PROJET RÉALISÉ PAR L'ÉQUIPE 11

RAPPORT DE GROUPE EN SCIENCES DES DONNÉES 2 + BASES DE DONNÉES

Akiki Chris, Sriri Samia, Laasili Omia, Souici Maïssane

Département MIASHS, UFR 6 Informatique, Mathématique et Statistique
Université Paul Valéry, Montpellier 3

Avril 2024

Soumis comme contribution partielle
pour le cours Science des données 2 et Bases de données

Déclaration de non plagiat

Nous déclarons que ce rapport est le fruit de notre seul travail, à part lorsque cela est indiqué explicitement.

Nous acceptons que la personne évaluant ce rapport puisse, pour les besoins de cette évaluation:

- la reproduire et en fournir une copie à un autre membre de l'université; et/ou,
- en communiquer une copie à un service en ligne de détection de plagiat (qui pourra en retenir une copie pour les besoins d'évaluation future).

Nous certifions que nous avons lu et compris les règles ci-dessus.

En signant cette déclaration, nous acceptons ce qui précède.

Remerciements

Nos plus sincères remerciements vont à nos encadrants pédagogiques, Mme Bringay et Mme Demangeot, pour les conseils avisés sur notre travail

28 avril 2024.

Résumé

Les Jeux Olympiques d'hiver de Beijing, qui se sont tenus en 2022, ont marqué un moment historique pour les compétitions internationales d'hiver. Cet événement a réuni des athlètes du monde entier, participant dans une variété de disciplines qui ont mis à l'épreuve leur force, agilité, et endurance dans des conditions hivernales. Beijing, devenant ainsi la première ville à avoir accueilli à la fois les Jeux Olympiques d'été et d'hiver, a mis en scène une série d'épreuves impressionnantes malgré les défis logistiques et sanitaires posés par la pandémie de COVID-19.

Table des matières

Chapitre 1

Introduction

1.1 Contenu d'une introduction

Depuis leur réinvention moderne à la fin du 19 siècle, les Jeux Olympiques ont été le théâtre où les athlètes les plus talentueux du monde se sont affrontés pour décrocher l'or, l'argent et le bronze. Ces compétitions intenses ont non seulement suscité l'admiration mondiale pour les exploits physiques des compétiteurs, mais ont également eu un impact profond sur la société, tant sur le plan sportif que sur le plan culturel, économique et politique.

cjjdn whw hejdnd dnn

Existe-t-il des différences significatives dans le nombre de médailles remportées par les athlètes aux Jeux Olympiques de Beijing de 2022 en fonction de leur pays de naissance dans des disciplines similaires ?

Nous allons chercher à savoir si il y a une différence significative dans les performance entre pays dans des disciplines similaires aux Jeux Olympiques. Nous préferons gardé les individus medaillés et non médaillés pour ne pas perdre d'informations. Par exemple, il pourrait y avoir un pays A avec plus de représentants mais avec x médailles et un pays B avec moins de représentants mais x médailles, ce pays serait alors considéré comme plus "performant".

Les résultats de cette recherche pourraient influencer la manière dont les ressources sont allouées pour le développement du sport, promouvoir l'équité et l'inclusion dans le domaine sportif, renforcer la crédibilité des compétitions internationales et favoriser des opportunités équitables pour tous les athlètes. En somme, cette question de recherche offre la possibilité d'identifier les disparités de performance entre les nations et de proposer des mesures visant à les réduire, contribuant ainsi à l'amélioration globale du paysage sportif international.

1.2 Responsabilités et composition de l'équipe

- Omia Laasili : Étudiant n°01 , Resp. du descriptif des tables, requêtes SQL et explications
- Maissane Souici : Étudiant n°02 , Resp. du rapport/ nettoyage et prétraitement des données, et de l'analyse
- Samia Sriri : Étudiant n°03 , Resp. de l'analyse des données, graphiques via R, des annexes et de la mise en page du rapport sur R
- Chris Akiki : Étudiant n°04 , Resp de l'analyse partielle des données

Chapitre 2

Base de données

2.1 Provenance des données

Notre jeu de données a été recupéré à partir du lien suivant :
https://www.kaggle.com/datasets/piterfm/beijing-2022-olympics./newline
Il contient l'ensemble des données des Jeux Olympiques d'hiver de Beijing en 2022. Le lien fournit 10 jeux de données au format .csv. Nous avons choisi de garder les fichiers:

- 'athletes.csv'(572.2 kB) qui contient les informations d'un athlète tels que le nom, l'âge , le sexe, la discipline exercée et le pays de jeu et de naissance.
- 'medals.csv'(127.05 kB) qui contient les informations sur les médailles obtenues par les athlètes tels que le type de médaille (or, argent ou bronze), sa date de remise, l'athlète l'ayant obtenu et la discipline correspondante.

Le premier fichier contient les informations personnelles des athlètes et le second indique les informations sur les médailles remportées c'est-à-dire par qui, pour quelle épreuve etc. Nous allons créer 4 tables grâce à ces 2 bases de données: Athletes, Pays, Disciplines, Medailles

2.2 Descriptif des tables

Table 2.1: Athletes (2631 × 7)

Nom colonne	Type	Signification	Caractéristique
idAthlete	int	identifiant	clé primaire
nom	varchar	nom	champs obligatoire
genre	varchar	Male/Female	champs obligatoire
DateNaiss	int	Date de naissance	champs obligatoire
idPaysnaissance	varchar	identifiant	clé étrangère
idPaysjeu	varchar	identifiant	clé étrangère
idDiscipline	varchar	identifiant	clé étrangère

Les données sont stockées dans un document CSV : "athletes.csv" (123 ko)

Table 2.2: Pays (101 × 2)

Nom colonne	Type	Signification	Caractéristique
idPays	int	identifiant	clé primaire
nompays	varchar	nom	champs obligatoire

Les données sont stockées dans un document CSV : “pays.csv” (4 ko)

Table 2.3: Disciplines (15 × 2)

Nom colonne	Type	Signification	Caractéristique
idDis	int	identifiant	clé primaire
nomDiscipline	varchar	nom	champs obligatoire

Les données sont stockées dans un document CSV : “disciplines.csv” (4 ko)

Table 2.4: Medailles (668 × 5)

Nom colonne	Type	Signification	Caractéristique
idMedaille	int	identifiant	clé primaire
type	varchar	Gold/Silver/Bronze	champs obligatoire
date	varchar	date de remise	champs obligatoire
idAthlete	varchar	identifiant	clé étrangère
idDiscipline	varchar	identifiant	clé étrangère

Les données sont stockées dans un document CSV : “medailles.csv” (29 ko)

2.3 Modèles MCD et MOD

- MCD, réalisé avec le logiciel Mocodo [https://www.mocodo.net/]:

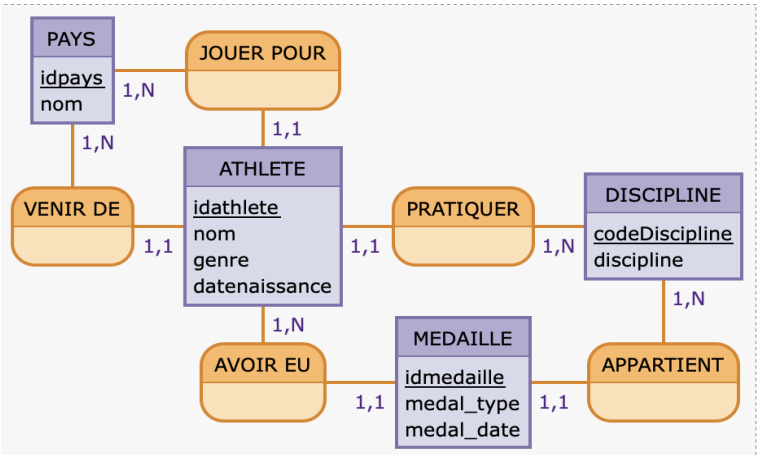

Figure 2.1: MCD

- MOD, réalisé avec le designer phpmyAdmin:

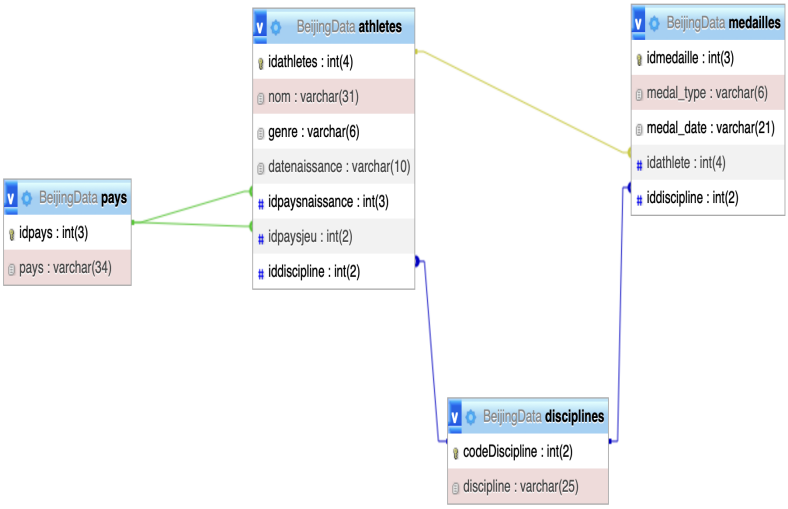

Figure 2.2: MOD

ATHLETES(idathletes, nom, genre, datenaissance, idpays, iddiscpline,idpaysnaissance, idpaysjeu)
MEDAILLES(idmedaille, medal_type,medal_date, idathlete,iddiscipline)
PAYS(idpays, nom)
DISCIPLINES (codediscipline, discipline)

2.4 Import des données

Source de données athletes et medailles:

On a du enlever tous les underscores des titres(_), les colonnes avec beaucoup de cases vides, comme la taille ont été supprimées, ainsi que les codes pays/disiplines (exemple: FR pour la France) devenus inutile après l'ajout des clés étrangères. Notamment, on a supprimé les lignes vides (sans informations). Pour une question de pratique, on a modifié certains noms d'attributs lors de l'importation sur phpmyAdmin.

On a divisé la base de données athletes en 3 tables différentes: disciplines, pays et athletes Nous avons décidé de garder les pays de naissance et de jeu de chaque joueur pour pouvoir comparer leur performance selon leur pays de naissance.

2.5 Requêtes réalisées

Requête n°1 : le nombre d'athlètes, de diciplines et de pays concernés par le jeu de données.

```
r1 <- DBI::dbFetch(DBI::dbSendQuery(conn = con, statement ="
SELECT 'nombre_athletes' AS recapitulatif, COUNT(DISTINCT athletes.idAthlete) AS nb
FROM athletes
UNION
SELECT 'nombre_disciplines'AS recapitulatif , COUNT(DISTINCT disciplines.idDiscipline) AS nb
FROM disciplines
UNION
SELECT 'nombre_pays' as recapitulatif, COUNT(DISTINCT pays.idPays) AS nb
FROM pays;
"))
```

recapitulatif	nb
nombre_athletes	2631
nombre_disciplines	15
nombre_pays	101

Figure 2.3: Requête n°1

Cette requête nous indique qu'il y a au total 2631 athlètes, 15 disciplines différentes et 101 pays.

Requête n°2 : ajout de la colonne 'nb_athletes' sur la table pays, cette colonne représente le nombre d'athlètes par pays.

```
" ALTER TABLE pays
ADD COLUMN nb_athletes int(4);
UPDATE pays
SET nb_athletes = ( SELECT COUNT( DISTINCT athletes.idAthlete)
FROM athletes
WHERE athletes.idPNaissance=pays.idPays); "
```

Cette requête nous permet d'avoir une meilleure idée du nombre d'athlètes qui sont nés dans chacun des 101 pays, directement à partir de la table pays.

Requête n°3 : affichage des 10 premiers enregistrements de la table pays

```
r3 <- DBI::dbFetch(DBI::dbSendQuery(conn = con, statement ="
SELECT * from pays limit 10;
"))
```

idPays	nomPays	nb_athletes
1	Denmark	47
2	Finland	78
4	Canada	232
5	Germany	139
7	Ukraine	37
8	Netherlands	42
9	People's Republic of China	142
10	France	93
11	Switzerland	168
12	Belgium	13

Figure 2.4: Requête n°3

L'extrait de cette requête nous fait remarquer que le nombre d'athlètes nés dans chacun des pays semble varier significativement d'un pays à l'autre.

Requête n°4 : les pays et les disciplines où un ou plusieurs athlètes nés dans ces pays ont remporté plus de 5 médailles d'or

```
r4 <- DBI::dbFetch(DBI::dbSendQuery(conn = con, statement ="
SELECT pays.nomPays,disciplines.nomDiscipline
FROM pays,medailles,athletes,disciplines
WHERE pays.idPays=athletes.idPNaissance
AND athletes.idAthlete=medailles.idAthlete
AND athletes.idDiscipline=disciplines.idDiscipline
AND medailles.type LIKE '%Gold%'
GROUP BY nomPays,nomDiscipline
HAVING COUNT(medailles.type)>5;
"))
```

nomPays	nomDiscipline
Austria	Alpine Skiing
Canada	Ice Hockey
Federal Republic of Germany	Luge
Finland	Ice Hockey
Germany	Bobsleigh
Norway	Biathlon
Norway	Cross-Country Skiing
People's Republic of China	Short Track Speed Skating
Russian Federation	Cross-Country Skiing
Russian Federation	Figure Skating
United States of America	Freestyle Skiing

Figure 2.5: Requête n°4

Sur les 101 pays, 9 pays correspondant au pays de naissance des athlètes comptabilisent plus de 5 médailles d'or pour une même discipline, certains pays sont cités

plusieurs fois tel que la Norvège où des athlètes nés dans ce pays ont remporté plus de 5 médailles d'or dans le 'Biathlon' ainsi que le 'Cross-Country Skiing'. Cela signifie que les athlètes d'origine norvégienne ont remporté plus de 5 médailles dans le Biathlon et plus de 5 médailles dans le ski de fond.

Requête n°5 : les pays avec le nombre de médailles de chaque type remporté par les athlètes originaires de ces pays, rangés par ordre alphbétique des noms de pays

Les pays, classés par ordre alphabétique, correspondant aux lieux de naissance des athlètes, avec le nombre de médailles de chaque type (or, argent, bronze) que ces athlètes ont remportées, indépendamment du pays qu'ils représentent dans les compétitions.

```
r5 <- DBI::dbFetch(DBI::dbSendQuery(conn = con, statement ="
SELECT pays.nomPays, medailles.type, COUNT(*) AS nb_medailles FROM pays
JOIN athletes ON pays.idPays = athletes.idPNaissance
JOIN medailles ON athletes.idAthlete = medailles.idAthlete GROUP BY pays.nomPays, medailles.type
ORDER BY pays.nomPays ASC;
"))
```

nomPays	type	nb_medailles
Australia	Bronze	1
Australia	Gold	2
Australia	Silver	3
Austria	Bronze	6
Austria	Gold	14
Austria	Silver	10
Belarus	Bronze	3
Belarus	Silver	3
Belgium	Bronze	1
Belgium	Gold	1
Canada	Bronze	26
Canada	Gold	32
Canada	Silver	8
Croatia	Bronze	1
Czech Republic	Gold	1
Czechoslovakia	Bronze	4
Estonia	Bronze	1
Federal Republic of Germany	Gold	8
Federal Republic of Germany	Silver	1
Finland	Bronze	11
Finland	Gold	25

Figure 2.6: Requête n°5

s

Cet extrait de requête nous permet d'avoir un aperçu des performances des athlètes par pays de naissance, en fonction du nombre de médailles pour chaque type de médaille gagnée. Elle nous permettra ensuite d'illustrer ces informations à travers un histogramme qu'on décrira plus tard.

Requête n°6 : les pays, le nombre d'athlètes par pays et le nombre de médailles d'or par pays rangé dans l'ordre décroissant

```
r6 <- DBI::dbFetch(DBI::dbSendQuery(conn = con, statement ="
SELECT pays.nomPays,pays.nb_athletes, COUNT(*) as nb_medailles_d_or
FROM pays ,athletes, medailles
WHERE pays.idPays=athletes.idPNaissance
AND athletes.idAthlete=medailles.idAthlete
and medailles.type like 'Gold'
GROUP BY pays.nomPays,pays.nb_athletes
ORDER by nb_medailles_d_or desc;
"))
```

nomPays	nb_athletes	nb_medailles_d_or
Canada	232	32
Norway	86	28
Finland	78	25
Russian Federation	171	17
Sweden	98	15
Austria	100	14
Germany	139	14
United States of America	231	13
People's Republic of China	142	11
Federal Republic of Germany	11	8
Netherlands	42	8
Switzerland	168	8
France	93	6
Slovenia	41	5
Great Britain	52	4
Italy	112	3
Japan	97	3
Australia	45	2
Republic of Korea	52	2
New Zealand	11	2
Slovakia	42	1

Figure 2.7: Requête n°6

On remarque que certains pays, bien qu'ils aient un nombre d'athlètes natifs conséquent, ont gagné peu de médailles d'or, comme l'Italie qui pour 112 athlètes nés en Italie, a gagné 3 medailles d'or. A l'inverse, les athlètes nés en Norvège ont remporté 28 médailles d'or pour seulement 86 athlètes.

Requête n°7 : les pays, le nombre d'athlètes par pays et le nombre de médailles par pays

```
r7 <- DBI::dbFetch(DBI::dbSendQuery(conn = con, statement="
SELECT Pays.nomPays, COUNT(DISTINCT athletes.idAthlete)
AS nb_athletes, COUNT(Medailles.idMedaille) AS Nombre_Medailles
FROM Pays LEFT JOIN athletes ON Pays.idpays = Athletes.idPNaissance
LEFT JOIN Medailles ON athletes.idAthlete = Medailles.idAthlete
GROUP BY Pays.nomPays;
"))
```

nomDiscipline	nb_medailles
Ice Hockey	125
Short Track Speed Skating	60
Biathlon	60
Cross-Country Skiing	60
Speed Skating	52
Alpine Skiing	48
Freestyle Skiing	45
Figure Skating	40
Curling	36
Snowboard	35
Ski Jumping	33
Bobsleigh	27
Luge	24
Nordic Combined	18
Skeleton	5

Figure 2.8: Requête n°7

Dans cet extrait, on remarque que le nombre d'athlètes natifs pour chaque pays et le nombre de médailles ne suivent pas forcément la même tendance. On peut

citer la Norvège, pour 86 athlètes natifs, 57 médailles ont été remportées, tous types confondus. À l'inverse, la Suisse a remporté 14 médailles pour 168 athlètes natifs. Nous nous demanderons plus tard si le nombre d'athlètes et le nombre de médailles sont corrélés.

Requête n°8 : Les pays de naissance des athlètes qui ont remporté des médailles, avec le nombre classé dans l'ordre décroissant, pour la discipline Ice Hockey.

```
r8 <- DBI::dbFetch(DBI::dbSendQuery(conn = con, statement ="
SELECT pays.nomPays, COUNT(*) AS nb_medailles
FROM pays
JOIN athletes ON pays.idPays = athletes.idPNaissance
JOIN medailles ON athletes.idAthlete = medailles.idAthlete
JOIN disciplines ON athletes.idDiscipline = disciplines.idDiscipline
WHERE disciplines.nomDiscipline = 'Ice Hockey'
GROUP BY pays.nomPays
ORDER BY nb_medailles DESC;
"))
```

nomPays	nb_medailles
Finland	33
Canada	22
Slovakia	20
United States of America	20
Russian Federation	17
USSR	6
Czechoslovakia	3
Ukraine	1
Belarus	1
Austria	1
Croatia	1

Figure 2.9: Requête n°8

On obtient 11 enregistrements avec cette requête, on peut noter que seuls 11 pays sur 101 sont cités. De plus, il semblerait que les athlètes ayant gagné des medailles dans cette discipline sont majoritairement nés dans des pays avec un climat froid.

Requête n°9 : le nombre de disciplines pour lesquelles joue chacun des 101 pays.

```
r10 <- DBI::dbFetch(DBI::dbSendQuery(conn = con, statement ="
SELECT pays.nomPays, COUNT(DISTINCT athletes.idDiscipline) AS nb_disciplines
FROM pays
JOIN athletes ON pays.idPays = athletes.idPNaissance
GROUP BY pays.nomPays
ORDER BY nb_disciplines DESC;
"))
```

nomPays	nb_disciplines ▾ 1
United States of America	15
People's Republic of China	15
Russian Federation	15
Germany	14
Canada	14
Italy	14
Czech Republic	14
Switzerland	13

Figure 2.10: Requête n°9

Nous avons tiré deux extraits du résultat de cette requête, la première correspond à la première partie et la deuxième à la dernière. Nous avons trié le nombre de disciplines dans l'ordre décroissant. Sur les 101 pays, on remarque que sur la première partie pour les athlètes natifs de ces pays, les pays qui jouent dans le plus de disciplines proviennent de pays avec un climat modéré ou froid, tandis que pour la dernière partie de la requête, on remarque qu'il semble que les pays qui jouent dans le moins de disciplines proviennent de pays avec un climat modéré ou chaud. Cela pourrait être lié au fait que l'on étudie les données pour les Jeux Olympiques d'hiver.

Chapitre 3

Matériel et Méthodes

3.1 Logiciels

Logiciels utilisés

- Overleaf : Pour permettre à tous les membres de collaborer en temps réel sur le rapport du projet.
- PhpMyAdmin (Version : 5.2.0) : Pour permettre de visualiser la base de données et de réaliser des requêtes sur la base.
- Excel (Version : 2021): Pour permettre le nettoyage des données avant l'importation des bases de données dans phpMyAdmin.
- Libre Office (Version : 7.3): Pour permettre le nettoyage des données avant l'importation des bases de données dans phpMyAdmin.
- Google Docs et Collab: Pour permettre à tous les membres de collaborer en temps réel sur le rapport du projet.
- Mocodo (Version : 3.1.0) : Pour permettre de réaliser le MCD de la base de données.
- Whatsapp: Pour permettre de communiquer entre les membres du groupe.
- R Studio (Version : 4.2.2) : Pour l'écriture du code.
- Chat GPT: Pour la correction des fautes d'orthographe.

Chaque membre du groupe a utilisé son ordinateur personnel.

Informations techniques sur les ordinateurs utilisés:

- Ordinateur 1: Systeme d'exploitation : Windows, vitesse du processeur:11th Gen Intel(R) Core(TM) i5-1155G7 @ 2.50GHz 2.50 GHz
- Ordinateur 2: Systeme d'exploitation : Windows, vitess processeur: Intel(R) Core(TM) i5-10210U CPU @ 1.60GHz 2.11 GHz
- Ordinateur 3: Systeme d'exploitation : Windows, vitesse processeur: 12th Gen Intel(R) Core(TM) i5-1235U 1.30 GHz
- Ordinateur 4: Systeme d'exploitation: MacOS Big Sur, vitesse processeur: Apple M1 (CPU 8 coeurs / GPU 8 coeurs / Neural Engine 16 coeurs) cadencé à 3,2 GHz

3.2 Description des Données

Notre base de données contient 4 tables de données stockées au format csv.

La taille des tables est :

- Table athletes: 123 ko
- Table pays: 4 ko
- Table discipline: 4 ko
- Table medailles: 29 ko

Chaque table contient:

- Table athletes: 3 unités statistiques: nom, genre, Datenaissance
- Table pays: 1 unités statistiques: nompays
- Table discipline: 1 unités statistiques: nomdiscipline
- Table medailles: 2 unités statistiques: type, date

3.3 Nettoyage des données

Pour le nettoyage de données, nous avons décidé de garder toutes les lignes des fichiers afin d'avoir une analyse cohérente. Nous avons cependant enlevé certaines colonnes qui ne nous semblaient pas nécessaire après l'ajout des clés étrangères. Après avoir supprimé les doublons via Excel, nous avons utilisé la fonction INDEX(EQUIV) pour ajouter les identifiants en clés étrangères des pays et des disciplines correspondant à chaque athlète.

3.4 Modélisation statistique

Outils et méthodes de statistiques utilisés:

Notre modélisation statistique repose sur deux tests, le test d'indépendance du Chi-deux afin d'évaluer le lien entre deux variables; et le test de corrélation linéaire dans le but de trouver un lien entre deux variables différentes.

- Parmi les avantages, on pourrait souligner la facilité d'utiliser les codes qui nous ont permis de récuperer les requêtes SQL qu'on a pu effectuer sur phpmyAdmin, et que l'on a par la suite récuperé afin de les utiliser pour nos représentations graphiques. Mais encore il y a la partie graphique sur le même écran avec le logiciel R qui nous permet de visualiser nos graphiques à l'instant. Pour en revenir sur les graphiques, on peut notamment notifier les nombreux types de graphiques disponibles afin d'illustrer différentes types de requêtes.

- Néanmoins, nous avons aussi rencontré des difficultés comme des erreurs souvent apparues, ou encore le fait que l'on ne puisse pas se partager un même fichier RMD, et qu'il fonctionne par la suite sur chaque ordinateur pour chaque personne du groupe. D'autant plus, malgré qu'il y ait un grand nombres de graphiques disponibles, nous n'avons pas réussi à faire un graphique en particulier, comme le diagramme radar.

Chapitre 4

Analyse Exploratoire des Données

Afin d'étudier plus en profondeur nos données, répondre à des hypothèses et trouver un lien parmis elles, nous avons utilisé le logiciel R et des requêtes SQL.

4.1 Utiliser R

Premièrement, nous avons connecté le logiciel R à un compte phpmyAdmin afin d'y associer chaque requête effectuée dessus.

```
ggplot(r9, aes(x = pays, y = nb_medailles)) +
  geom_bar(stat = "identity") +
  labs(x = "Pays de naissance", y = "Nombre de médailles",
       title = "Nombre de médailles obtenues par pays en hockey sur glace") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

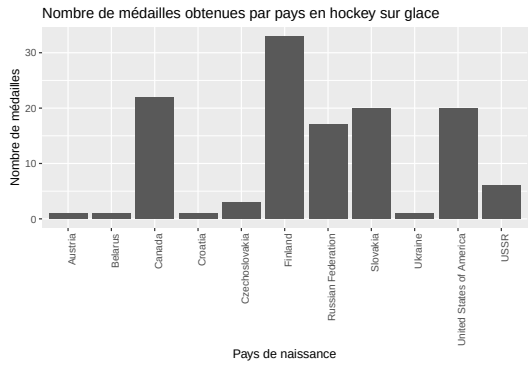

Figure 4.1: Nombre de médailles obtenues par pays pour le hockey sur glace.

La Finlande est le pays de naissance qui a remporté le plus de médailles, avec un total de 33 médailles. Le Canada est le deuxième pays de naissance avec le plus de médailles, pour un total de 22 médailles. Les États-Unis et la Slovaquie sont les troisièmes pays de naissance avec un total de 20 médailles chacun. Les autres pays ont une présence moins dominante dans les compétitions de hockey.

Nous pouvons voir que les athletes médaillés qui reviennent le plus sont ceux qui sont nés d'Europe de l'Est et d'Amérique du Nord. Nous pouvons donc nous demander quelles sont les régions du monde qui excellent le plus dans les JO d'hiver.

Afin de répondre à la problématique, nous cherchons alors à montrer les différences signicatives dans la répartition du type des médailles parmis les pays..

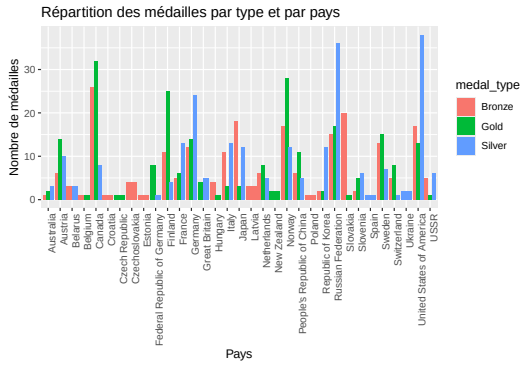

Figure 4.2: Répartition des médailles par type et par pays

Ce graphique illustre comment les médailles d'or, d'argent et de bronze sont distribuées parmi divers pays de naissance aux Jeux Olympiques. Les États-Unis et la Russie se démarquent nettement avec un total de 68 médailles chacun, illustrant leur position dominante. Le Canada, non loin derrière, rassemble 66 médailles, soulignant sa puissance et son excellence dans différentes disciplines. Le graphique inclut aussi des nations avec des performances plus discrètes, fournissant une perspective élargie sur l'étendue des réalisations olympiques parmi divers pays. Ces différences pourraient expliquer notammenent que certains pays ont plus de performances pour des disciplines particulières que d'autres, tout en sachant que ceux qui en ont obtenu 0 ne sont pas représentés sur le graphique.

Chapitre 5

Analyse et Résultats

Pour un troisième graphique, nous allons effectuer le test de corrélation linéaire avec des données centrées et réduites en cherchant à rejeter ou soutenir l'hypothèse suivante :

H1 : "Les performances des pays varient en fonction du type de discipline effectué."

```
suppressWarnings({
model <- lm(Nombre_Medailles ~ Nombre_Athletes, data = r7)
summary(model)
r7$Nombre_Athletes_CR <- scale(r7$Nombre_Athletes, center = TRUE, scale = TRUE)
r7$Nombre_Medailles_CR <- scale(r7$Nombre_Medailles, center = TRUE, scale = TRUE)
model_cr <- lm(Nombre_Medailles_CR ~ Nombre_Athletes_CR, data = r7)
summary(model_cr)
library(ggplot2)
ggplot(r7, aes(x = Nombre_Athletes_CR, y = Nombre_Medailles_CR)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, col = "blue") +
  labs(title = "Régression linéaire des données centrées et réduites",
       x = "Nombre d'athlètes (centré réduit)",
       y = "Nombre de médailles (centré réduit)")
correlation_coefficient <- cor(r7$Nombre_Athletes, r7$Nombre_Medailles,
                               use = "complete.obs", method = "pearson")
correlation_coefficient})
```

```
## [1] 0.8889788
```

Analyse des résultats de l'hypothèse 1 : Le nombre d'athlètes et le nombre de médailles sont corrélés de manière positive par la droite de régression, ce qui suggère que les pays de naissance avec le plus d'athlètes ont tendance à remporter plus de médailles. Une relation solide et positive entre ces deux variables est illustrée par un coefficient de corrélation de Pearson calculé et égal à 0.8889788 qui se rapproche de la valeur maximale 1. Cette forte corrélation linéaire montre que plus il y a d'athlètes, plus il y a de chances d'avoir de medaillés par pays de naissance.

Ainsi, le coefficient de corrélation de Pearson indique que l'hypothèse est vraie.

Pour finir, nous avons effectué le test d'indépendance du Chi-deux pour évaluer l'indépendance entre un pays de naissance et une discipline.

H2 : "Les performances des pays de naissance varient en fonction de la discipline."

```
library(ggplot2)
ggplot(r15, aes(x = pays, y = nb_medailles, fill = discipline)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title =
         "Nombre de médailles obtenues par pays et par discipline",
```

```
       x = "Pays",
       y = "Nombre de médailles") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

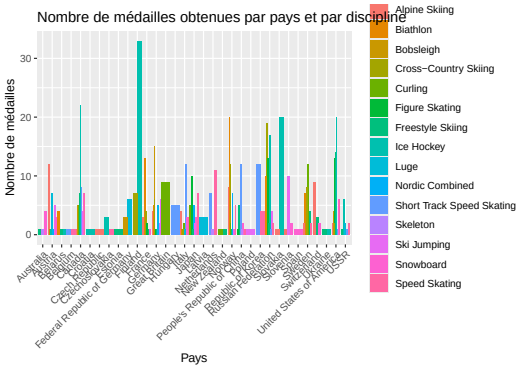

Figure 5.1: Nombre de médailles obtenues par pays et par discipline

```
table_contingence <- table(r15$pays, r15$discipline)
suppressWarnings({
  chi2_test <- chisq.test(table_contingence)
print(chi2_test)
})
```

```
##
##  Pearson's Chi-squared test
##
## data:  table_contingence
## X-squared = 343.38, df = 448, p-value = 0.9999
```

Analyse des résultats de l'hypothèse 2 : Ce graphique illustre la distribution des médailles par pays et par discipline lors des Jeux Olympiques d'hiver, couvrant des sports comme le ski alpin et le biathlon. L'analyse par le test de Chi2 a donné une valeur de 343.38 avec une valeur p de 0.9999, largement au-dessus du seuil accepté de 0.05.

Lorsque la statistique de Chi-deux est élevée, cela suggère que les variables ne sont pas indépendantes l'une de l'autre. Par ailleurs, avec une valeur de p trés élevée, on ne rejette pas l'hypothèse nulle qui décrit que la répartition des médailles entre les pays de naissance et les disciplines est indépendante. En revanche, on constate que le nombre élevé de ddl est dû à une multiplication du nombre de catégories de pays par le nombre de disciplines. Donc, la conclusion reste que la répartition des médailles n'est pas aléatoire et qu'il y a une relation significative entre le pays de naissance et la discipline en terme de répartition des médailles. La question serait-elle de voir pourquoi ce sont les athlètes nés dans les pays de l'Europe de l'Est et les Etats-Unis qui gagnent le plus de médailles.

Chapitre 6

Discussion

Certains graphiques mettent en évidence des pays de naissance clairement dominants dans certaines disciplines, comme les États-Unis en Snowboard ou la Finlande en Hockey sur glace, tandis que d'autres pays ont des succès plus mesurés dans des sports individuels.

Notre premiere hypothèse nous montre que les performances des pays varient en fonction du type de discipline effectué. En effet, grâce au coefficient de corrélation de Pearson, nous avons pu voir que plus il y a d'athlètes, plus il y aura de médailles, ce qui semble logique après réflexion donc on ne peux pas vraiment comparer les athlètes de chaque pays entre eux car chaque pays n'a pas le même nombre de participants pour chaque discipline. Par exemple, la France peut avoir 30 athlètes nés en France pour le Ski et la Russie 4, donc la France aura plus de probabilités de remporter des médailles.

Cette conclusion rend la réponse à notre problematique plus difficile. Par ailleurs, le test du Chi-deux que nous avons effectué lors de notre 2ème hypothèse nous approuve que les pays de naissance ayant le plus de médaille malgré la différence de disciplines reste les pays ou le nombre d'athlètes reste élevé. Lorsque nous avons effectué nos requêtes nous avons pu voir que les 3 pays avec le plus d'athlètes sont le Canada, les Etats- Unis et la Russie et ce sont les 3 pays qui ont le plus de médailles, ce qui prouve d'autant plus que nous ne pouvons pas complètement répondre à la question car le nombre d'athletes n'est pas proportionnel dans chaque pays ce qui fausse grandement nos calcules ou encore nos tests.

On constate que ces résultats insinuent que même si certaines tendances générales ressortent, des exceptions existent également, soulignant la complexité de la performance olympique dans toutes les disciplines et dans tous les pays. Ces analyses peuvent fournir une compréhension approfondie de la dynamique de la compétition olympique et guider les futures stratégies de développement des athlètes.

Chapitre 7

Conclusion et perspectives

En résumé, nous ne pouvons montrer qu'il y a des différences entre les disciplines pratiquées des athlètes selon leur pays de naissance car le nombre d'athlètes de chaque pays n'est pas le même.

De plus, les athlètes faisant partis d'un pays de naissance n'a pas forcément gagné de médailles donc ce n'est pas assez représentatif. Il aurait fallu avoir les classements des joueurs dans chaque discipline et le même nombre de joueurs dans chaque pays pour pouvoir faire des calculs et une analyse cohérente.

Etude sur le court terme :

notre analyse de réponse a été limité car il y a plusieurs paramètres à prendre en compte.

Etude sur le long terme

on pourrait ajouter une table tels que les continents pour faire des hypothèses selon la provenance des athlètes par continent ce qui serait déjà plus representatif. De plus, on pourrait pour chaque discipline faire une table épreuve avec en attributs le nom de chaque épreuve, le moyen de jeu (en équipe ou seul) pour que la remise de médaille soit beaucoup plus représentatif.

Difficultés :

- Importation des données difficiles surtout pour la création des différentes clés étrangères
- Problématiques des base de données entre Windows et MacOs
- Grand nombre de données manquantes

Annexes

Malgré que nous ayons effectué différents graphiques pour illustrer notre réponse à la problématique, nous pouvons en ajouter afin de montrer pour 3 disciplines différentes, une répartition bien aléatoire de réussite et d'obtention des médailles selon les pays.

Codes

```
library(DBI)
con <- DBI::dbConnect(
  drv = RMySQL::MySQL(),
  host = "localhost",
  port = 3306,
  username = "root",
  password = "root",
  dbname = "beijingdata")
d2 <- DBI::dbFetch(DBI::dbSendQuery(conn = con, statement ="
  SELECT pays.pays, COUNT(*) AS nb_medailles
  FROM pays
  JOIN athletes ON pays.idpays = athletes.idpaysnaissance
  JOIN medailles ON athletes.idathletes = medailles.idathlete
  JOIN disciplines
  ON athletes.iddiscipline = disciplines.codeDiscipline
  WHERE disciplines.discipline = 'Freestyle Skiing'
  GROUP BY pays.pays
  ORDER BY nb_medailles DESC;"))
library(ggplot2)
ggplot(d2, aes(x = "", y = nb_medailles, fill = pays)) +
  geom_bar(stat = "identity") +
  coord_polar("y", start = 0) +
  labs(title =
         "Répartition des médailles pour le Freestyle Skiing par pays") +
  theme_void() +
  theme(legend.position = "bottom",
        plot.title.position = "plot")
```

Figure 7.1: Répartition des médailles pour le Freestyle Skiing par pays

La distribution des médailles en Freestyle Skiing par pays est présentée sur le diagramme, avec les États-Unis, la Norvège et le Canada comme principaux champions de cette discipline. Il y a aussi d'autres pays tels que la France, l'Italie et la Suisse, mais avec moins de médailles, tandis que des pays tels que l'Australie, l'Autriche et le Japon sont plus modestes.

```
library(DBI)
con <- DBI::dbConnect(
  drv = RMySQL::MySQL(),
  host = "localhost",
  port = 3306,
  username = "root",
  password = "root",
  dbname = "beijingdata")
d3 <- DBI::dbFetch(DBI::dbSendQuery(conn = con, statement ="
SELECT pays.pays, COUNT(*) AS nb_medailles
FROM pays
JOIN athletes ON pays.idpays = athletes.idpaysnaissance
JOIN medailles
ON athletes.idathletes = medailles.idathlete
JOIN disciplines
ON athletes.iddiscipline = disciplines.codeDiscipline
WHERE disciplines.discipline = 'Snowboard'
GROUP BY pays.pays
ORDER BY nb_medailles DESC;"))
library(ggplot2)
ggplot(d3, aes(x = "", y = nb_medailles, fill = pays)) +
  geom_bar(stat = "identity") +
  coord_polar("y", start = 0) +
  labs(title
       = "Répartition des médailles pour la discipline Snowboard par pays") +
  theme_void()
```

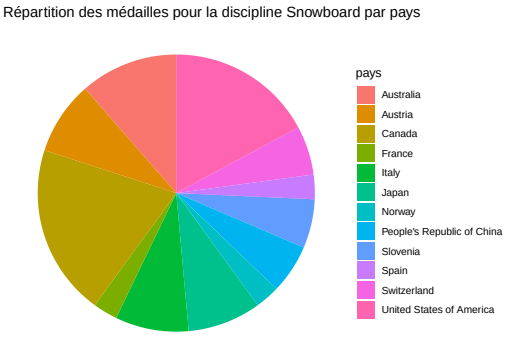

Figure 7.2: Répartition des médailles pour la discipline Snowboard par pays

Ce diagramme circulaire présente la distribution des médailles en Snowboard de vitesse par pays, mettant en avant les meilleures performances de pays tels que les États-Unis, la Russie et les Pays-Bas. Il présente également une forte présence canadienne et norvégienne. Des résultats plus mesurés sont obtenus par d'autres pays tels que la Suède, l'Italie et le Japon. Ce schéma résume de manière très efficace les compétences en Snowboard de vitesse à l'échelle mondiale, mettant en évidence les pays qui ont connu une perfection historique dans cette discipline.

```
library(DBI)
con <- DBI::dbConnect(
  drv = RMySQL::MySQL(),
  host = "localhost",
  port = 3306,
```

```
  username = "root",
  password = "root",
  dbname = "beijingdata")
d4 <- DBI::dbFetch(DBI::dbSendQuery(conn = con, statement ="
SELECT pays.pays, COUNT(*) AS nb_medailles
FROM pays
JOIN athletes ON pays.idpays = athletes.idpaysnaissance
JOIN medailles
ON athletes.idathletes = medailles.idathlete
JOIN disciplines
ON athletes.iddiscipline = disciplines.codeDiscipline
WHERE disciplines.discipline = 'Speed Skating'
GROUP BY pays.pays
ORDER BY nb_medailles DESC;"))
library(ggplot2)
ggplot(d4, aes(x = "", y = nb_medailles, fill = pays)) +
  geom_bar(width = 1, stat = "identity") +
  coord_polar("y", start = 0) +
  labs(title =
         "Répartition des médailles pour la discipline 'Speed Skating' par pays",
       fill = "Pays") +
  theme_void() +
  theme(legend.position = "right")
```

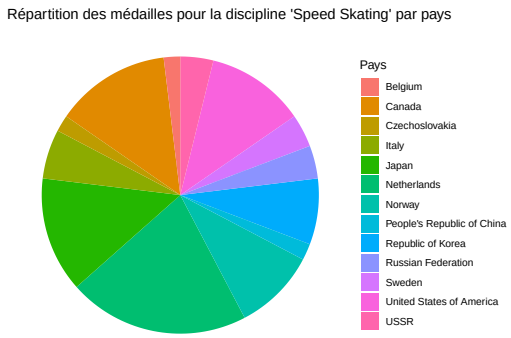

Figure 7.3: Répartition des médailles pour la discipline 'Speed Skating par pays

Ce diagramme circulaire représente la distribution des médailles pour la discipline " Speed Skating" par pays, mettant en évidence la bonne performance des États-Unis, de la Russie et du Canada. Il met également en lumière des apports importants de la France, de la Suisse et de la Norvège. D'autres nations telles que l'Allemagne, le Japon et la Suède affichent des résultats plus modestes dans ce domaine. Ce graphique facilite la détection rapide des principaux acteurs mondiaux du Speed Skating et permet de comparer leur réussite relative.

Tables

Table 7.1: Epreuve

Nom colonne	Type	Signification	Caractéristique
idEpreuve	int	identifiant	clé primaire
nomEpreuve	varchar	nom	champs obligatoire
typeEpreuve	varchar	seul/équipe	champs obligatoire

Table 7.2: Continent

Nom colonne	Type	Signification	Caractéristique
idContinent	int	identifiant	clé primaire
nomContinent	varchar	nom	champs obligatoire