MÃfÂ©dailles Jeux Olympiques d'hiver de 1924 ÃfÆ'Ã,Â  2010

Kathleen GALVIN & Marie BERGERET

30 mars 2017

Dans le cadre du cours de Modèles Statistiques, nous allons analyser un jeu de données, selon les méthodes étudiées en TD. Nous avons donc cherché des données sur un sujet qui nous intéressait. Grâce au site www.data.gouv.fr, nous avons trouvé une base de données contenant des informations sur les nations médaillées au Jeux Olympiques (JO) d'hiver entre les années 1924 et 2010.

Dans un premier temps il a été nécessaire d'installer plusieurs packages :

library(readr)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(ggplot2)

# Présention des données

Les données sources se situent dans le fichier 'winter\_olympics\_medals.csv', nous allons donc les importer.

## Importation

Ci-dessous se trouve un extrait des données, à savoir les 7 premières lignes du fichier (Intitilé des colonnes et lignes de données).

donnees<- read\_csv(file = "winter\_olympics\_medals.csv")

## Parsed with column specification:  
## cols(  
## id = col\_integer(),  
## year = col\_integer(),  
## sport = col\_character(),  
## medal = col\_character(),  
## code = col\_character(),  
## country = col\_character(),  
## host = col\_logical()  
## )

head(donnees)

## # A tibble: 6 × 7  
## id year sport medal code country host  
## <int> <int> <chr> <chr> <chr> <chr> <lgl>  
## 1 31666 1924 biathlon gold SUI Switzerland FALSE  
## 2 31666 1924 biathlon silver FIN Finland FALSE  
## 3 31666 1924 biathlon bronze FRA France TRUE  
## 4 31688 1924 bobsleigh gold SUI Switzerland FALSE  
## 5 31688 1924 bobsleigh silver GBR GreatBritain FALSE  
## 6 31688 1924 bobsleigh bronze BEL Belgium FALSE

Cette base de données est contituée de 2 variables quantitatives ('int') et de 5 variables qualtatives ('chr' et 'lgt') : - id : Identifiant (il est le même pour un même sport et une même année, nous ne possédons pas le nom des différentes épreuves)  
- year : Année des JO  
- sport : Nom du sport (et pas celui de l'épreuve)  
- medal : Couleur de la médaille obtenue - gold, silver, bronze -  
- code : Code de la nation  
- country : Nom de la nation  
- host : Booléen qui indique si le pays ayan gagné la médaille était le pays d'acceuil ou non - TRUE, FALSE -

Une ligne correspond à la victoire d'une nation dans un sport lors d'une année de jeux.

## Dates

Nous possédons les données des Jeux Olympiques d'hivers qui ont eu lieux entre les deux années suivantes.

min(donnees$year)

## [1] 1924

max(donnees$year)

## [1] 2010

Les JO d'hiver ayant lieu tous les quatres ans, il n'est pas nécessaire d'afficher toutes les années.

## Sports

Voici la liste des différents sports représentés de 1924 à 2010 :

donnees%>%select(sport)%>%unique()

## # A tibble: 15 × 1  
## sport  
## <chr>  
## 1 biathlon  
## 2 bobsleigh  
## 3 crosscountry  
## 4 figureskating  
## 5 icehockey  
## 6 nordiccombined  
## 7 skijumping  
## 8 speedskating  
## 9 curling  
## 10 skeleton  
## 11 alpineskiing  
## 12 luge  
## 13 freestyleskiing  
## 14 shorttrack  
## 15 snowboard

Il y a donc 15 sport différents aux JO d'hiver (du moins jusqu'en 2010).

## Pays

Lors de tous ces jeux, de nombreux pays ont participé.

### Participants

Ces données se situent dans un autre fichier csv que nous allons importer : 'pays.csv'. Ci-dessous un extrait de ces données (Code et nom de nation).

dpays<- read\_csv(file = "pays.csv")

## Parsed with column specification:  
## cols(  
## code = col\_character(),  
## country = col\_character()  
## )

head(dpays)

## # A tibble: 6 × 2  
## code country  
## <chr> <chr>  
## 1 AFG Afghanistan  
## 2 RSA South Africa  
## 3 ALB Albania  
## 4 ALG Algeria  
## 5 GER Germany  
## 6 AND Andorra

Nous allons maintenant compter le nombre de nations ayant participé aus JO.

dpays%>%select(code)%>%unique()%>%summarise(N=n())

## # A tibble: 1 × 1  
## N  
## <int>  
## 1 228

Ainsi, 228 pays ont participé au moins un sport d'un JO entre ces deux dates.

### Gagnants

Nous allons désormais nous concentrer sur les pays ayant célébré au moins une victoire, c'est à dire ayant gagné médaille.

donnees%>%select(country)%>%unique()%>%summarise(N=n())

## # A tibble: 1 × 1  
## N  
## <int>  
## 1 45

donnees%>%select(code)%>%unique()%>%summarise(N=n())

## # A tibble: 1 × 1  
## N  
## <int>  
## 1 42

Ainsi, 42 nations ont gagné des médailles entre 1924 et 2010. Nous avons remarqué qu'en comptant par nom de pays ('country') plutôt que par le code ('code'), nous obtenions plus de résultats. Cela s'explique par le fait que certains pays ont changé de nom au cours du temps, comme l'Allemagne : le code est resté'GER' alors que le nom est passé de 'Federal Republic of Germany (1950-1990)' et 'German Democratic Republic (1955-1990)' à 'Germany'. Il nous ait paru plus cohérent de travailler avec le code des pays.

Maintenant que nous avons présenté les données, nous allons les analyser selon une certaine ligne de reflexion. Dans un premier temps, nous allons essayer de trouver les pays qui ont gagné le plus (et réciproquement le moins) de médailles sur l'ensemble de ces jeux, puis par année de Jo et enfin par sport. Ensuite, nous allons regarder si ce sont toujours les mêmes pays qui restent "au sommet" ou si, au cours des années, de nouveaux pays "compétitifs" ont émergé. Enfin, dans une dernière partie qui résulte surtout de notre curiostité, nous étudierons le cas de la France.

# Quel pays gagne le plus/ le moins ?

Nous allons maintenant regarder la répartition du nombre de médailles gagnées par pays sur l'ensembles de JO.

## Peu importe la couleur de la médaille

On va donc compter le nombre de médaille par pays (nous raisonnons toujours sur le code) sur les 86 ans de données que nous possédons.

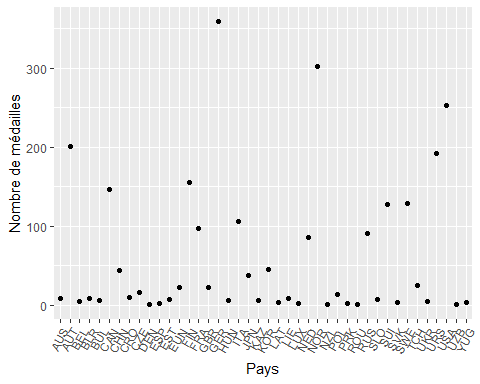
donnees%>%group\_by(code)%>%summarise(N=n())

## # A tibble: 42 × 2  
## code N  
## <chr> <int>  
## 1 AUS 9  
## 2 AUT 201  
## 3 BEL 5  
## 4 BLR 9  
## 5 BUL 6  
## 6 CAN 146  
## 7 CHN 44  
## 8 CRO 10  
## 9 CZE 16  
## 10 DEN 1  
## # ... with 32 more rows

Nous allons mainteant transformer ces données en graphique afin de mieux les visuliser.

donnees%>%group\_by(code)%>%summarise(N=n())%>%arrange(-N)%>%ggplot(aes(x=code, y=N))+geom\_point(width=0.3)+xlab("Pays")+ylim(0,NA)+ylab("Nombre de médailles")+ theme(axis.text.x = element\_text(angle = 60, vjust = 1, hjust=1))

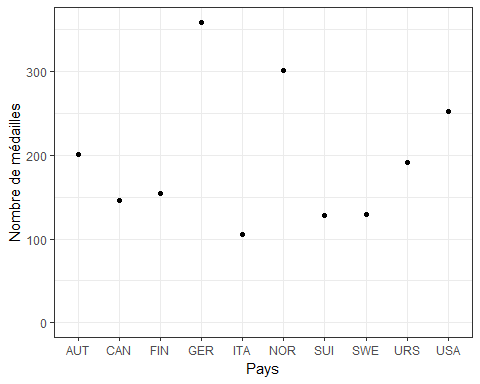
## Warning: Ignoring unknown parameters: width

 Le nombre de pays étant grand il n'est pas forcément aisé de lire ce graphique mais on voit quand même très bien les nations qui ont le plus gagné: Autriche, Allemagne, Norvège... De plus il est égalemnet possible de voir en un coup d'oeil les pays qui ont au contraire, gagné le moins de médailles : Roumanie, Danemark...

Pour une meilleur lisibilité, nous allons maintenant extraire les 10 pays qui ont gagné le plus de médailles sur ces années

MedailleGeneral<-donnees%>%group\_by(code)%>%summarise(N=n())%>%arrange(-N)%>%slice(1:10)  
MedailleGeneral%>%ggplot(aes(x=code, y=N))+geom\_point(width=0.3)+xlab("Pays")+ylim(0,NA)+ylab("Nombre de médailles")+theme\_bw()

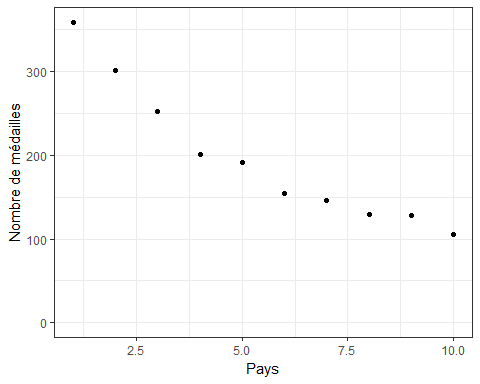
## Warning: Ignoring unknown parameters: width

 Nous pouvons donc voir les 10 "meilleurs" pays, enfin, plus rigoureusement les pays qui ont gagné le plus de médailles (toutes couleurs confondues) lors de l'ensemble des JO. Comme nous le pensions les pays tels que l'Allemagne, les pays scandinaves ou encore les Etats Unis y figurent. Cependant la france n'entre pas dans le top 10, mais nous verron sont cas dans la dernière partie du rapport.

Nous n'avons pas d'un coup d'oeil un classement par ordre croissant ou décroissant, nous allons donc essayer d'ordonner le graphique.

donnees%>%group\_by(code)%>%summarise(N=n())%>%arrange(-N)%>%slice(1:10)%>%mutate(ordre=1:10)%>%ggplot(aes(x=ordre, y=N))+geom\_point(width=0.3)+xlab("Pays")+ylim(0,NA)+ylab("Nombre de médailles")+theme\_bw()

## Warning: Ignoring unknown parameters: width

 Ainsi, nous visualisons beaucoup plus facilement le "classement" et le nombre de médaille de chaque nation : Allemagne, Norvège...

### Le moins ?

Nous allons ensuite extraire les 10 pays qui ont gagné le moins de médailles sur ces 86 années.

donnees%>%group\_by(code)%>%summarise(N=n())%>%arrange(N)%>%slice(1:10)

## # A tibble: 10 × 2  
## code N  
## <chr> <int>  
## 1 DEN 1  
## 2 NZL 1  
## 3 ROU 1  
## 4 UZB 1  
## 5 ESP 2  
## 6 LUX 2  
## 7 PRK 2  
## 8 LAT 3  
## 9 SVK 4  
## 10 YUG 4

On peut voir que tous ces pays ont gagné moins de 5 médailles au JO d'hiver.

Maintenant que nous avons vu les pays qui gagnent le plus et le moins de médailles couleurs confondues, nous allons regarder pour chaque couleur de médaille afin de voir si ce sont toujours les mêmes pays et dans le même ordre.

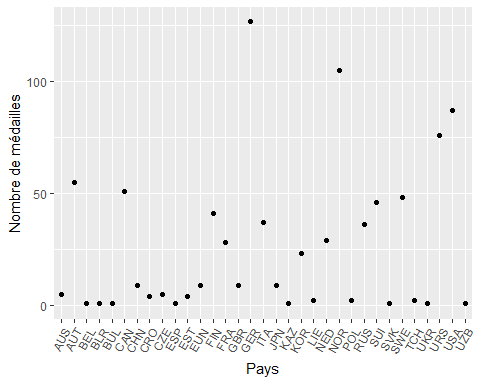
## Selon la couleur de la médaille

### Or

Nous allons faire comme précédemment mais en ne sélectionnant que les médailles d'or.

MedailleOr<-donnees%>%group\_by(code)%>%filter(medal=='gold')%>%summarise(N=n())%>%arrange(-N)%>%slice(1:10)  
donnees%>%group\_by(code)%>%filter(medal=='gold')%>%summarise(N=n())%>%ggplot(aes(x=code, y=N))+geom\_point(width=0.3)+xlab("Pays")+ylim(0,NA)+ylab("Nombre de médailles")+ theme(axis.text.x = element\_text(angle = 60, vjust = 1, hjust=1))

## Warning: Ignoring unknown parameters: width

 Comme il n'est pas très facile de confronter le plot des médailles toutes couleurs confondues et des médailles d'or, et afin de répondre à notre question, nous allons confronter les liste des 10 pays qui ont gagné le plus dans les deux cas.

tout<-MedailleGeneral%>%mutate(ordre=1:10)%>%select(code,ordre)%>%arrange(code)

or<-MedailleOr%>%mutate(ordre=1:10)%>%select(code,ordre)%>%arrange(code)

tout$code

## [1] "AUT" "CAN" "FIN" "GER" "ITA" "NOR" "SUI" "SWE" "URS" "USA"

tout$ordre

## [1] 4 7 6 1 10 2 9 8 5 3

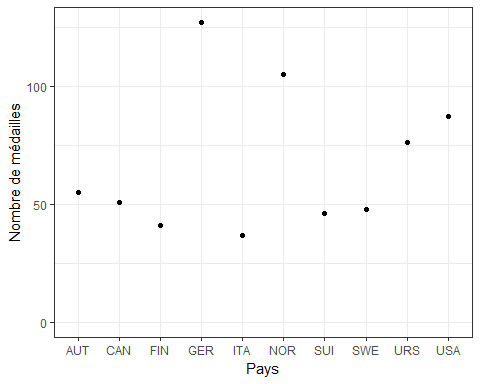
or$ordre

## [1] 5 6 9 1 10 2 8 7 4 3

Donc même si l'affichage n'est pas optimal, on peut déjà voir que ce sont la même liste de pays qui gagnent le plus de médailles en général et les médailles d'or. seul l'ordre varie légèrement. En effet les trois premiers pays (Allemagne, Norvège et Finlande) restent classés dans le même ordre. Pour la plupart des nations nous observons une variation d'un rang sauf pour la Finlande qui est trois places derrière quand on regarde les médailles d'or.

donnees%>%group\_by(code)%>%filter(medal=='gold')%>%summarise(N=n())%>%arrange(-N)%>%slice(1:10)%>%ggplot(aes(x=code, y=N))+geom\_point(width=0.3)+xlab("Pays")+ylim(0,NA)+ylab("Nombre de médailles")+theme\_bw()

## Warning: Ignoring unknown parameters: width

 L'allure de ce graphique est similaire au précédent concernant les médailles en général, les écart entre les différents pays sont respectés, par exemple l'Allemagne et la Norvège se distinguent toujours autant.

### Argent

Nous allons rapidement regarder si c'est la même chosse pour les médailles d'argent

MedailleArgent<-donnees%>%group\_by(code)%>%filter(medal=='silver')%>%summarise(N=n())%>%arrange(-N)%>%slice(1:10)%>%arrange(code)  
MedailleArgent

## # A tibble: 10 × 2  
## code N  
## <chr> <int>  
## 1 AUT 70  
## 2 CAN 48  
## 3 FIN 59  
## 4 GER 129  
## 5 ITA 32  
## 6 NOR 107  
## 7 SUI 36  
## 8 SWE 34  
## 9 URS 58  
## 10 USA 90

Ce sont toujours les 10 mêmes pays.

argent<-MedailleOr%>%mutate(ordre=1:10)%>%select(code,ordre)%>%arrange(code)  
tout$code

## [1] "AUT" "CAN" "FIN" "GER" "ITA" "NOR" "SUI" "SWE" "URS" "USA"

tout$ordre

## [1] 4 7 6 1 10 2 9 8 5 3

or$ordre

## [1] 5 6 9 1 10 2 8 7 4 3

argent$ordre

## [1] 5 6 9 1 10 2 8 7 4 3

Encore une fois l'ordre du podium est conservé et comme pour les médailles d'or, très peu de mouvement de rangs.

### Bronze

Et enfin pour les medailles de bronze :

MedailleBronze<-donnees%>%group\_by(code)%>%filter(medal=='bronze')%>%summarise(N=n())%>%arrange(-N)%>%slice(1:10)  
MedailleBronze%>%mutate(ordre=1:10)

## # A tibble: 10 × 3  
## code N ordre  
## <chr> <int> <int>  
## 1 GER 103 1  
## 2 NOR 90 2  
## 3 AUT 76 3  
## 4 USA 76 4  
## 5 URS 58 5  
## 6 FIN 55 6  
## 7 CAN 47 7  
## 8 SWE 47 8  
## 9 SUI 46 9  
## 10 FRA 42 10

Cette fois nous ne pouvons pas faire la même méthode car un pays a été remplace : la France a plus gagné de médaille que l'Italie ce qui l'a fait sortir du top 10. L'Autriche a également fait sa place au détriment des Etats Unis

# Analyse par Jeux Olympiques

Peu importe la couleur de la médaille

donnees%>%group\_by(year,code)%>%summarise(N=n())%>%arrange(-N)%>%slice(1:3)

## Source: local data frame [63 x 3]  
## Groups: year [21]  
##   
## year code N  
## <int> <chr> <int>  
## 1 1924 NOR 16  
## 2 1924 FIN 11  
## 3 1924 GBR 4  
## 4 1928 NOR 14  
## 5 1928 USA 6  
## 6 1928 SWE 5  
## 7 1932 USA 12  
## 8 1932 NOR 10  
## 9 1932 CAN 7  
## 10 1936 NOR 15  
## # ... with 53 more rows

# Analyse par sport

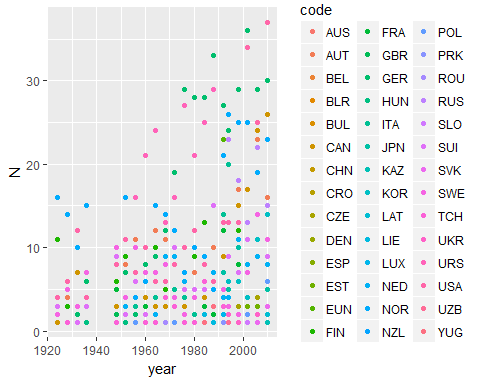
Peu importe la couleur de la mÃ©daille

donnees%>%group\_by(sport,code)%>%summarise(N=n())%>%arrange(-N)%>%slice(1:3)

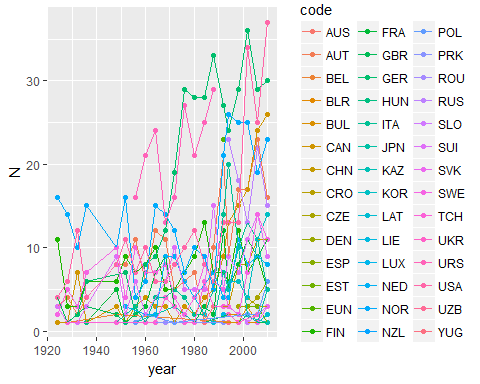
## Source: local data frame [45 x 3]  
## Groups: sport [15]  
##   
## sport code N  
## <chr> <chr> <int>  
## 1 alpineskiing AUT 105  
## 2 alpineskiing SUI 55  
## 3 alpineskiing FRA 43  
## 4 biathlon GER 59  
## 5 biathlon NOR 29  
## 6 biathlon RUS 20  
## 7 bobsleigh GER 40  
## 8 bobsleigh SUI 30  
## 9 bobsleigh USA 20  
## 10 crosscountry NOR 96  
## # ... with 35 more rows

# Evolution dans le remps

donnees%>%group\_by(year,code)%>%summarise(N=n())%>%ggplot(aes(x=year, y=N, color=code))+geom\_point()

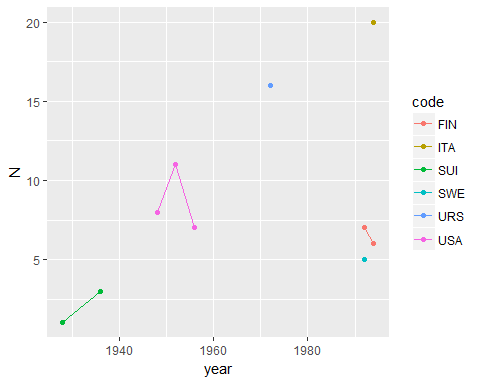


donnees%>%group\_by(year,code)%>%summarise(N=n())%>%ggplot(aes(x=year, y=N, color=code))+geom\_point()+geom\_line()

 ATTENTION

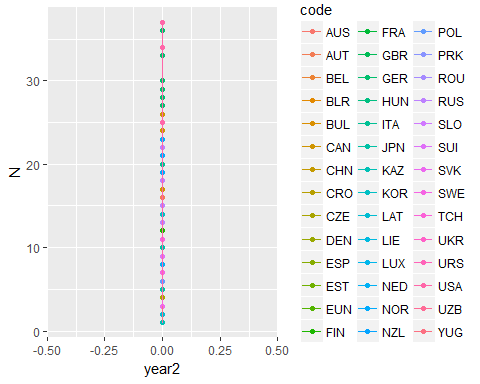
donnees%>%group\_by(year,code)%>%summarise(N=n())%>%filter(code==c('GER','NOR','USA','AUT','URS','FIN','CAN','SWE','SUI','ITA'))%>%ggplot(aes(x=year, y=N, color=code))+geom\_point()+geom\_line()

## Warning in c("AUS", "AUT", "BLR", "CAN", "CHN", "CRO", "CZE", "EST",  
## "FIN", : la taille d'un objet plus long n'est pas multiple de la taille  
## d'un objet plus court  
  
## Warning in c("AUS", "AUT", "BLR", "CAN", "CHN", "CRO", "CZE", "EST",  
## "FIN", : la taille d'un objet plus long n'est pas multiple de la taille  
## d'un objet plus court  
  
## Warning in c("AUS", "AUT", "BLR", "CAN", "CHN", "CRO", "CZE", "EST",  
## "FIN", : la taille d'un objet plus long n'est pas multiple de la taille  
## d'un objet plus court  
  
## Warning in c("AUS", "AUT", "BLR", "CAN", "CHN", "CRO", "CZE", "EST",  
## "FIN", : la taille d'un objet plus long n'est pas multiple de la taille  
## d'un objet plus court  
  
## Warning in c("AUS", "AUT", "BLR", "CAN", "CHN", "CRO", "CZE", "EST",  
## "FIN", : la taille d'un objet plus long n'est pas multiple de la taille  
## d'un objet plus court  
  
## Warning in c("AUS", "AUT", "BLR", "CAN", "CHN", "CRO", "CZE", "EST",  
## "FIN", : la taille d'un objet plus long n'est pas multiple de la taille  
## d'un objet plus court  
  
## Warning in c("AUS", "AUT", "BLR", "CAN", "CHN", "CRO", "CZE", "EST",  
## "FIN", : la taille d'un objet plus long n'est pas multiple de la taille  
## d'un objet plus court  
  
## Warning in c("AUS", "AUT", "BLR", "CAN", "CHN", "CRO", "CZE", "EST",  
## "FIN", : la taille d'un objet plus long n'est pas multiple de la taille  
## d'un objet plus court  
  
## Warning in c("AUS", "AUT", "BLR", "CAN", "CHN", "CRO", "CZE", "EST",  
## "FIN", : la taille d'un objet plus long n'est pas multiple de la taille  
## d'un objet plus court  
  
## Warning in c("AUS", "AUT", "BLR", "CAN", "CHN", "CRO", "CZE", "EST",  
## "FIN", : la taille d'un objet plus long n'est pas multiple de la taille  
## d'un objet plus court  
  
## Warning in c("AUS", "AUT", "BLR", "CAN", "CHN", "CRO", "CZE", "EST",  
## "FIN", : la taille d'un objet plus long n'est pas multiple de la taille  
## d'un objet plus court  
  
## Warning in c("AUS", "AUT", "BLR", "CAN", "CHN", "CRO", "CZE", "EST",  
## "FIN", : la taille d'un objet plus long n'est pas multiple de la taille  
## d'un objet plus court  
  
## Warning in c("AUS", "AUT", "BLR", "CAN", "CHN", "CRO", "CZE", "EST",  
## "FIN", : la taille d'un objet plus long n'est pas multiple de la taille  
## d'un objet plus court  
  
## Warning in c("AUS", "AUT", "BLR", "CAN", "CHN", "CRO", "CZE", "EST",  
## "FIN", : la taille d'un objet plus long n'est pas multiple de la taille  
## d'un objet plus court  
  
## Warning in c("AUS", "AUT", "BLR", "CAN", "CHN", "CRO", "CZE", "EST",  
## "FIN", : la taille d'un objet plus long n'est pas multiple de la taille  
## d'un objet plus court  
  
## Warning in c("AUS", "AUT", "BLR", "CAN", "CHN", "CRO", "CZE", "EST",  
## "FIN", : la taille d'un objet plus long n'est pas multiple de la taille  
## d'un objet plus court  
  
## Warning in c("AUS", "AUT", "BLR", "CAN", "CHN", "CRO", "CZE", "EST",  
## "FIN", : la taille d'un objet plus long n'est pas multiple de la taille  
## d'un objet plus court  
  
## Warning in c("AUS", "AUT", "BLR", "CAN", "CHN", "CRO", "CZE", "EST",  
## "FIN", : la taille d'un objet plus long n'est pas multiple de la taille  
## d'un objet plus court



#(donnees$year - min( donnees$year))/4  
  
#donnees %>% mutate(year2=(year-min(year))/4)  
  
#donnees%>%group\_by(year,code)%>%summarise(N=n())%>%mutate(year2=(year-min(year))/4)

donnees%>%group\_by(year,code)%>%summarise(N=n())%>%mutate(year2=(year-min(year))/4)%>%ggplot(aes(x=year2, y=N, color=code))+geom\_point()+geom\_line()



# Position de la France

donnees%>%group\_by(year,sport,code)%>%summarise(N=n())%>%arrange(-N)%>%slice(1)%>%filter(code=='FRA')

## Source: local data frame [4 x 4]  
## Groups: year, sport [4]  
##   
## year sport code N  
## <int> <chr> <chr> <int>  
## 1 1924 curling FRA 1  
## 2 1968 alpineskiing FRA 8  
## 3 1992 freestyleskiing FRA 5  
## 4 2010 biathlon FRA 6

donnees%>%group\_by(year,sport,code)%>%summarise(N=n())%>%arrange(-N)%>%slice(1:3)%>%filter(code=='FRA')

## Source: local data frame [22 x 4]  
## Groups: year, sport [22]  
##   
## year sport code N  
## <int> <chr> <chr> <int>  
## 1 1924 biathlon FRA 1  
## 2 1924 curling FRA 1  
## 3 1924 figureskating FRA 1  
## 4 1928 figureskating FRA 1  
## 5 1936 alpineskiing FRA 1  
## 6 1948 alpineskiing FRA 5  
## 7 1952 figureskating FRA 1  
## 8 1960 alpineskiing FRA 3  
## 9 1964 alpineskiing FRA 6  
## 10 1968 alpineskiing FRA 8  
## # ... with 12 more rows