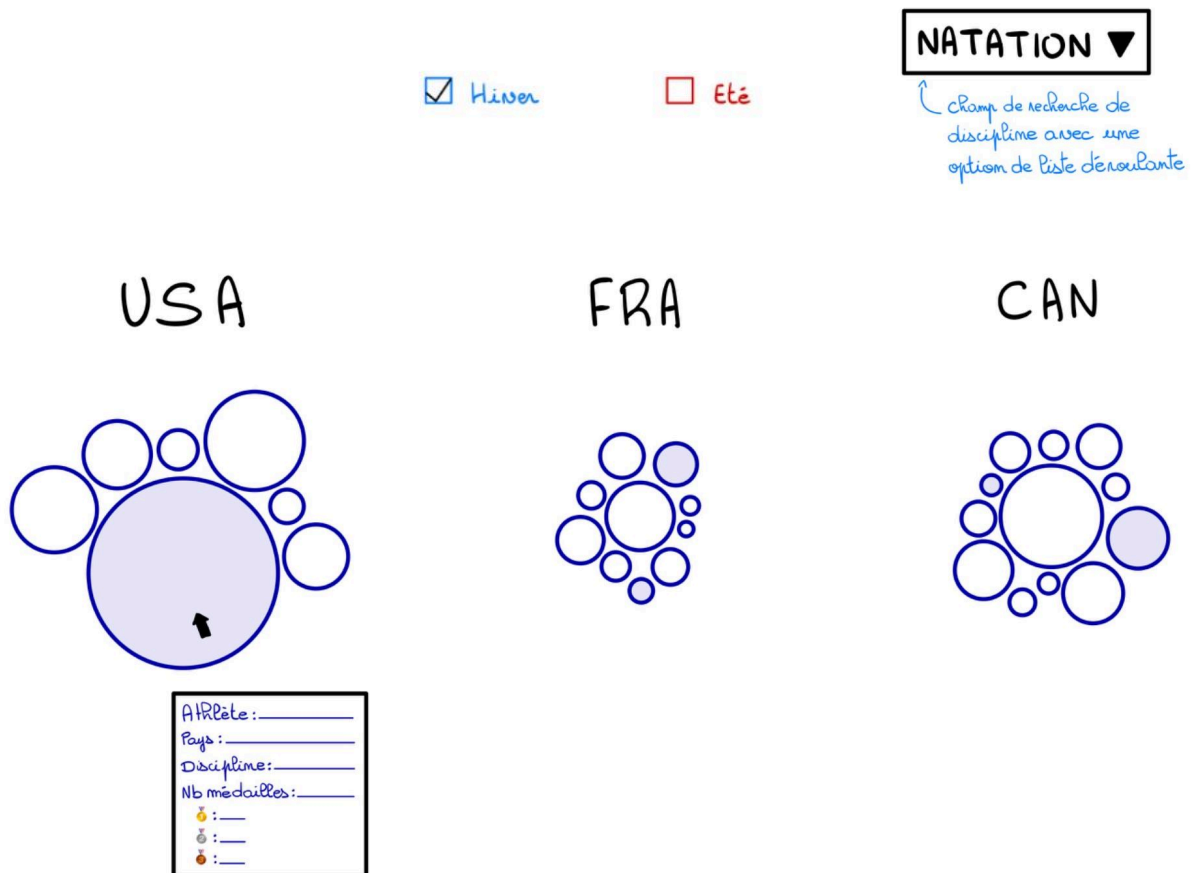
Interaction

Le fait de passer la souris sur un cercle affiche le nom de l'athlète, son pays, sa discipline ainsi que le détail de son nombre de médailles gagnées.

Il est possible pour l'utilisateur de colorer les bulles correspondants aux athlètes pratiquant la discipline sélectionnée dans un menu déroulant.

Enfin, l'utilisateur peut filtrer en cochant soit les jeux d'hiver ou soit les jeux d'été. Cela influence les pays affichés car ne seront affichés que les pays ayant gagné le plus de points cumulés sur la période 1992-2020 pour le type d'édition des JO sélectionné par l'utilisateur (été ou hiver).

Prévisualisation



Ce graphique permet donc bien de distinguer les plus grands athlètes de tous les temps qui se démarquent par leur grandeur (nombre de médailles gagnées au long de leur carrière) tout en distinguant quels pays produisent le plus de grands athlètes.

Pouvoir sélectionner soit les JO d'hiver ou soit ceux d'été permet de visualiser les pays comme la Suède ou la Norvège qui ne gagnent généralement des points que lors des JO d'hiver.

4.6 Visualisation 5

Cette visualisation répond aux questions cibles suivantes :

13	A quel point les meilleurs athlètes impactent-ils le classement de leur pays ?
15	Existe-t-il des périodes où un pays a bénéficié d'un "âge d'or" porté par quelques talents exceptionnels ?

Description

Cette visualisation est un *small multiple* d'un graphique de type "slope chart".

Chaque multiple représente le classement d'un certain pays au jeux olympiques pour une certaine année selon que l'on compte les multi-médaillés (athlète ayant gagné plus de 2 médailles lors d'une édition) dans le classement (axe des ordonnées de gauche) ou qu'on les retire du classement (axe des ordonnées de droite). Ces deux valeurs sont reliées par une ligne, ce qui crée visuellement une pente plus ou moins inclinée vers le bas.

Chaque slope chart du small multiple représente une édition des JO sur la période considérée (1992-2020). Ainsi, l'ensemble des multiples visibles représente l'impact des athlètes multi-médaillés sur le classement d'un pays lors des 8 dernières éditions des JO sélectionnés (hiver ou été).

L'axe des ordonnées de gauche représente le classement du pays sélectionné dans le menu déroulant pour l'année en question, toute discipline confondue. L'axe des ordonnées de droite représente cette même statistique tout en excluant du total des points, servant à établir le classement, les athlètes qui ont remporté plus de 2 médailles dans l'édition en question.

Pour réaliser le classement, trois points seront accordés pour une médaille d'or, deux pour une médaille d'argent et un pour une médaille de bronze.

Il sera possible de voir les graphiques des 20 pays les plus performants en termes de points en sélectionnant un pays en particulier via un menu déroulant.

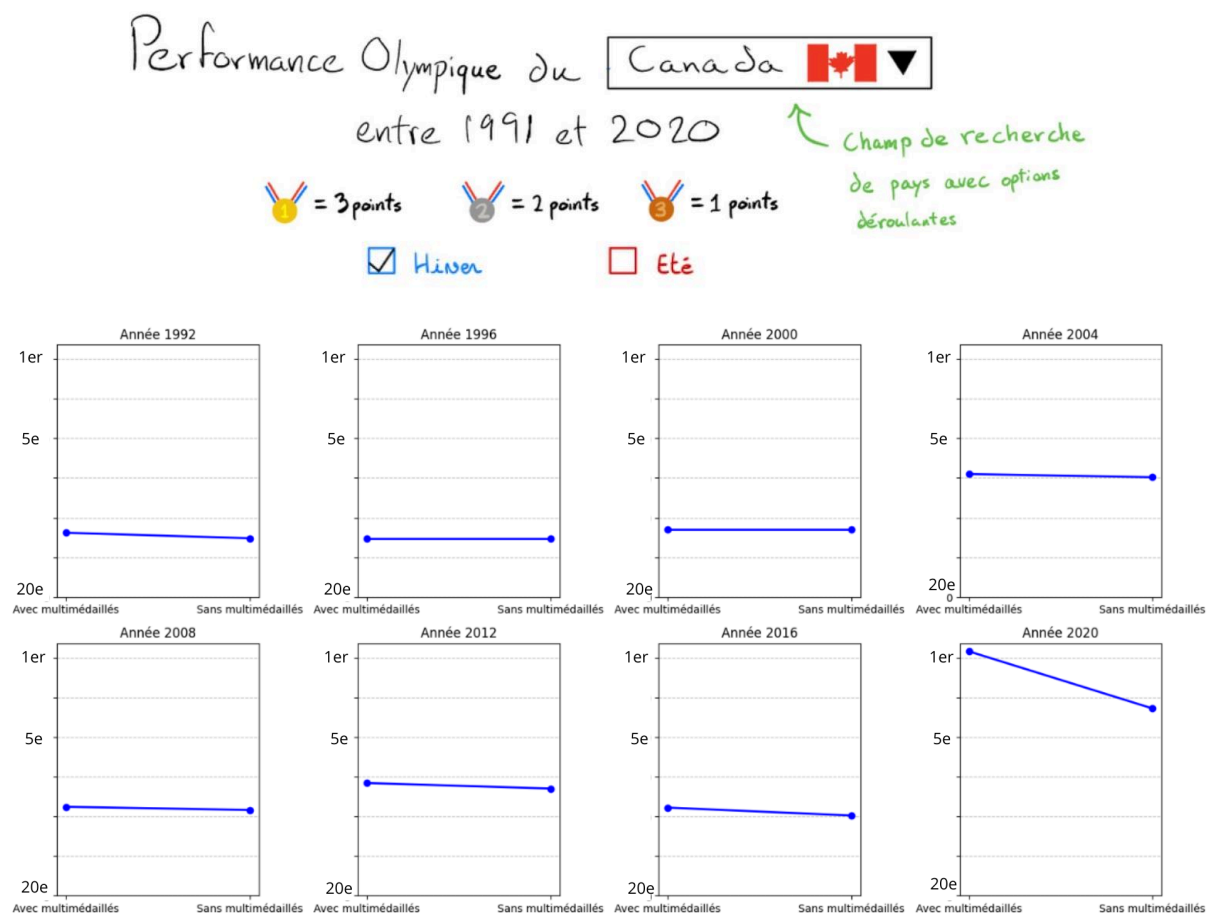
Interaction

Le fait de passer la souris sur une des pentes affiche le classement du pays à l'année en question en incluant ou non les multi-médaillés.

Un menu déroulant permet de choisir un pays en particulier parmi les 20 pays ayant gagnés le plus de points cumulés sur la période 1992-2020. Toute discipline confondue.

Enfin, l'utilisateur peut venir cocher les cases relatives aux JO d'hiver ou d'été pour venir filtrer les éditions affichées sur les *small multiple*. S'il coche "JO d'hiver", alors les dernières éditions des JO d'hiver remontant au maximum jusqu'à 1992 seront affichées. Il ne sera pas possible d'afficher à la fois les jeux d'hiver et d'été.

Prévisualisation



Ce graphique permet donc d'observer l'impact des talents individuels sur le classement des pays les plus performants. En effet, plus l'ensemble des pentes d'un pays relatives aux 8 dernières éditions des JO est raide, plus le pays semble porté par des athlètes

sur-performants. Si au contraire les pentes sont pratiquement plates, alors cela montre que le classement du pays en question n'est pas dû à des athlètes particulièrement sur-performants mais à un ensemble d'athlètes performants.

De plus, ce détail par édition permet de remarquer d'éventuelles périodes "d'âge d'or" avec des performances supérieures à l'habitude pour le pays étudié. Cet âge d'or est d'autant plus porté par des athlètes performants si la pente est fortement inclinée.