Cours d’analyse exploratoire

1. Compréhension du problème
2. Données disponibles
   * Exploitation
     + SQL
     + Python
     + BI
3. Voir si j’obtiens un élément de réponse -X
4. Préparer les données
   * PPT
   * Dashboard
   * Modèle ML (modèle de prévision)
5. Mettre la donnée à disposition

Exercice

1. Est-ce qu’un pays qui héberge les Jeux olympiques est meilleur qu’habituellement ?
2. Combien de médailles vont être remportées par la France ?

**Analyse des données du Jeux Olympique Tokyo pour prédire celui de la France**

**Période d’étude : 1990 - 2016**

|  |  |  |  |
| --- | --- | --- | --- |
| **Édition été** | **Pays** | **Édition Hiver** | **Pays** |
| **2016** | **Brésil** | **2018** | **Corée du sud** |
| **2012** | **Grande Bretagne** | **2014** | **Russie** |
| **2008** | **Chine** | **2010** | **Canada** |
| **2004** | **Grèce** | **2006** | **Italie** |
| **2000** | **Australie** | **2002** | **USA** |

1. **Nombre de discipline par édition**

**SELECT count(distinct Sport) as nombre\_discipline,Season,Year**

**FROM `jeux-olympiques-371410.Tokyo\_2024.athlete\_events`**

**where Year = à compléter**

**Group by Season, Year**

**Having Season = "à compléter"**

**order by Year**

1. **Nombre de disciplines auxquelles participe un pays par édition**

**with athlete as (**

**SELECT  ID, Name, NOC, Year, Season, City, Sport, Event, Medal**

**FROM `jeux-olympiques-371410.Tokyo\_2024.athlete\_events`**

**),**

**region as (**

**SELECT NOC, Region**

**FROM `jeux-olympiques-371410.Tokyo\_2024.noc\_regions\_1`**

**),**

**palmares as (**

**select \***

**from athlete**

**full join region**

**using (NOC)**

**)**

**SELECT count(distinct Sport)  as nombre\_discipline,Season,Year**

**FROM palmares**

**Where NOC = "à compléter" and Year = à compléter**

**Group by Season, Year, NOC**

**Having Season = "à compléter"**

1. **Nombre de compétions par édition**

**SELECT count(distinct Event)  as nombre\_ discipline,Season,Year**

**FROM `jeux-olympiques-371410.Tokyo\_2024.athlete\_events`**

**where Year = à compléter**

**Group by Season, Year**

**Having Season = "à compléter"**

**order by Year**

**nombre\_discipline Season Year**

**29 Summer 1992**

**31 Summer 1996**

**34 Summer 2000**

**34 Summer 2004**

**34 Summer 2008**

**32 Summer 2012**

**34 Summer 2016**

1. **Nombre de compétitions auxquelles un pays participe par édition**

**with athlete as (**

**SELECT  ID, Name, NOC, Year, Season, City, Sport, Event, Medal**

**FROM `jeux-olympiques-371410.Tokyo\_2024.athlete\_events`**

**),**

**region as (**

**SELECT NOC, Region**

**FROM `jeux-olympiques-371410.Tokyo\_2024.noc\_regions\_1`**

**),**

**palmares as (**

**select \***

**from athlete**

**full join region**

**using (NOC)**

**)**

**SELECT count(distinct Event)  as nombre\_competion,Season,Year**

**FROM palmares**

**Where NOC = "à compléter" and Year = à compléter**

**Group by Season, Year, NOC**

**Having Season = "à compléter"**

1. **Classement médaille par pays sur une édition**

**with athlete as (**

**SELECT  ID, Name, NOC, Year, Season, City, Sport, Event, Medal**

**FROM `jeux-olympiques-371410.Tokyo\_2024.athlete\_events`**

**),**

**region as (**

**SELECT NOC, Region**

**FROM `jeux-olympiques-371410.Tokyo\_2024.noc\_regions\_1`**

**),**

**palmares as (**

**select \***

**from athlete**

**full join region**

**using (NOC)**

**)**

**SELECT count(distinct Event)  as nombre\_medaille,Season,Year, pays**

**FROM palmares**

**Where Medal != "NA" AND Year = à compléter**

**Group by Season, Year, NOC, pays**

**order by nombre\_medaille desc**

1. **classement par médaille d’or sur une édition**

**with athlete as (**

**SELECT  ID, Name, NOC, Year, Season, City, Sport, Event, Medal**

**FROM `jeux-olympiques-371410.Tokyo\_2024.athlete\_events`**

**),**

**region as (**

**SELECT NOC,**Region

**FROM `jeux-olympiques-371410.Tokyo\_2024.noc\_regions\_1`**

**),**

**palmares as (**

**select \***

**from athlete**

**full join region**

**using (NOC)**

**)**

**SELECT count(distinct Event)  as nombre\_medaille,Season,Year, pays**

**FROM palmares**

**Where Medal = "Gold" AND Year = à compléter**

**Group by Season, Year, NOC, pays**

**order by nombre\_medaille desc**

1. **Classement par nombre total des médailles en utilisant le coefficient par type de médaille**

**with athlete as (**

**SELECT  ID, Name, NOC, Year, Season, City, Sport, Event, Medal**

**FROM `jeux-olympiques-371410.Tokyo\_2024.athlete\_events`**

**),**

**region as (**

**SELECT NOC, Region**

**FROM `jeux-olympiques-371410.Tokyo\_2024.regions`**

**),**

**palmares as (**

**select \***

**from athlete**

**full join region**

**using (NOC)**

**),**

**t1 as (**

**SELECT count(distinct Event)  as Or\_medaille,Season,Year, NOC**

**FROM palmares**

**Where Medal = "Gold" AND Year = à compléter**

**Group by Season, Year, NOC**

**),**

**t2 as (**

**SELECT count(distinct Event)  as Silver,Season,Year, NOC**

**FROM palmares**

**Where Medal = "Silver" AND Year = à compléter**

**Group by Season, Year, NOC**

**),**

**t3 as (**

**SELECT count(distinct Event)  as Bronze,Season,Year, NOC**

**FROM palmares**

**Where Medal = "Bronze" AND Year = à compléter**

**Group by Season, Year, NOC**

**),**

**t1\_t2 as(**

**select \***

**from t1 full join t2**

**using (Season, NOC, Year)**

**)**

**Select \*,**

**0.5\* ifnull(Or\_medaille,0) + 0.3\*ifnull(Silver,0) + 0.2\*ifnull(Bronze,0) as coeff**

**from t3 full join t1\_t2**

**using (Season, NOC, Year)**

**Order by coeff desc**

1. **Table récapitulative par édition**

**with athlete as (**

**SELECT  ID, Name, NOC, Year, Season, City, Sport, Event, Medal**

**FROM `jeux-olympiques-371410.Tokyo\_2024.athlete\_events`**

**),**

**region as (**

**SELECT NOC,**

**Region**

**FROM `jeux-olympiques-371410.Tokyo\_2024.regions`**

**),**

**palmares as (**

**select \***

**from athlete**

**full join region**

**using (NOC)**

**),**

**t1 as (**

**SELECT count(distinct Event)  as Or\_medaille,Season,Year, NOC**

**FROM palmares**

**Where Medal = "Gold" AND Year = à compléter**

**Group by Season, Year, NOC**

**),**

**t2 as (**

**SELECT count(distinct Event)  as Silver,Season,Year, NOC**

**FROM palmares**

**Where Medal = "Silver" AND Year = à compléter**

**Group by Season, Year, NOC**

**),**

**t3 as (**

**SELECT count(distinct Event)  as Bronze,Season,Year, NOC**

**FROM palmares**

**Where Medal = "Bronze" AND Year = à compléter**

**Group by Season, Year, NOC**

**),**

**t1\_t2 as(**

**select \***

**from t1 full join t2**

**using (Season, NOC, Year)**

**)**

**Select \*,**

**0.5\* ifnull(Or\_medaille,0) + 0.3\*ifnull(Silver,0) + 0.2\*ifnull(Bronze,0) as coeff,**

**ifnull(Or\_medaille,0) + ifnull(Silver,0) + ifnull(Bronze,0) as total**

**from t3 full join t1\_t2**

**using (Season, NOC, Year)**

**Order by coeff desc**

**Groupez cette étude de type des jeux soit winter soit summer; ne mélangez pas**

**Après ça, faites une étude comparative pour vérifier l’évolution des différentes performances par rapport à chaque pays entre 1990 et 2016**

**Est-ce que les performances sont meilleures quand on est le pays organisateur ?**

**Donner le taux d’accroissement des performances entre les éditions non organisateurs et les éditions organisateurs**

**Donnez :**

1. **max médaille obtenu sur une édition pour**

**First Answer**

**1→ SELECT Year,count(Medal) as Medailles, NOC**

**FROM `projetdataeexploreur.Donnees\_JO.athlete\_events`**

**where Medal in('Gold', 'Silver', 'Bronse')**

**group by NOC,Year**

**having count(Medal) is not NULL**

**LIMIT 50**

**but we can used query with to extract between 1990 –2021**

**Option two :**

**SELECT Year,count(Medal) as Medailles, NOC**

**FROM `projetdataeexploreur.Donnees\_JO.athlete\_events`**

**where Medal in('Gold', 'Silver', 'Bronse') and Year between 1990 and 2021**

**group by NOC,Year**

**2→ Question two**

**SELECT Year,count(Medal) as Medailles, NOC**

**FROM `projetdataeexploreur.Donnees\_JO.athlete\_events`**

**where Medal in( 'Silver')**

**group by NOC,Year**

**Option Two**

**SELECT Year,count(Medal) as Medailles, NOC**

**FROM `projetdataeexploreur.Donnees\_JO.athlete\_events`**

**where Medal in( 'Silver') and Year between 1990 and 2021**

**group by NOC,Year**

**For each Country**

**SELECT count(distinct Event) as nbre\_event,NOC,Year**

**FROM `projetdataeexploreur.Donnees\_JO.athlete\_events`**

**WHERE  Medal!='NA'AND NOC='USA' AND Year BETWEEN 1990 AND 2021 AND Season ='Summer'**

**group by NOC,Year**

**For each Country**

**SELECT count(distinct Event) as nbre\_event,NOC,Year**

**FROM `projetdataeexploreur.Donnees\_JO.athlete\_events`**

**WHERE  Medal!='NA'AND NOC='CAN' AND Year BETWEEN 1990 AND 2021 AND Season ='Summer'**

**group by NOC,Year**

**Médaille pour le canada en hiver**

**nbre\_event NOC Year**

7 CAN 1992

12 CAN 1994

13 CAN 1998

15 CAN 2002

21 CAN 2006

24 CAN 2010

21 CAN 2014

**Médaille pour la grêce en été**

**nbre\_event NOC Year**

2 GRE 1992

8 GRE 1996

13 GRE 2000

16 GRE 2004

4 GRE 2008

2 GRE 2012

6 GRE 2016

**Médaille** d’or **pour la grêce en été**

**nbre\_event NOC Year**

2 GRE 1992

4 GRE 1996

4 GRE 2000

6 GRE 2004

3 GRE 2016

**Médaille** d’argent **pour la grêce en été**

**nbre\_event NOC Year**

4 GRE 1996

6 GRE 2000

6 GRE 2004

2 GRE 2008

1 GRE 2016

**Médaille** de bronze **pour la grêce en été**

**nbre\_event NOC Year**

3 GRE 2000

4 GRE 2004

2 GRE 2008

2 GRE 2012

2 GRE 2016

**Médaille pour l’Australie en été :**

|  |  |
| --- | --- |
| **26** | **1992** |

|  |  |
| --- | --- |
| **39** | **1996** |

|  |  |
| --- | --- |
| **55** | **2000** |

|  |  |
| --- | --- |
| **46** | **2004** |

|  |  |
| --- | --- |
| **44** | **2008** |

|  |  |
| --- | --- |
| **35** | **2012** |

|  |  |
| --- | --- |
| **29** | **2016** |

**médaille d’Italie en hiver**

**\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***

**Après ça, faites une étude comparative pour vérifier l’évolution des différentes performances par rapport à chaque pays entre 1990 et 2016**

**Est-ce que les performances sont meilleures quand on est le pays organisateur ?**

**Donner le taux d’accroissement des performances entre les éditions non organisateurs et les éditions organisateurs**

**Graphe Page 2**

**Indicateur de performance basé sur le taux de victoire (ou réussite) calcul basé sur le ((nombre de médailles gagnés \* 100) / le nombre de compétitions auxquelles elle a participé).**