

# **APPRENDRE PAR LA PRATIQUE: MANIPULATION DE DONNÉES AVEC R**

## **SESSION 1 - AVRIL 2022**

Ce document est un compilé des travaux réalisés par les participants durant la première session d'apprendre par la pratique. Ils étaient plus de 50 à s'engager, seuls les 5 participants suivants ont fait l'exercice jusqu'au bout :

**Dr. Cheikh Ibrahima Fall DIOP - mamecheikhdiop@gmail.com**

**Daniella Lowa - daniellalowa@gmail.com**

**Ibrahima Diallo - dialloibrahima46@gmail.com**

**Armel TINDO - armeltindo@gmail.com**

**Andrianavalondrahona Mariel - valofils2@gmail.com**

# Apprendre par la pratique - Session 1 : Manipulation de données avec R

LSTP - Statistics & Coding

2022-04-21

## Introduction

Bienvenue à la première session d'apprendre par la pratique. Vous avez majoritairement sélectionné, comme sujet d'intérêt, la manipulation de données avec R. Pour cette première session de travail, nous allons donc nous concentrer sur cette thématique .

Depuis sa création en Août 1993 par Ross Ihaka et Robert Gentleman, le langage de programmation R est devenu populaire à travers le monde. R est un langage de programmation puissant pour traiter et analyser rigoureusement des données. La force de R vient de son interface très convivial R studio et d'une communauté très active de développeurs. Peu importe la discipline dans laquelle vous évoluez, vous pouvez trouver un avantage à apprendre et à utiliser R. Gratuit, le logiciel R est accessible à tous. Pour ceux qui ne l'ont pas encore installé, veuillez visiter: <https://larmarange.github.io/analyse-R/installation-de-R-et-RStudio.html>.

## Instructions

Ces instructions constituent un guide non exhaustif mais illustrent le travail minimal requis pour obtenir un certificat de participation. Vous êtes libre de faire appel à votre imagination pour réaliser des manipulations supplémentaires.

Vous devez choisir un jeu de données parmi les 4 jeux de données suivants:

- TED Talks
- Commentaires sur Twitter
- Données portant sur les Fraudes
- Performances des étudiants

Ensuite, il est obligatoire de travailler avec les 3 jeux de données portant sur l'agriculture dans le dossier archives. Il est aussi obligatoire de travailler avec le jeu de données portant sur les commandes Amazon. Note : les données ont été prises de Kaggle.

Il est préférable de faire la rédaction avec Rmarkdown mais ce n'est pas obligatoire. Tout autre format lisible est permis (Word, Latex, Power Point etc.).

La durée de cette première session est de 4 semaines. La correction se fera par les membres d'un même groupe à chaque semaine. Vous pouvez nous écrire au besoin pour poser des questions.

## Semaine 1 à 2

Étape 1 : Expliquez en quoi consiste une manipulation de données

Étape 2 : Expliquez le contexte des données

Étape 3 : Décrivez vos données (type de variables, statistiques descriptives, données aberrantes, présence de doublons)

Étape 4 : Réalisez quelques visualisations simples de vos données: diagramme en boite, diagramme en barre, nuage de points, nuage de mots.

## Semaine 3 à 4

Fusionner les données du dossier archive

Question 1 : Quels sont les pays exclus à la suite de cette fusion?

Question 2 : Quelle est la quantité totale de pesticide utilisée en Amérique, en Afrique et en Australie?

Question 3 : Quel est la moyenne et médiane de rendement de céréales (hectogramme par hectare (Hg/Ha)) pour l'Amérique, l'Afrique et l'Australie?

Travailler avec la base de données portant sur les commandes Amazon.

Quelle est la ville où le nombre d'envoie (shipping) est le plus élevé?

Convertissez la variable order\_date en années (Recommandation : utilisez le package lubridate)

Calculez la moyenne de frais d'envois par année (Recommandation : utilisez le package stringr).

```
#-----
# Les jeux de données
#-----

#Répertoire de travail
setwd("~/Desktop/Statistics&Coding/L'essentiel de la stat/Production/data analysis LSTP/data session 1")

#Charger les données

dir() # pour voir les fichiers dans le dossier de travail

## [1] "archive"                 "data.csv"
## [3] "Exemples.R"               "Fraud.csv"
## [5] "learnstat_session1.md"    "learnstat_session1.pdf"
## [7] "learnstat_session1.Rmd"   "orders_data.xlsx"
## [9] "StudentsPerformance.csv"  "Tweets.csv"

#TED Talks
data_ted <- read.csv("data.csv", sep = ",", header = TRUE)
str(data_ted)
```

```

## 'data.frame': 5440 obs. of 6 variables:
## $ title : chr "Climate action needs new frontline leadership" "The dark history of the overthrow o...
## $ author: chr "Ozawa Bineshi Albert" "Sydney Iaukea" "Martin Reeves" "James K. Thornton" ...
## $ date : chr "December 2021" "February 2022" "September 2021" "October 2021" ...
## $ views : int 404000 214000 412000 427000 2400 422000 412000 455000 66000 584000 ...
## $ likes : int 12000 6400 12000 12000 72 12000 12000 13000 1900 17000 ...
## $ link : chr "https://ted.com/talks/ozawa_bineshi_albert_climate_action_needs_new_frontline_leade...
# Commentaires sur Twitter
data_tweet <- read.csv("Tweets.csv", sep = ",", header = TRUE)
str(data_tweet)

## 'data.frame': 27481 obs. of 4 variables:
## $ textID      : chr "cb774db0d1" "549e992a42" "088c60f138" "9642c003ef" ...
## $ text        : chr "I`d have responded, if I were going" "Sooo SAD I will miss you here in San...
## $ selected_text: chr "I`d have responded, if I were going" "Sooo SAD" "bullying me" "leave me alone...
## $ sentiment   : chr "neutral" "negative" "negative" "negative" ...
# Données portant sur les Fraudes

data_fraud <- read.csv("Fraud.csv", sep = ",", header = TRUE)
str(data_fraud)

## 'data.frame': 6362620 obs. of 11 variables:
## $ step       : int 1 1 1 1 1 1 1 1 1 1 ...
## $ type       : chr "PAYMENT" "PAYMENT" "TRANSFER" "CASH_OUT" ...
## $ amount     : num 9840 1864 181 181 11668 ...
## $ nameOrig   : chr "C1231006815" "C1666544295" "C1305486145" "C840083671" ...
## $ oldbalanceOrg: num 170136 21249 181 181 41554 ...
## $ newbalanceOrig: num 160296 19385 0 0 29886 ...
## $ nameDest   : chr "M1979787155" "M2044282225" "C553264065" "C38997010" ...
## $ oldbalanceDest: num 0 0 0 21182 0 ...
## $ newbalanceDest: num 0 0 0 0 0 ...
## $ isFraud    : int 0 0 1 1 0 0 0 0 0 0 ...
## $ isFlaggedFraud: int 0 0 0 0 0 0 0 0 0 0 ...

# Performances des étudiants

data_student_perf <- read.csv("StudentsPerformance.csv", sep = ",", header = TRUE)
str(data_student_perf)

## 'data.frame': 1000 obs. of 8 variables:
## $ gender           : chr "female" "female" "female" "male" ...
## $ race.ethnicity   : chr "group B" "group C" "group B" "group A" ...
## $ parental.level.of.education: chr "bachelor's degree" "some college" "master's degree" "associate...
## $ lunch            : chr "standard" "standard" "standard" "free/reduced" ...
## $ test.preparation.course: chr "none" "completed" "none" "none" ...
## $ math.score       : int 72 69 90 47 76 71 88 40 64 38 ...
## $ reading.score   : int 72 90 95 57 78 83 95 43 64 60 ...
## $ writing.score   : int 74 88 93 44 75 78 92 39 67 50 ...
#-----
#Multiples jeux de données : données agricoles
#-----
#Répertoire de travail

setwd("~/Desktop/Statistics&Coding/L'essentiel de la stat/Production/data analysis LSTP/data session 1/")


```

```

dir()

## [1] "CerealCropYield_1961-2018.csv"      "FertilizerConsumption_1961-2018.csv"
## [3] "PesticideUsage_1990-2017.csv"

data_cereal <- read.csv("CerealCropYield_1961-2018.csv", sep = ",", header = TRUE)

data_ferti <- read.csv("FertilizerConsumption_1961-2018.csv", sep = ",", header = TRUE)

data_pest <- read.csv("PesticideUsage_1990-2017.csv", sep = ",", header = TRUE)

str(data_cereal)

## 'data.frame': 202 obs. of 3 variables:
## $ Country : chr "Afghanistan" "Africa" "Albania" "Algeria" ...
## $ Yield..hg.ha. : int 808952 692014 1651049 534659 2135761 362959 959479 1664144 580120 92410...
## $ Area.harvested..ha.: num 1.68e+08 4.83e+09 1.50e+07 1.58e+08 7.50e+09 ...

str(data_ferti)

## 'data.frame': 184 obs. of 2 variables:
## $ Country : chr "Afghanistan" "Albania" "Algeria" "Angola" ...
## $ FertilizerQuantity: num 5720345 4316765 8578144 2030699 43153465 ...

str(data_pest)

## 'data.frame': 165 obs. of 2 variables:
## $ Country : chr "Albania" "Algeria" "Angola" "Antigua and Barbuda" ...
## $ Total.Pesticides.use.per.area.of.land..kg.ha.: num 14.25 12.88 0.43 72.79 98.15 ...

#-----#Répertoire de travail
setwd("~/Desktop/Statistics&Coding/L'essentiel de la stat/Production/data analysis LSTP/data session 1")
# Commandes Amazon
library("readxl")
data_amazon <- read_excel("orders_data.xlsx")
summary(data_amazon)

##   order_no          order_date        buyer       ship_city
##  Length:171    Length:171    Length:171    Length:171
##  Class :character Class :character Class :character Class :character
##  Mode  :character Mode  :character Mode  :character Mode  :character
##   ship_state        sku      description     quantity
##  Length:171    Length:171    Length:171    Length:171
##  Class :character Class :character Class :character Class :character
##  Mode  :character Mode  :character Mode  :character Mode  :character
##   item_total      shipping_fee        cod   order_status
##  Length:171    Length:171    Length:171    Length:171
##  Class :character Class :character Class :character Class :character
##  Mode  :character Mode  :character Mode  :character Mode  :character

#Exemple d'extraction

stringr::str_sub(data_amazon$shipping_fee, 2, 4)

## [1] NA    "60." "60." NA    "84." NA    NA    "84." "84." NA    "114" "60."
## [13] "84." "84." "84." "62." "81." "84." "84." "84." "84." "84." "60." "84." "84."

```

```
## [25] NA      "60." NA      "84." "84." NA      "60." "60." "84." "84." "47." "84."
## [37] "84." "84." "84." "84." "84." "84." "47." NA      "60." NA      "60."
## [49] NA      "84." "84." NA      NA      "84." "47." "47." NA      "84." "178" NA
## [61] NA      "84." "84." NA      "84." "84." "84." NA      "47." "60." NA
## [73] NA      NA      "210" "84." NA      "84." "60." "84." "84." "47." "84." "114"
## [85] "84." "81." "84." "84." "210" NA      NA      "84." "60." NA      "84." "84."
## [97] "84." "84." "84." NA      "60." "84." "84." "84." "60." "84." "84." "84."
## [109] "84." "80." "84." "146" "60." "84." "84." "84." "114" "133" "84." "84."
## [121] "114" "241" "84." "84." "84." "84." "84." "84." "114" "84." "84." "84."
## [133] "84." "84." "84." "84." "84." "84." "84." "47." "47." "84." "84." "47."
## [145] "84." "84." "84." "60." "84." "84." "84." "60." "84." "84." "84." "84."
## [157] "84." "84." "47." "84." "84." "84." "84." "84." "84." "84." "84." "114"
## [169] "105" "80." "84."
```

Merci de participer et bonne pratique!

LSTP - Statistics & Coding

# Quick tips: tableaux de fréquences et nuage de mots avec R

LSTP - Statistics & Coding

2022-04-22

Pour la session 1 de manipulation de données avec R, vous êtes appelés à traiter du texte. Ci-dessous, un exemple de traitement de texte avec des données de TED. À noter que ce n'est pas la même base de données que vous devez utiliser.

```
#Chargement des librairies
library(wordcloud2) #pour créer un nuage de mots
library(quanteda) #Pour l'analyse de texte
#Pour enregistrer le nuage de mots sous différents formats.
library(htmlwidgets)
webshot::install_phantomjs()
#Répertoire de travail
setwd("~/Desktop/Statistics&Coding/L'essentiel de la stat/Production/data analysis LSTP/data session 1")

#Les données
data_ted <- read.csv("ted_main.csv", sep = ",", header = TRUE)

#Les étiquettes

head(data_ted$tags)

## [1] "[['children', 'creativity', 'culture', 'dance', 'education', 'parenting', 'teaching']]"
## [2] "[['alternative energy', 'cars', 'climate change', 'culture', 'environment', 'global issues', 'sc]"
## [3] "[['computers', 'entertainment', 'interface design', 'media', 'music', 'performance', 'simplicity']"
## [4] "[['MacArthur grant', 'activism', 'business', 'cities', 'environment', 'green', 'inequality', 'po']"
## [5] "[['Africa', 'Asia', 'Google', 'demo', 'economics', 'global development', 'global issues', 'healt]"
## [6] "[['business', 'culture', 'entertainment', 'goal-setting', 'motivation', 'potential', 'psychology']"

#Description des étiquettes (tags)
tag <- as.vector(data_ted$tags)
tags2 <- corpus(tag)
tags3 <- tokens(tags2, remove_punct = TRUE, remove_hyphens = TRUE)
dfm_tags <- dfm(tags3, tolower = TRUE)
names_tags <- colnames(dfm_tags)
dat_tags <- data.frame(Noms=names_tags, freq=colSums(dfm_tags))

#10 premières lignes
head(dat_tags, 10)

##          Noms freq
## children    children 143
## creativity   creativity 189
## culture      culture  486
## dance        dance   25
## education    education 153
```

```

## parenting      parenting   50
## teaching      teaching    43
## alternative  alternative  37
## energy        energy     122
## cars          cars       29

#On sélectionne les mots qui apparaissent plus de 100 fois
dat_tags2<-subset(dat_tags,freq>100)
nuage_mots= wordcloud2(dat_tags,size = 2)
saveWidget(nuage_mots,"nuage_mots.html",selfcontained = F)
webshot::webshot("nuage_mots.html","nuage_mots.png",vwidth = 700, vheight = 500, delay =10)

```



# Quick tips 2 : visualisation simple des données sur la carte du monde

LSTP

2022-04-27

Dans le cadre de l'exercice pratique, manipulation de données avec R, vous êtes appelés à produire des visualisations simples avec vos données. On partage avec vous deux exemples simples de représentation de données sur la carte du monde. On utilise ici les données sur l'utilisation de pesticides dans le monde.

```
#Charger les données
data_pest <- read.csv("PesticideUsage_1990-2017.csv", sep = ",", header = TRUE)
colnames(data_pest) <- c("Country", "Pesticide")
summary(data_pest)

##      Country          Pesticide
##  Length:165      Min.   :  0.02
##  Class :character 1st Qu.:  4.72
##  Mode  :character Median :31.24
##                  Mean   :62.89
##                  3rd Qu.:85.28
##                  Max.  :406.43

#Option 1: Avec rworldmap
library(rworldmap)

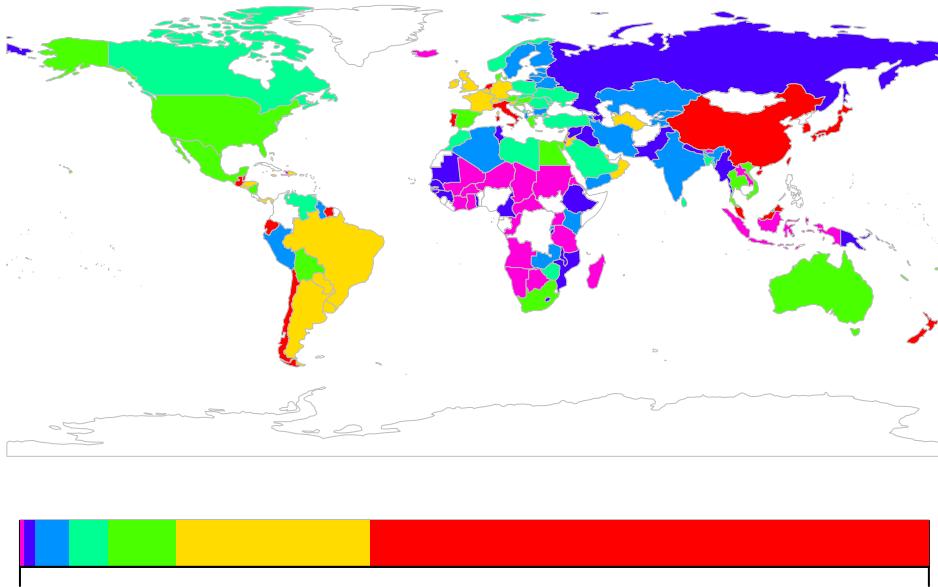
## Loading required package: sp
## ### Welcome to rworldmap ####
## For a short introduction type : vignette('rworldmap')
#Utilisez la fonction joinCountryData pour fusionner les données sur l'utilisation de pesticide avec le

data_pest_map <- joinCountryData2Map(data_pest
                                      , joinCode = "NAME"
                                      , nameJoinColumn = "Country")

## 151 codes from your data successfully matched countries in the map
## 14 codes from your data failed to match with a country code in the map
## 92 codes from the map weren't represented in your data
#Utilisez la fonction mapCountryData pour représenter les données.

mapCountryData(data_pest_map, nameColumnToPlot="Pesticide",
               colourPalette = "rainbow")
```

## Pesticide



0.02

406

#Option 2 : Carte avec ggplot2 et rnaturalearth

```
library(rnaturalearth) #contient les cartes des pays du monde.
library(rnaturalearthdata)
library(ggplot2)

monde <- ne_countries(returnclass = "sf") # Pour récupérer les données sur les pays.
str(monde) # Pour voir la structure des données.

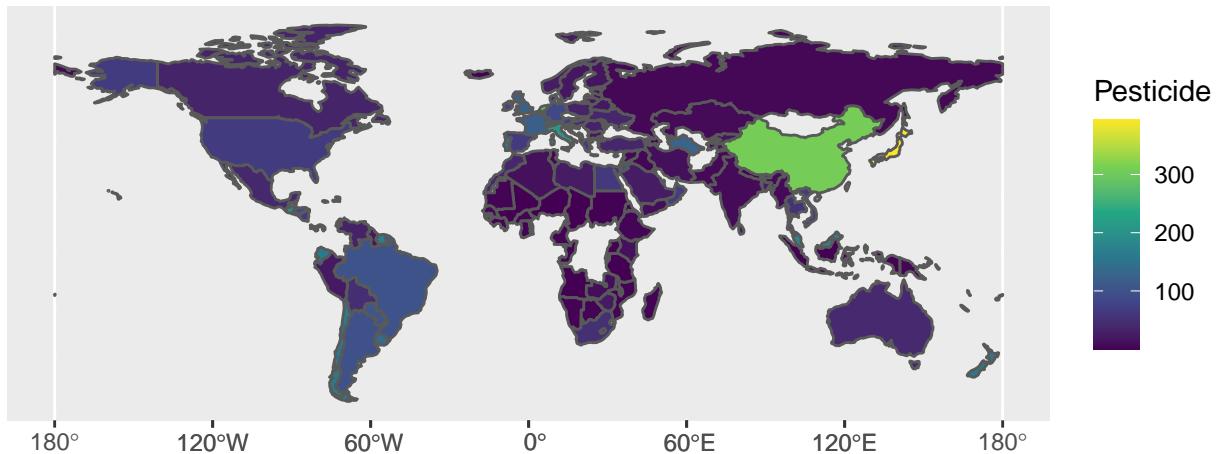
## Classes 'sf' and 'data.frame': 177 obs. of 64 variables:
## $ scalerank : int 1 1 1 1 1 1 1 3 1 1 ...
## $ featurecla: chr "Admin-0 country" "Admin-0 country" "Admin-0 country" "Admin-0 country" ...
## $ labelrank : num 3 3 6 4 2 6 4 6 2 4 ...
## $ sovereign: chr "Afghanistan" "Angola" "Albania" "United Arab Emirates" ...
## $ sov_a3 : chr "AFG" "AGO" "ALB" "ARE" ...
## $ adm0_dif : num 0 0 0 0 0 0 1 1 0 ...
## $ level : num 2 2 2 2 2 2 2 2 2 2 ...
## $ type : chr "Sovereign country" "Sovereign country" "Sovereign country" "Sovereign country"
## $ admin : chr "Afghanistan" "Angola" "Albania" "United Arab Emirates" ...
## $ adm0_a3 : chr "AFG" "AGO" "ALB" "ARE" ...
## $ geou_dif : num 0 0 0 0 0 0 0 0 0 ...
## $ geounit : chr "Afghanistan" "Angola" "Albania" "United Arab Emirates" ...
## $ gu_a3 : chr "AFG" "AGO" "ALB" "ARE" ...
## $ su_dif : num 0 0 0 0 0 0 0 0 0 ...
## $ subunit : chr "Afghanistan" "Angola" "Albania" "United Arab Emirates" ...
## $ su_a3 : chr "AFG" "AGO" "ALB" "ARE" ...
## $ brk_diff : num 0 0 0 0 0 0 0 0 0 ...
## $ name : chr "Afghanistan" "Angola" "Albania" "United Arab Emirates" ...
## $ name_long : chr "Afghanistan" "Angola" "Albania" "United Arab Emirates" ...
## $ brk_a3 : chr "AFG" "AGO" "ALB" "ARE" ...
```

```

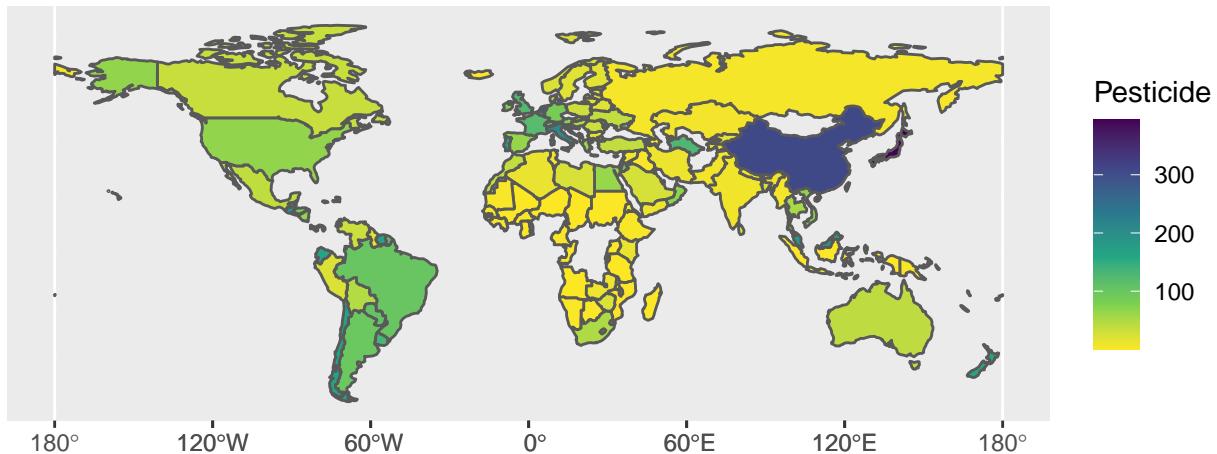
## $ brk_name : chr "Afghanistan" "Angola" "Albania" "United Arab Emirates" ...
## $ brk_group : chr NA NA NA NA ...
## $ abbrev : chr "Afg." "Ang." "Alb." "U.A.E." ...
## $ postal : chr "AF" "AO" "AL" "AE" ...
## $ formal_en : chr "Islamic State of Afghanistan" "People's Republic of Angola" "Republic of Albania"
## $ formal_fr : chr NA NA NA NA ...
## $ note_adm0 : chr NA NA NA NA ...
## $ note_brk : chr NA NA NA NA ...
## $ name_sort : chr "Afghanistan" "Angola" "Albania" "United Arab Emirates" ...
## $ name_alt : chr NA NA NA NA ...
## $ mapcolor7 : num 5 3 1 2 3 3 4 7 1 3 ...
## $ mapcolor8 : num 6 2 4 1 1 1 5 5 2 1 ...
## $ mapcolor9 : num 8 6 1 3 3 2 1 9 2 3 ...
## $ mapcolor13: num 7 1 6 3 13 10 NA 11 7 4 ...
## $ pop_est : num 28400000 12799293 3639453 4798491 40913584 ...
## $ gdp_md_est: num 22270 110300 21810 184300 573900 ...
## $ pop_year : num NA NA NA NA NA NA NA NA NA ...
## $ lastcensus: num 1979 1970 2001 2010 2010 ...
## $ gdp_year : num NA NA NA NA NA NA NA NA NA ...
## $ economy : chr "7. Least developed region" "7. Least developed region" "6. Developing region" "6. ...
## $ income_grp: chr "5. Low income" "3. Upper middle income" "4. Lower middle income" "2. High income ...
## $ wikipedia : num NA NA NA NA NA NA NA NA NA ...
## $ fips_10 : chr NA NA NA NA ...
## $ iso_a2 : chr "AF" "AO" "AL" "AE" ...
## $ iso_a3 : chr "AFG" "AGO" "ALB" "ARE" ...
## $ iso_n3 : chr "004" "024" "008" "784" ...
## $ un_a3 : chr "004" "024" "008" "784" ...
## $ wb_a2 : chr "AF" "AO" "AL" "AE" ...
## $ wb_a3 : chr "AFG" "AGO" "ALB" "ARE" ...
## $ woe_id : num NA NA NA NA NA NA NA NA ...
## $ adm0_a3_is: chr "AFG" "AGO" "ALB" "ARE" ...
## $ adm0_a3_us: chr "AFG" "AGO" "ALB" "ARE" ...
## $ adm0_a3_un: num NA NA NA NA NA NA NA NA ...
## $ adm0_a3_wb: num NA NA NA NA NA NA NA NA ...
## $ continent : chr "Asia" "Africa" "Europe" "Asia" ...
## $ region_un : chr "Asia" "Africa" "Europe" "Asia" ...
## $ subregion : chr "Southern Asia" "Middle Africa" "Southern Europe" "Western Asia" ...
## $ region_wb : chr "South Asia" "Sub-Saharan Africa" "Europe & Central Asia" "Middle East & North Africa" ...
## $ name_len : num 11 6 7 20 9 7 10 22 9 7 ...
## $ long_len : num 11 6 7 20 9 7 10 35 9 7 ...
## $ abbrev_len: num 4 4 4 6 4 4 4 10 4 5 ...
## $ tiny : num NA NA NA NA NA NA 2 NA NA ...
## $ homepart : num 1 1 1 1 1 1 1 NA 1 1 ...
## $ geometry :sfc_MULTIPOLYGON of length 177; first list element: List of 1
##   ..$ :List of 1
##     ... .$. : num [1:69, 1:2] 61.2 62.2 63 63.2 64 ...
##     ... - attr(*, "class")= chr [1:3] "XY" "MULTIPOLYGON" "sfg"
##     - attr(*, "sf_column")= chr "geometry"
##     - attr(*, "agr")= Factor w/ 3 levels "constant","aggregate",...: NA NA NA NA NA NA NA NA NA ...
##     .. - attr(*, "names")= chr [1:63] "scalerank" "featurecla" "labelrank" "sovereignty" ...
#On fusionne les données sur les pays avec nos données sur l'utilisation de pesticide.
data_pest_map <- merge(monde, data_pest, by.x = "name", by.y = "Country")

```

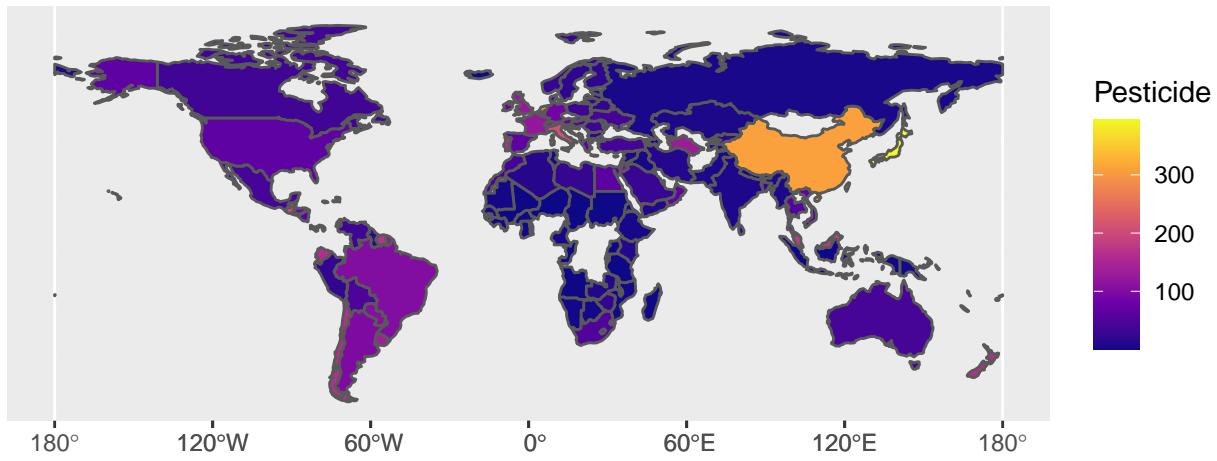
```
# 1  
ggplot(data = data_pest_map) +  
  geom_sf(aes(fill = Pesticide)) +  
  scale_fill_viridis_c() # Pour ajouter les couleurs
```



```
# 2  
ggplot(data = data_pest_map) +  
  geom_sf(aes(fill = Pesticide)) +  
  scale_fill_viridis_c(direction = -1) # Pour changer l'ordre d'intensité des couleurs
```



```
# 3  
ggplot(data = data_pest_map) +  
  geom_sf(aes(fill = Pesticide)) +  
  scale_fill_viridis_c(option = "plasma") #Option pour changer la couleur
```



# Quick Tips 3 : diagramme en barre avec pourcentage

LSTP - Statistics & Coding

2022-04-28

## Étape 1: Calculer les pourcentages et ajouter les labels associés avec dplyr

```
#Chargez les données
data_fraud <- read.csv("Fraud.csv", sep = ",", header = TRUE)

#Calcul des pourcentages et ajout des labels associés avec dplyr

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
plot_data_fraud <- data_fraud %>%
  count(type) %>%
  mutate(pct = n / sum(n),
        pctlabel = paste0(round(pct*100), "%"))

head(plot_data_fraud)

##      type     n      pct pctlabel
## 1 CASH_IN 1399284 0.219922610    22%
## 2 CASH_OUT 2237500 0.351663309    35%
## 3 DEBIT    41432 0.006511783     1%
## 4 PAYMENT  2151495 0.338146078    34%
## 5 TRANSFER 532909 0.083756220     8%
```

## Étape 2: Utiliser ggplot2 pour réaliser le diagramme en barres

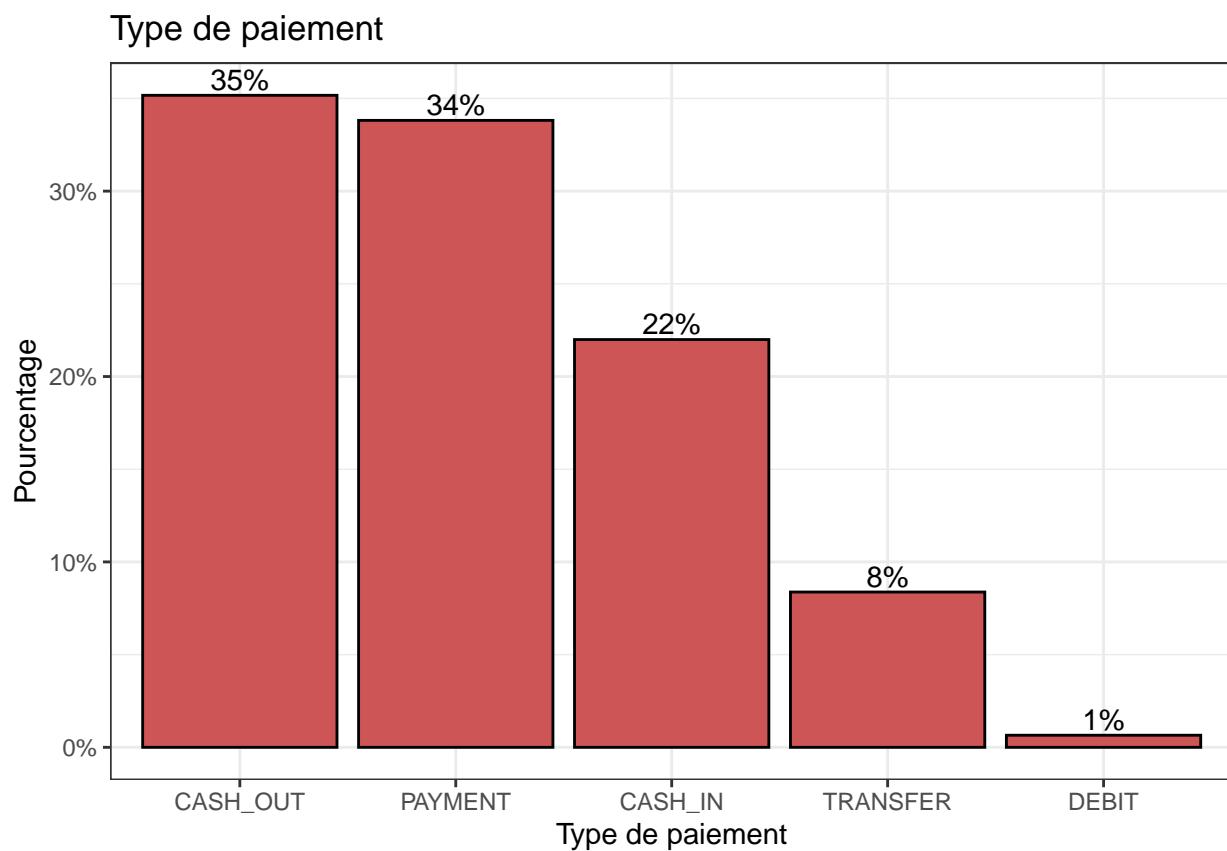
```
# diagramme en barre avec pourcentage et mettre les barres en ordre décroissant.
library(ggplot2)

ggplot(plot_data_fraud,
```

```

aes(x = reorder(type, -pct),
    y = pct)) +
geom_bar(stat = "identity",
         fill = "indianred3",
         color = "black") +
geom_text(aes(label = pctlabel),
          vjust = -0.25) +
scale_y_continuous(labels = scales::percent) +
labs(x = "Type de paiement",
     y = "Pourcentage",
     title = "Type de paiement") + theme_bw()

```



# Apprendre par la pratique - Session 1 : Manipulation de données avec R

Cheikh Ibrahima Fall DIOP

2022-05-11

---

## Table des matières

<b>1 Étape 1 : Explication du concept de manipulation de données</b>	<b>1</b>
<b>2 Étape 2 : Contexte des données</b>	<b>2</b>
2.1 Données choisies . . . . .	2
2.2 Chargement des données . . . . .	2
<b>3 Étape 3 : Description des données</b>	<b>2</b>
3.1 Type de variables . . . . .	2
3.1.1 Données sur les performances des étudiants . . . . .	2
3.1.2 Données sur les céréales cultivées de 1961 à 2018 . . . . .	2
3.1.3 Données sur la consommation de fertilisants de 1961 à 2018 . . . . .	3
3.1.4 Données sur l'usage de pesticides de 1990 à 2017 . . . . .	3
3.1.5 Données sur les commandes sur Amazon . . . . .	3
3.2 Statistiques descriptives . . . . .	4
3.2.1 Données sur les performances des étudiants . . . . .	4
3.2.2 Données sur les céréales cultivées de 1961 à 2018 . . . . .	4
3.2.3 Données sur la consommation de fertilisants de 1961 à 2018 . . . . .	4
3.2.4 Données sur l'usage de pesticides de 1990 à 2017 . . . . .	5
3.2.5 Données sur les commandes sur Amazon . . . . .	5
3.3 Données aberrantes . . . . .	5
3.4 Présence de doublons . . . . .	5
3.4.1 Données sur les performances des étudiants . . . . .	5
3.4.2 Données sur les céréales cultivées de 1961 à 2018 . . . . .	5
3.4.3 Données sur la consommation de fertilisants de 1961 à 2018 . . . . .	5
3.4.4 Données sur l'usage de pesticides de 1990 à 2017 . . . . .	5
3.4.5 Données sur les commandes sur Amazon . . . . .	6
<b>4 Étape 4 : Quelques visualisations simples des données : diagramme en boîte, diagramme en barre, nuage de points, nuage de mots.</b>	<b>6</b>

---

## 1 Étape 1 : Explication du concept de manipulation de données

La manipulation des données est l'étape qui consiste à les importer et à les transformer. Ces opérations incluent entre autre :

- le tri,
- la sélection (de variables, de lignes ou de colonnes),
- la création, le renommage ou la suppression de variables,

- la fusion de bases,
  - etc.

## 2 Étape 2 : Contexte des données

## 2.1 Données choisies

Pour ce travail, je choisis les données sur les performances des étudiants.

## 2.2 Chargement des données

```
setwd("D:/Downloads/LSTP")
# Base sur les Performances des étudiants
library(readr)
data_student_perf <- read_csv("./StudentsPerformance.csv")
# Jeux de données portant sur agriculture
data_cereal <- read_csv("./archive/CerealCropYield_1961-2018.csv")
data_ferti <- read_csv("./archive/FertilizerConsumption_1961-2018.csv")
data_pest <- read_csv("./archive/PesticideUsage_1990-2017.csv")
# Jeu de données portant sur les commandes Amazon
library(readxl)
orders_data <- read_excel("./orders_data.xlsx")
```

### 3 Étape 3 : Description des données

### 3.1 Type de variables

### **3.1.1 Données sur les performances des étudiants**

```
str(data_student_perf)

## # spec_tbl_df [1,000 x 8] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
## # $ gender : chr [1:1000] "female" "female" "female" "male" ...
## # $ race/ethnicity : chr [1:1000] "group B" "group C" "group B" "group A" ...
## # $ parental level of education: chr [1:1000] "bachelor's degree" "some college" "master's degree" "high school" ...
## # $ lunch : chr [1:1000] "standard" "standard" "standard" "free/reduced" ...
## # $ test preparation course : chr [1:1000] "none" "completed" "none" "none" ...
## # $ math score : num [1:1000] 72 69 90 47 76 71 88 40 64 38 ...
## # $ reading score : num [1:1000] 72 90 95 57 78 83 95 43 64 60 ...
## # $ writing score : num [1:1000] 74 88 93 44 75 78 92 39 67 50 ...
## # - attr(*, "spec")=
## .. cols(
## ..   gender = col_character(),
## ..   `race/ethnicity` = col_character(),
## ..   `parental level of education` = col_character(),
## ..   lunch = col_character(),
## ..   `test preparation course` = col_character(),
## ..   `math score` = col_double(),
## ..   `reading score` = col_double(),
## ..   `writing score` = col_double()
## .. )
## # - attr(*, "problems")=<externalptr>
```

### 3.1.2 Données sur les céréales cultivées de 1961 à 2018

```
str(data_cereal)
```

```

## spec_tbl_df [202 x 3] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
## $ Country : chr [1:202] "Afghanistan" "Africa" "Albania" "Algeria" ...
## $ Yield (hg/ha) : num [1:202] 808952 692014 1651049 534659 2135761 ...
## $ Area harvested (ha): num [1:202] 1.68e+08 4.83e+09 1.50e+07 1.58e+08 7.50e+09 ...
## - attr(*, "spec")=
##   .. cols(
##     .. `Country` = col_character(),
##     .. `Yield (hg/ha)` = col_double(),
##     .. `Area harvested (ha)` = col_double()
##   )
## - attr(*, "problems")=<externalptr>

```

### 3.1.3 Données sur la consommation de fertilisants de 1961 à 2018

```
str(data_ferti)
```

```

## spec_tbl_df [184 x 2] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
## $ Country : chr [1:184] "Afghanistan" "Albania" "Algeria" "Angola" ...
## $ FertilizerQuantity: num [1:184] 5720345 4316765 8578144 2030699 43153465 ...
## - attr(*, "spec")=
##   .. cols(
##     .. `Country` = col_character(),
##     .. `FertilizerQuantity` = col_double()
##   )
## - attr(*, "problems")=<externalptr>

```

### 3.1.4 Données sur l'usage de pesticides de 1990 à 2017

```
str(data_pest)
```

```

## spec_tbl_df [165 x 2] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
## $ Country : chr [1:165] "Albania" "Algeria" ...
## $ Total Pesticides use per area of land (kg/ha): num [1:165] 14.25 12.88 0.43 72.7 ...
## - attr(*, "spec")=
##   .. cols(
##     .. `Country` = col_character(),
##     .. `Total Pesticides use per area of land (kg/ha)` = col_double()
##   )
## - attr(*, "problems")=<externalptr>

```

### 3.1.5 Données sur les commandes sur Amazon

```
str(orders_data)
```

```

## tibble [171 x 12] (S3: tbl_df/tbl/data.frame)
## $ order_no : chr [1:171] "405-9763961-5211537" "404-3964908-7850720" "171-81033...
## $ order_date : chr [1:171] "Sun, 18 Jul, 2021, 10:38 pm IST" "Tue, 19 Oct, 2021, ...
## $ buyer : chr [1:171] "Mr." "Minam" "yatipertin" "aciya" ...
## $ ship_city : chr [1:171] "CHANDIGARH," "PASIGHAT," "PASIGHAT," "DEVARAKONDA," ...
## $ ship_state : chr [1:171] "CHANDIGARH" "ARUNACHAL PRADESH" "ARUNACHAL PRADESH" ...
## $ sku : chr [1:171] "SKU: 2X-3C0F-KNJE" "SKU: DN-0wdx-vyot" "SKU: DN-0w...
## $ description : chr [1:171] "100% Leather Elephant Shaped Piggy Coin Bank | Block ...
## $ quantity : chr [1:171] "1" "1" "1" "1" ...
## $ item_total : chr [1:171] "<U+20B9>449.00" "<U+20B9>449.00" "<U+20B9>449.00" NA ...
## $ shipping_fee: chr [1:171] NA "<U+20B9>60.18" "<U+20B9>60.18" NA ...
## $ cod : chr [1:171] NA NA NA "Cash On Delivery" ...
## $ order_status: chr [1:171] "Delivered to buyer" "Delivered to buyer" "Delivered to ...

```

## 3.2 Statistiques descriptives

### 3.2.1 Données sur les performances des étudiants

```
library(gtsummary)
tbl_summary(data_student_perf)
```

Characteristic	N = 1,000
gender	
female	518 (52%)
male	482 (48%)
race/ethnicity	
group A	89 (8.9%)
group B	190 (19%)
group C	319 (32%)
group D	262 (26%)
group E	140 (14%)
parental level of education	
associate's degree	222 (22%)
bachelor's degree	118 (12%)
high school	196 (20%)
master's degree	59 (5.9%)
some college	226 (23%)
some high school	179 (18%)
lunch	
free/reduced	355 (36%)
standard	645 (64%)
test preparation course	
completed	358 (36%)
none	642 (64%)
math score	66 (57, 77)
reading score	70 (59, 79)
writing score	69 (58, 79)

### 3.2.2 Données sur les céréales cultivées de 1961 à 2018

```
summary(data_cereal)
```

```
##      Country          Yield (hg/ha)      Area harvested (ha)
##  Length:202        Min.   : 43945      Min.   :2.340e+02
##  Class  :character  1st Qu.: 624933    1st Qu.:4.732e+06
##  Mode   :character  Median  : 993042    Median  :3.032e+07
##                  Mean   :1238853     Mean   :2.667e+08
##                  3rd Qu.:1643923    3rd Qu.:1.343e+08
##                  Max.   :5351936     Max.   :7.503e+09
```

### 3.2.3 Données sur la consommation de fertilisants de 1961 à 2018

```
summary(data_ferti)
```

```
##      Country          FertilizerQuantity
##  Length:184        Min.   :6.312e+03
##  Class  :character  1st Qu.:9.429e+05
##  Mode   :character  Median :5.654e+06
##                  Mean   :6.434e+07
##                  3rd Qu.:2.649e+07
##                  Max.   :2.430e+09
```

### **3.2.4 Données sur l'usage de pестициdes de 1990 à 2017**

```
summary(data_pest)

##      Country          Total Pesticides use per area of land (kg/ha)
##  Length:165      Min.   :  0.02
##  Class :character 1st Qu.:  4.72
##  Mode  :character Median : 31.24
##                  Mean   : 62.89
##                  3rd Qu.: 85.28
##                  Max.   :406.43
```

### **3.2.5 Données sur les commandes sur Amazon**

```
summary(orders_data)
```

```
##      order_no        order_date       buyer      ship_city
##  Length:171      Length:171      Length:171      Length:171
##  Class :character Class :character Class :character Class :character
##  Mode  :character Mode  :character Mode  :character Mode  :character
##      ship_state      sku      description      quantity
##  Length:171      Length:171      Length:171      Length:171
##  Class :character Class :character Class :character Class :character
##  Mode  :character Mode  :character Mode  :character Mode  :character
##      item_total      shipping_fee      cod      order_status
##  Length:171      Length:171      Length:171      Length:171
##  Class :character Class :character Class :character Class :character
##  Mode  :character Mode  :character Mode  :character Mode  :character
```

## **3.3 Données aberrantes**

## **3.4 Présence de doublons**

### **3.4.1 Données sur les performances des étudiants**

```
sum(duplicated(data_student_perf))

## [1] 0
```

### **3.4.2 Données sur les céréales cultivées de 1961 à 2018**

```
sum(duplicated(data_cereal))

## [1] 0
```

### **3.4.3 Données sur la consommation de fertilisants de 1961 à 2018**

```
sum(duplicated(data_ferti))

## [1] 0
```

### **3.4.4 Données sur l'usage de pестициdes de 1990 à 2017**

```
sum(duplicated(data_pest))

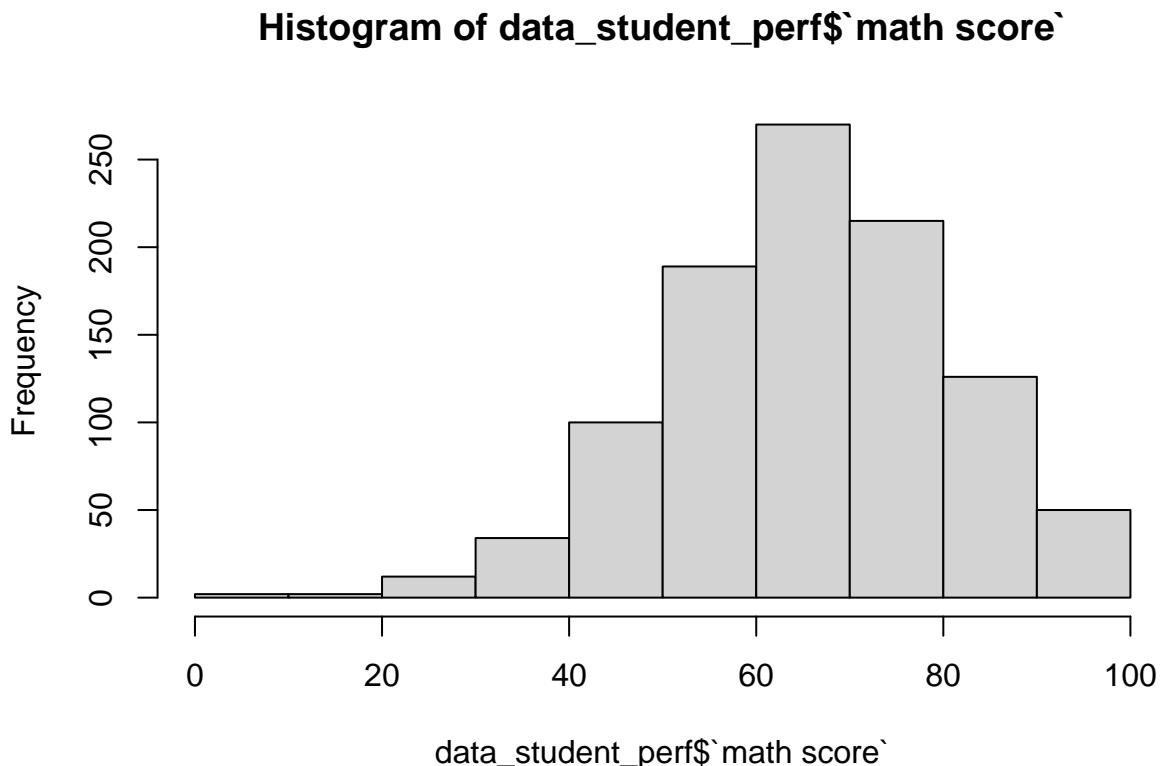
## [1] 0
```

### 3.4.5 Données sur les commandes sur Amazon

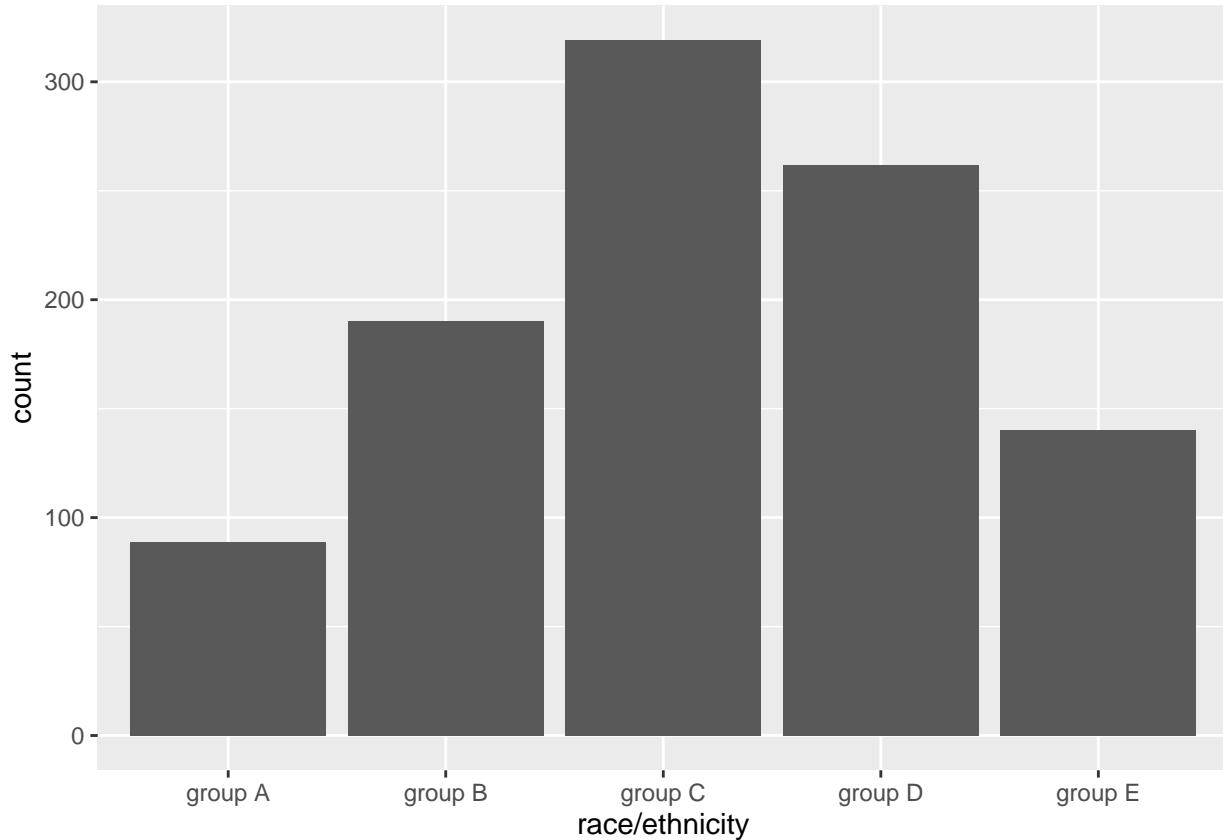
```
sum(duplicated(orders_data))  
## [1] 0  
Aucune des bases ne comporte de doublons
```

## 4 Étape 4 : Quelques visualisations simples des données : diagramme en boîte, diagramme en barre, nuage de points, nuage de mots.

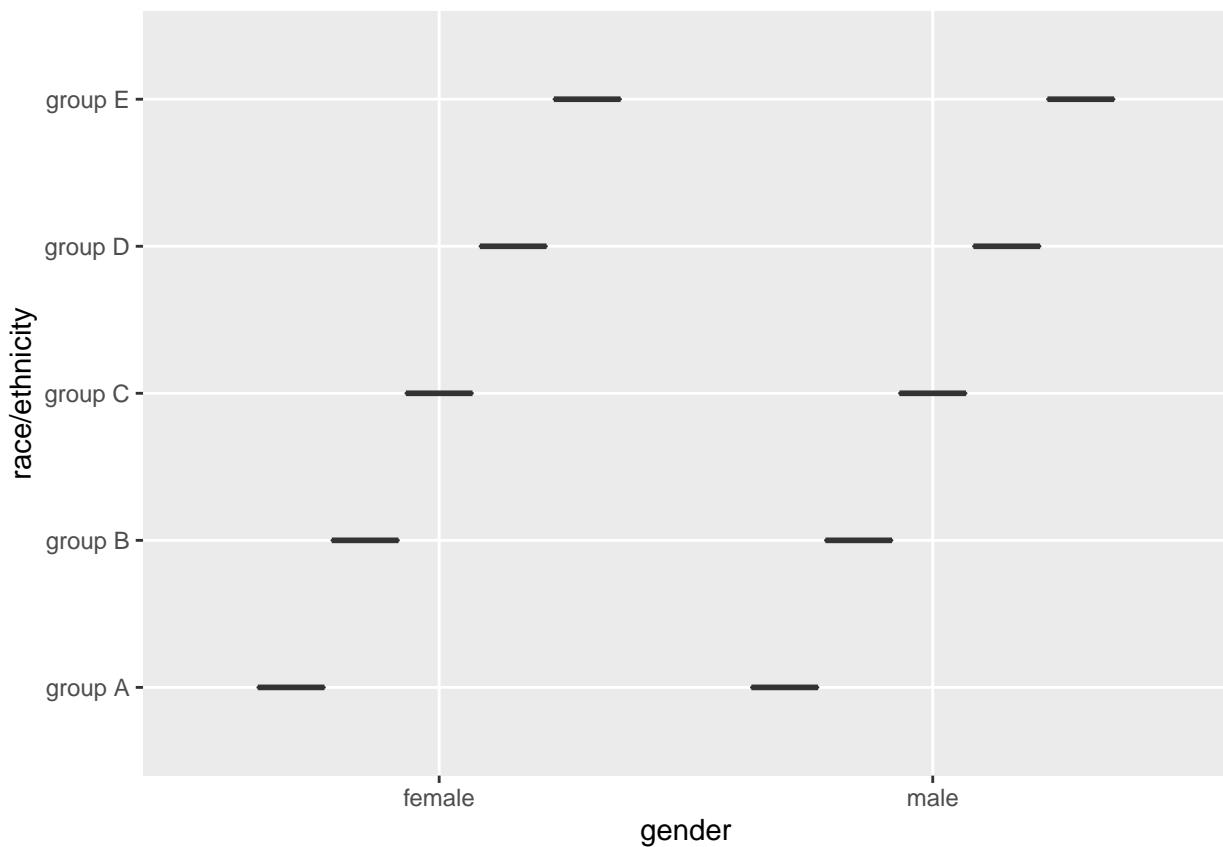
```
hist(data_student_perf$`math score`)
```



```
library(tidyverse)  
ggplot(data_student_perf) + geom_bar(aes(x = `race/ethnicity`))
```



```
ggplot(data_student_perf) + geom_boxplot(aes(x = gender, y = `race/ethnicity`))
```



# Apprendre par la pratique

## Manipulation de données avec R

### Semaines 3 et 4

Dr. Cheikh Ibrahima Fall DIOP\*

2022-05-20

## 1 Fusionner les données du dossier archive

```
# Chargement des bases
setwd("D:/Downloads/LSTP/archive") # Répertoire de travail
library(readr)
# Jeux de données portant sur agriculture
data_cereal <- read_csv("./CerealCropYield_1961-2018.csv")
data_ferti <- read_csv("./FertilizerConsumption_1961-2018.csv")
data_pest <- read_csv("./PesticideUsage_1990-2017.csv")
# Fusion des bases
library(dplyr)
bases <- left_join(data_cereal, data_ferti, by='Country') %>%
    left_join(., data_pest, by='Country')
```

### 1.1 Question 1 : Quels sont les pays exclus à la suite de cette fusion ?

Pour déterminer les pays exclus à la suite de cette fusion, on réalise un *anti\_join*, ce qui permet d'identifier les pays qui ne sont pas présents sur toutes les trois bases.

```
# Fusion des bases
residuals_countries <- anti_join(data_pest, data_ferti, by='Country') %>%
    anti_join(., data_cereal, by='Country')
knitr::kable(residuals_countries[["Country"]], caption = 'Pays exclus à la suite de la fusion')
```

TABLE 1: Pays exclus à la suite de la fusion

Country
Australia & New Zealand
Cook Islands
French Polynesia
Saint Kitts and Nevis
Samoa
Seychelles
Southern Europe
Tonga
Western Asia

\*mamecheikhdiop@gmail.com - cheikh-ibrahima.diop@ugb.edu.sn

†(+221)77-543-01-00 - (+221)76-543-01-00

## 1.2 Question 2 : Quelle est la quantité totale de pesticide utilisée en Amérique, en Afrique et en Australie ?

```
tot_pesto <- bases %>%
  filter(Country == "Americas" | Country == "Africa" | Country == "Australia") %>%
  summarise(tot_pest = sum(`Total Pesticides use per area of land (kg/ha)`,na.rm = TRUE))
```

La quantité totale de pesticide utilisée en Amérique, en Afrique et en Australie est de 41.11 (kg/ha).

## 1.3 Question 3 : Quel est la moyenne et médiane de rendement de céréales (hectogramme par hectare (Hg/Ha)) pour l'Amérique, l'Afrique et l'Australie ?

```
moy_pesto <- bases %>%
  filter(Country == "Americas" | Country == "Africa" | Country == "Australia") %>%
  summarise(moy_pest = mean(`Yield (hg/ha)`,na.rm = TRUE))
med_pesto <- bases %>%
  filter(Country == "Americas" | Country == "Africa" | Country == "Australia") %>%
  summarise(med_pest = median(`Yield (hg/ha)`,na.rm = TRUE))
```

Le rendement moyen de céréales pour l'Amérique, l'Afrique et l'Australie est de 1250626 (hg/ha) ; la médiane pour cette variable est de 924103(hg/ha).

## 2 Travailler avec la base de données portant sur les commande Amazon.

```
# Chargement de la base
setwd("D:/Downloads/LSTP") # Répertoire de travail
library(readr)
# Jeu de données portant sur les commandes Amazon
library(readxl)
orders_data <- read_excel("./orders_data.xlsx")
```

### 2.1 Quelle est la ville où le nombre d'envoie (shipping) est le plus élevé ?

```
max_ship <- orders_data %>%
  group_by(ship_city) %>%
  summarise(nb=n(),prop = round(nb/nrow(orders_data)*100,digits = 2)) %>%
  slice_max(nb)
```

MUMBAI, est la ville d'où vient le plus grand nombre d'envoi. De Cette ville sont partis 17 envois soit 9.94 %.

### 2.2 Convertissez la variable order\_date en années.

```
library(lubridate)
library(stringr)
a <- str_sub(orders_data$order_date, 6, -5)
orders_data$order_date <- year(parse_date_time(a, '%d%m%y, %I:%M %p'))
```

### 2.3 Calculez la moyenne de frais d'envois par année.

```
# On supprime le symbole de la roupie indienne
orders_data$shipping_fee <-str_sub(orders_data$shipping_fee,2,nchar(orders_data$shipping_fee))
# On convertie en nombre
orders_data$shipping_fee <- as.double(orders_data$shipping_fee)
# On calcul la dépense moyenne par années
d<- orders_data %>%
  group_by(order_date) %>%
```

```
summarise(frais_moyen = mean(shipping_fee,na.rm = TRUE))  
knitr::kable(d, caption = 'Moyenne de frais d'envois par année.')
```

TABLE 2: Moyenne de frais d'envois par année.

order_date	frais_moyen
2021	85.79540
2022	81.05125

# Apprendre par la pratique - Session 1: Manipulation de données avec R [Rapport complet]

Daniella Lowa

22/05/2022

## Contents

Qu'est-ce que la manipulation de données ? . . . . .	1
Prédire la fraude dans les services de paiement financier . . . . .	2
Etape 2: Expliquez le contexte des données . . . . .	2
Etape 3: Décrivez vos données (type de variables, statistiques descriptives, données aberrantes, présence de doublons) . . . . .	2
Effet des engrains et des pesticides sur les céréales ? . . . . .	8
Analyse des effets de l'utilisation d'engrais et de pesticides sur les rendements des cultures céréalier dans le monde entier. . . . .	8
<b>Semaines 3 à 4</b>	<b>9</b>
Fusionnage des différentes tables. . . . .	9
Le rôle des pesticides sur les cultures agricoles . . . . .	10
Le rôle des engrains sur les cultures agricoles . . . . .	12
La quantité totale de pesticides . . . . .	14
Amazon . . . . .	15
Montant des commandes et coûts de livraison . . . . .	15
Quelles sont les villes (par ordre décroissant) ayant le plus d'envois ? . . . . .	17
Transformation de la variable order_date en format date . . . . .	20
Détermination des frais d'envoi moyen par année . . . . .	20
#Semaines 1 à 2	

## Qu'est-ce que la manipulation de données ?

A titre d'exemple, en 2018, le marché du Data Marketing représentait *2 milliards d'euros de chiffres d'affaire*, et ce, malgré la mise en application du *Règlement Général sur la Protection des Données*. Or pour profiter de la richesse d'informations des données; il est nécessaire de transformer les données brutes en informations utiles. La manipulation des données est donc, la transformation des données dans le format requis afin

qu'elles puissent être facilement nettoyées et mappées pour extraire des informations. Source:<https://www.astera.com/fr/type/blog/data-manipulation-tools/>

Concrètement sur R, la manipulation des données s'apparente aux outils de visualisation, recodage, tris, fusionnement, scrapping, l'appel des fonctions telles que **dplyr**, **data.table**, **stringr**,**tidyR** etc...

Par soucis d'homogénéité, les cinq jeux de données seront présentés individuellement selon l'ordre des étapes 2, 3 et 4.

## Prédire la fraude dans les services de paiement financier

### Etape 2: Expliquez le contexte des données

Mis en ligne par ARJUN Joshua sur Kaggle il y a quatre ans, ce jeu de données est -du moins à son époque- l'un des quatre jeux de Kaggle contenant *des informations sur le risque croissant de fraude financière numérique*, ce qui souligne la difficulté d'obtenir de telles données.

Il s'agit d'un ensemble de données synthétiques *Paysim* de transactions d'argent mobile. Cet ensemble de données est réduit d'un quart par rapport à l'ensemble de données original qui est présenté dans l'article "**PaySim : A financial mobile money simulator for fraud detection**".

### Etape 3: Décrivez vos données (type de variables, statistiques descriptives, données aberrantes, présence de doublons)

Nous avons 11 variables telles que :

- step, caractéristique *int* : Cartographie une unité de temps dans le monde réel. Dans ce cas, une étape correspond à une heure de temps.
- type, caractéristique *chr* : ceux sont les types de paiement, on retrouve CASH-IN, CASH-OUT, DEBIT, PAYMENT and TRANSFER
- amount, caractéristique *num* : montant de la transaction en monnaie locale
- nameOrig, caractéristique *chr* : le client qui a lancé la transaction
- oldbalanceOrg, caractéristique *num* : solde initial avant la transaction
- newbalanceOrig, caractéristique *num* : le solde du client après la transaction.
- nameDest, caractéristique *chr* : ID du destinataire de la transaction.
- oldbalanceDest, caractéristique *num* : le solde initial du bénéficiaire avant la transaction.
- newbalanceDest, caractéristique *num* : le solde du bénéficiaire après la transaction.
- isFraud, caractéristique *int* : identifie une transaction frauduleuse (1) et non frauduleuse (0)
- isFlaggedFraud, caractéristique *int* : l'origine de isFlaggedFraud n'est pas claire

Le jeu de données contient-il des données abérrantes ou doublons ?

```
summary(duplicated(data.fraud))
```

```
##      Mode   FALSE
## logical 6362620
```

```

d.aberrantes_amount <- which(data.fraud$amount < quantile(data.fraud$amount, 0.25) - 1.5*IQR(data.fraud$amount))
d.aberrantes_oldOrg <- which(data.fraud$oldbalanceOrg < quantile(data.fraud$oldbalanceOrg, 0.25) - 1.5*IQR(data.fraud$oldbalanceOrg))
d.aberrantes_neworig <- which(data.fraud$newbalanceOrig < quantile(data.fraud$newbalanceOrig, 0.25) - 1.5*IQR(data.fraud$newbalanceOrig))
d.aberrantes_newDest <- which(data.fraud$newbalanceDest < quantile(data.fraud$newbalanceDest, 0.25) - 1.5*IQR(data.fraud$newbalanceDest))
d.aberrantes_oldDest <- which(data.fraud$oldbalanceDest < quantile(data.fraud$oldbalanceDest, 0.25) - 1.5*IQR(data.fraud$oldbalanceDest))

Il n'y a ni doublons, ni données abbérantes.

table(data.fraud$type, useNA="always")

##
##   CASH_IN CASH_OUT      DEBIT    PAYMENT TRANSFER      <NA>
## 1399284 2237500     41432  2151495   532909        0

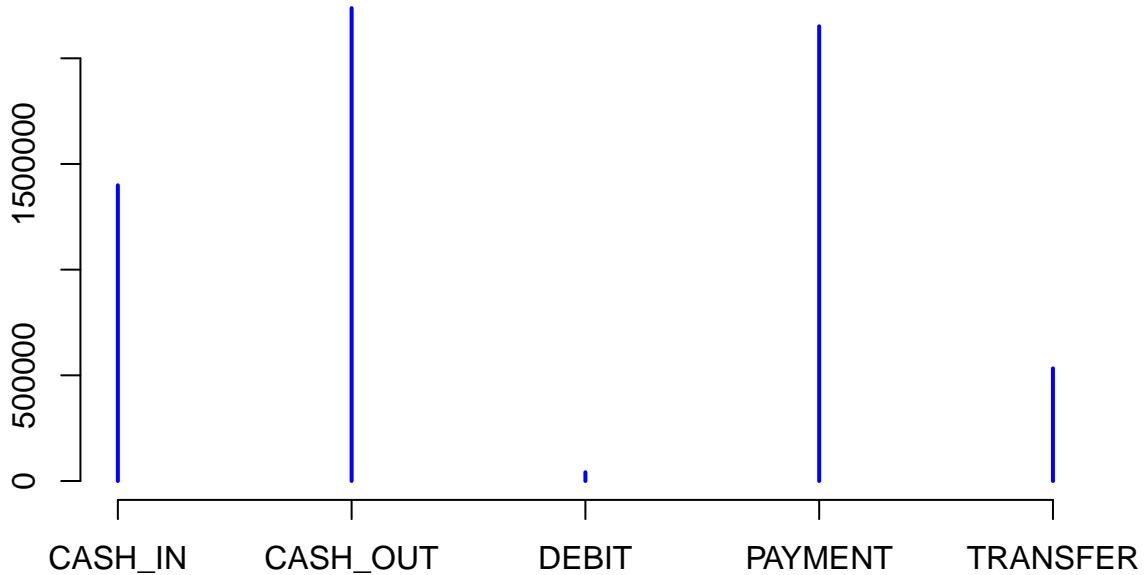
prop.table(table(data.fraud$type))

##
##      CASH_IN      CASH_OUT      DEBIT      PAYMENT      TRANSFER
## 0.219922610 0.351663309 0.006511783 0.338146078 0.083756220

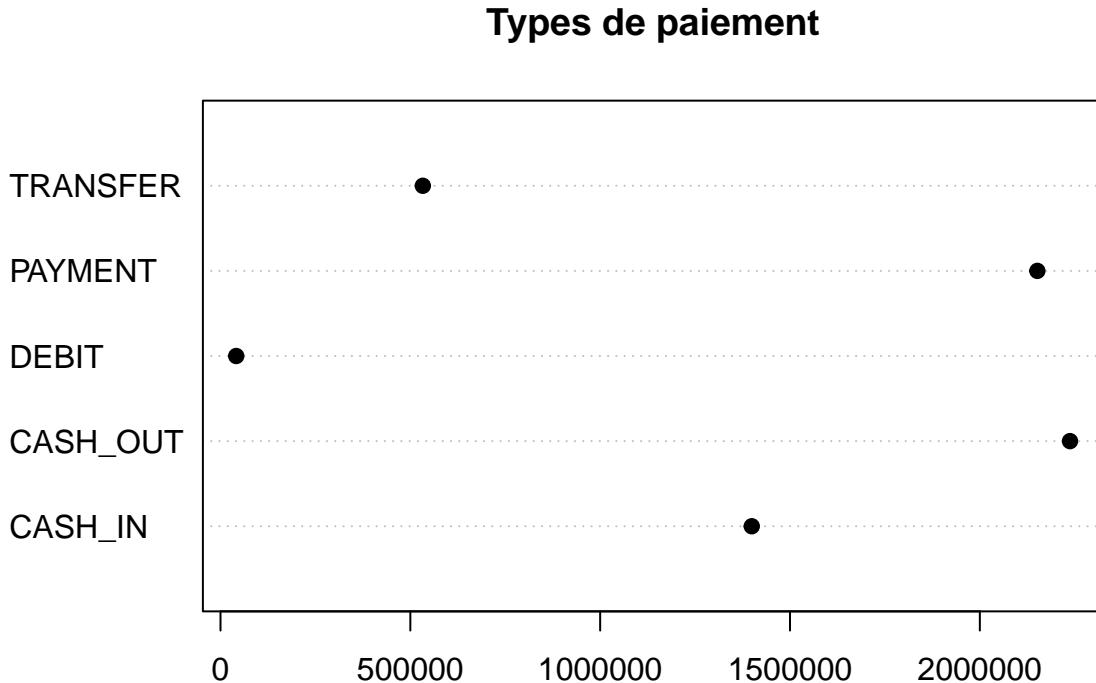
plot(table(data.fraud$type), col="blue", main="paiement", ylab="")

```

## paiement



```
dotchart(as.matrix(table(data.fraud$type))[, 1], main = "Types de paiement", pch = 19)
```

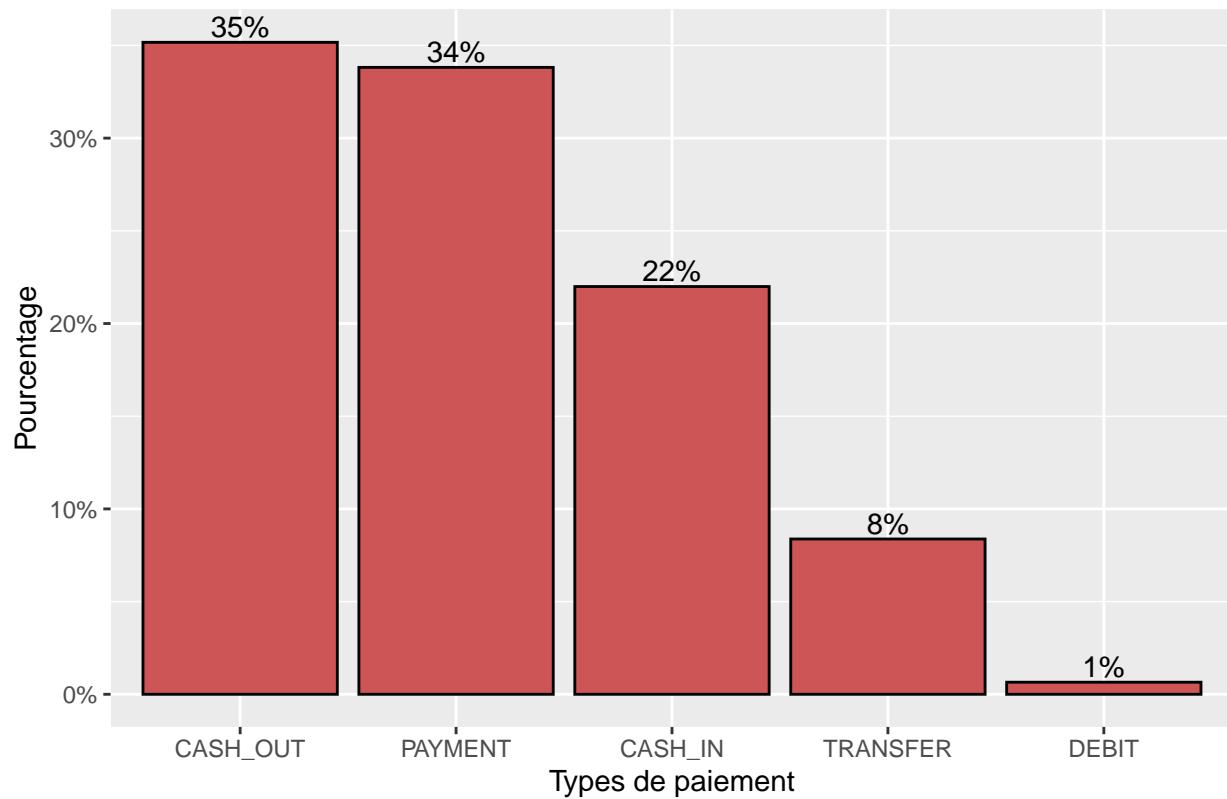


On comptabilise 1399284 d'opérations de dépôt d'argent (Cash-in) Tandis que le nombre de retrait d'argent (cash-out) s'élève à 2237500 Les transactions relatives au paiement, transfert et au débit sont respectivement de 2151495, 532909 et 41432 Les cash-out représentent environ 35%, les paiements 34%, les cash-in 22%, les transferts 8% et seulement 0,6% pour les débits.

Voici une représentation plus qualitative des types de paiement.

```
plotdata <- data.fraud %>%
  count(type) %>%
  mutate(pct = n / sum(n))
  ,
  pctlabel = paste0(round(pct*100), "%"))
ggplot(plotdata,
  aes(x = reorder(type, -pct),
      y = pct)) +
  geom_bar(stat = "identity",
            fill = "indianred3",
            color = "black") +
  geom_text(aes(label = pctlabel),
            vjust = -0.25) +
  scale_y_continuous(labels = scales::percent) +
  labs(x = "Types de paiement",
       y = "Pourcentage",
       title = "Types de paiement")
```

## Types de paiement



On souhaite maintenant avoir une idée sur les caractéristiques générales des montants des transactions.

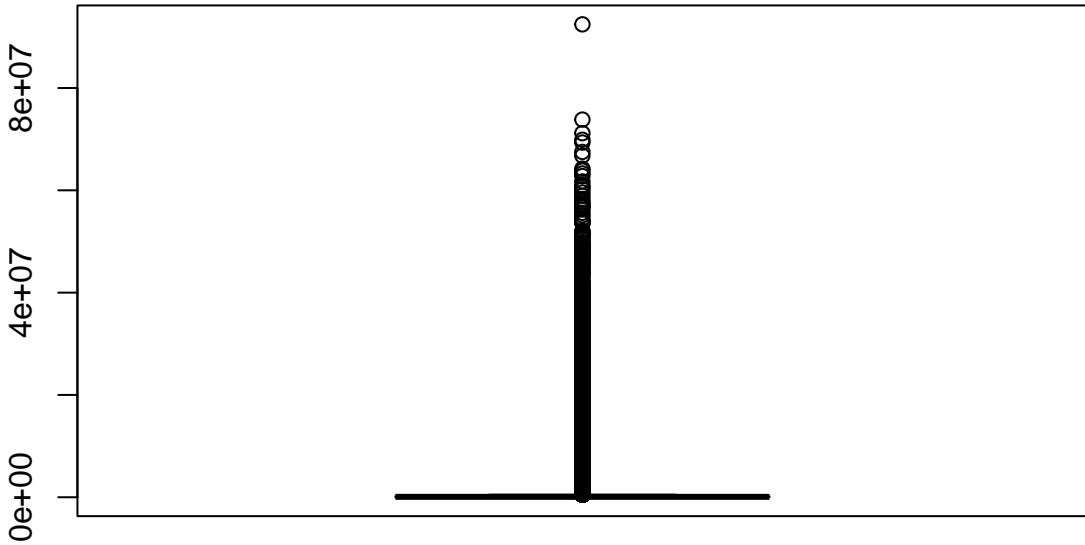
```
summary(data.fraud$amount)
```

```
##      Min.    1st Qu.     Median      Mean    3rd Qu.      Max.
##        0     13390     74872    179862   208721  92445517
```

```
sd(data.fraud$amount, na.rm = TRUE)
```

```
## [1] 603858.2
```

```
boxplot(data.fraud$amount)
```



Les montants des transactions sont au minimum nuls et au maximum de 92 445 517 La moyenne du montant de transaction est de 179 862, or l'écart-type est de 603858.2 donc les valeurs sont dispersées autour de la moyenne

Existe-t-il des personnes étant à la fois donneur et receveur des transactions ?

```
summary(data.fraud$nameOrig==data.fraud$nameDest)
```

```
##      Mode      FALSE
## logical 6362620
```

Non, on ne retrouve pas cette situation.

Vérifions dans un espace relativement court, le solde initial et solde final de tous les clients.

```
sum(data.fraud$oldbalanceOrg - data.fraud$newbalanceOrig)
```

```
## [1] -1.35082e+11
```

Visiblement, la différence est négative. Toutefois, l'unité de temps étant une heure, il est possible que d'autres transactions interviennent entre temps mais nous pouvons également suspecter une potentielle fraude.

```
summary(data.fraud$newbalanceOrig<data.fraud$oldbalanceOrg)
```

```
##      Mode      FALSE      TRUE
## logical 3488290 2874330
```

Lorsqu'on regarde de plus près, seulement 2 874 330 soldes initiaux sont supérieurs aux soldes finaux. C'est contre-intuitif et on peut se demander s'il n'y a pas un réseau financier entre clients où de multiples transactions transitent entre comptes.

Du côté des bénéficiaires, regardons si le solde augmente globalment après les transactions.

```
sum(data.fraud$newbalanceDest - data.fraud$oldbalanceDest)
```

```
## [1] 790840145696
```

Les nouveaux soldes sont globalement en hausse

Toutefois.. Vérifions également si les soldes finaux sont tout le temps supérieurs aux soldes initiaux

```
summary(data.fraud$oldbalanceDest<data.fraud$newbalanceDest)
```

```
##      Mode   FALSE    TRUE  
## logical 3556156 2806464
```

Quel choc! 2 806 464 des soldes finaux sont inférieurs aux soldes initiaux. Illogique.

Reste à découvrir si les détections de fraude sauront mettre en relief les premiers alertes révélées ci-dessus:

```
summary(data.fraud$isFraud=="1")
```

```
##      Mode   FALSE    TRUE  
## logical 6354407     8213
```

Visiblement pas car sur l'ensemble des transactions soit plus de 6 millions, on recense que 8 213 cas de fraude.

La variable ci-dessous n'est pas identifiée mais regardons malgré tout la fréquence des modalités.

```
summary(data.fraud$isFlaggedFraud=="0")
```

```
##      Mode   FALSE    TRUE  
## logical        16 6362604
```

```
head(summary(data.fraud$isFlaggedFraud=="1"))
```

```
##      Mode   FALSE    TRUE  
## logical 6362604      16
```

Seulement 11 "isFlaggedFraud" ont une modalité "1". Ni information ni lien ne ressort.

Une opération frauduleuse a-t-elle un lien avec la variable inconnue de type "1" ?

```
cor(data.fraud$isFraud=="1",data.fraud$isFlaggedFraud=="1")
```

```
## [1] 0.0441092
```

Non, il n'y a pas de résultat concluant entre ces deux variables.

## Effet des engrais et des pesticides sur les céréales ?

Analyse des effets de l'utilisation d'engrais et de pesticides sur les rendements des cultures céréaliers dans le monde entier.

```
library(questionr)

## Warning: package 'questionr' was built under R version 4.1.3

pesticide <- readxl::read_xlsx("C:/Users/danie/Desktop/Pesticide.xlsx")
pesticide <- rename.variable(pesticide,"Total Pesticides (kg/ha)","pesticides_kg/ha")
fertilizer <- readxl::read_xlsx("C:/Users/danie/Desktop/Fertilizer.xlsx")
fertilizer <- rename.variable(fertilizer,"FertilizerQuantity","Quantité_engrais")
cereal <- readxl::read_xlsx("C:/Users/danie/Desktop/Cereal.xlsx")
cereal <- rename.variable(cereal,"Area_harvested_ha","Superficie_récoltée")
cereal <- rename.variable(cereal, "Yield_hg/ha","Rendement_hg/ha")
```

Le jeu de données cereal : rendement mondial des cultures céréaliers de 1961 à 2018, on retrouve notamment les variables: superficie récoltée et rendement par hg/ha Le jeu de données pesticide : utilisation mondiale de pesticides de 1990 à 2017, on retrouve l'utilisation de pesticides en kg/ha Le jeu de données fertilizer : consommation mondiale d'engrais de 1961 à 2018, ici c'est la quantité d'engrais qui est représentée.

Nous sommes-nous en présence de doublons ou données abérrantes ?

```
données_aberrantes_pesticides <- which(pesticide$`pesticides_kg/ha` < quantile(pesticide$`pesticides_kg/ha`, 0.25))
données_aberrantes_engrais <- which(fertilizer$Quantité_engrais < quantile(fertilizer$Quantité_engrais, 0.25))
données_aberrantes_superficie <- which(cereal$Superficie_récoltée < quantile(cereal$Superficie_récoltée, 0.25))
données_aberrantes_rendement <- which(cereal$`Rendement_hg/ha` < quantile(cereal$`Rendement_hg/ha`, 0.25))
summary(duplicated(pesticide))
```

```
##      Mode   FALSE
## logical      165
```

```
summary(duplicated(cereal))
```

```
##      Mode   FALSE
## logical      202
```

```
summary(duplicated(fertilizer))
```

```
##      Mode   FALSE
## logical      184
```

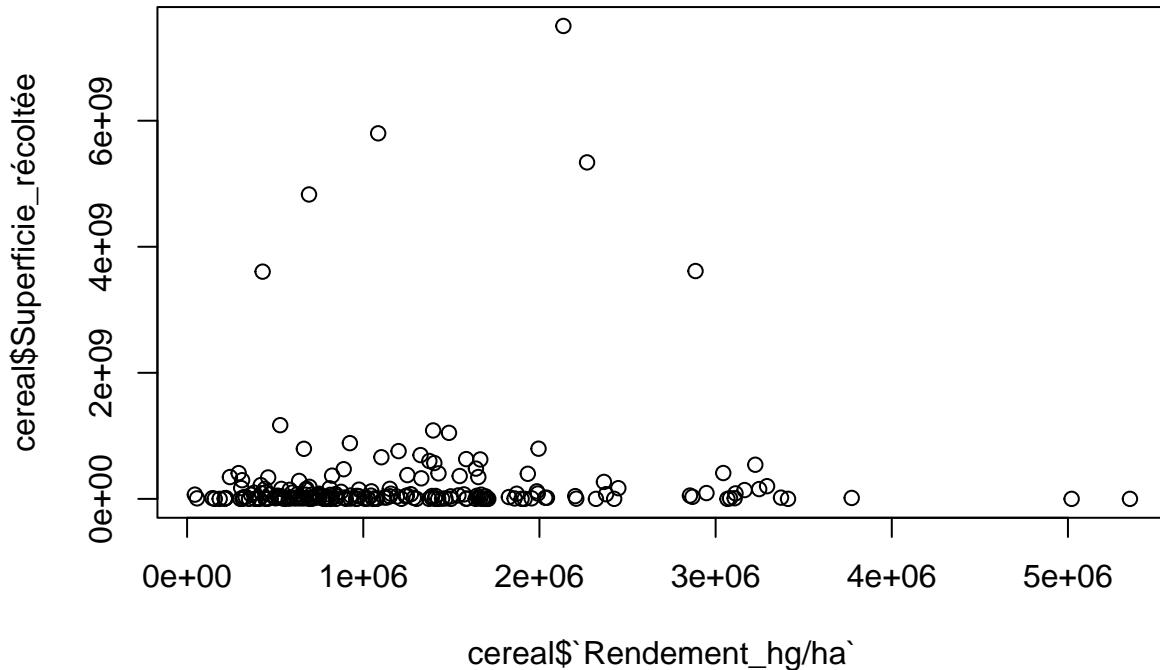
Il n'y a ni données abbérantes ni doublons dans les variables agricoles.

####Lien entre superficie récoltée et rendement ?

```
cor(cereal$`Rendement_hg/ha`,cereal$Superficie_récoltée)
```

```
## [1] 0.07948626
```

```
plot(x = cereal$`Rendement_hg/ha`, y = cereal$Superficie_récoltée )
```



Premier constat: Une grande superficie de récolte n'induit pas un rendement élevé. Il existe une très faible corrélation entre les deux variables (0.08).

## Semaines 3 à 4

Fusionnage des différentes tables.

```
agr <- merge(pesticide, fertilizer, by = "Country")
agri <- merge(agr, cereal, by = "Country")
exclus <- anti_join(cereal, agri, by = "Country")
exclus$Country
```

```
## [1] "Afghanistan"          "Africa"
## [3] "Americas"             "Antigua and Barbuda"
## [5] "Bahamas"               "Barbados"
## [7] "Belgium-Luxembourg"   "Benin"
## [9] "Bosnia and Herzegovina" "Cambodia"
## [11] "Central America"       "Colombia"
## [13] "Costa Rica"            "Cote d'Ivoire"
## [15] "Cuba"                  "Czechoslovakia"
## [17] "Democratic Republic of Congo" "Djibouti"
```

```

## [19] "Dominica"
## [21] "Ethiopia PDR"
## [23] "Gabon"
## [25] "Grenada"
## [27] "Guam"
## [29] "Liberia"
## [31] "Malta"
## [33] "Micronesia (country)"
## [35] "Middle Africa"
## [37] "Montserrat"
## [39] "North Korea"
## [41] "Puerto Rico"
## [43] "Saint Lucia"
## [45] "Sao Tome and Principe"
## [47] "Sierra Leone"
## [49] "Somalia"
## [51] "Sudan (former)"
## [53] "United Arab Emirates"
## [55] "Western Sahara"

## [19] "Eswatini"
## [21] "French Guiana"
## [23] "Georgia"
## [25] "Guadeloupe"
## [27] "Hong Kong"
## [29] "Maldives"
## [31] "Melanesia"
## [33] "Micronesia (region)"
## [35] "Mongolia"
## [37] "Nigeria"
## [39] "Philippines"
## [41] "Reunion"
## [43] "Saint Vincent and the Grenadines"
## [45] "Serbia"
## [47] "Solomon Islands"
## [49] "South Sudan"
## [51] "Trinidad and Tobago"
## [53] "Uzbekistan"

```

Il y a 55 pays et régions du monde exclus suite à la fusion. Ceux sont les suivants: “Afghanistan” “Africa” “Americas”

“Antigua and Barbuda” “Bahamas” “Barbados”  
 “Belgium-Luxembourg” “Benin” “Bosnia and Herzegovina”  
 “Cambodia” “Central America” “Colombia”  
 “Costa Rica” “Cote d’Ivoire” “Cuba”  
 “Czechoslovakia” “Democratic Republic of Congo” “Djibouti”  
 “Dominica” “Eswatini” “Ethiopia PDR”  
 “French Guiana” “Gabon” “Georgia”  
 “Grenada” “Guadeloupe” “Guam”  
 “Hong Kong” “Liberia” “Maldives”  
 “Malta” “Melanesia” “Micronesia (country)”  
 “Micronesia (region)” “Middle Africa” “Mongolia”  
 “Montserrat” “Nigeria” “North Korea”  
 “Philippines” “Puerto Rico” “Reunion”  
 “Saint Lucia” “Saint Vincent and the Grenadines” “Sao Tome and Principe”  
 “Serbia” “Sierra Leone” “Solomon Islands”  
 “Somalia” “South Sudan” “Sudan (former)”  
 “Trinidad and Tobago” “United Arab Emirates” “Uzbekistan”  
 “Western Sahara”

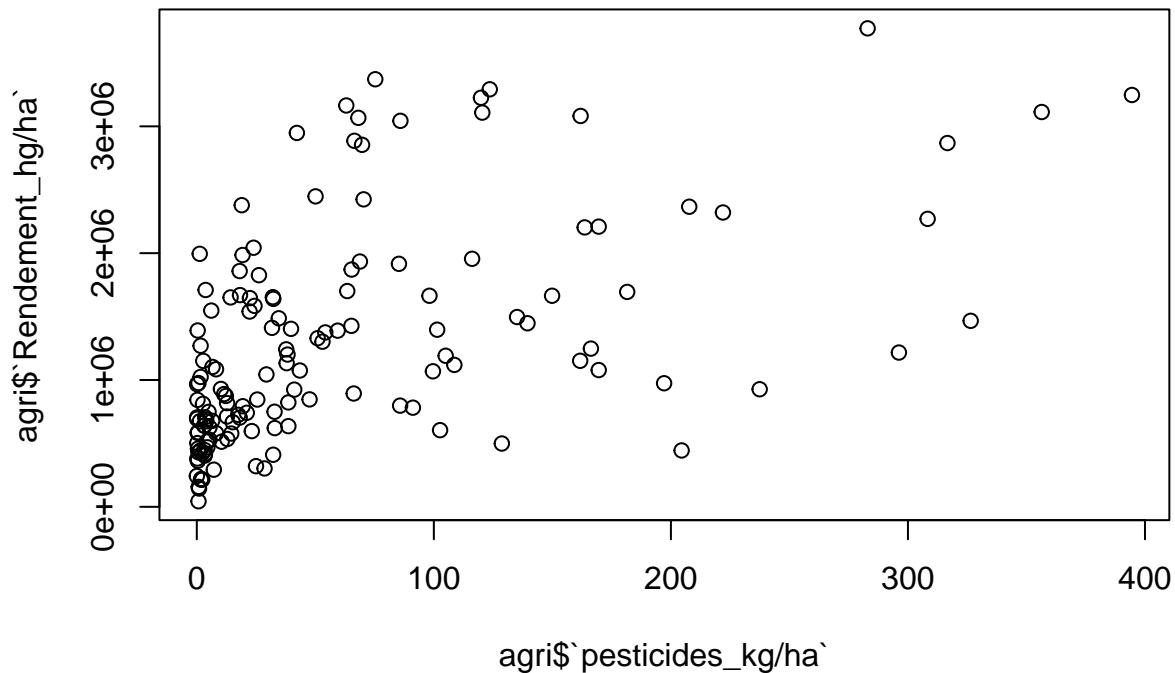
**Le rôle des pesticides sur les cultures agricoles** Quel impact ont les pesticides sur les céréales ?

```
cor(agri$`pesticides_kg/ha`, agri$`Rendement_hg/ha`)
```

```
## [1] 0.5358773
```

```
plot(agri$`pesticides_kg/ha`, agri$`Rendement_hg/ha` , main="Pesticides vs Récoltes")
```

## Pesticides vs Récoltes

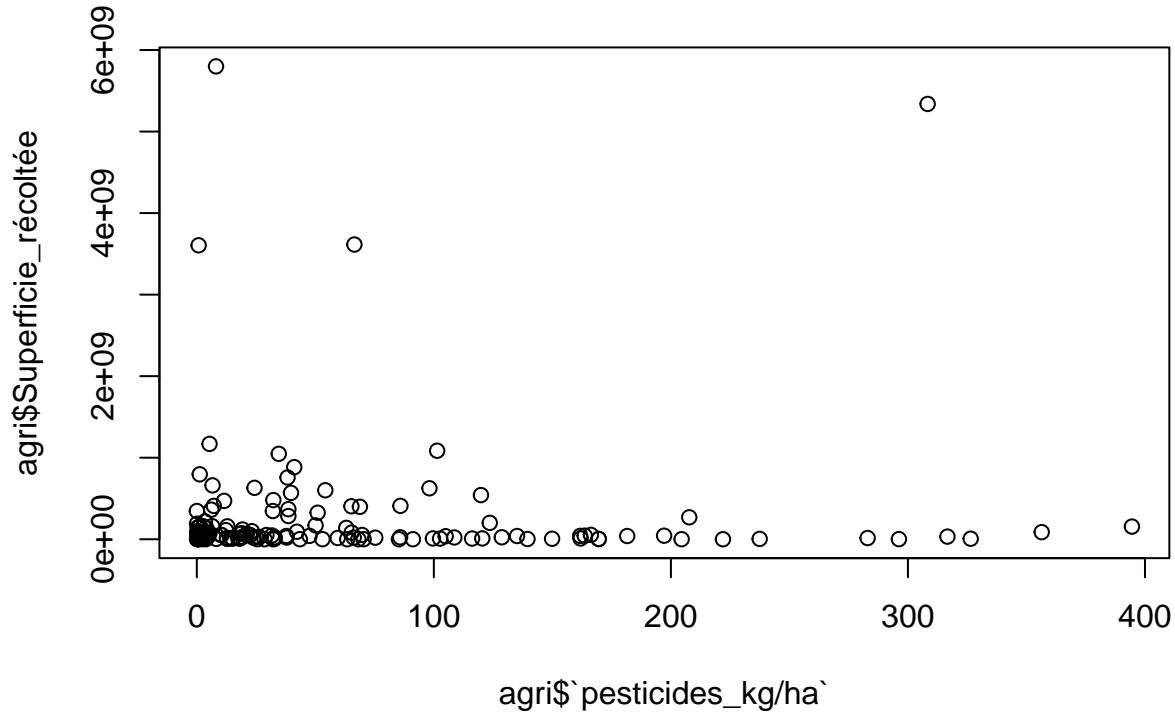


Il y a une faible corrélation positive (0.54) entre les variables. L'ajout de pesticides dans les cultures céréalières, accompagnent de façon modérée l'augmentation des récoltes.

Quel impact ont les pesticides sur les surface récoltées ?

```
cor( agri$`pesticides_kg/ha` , agri$Superficie_récoltée)  
## [1] 0.06116242  
plot(agri$`pesticides_kg/ha` , agri$Superficie_récoltée, main="Pesticides vs Surface récoltée")
```

## Pesticides vs Surface récoltée



Il y a une très faible corrélation (0.06) entre les pesticides et la surface récoltée, on peut penser qu'une faible quantité de pesticides couvre initialement un large périmètre de surface récoltée.

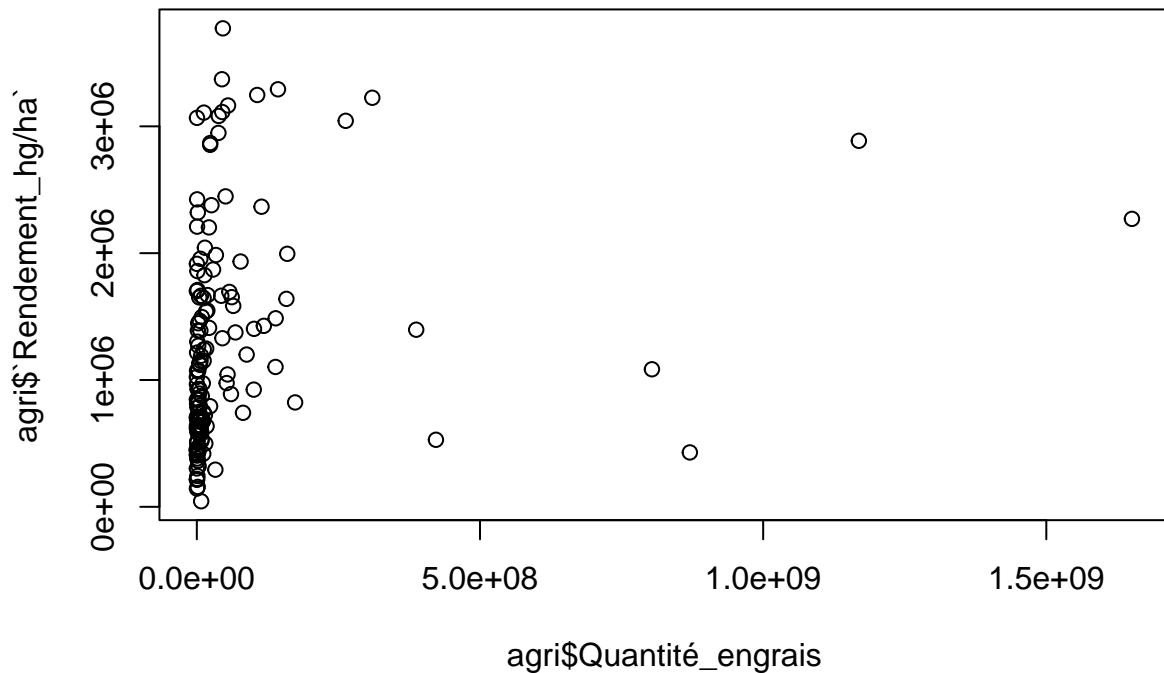
**Le rôle des engrains sur les cultures agricoles** Quel impact ont les engrains sur les céréales ?

```
cor(agri$Quantité_engrais,agri$`Rendement_hg/ha`)
```

```
## [1] 0.2032389
```

```
plot(agri$Quantité_engrais,agri$`Rendement_hg/ha` , main="Engrais vs Récoltes céréalier")
```

## Engrais vs Récoltes céréalières



La faible relation positive (0.20) entre la quantité d'engrais et les récoltes céréalières suggère que la quantité d'engrais n'est pas nécessaire à l'obtention d'une grande récolte céréalière. On peut s'interroger alors, sur la nécessité d'une terre fertile/qualitative et/ou de la qualité des engrais.

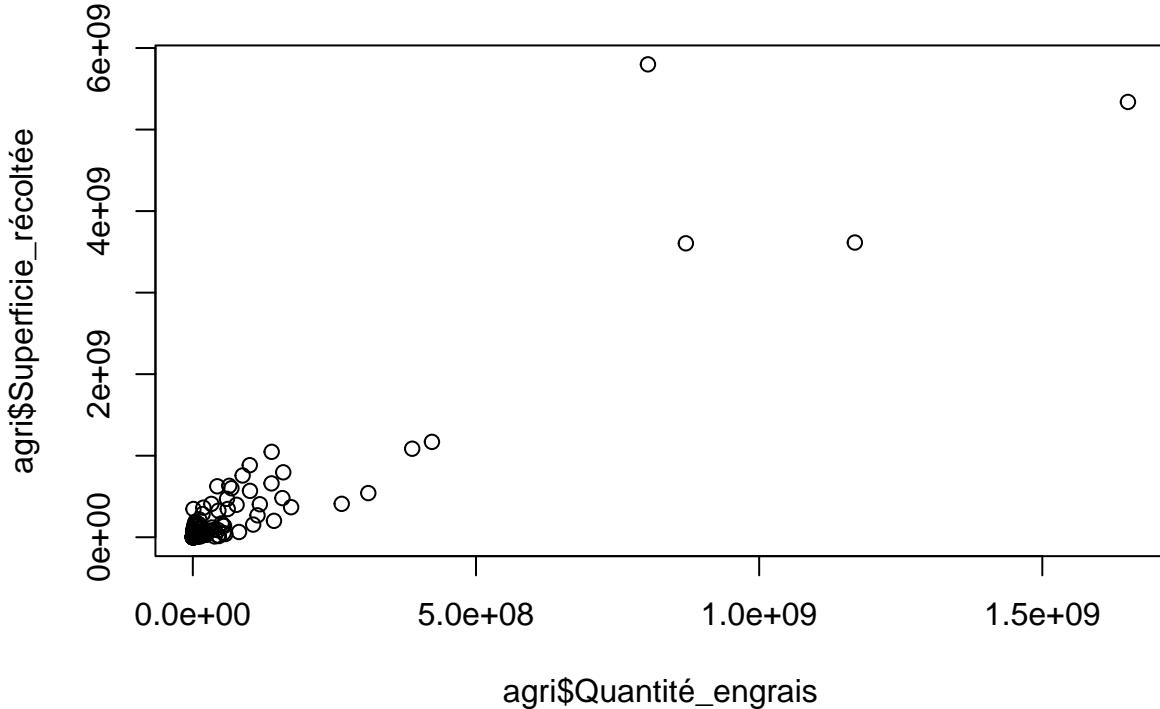
Quel impact ont les engrais sur les surface récoltées ?

```
cor(agri$Quantité_engrais, agri$Superficie_récoltée)
```

```
## [1] 0.926566
```

```
plot(agri$Quantité_engrais, agri$Superficie_récoltée, main="Engrais vs Surface récoltée")
```

## Engrais vs Surface récoltée



Il y a une corrélation élevée (0.93) entre les engrais et la surface récoltée, ce qui peut-être logique vu qu'il faut répartir les engrains selon la surface d'une part, et d'autre part, on peut se poser la question quant à la possibilité d'une terre agricole de produire sans quelconque engrais..

### La quantité totale de pesticides

```
pestiAfri <- agri %>% filter(Country=="Africa") %>% select(`pesticides_kg/ha`)
pestiAme <- agri %>% filter(Country=="Americas") %>% select(`pesticides_kg/ha`)
```

Il est impossible de savoir la quantité totale de pesticides pour l'Amérique et l'Afrique car nous n'avons pas les données. Regardons pour l'Australie.

```
pestiAus <- agri %>% filter(Country=="Australia") %>% select(`pesticides_kg/ha`)
superfiAus <- agri %>% filter(Country=="Australia") %>% select(`Superficie_récoltée`)
prod(pestiAus$`pesticides_kg/ha`, superfiAus$Superficie_récoltée)
```

```
## [1] 36339211263
```

La quantité totale de pesticides déployée en Australie aux vues de la superficie récoltée, est de 36 339 211 263 kilogrammes.

```
rendAfri <- cereal %>% filter(Country=="Africa") %>% select(`Rendement_hg/ha`)
rendAme <- cereal %>% filter(Country=="Americas") %>% select(`Rendement_hg/ha`)
```

```

rendAus <- cereal %>% filter(Country=="Australia") %>% select(`Rendement_hg/ha`)
rendAfri <- as.numeric(rendAfri)
rendAus <- as.numeric(rendAus)
rendAme <- as.numeric(rendAme)
median(rendAus,rendAme,rendAfri)

## [1] 924103

mean(rendAus,rendAme,rendAfri)

## [1] 924103

(rendAus+rendAfri+rendAme)/3

## [1] 1250626

```

Les fonctions mean et median donnent la même valeur, on peut suspecter que les données soient symétriques. Néanmoins, une autre méthode me donne une moyenne différente. Pour être sûre, je vais revérifier le résultat d'une autre façon. *Problème de sortie sur Rmardown, toutefois avec les formules ci-dessous, on règle le problème* verif <- rbind(rendAfri,rendAus) veriff <- rbind(verif,rendAme) var <- veriff\$Rendement\_hg/ha mean(var) median(var) Heureusement que nous avons vérifié; la moyenne du rendement en hg par ha pour l'Afrique, l'Australie et les Amériques est de 1250626, tandis que leur médiane est de 924103.

## Amazon

```

AMZN <- read.csv("C:/Users/danie/Desktop/amazon.csv", sep=",", header=TRUE)
AMZN2 <- na.omit(AMZN)

```

Nous sommes en possession de plusieurs données concernant des commandes. On aimerait se faire une image objective du contenu. Dans ce jeu de données, nous avons les variables suivantes: Order\_no : numéro de commande Order\_date : Date de commande Buyer : Acheteur Ship\_city : Ville où l'achat se fait Ship\_state : région où l'achat se fait SKU : numéro de série de l'article Description: Description de l'article Quantity : quantité achetée Item\_total : Montant total de la commande Shipping\_fee : Coût de livraison Cod : Règlement lors de la commande Order\_status : statut de la livraison

### Montant des commandes et coûts de livraison

```

AMZN2$item_total <- as.numeric(AMZN2$item_total)

sum(AMZN2$item_total, na.rm=TRUE)

## [1] 78343

sum(AMZN2$shipping_fee, na.rm=TRUE)

## [1] 11076.66

```

Les sommes des montants totaux de commandes et coûts de livraisons sont respectivement de 78 343 et 11 076.66

```
cor(AMZN2$item_total,AMZN2$shipping_fee)
```

```
## [1] 0.5578375
```

```
AMZN2$cod
```

```
## [1] ""          ""          ""  
## [4] "Cash On Delivery" ""          ""  
## [7] ""          "Cash On Delivery" ""  
## [10] ""         "Cash On Delivery" ""  
## [13] ""         "Cash On Delivery" ""  
## [16] ""         ""          "Cash On Delivery"  
## [19] "Cash On Delivery" "Cash On Delivery" ""  
## [22] ""         "Cash On Delivery" ""  
## [25] "Cash On Delivery" ""          ""  
## [28] ""         "Cash On Delivery" ""  
## [31] ""         ""          "Cash On Delivery"  
## [34] ""         "Cash On Delivery" "Cash On Delivery"  
## [37] ""         ""          ""  
## [40] "Cash On Delivery" ""          ""  
## [43] ""         "Cash On Delivery" ""  
## [46] ""         ""          ""  
## [49] ""         "Cash On Delivery" ""  
## [52] ""         ""          ""  
## [55] ""         ""          ""  
## [58] "Cash On Delivery" ""          ""  
## [61] ""         "Cash On Delivery" ""  
## [64] ""         ""          ""  
## [67] ""         "Cash On Delivery" ""  
## [70] "Cash On Delivery" ""          ""  
## [73] "Cash On Delivery" "Cash On Delivery" ""  
## [76] ""         ""          "Cash On Delivery"  
## [79] ""         "Cash On Delivery" ""  
## [82] ""         ""          ""  
## [85] ""         ""          ""  
## [88] ""         ""          ""  
## [91] ""         ""          ""  
## [94] ""         ""          ""  
## [97] ""         "Cash On Delivery" ""  
## [100] ""        ""          ""  
## [103] ""        ""          "Cash On Delivery"  
## [106] ""        "Cash On Delivery" ""  
## [109] ""        ""          ""  
## [112] "Cash On Delivery" ""          ""  
## [115] ""        ""          ""  
## [118] "Cash On Delivery" "Cash On Delivery" ""  
## [121] ""        ""          "Cash On Delivery"  
## [124] ""        ""          "Cash On Delivery"  
## [127] ""        ""          "Cash On Delivery"  
## [130] ""
```

```
prop.table(table(AMZN2$cod))
```

```
##  
##           Cash On Delivery  
##      0.7461538      0.2538462
```

```
summary(AMZN2$cod=="Cash On Delivery")
```

```
##     Mode    FALSE    TRUE  
## logical      97      33
```

```
summary(AMZN2$order_status=="Delivered to buyer")
```

```
##     Mode    TRUE  
## logical     130
```

Le montant de la commande est moyennement corrélé (0.56) au coût de livraison. Pour le moment, on sait que le paiement à la livraison (Cash on Delivry) est une pratique assez faible dans l'ensemble des transactions, elle représente O.25% des commandes soit 33 commandes. Finalement, on est curieux de savoir si le paiement à la livraison dépend du statut de la livraison ou du montant de la commande..

```
cor(AMZN2$cod=="Cash On Delivery",AMZN2$shipping_fee)
```

```
## [1] -0.04586206
```

```
cor(AMZN2$cod=="Cash On Delivery",AMZN2$item_total)
```

```
## [1] -0.06046965
```

On peut confirmer que le paiement à la livraison ne dépend pas du statut de la livraison, ni du montant de la commande.

**Quelles sont les villes (par ordre décroissant) ayant le plus d'envois ?**

```
AMZN2$ship_city <- gsub(",","", AMZN2$ship_city)  
AMZN2$ship_city
```

```
## [1] "PASIGHAT"          "PASIGHAT"          "MUMBAI"  
## [4] "BAREILLY"          "BENGALURU"          "FARIDABAD"  
## [7] "AGARTALA"          "COONOOR"            "PUNE"  
## [10] "KOLKATA"           "MAHALINGPUR"        "MUMBAI"  
## [13] "HYDERABAD"          "MUMBAI"              "MUMBAI 400 026"  
## [16] "CUTTACK"            "BENGALURU"          "JALESWAR"  
## [19] "PUNEpune"          "NEW DELHI"           "Bhubaneswar"  
## [22] "JAGDALPUR"         "HYDERABAD"           "BENGALURU"  
## [25] "KOLKATA"            "BENGALURU"           "SALEM"  
## [28] "PUNE"                "JAMMU"               "HYDERABAD"
```

```

## [31] "AHMEDABAD"
## [34] "HYDERABAD"
## [37] "KOLKATA"
## [40] "Surat"
## [43] "BAREILLY"
## [46] "BIDHAN NAGAR"
## [49] "CHENNAI"
## [52] "BENGALURU"
## [55] "MUMBAI"
## [58] "CHENNAI"
## [61] "GURUGRAM"
## [64] "AHMEDABAD"
## [67] "KATWA"
## [70] "JODHPUR"
## [73] "CHANDIGARH"
## [76] "CHANDIGARH"
## [79] "GURUGRAM"
## [82] "Mumbai"
## [85] "CHENNAI"
## [88] "INDORE"
## [91] "NEW DELHI"
## [94] "BENGALURU"
## [97] "GHAZIABAD"
## [100] "SECUNDERABAD"
## [103] "KOLKATA"
## [106] "KOLKATA"
## [109] "LUCKNOW"
## [112] "Kodambakkam Chennai"
## [115] "CHENNAI"
## [118] "BILIMORA"
## [121] "KANPUR"
## [124] "Visakhapatnam"
## [127] "DEHRADUN"
## [130] "MUMBAI"

"GURUGRAM"
"KARAIKKUDI"
"KOLKATA"
"Pune"
"MUMBAI"
"JAGDALPUR"
"PALAI"
"HYDERABAD"
"BENGALURU"
"Mumbai"
"GUWAHATI"
"GURUGRAM"
"NEW DELHI"
"KATWA"
"JALANDHAR"
"JAIPUR"
"KORBA"
"KOLKATA"
"MUMBAI"
"THAMARASSERY"
"HYDERABAD"
"NEW DELHI"
"NEW DELHI"
"NOIDA"
"Kolkata"
"SIWAN"
"Kolkata"
"chennai"
"HYDERABAD"
"BENGALURU"
"SALEM"
"CHENNAI"
"PUNE"
"Kolkata"
"PUNJAB"
"KUMBAI"
"chennai"
"BENGALURU"
"Visakhapatnam"
"BENGALURU"
"AMROHA"
"BURDWAN"
"ALLAHABAD"
"NEW DELHI"
"VADODARA"
"AMROHA"
"LUCKNOW"
"KOLKATA"
"NAVY MUMBAI"
"Solan"
"Durg"
"KOLKATA"

```

```
prop.table(table(AMZN2$ship_city))
```

	AGARTALA	AHMEDABAD	ALLAHABAD	AMROHA
##	0.007692308	0.015384615	0.007692308	0.007692308
##	Bardez	BAREILLY	BENGALURU	Bhubaneswar
##	0.007692308	0.015384615	0.092307692	0.007692308
##	BIDHAN NAGAR	BILIMORA	BURDWAN	CHANDIGARH
##	0.007692308	0.007692308	0.007692308	0.015384615
##	chennai	CHENNAI	COONOOR	CUTTACK
##	0.015384615	0.038461538	0.007692308	0.007692308
##	DEHRADUN	Durg	FARIDABAD	GHAZIABAD
##	0.007692308	0.007692308	0.007692308	0.007692308
##	GURUGRAM	GUWAHATI	HYDERABAD	INDORE
##	0.038461538	0.007692308	0.053846154	0.007692308
##	JAGDALPUR	JAIPUR	JALANDHAR	JALESWAR
##	0.015384615	0.007692308	0.007692308	0.007692308
##	JAMMU	JODHPUR	KANPUR	KARAIKKUDI
##	0.007692308	0.007692308	0.007692308	0.007692308

```

##          KATWA Kodambakkam Chennai           Kolkata      KOLKATA
## 0.007692308          0.007692308 0.023076923 0.069230769
##          KORBA             LUCKNOW           MAHALINGPUR    MALDA
## 0.007692308          0.007692308 0.007692308 0.007692308
##          Mumbai            MUMBAI          MUMBAI 400 026 NAVI MUMBAI
## 0.015384615          0.107692308 0.007692308 0.015384615
##          NEW DELHI          NOIDA            PALAI        PASIGHAT
## 0.046153846          0.015384615 0.007692308 0.015384615
##          Pune              PUNE          PUNEpune      PUNJAB
## 0.007692308          0.023076923 0.007692308 0.007692308
##          SALEM          SECUNDERABAD          SILCHAR      SIWAN
## 0.015384615          0.015384615 0.007692308 0.007692308
##          Solan              Surat          THAMARASSERY Thane District
## 0.007692308          0.007692308 0.007692308 0.007692308
##          VADODARA          Visakhapatnam
## 0.007692308          0.023076923

```

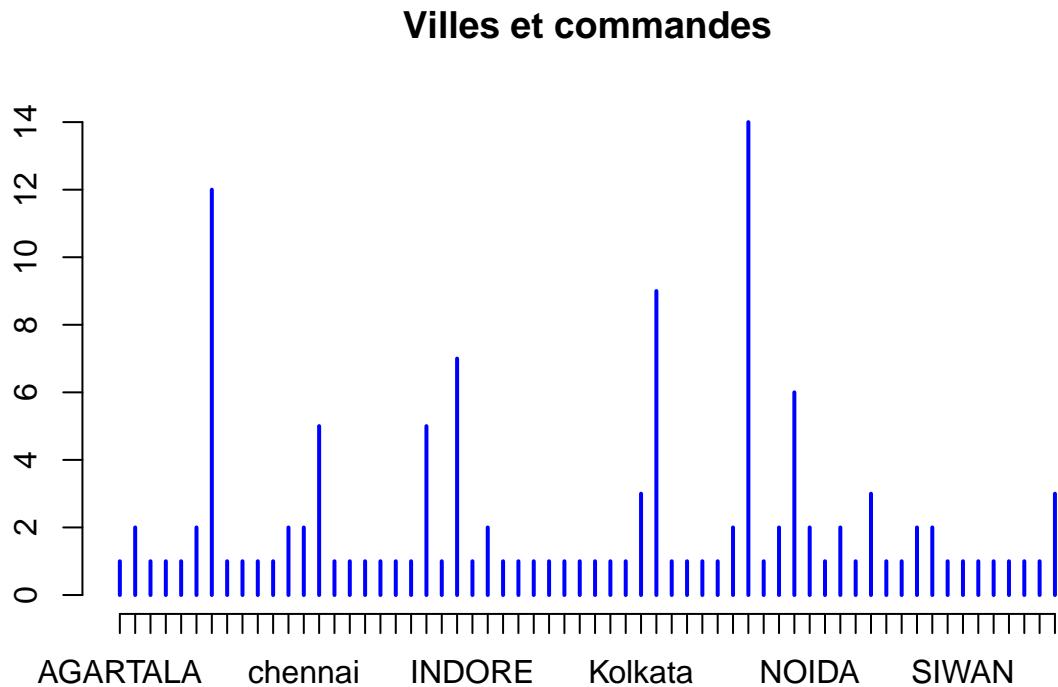
```
sort(prop.table(table(AMZN2$ship_city)))
```

```

##          AGARTALA          ALLAHABAD          AMROHA      Bardez
## 0.007692308          0.007692308 0.007692308 0.007692308
##          Bhubaneswar          BIDHAN NAGAR          BILIMORA    BURDWAN
## 0.007692308          0.007692308 0.007692308 0.007692308
##          COONOOR            CUTTACK            DEHRADUN     Durg
## 0.007692308          0.007692308 0.007692308 0.007692308
##          FARIDABAD          GHAZIABAD          GUWAHATI      INDORE
## 0.007692308          0.007692308 0.007692308 0.007692308
##          JAIPUR            JALANDHAR          JALESWAR     JAMMU
## 0.007692308          0.007692308 0.007692308 0.007692308
##          JODHPUR            KANPUR            KARAIKKUDI   KATWA
## 0.007692308          0.007692308 0.007692308 0.007692308
##          Kodambakkam Chennai          KORBA          LUCKNOW      MAHALINGPUR
## 0.007692308          0.007692308 0.007692308 0.007692308
##          MALDA          MUMBAI 400 026          PALAI        Pune
## 0.007692308          0.007692308 0.007692308 0.007692308
##          PUNEpune          PUNJAB            SILCHAR      SIWAN
## 0.007692308          0.007692308 0.007692308 0.007692308
##          Solan              Surat          THAMARASSERY Thane District
## 0.007692308          0.007692308 0.007692308 0.007692308
##          VADODARA          AHMEDABAD          BAREILLY     CHANDIGARH
## 0.007692308          0.015384615 0.015384615 0.015384615
##          chennai            JAGDALPUR          Mumbai      NAVI MUMBAI
## 0.015384615          0.015384615 0.015384615 0.015384615
##          NOIDA            PASIGHAT            SALEM        SECUNDERABAD
## 0.015384615          0.015384615 0.015384615 0.015384615
##          Kolkata            PUNE          Visakhapatnam CHENNAI
## 0.023076923          0.023076923 0.023076923 0.038461538
##          GURUGRAM          NEW DELHI          HYDERABAD      KOLKATA
## 0.038461538          0.046153846 0.053846154 0.069230769
##          BENGALURU          MUMBAI
## 0.092307692          0.107692308

```

```
plot(table(AMZN2$ship_city), col="blue", main="Villes et commandes", ylab="")
```



La ville où le nombre d'envois est le plus élevé est Mumbai (10%), suivie de Bengaluru (9%) puis de Kolkata (6%).

*Problème de sortie sur Rmarkdown, toutefois voici les formules adéquates pour les prochaines questions*

#### Transformation de la variable `order_date` en format date

```
v2 <- stringr::str_sub(AMZN2$order_date, 6, 26) AMZN2$order_date <- lubridate::parse_date_time(v2, '%d%m%y, %H:%M') summary(order_date)
```

#### Détermination des frais d'envoi moyen par année

```
aggregate(AMZN2$shipping_fee, by = list(lubridate :: year(AMZN2$order_date)), FUN=mean)
```

On constate qu'en moyenne les frais d'envoi pour l'année 2021 sont plus élevés que ceux de 2022. Ils sont respectivement d'environ 86. 88 et 80.50 On a vu précédemment que les frais d'envoi étaient moyennement corrélés aux montant de la commande. Pour autant, pouvons-nous tirer des conclusions ?...

Daniella Lowa

# Apprendre la manipulation de données avec R par la pratique - Session 1

Ibrahima Diallo

2022-05-07

- 1 Étape 1 : Expliquons en quoi consiste une manipulation de données.
- 2 Étape 2 : Expliquons le contexte des données.
- 3 Étape 3 : Décrivons nos données (type de variables, statistiques descriptives, données aberrantes, présence de doublons).
  - 3.1 Description type de variables
  - 3.2 Statistiques descriptives
    - 3.2.1 Tableau-synthèse
  - 3.3 Détection de données aberrantes
  - 3.4 Teste de présence de doublons
- 4 Étape 4 : Réalisons quelques visualisations simples de nos données: diagramme en boite, diagramme en barre, nuage de points, nuage de mots.
  - 4.0.1 Tableau de fréquences
- 5 Modèle de régression logit-binomial.
  - 5.1 Estimation des déterminants de performance de l'étudiant
  - 5.2 Estimation des risques relatifs ou rapport des chances (Odds ratios)
  - 5.3 Estimation des effets marginaux
  - 5.4 Test de corrélation : Corrélogramme
  - 5.5 Tests de significativité et de la qualité du modèle
    - 5.5.1 Test de significativité
    - 5.5.2 Test de la Qualité de l'estimation

## Semaine 1 à 2:

## 1 Étape 1 : Expliquons en quoi consiste une manipulation de données.

Une manipulation de données consiste tout simplement à utiliser les variables d'une base données qui à travers, des calculs minutieuses ou des représentations graphiques sur ces dernières de façon univariée, bivariée ou multivariée on en déduit des informations pertinentes permettant de répondre à une problématique bien donnée .

## 2 Étape 2 : Expliquons le contexte des données.

Nous vivons dans un monde dans lequel les ressources économiques se rarifient de plus en plus, incitant les entreprises à trouver des alternatives innovantes pour maximiser leurs profits en cherchant à bien se positionner sur le marché des biens et des services et l'une des meilleures alternatives qui s'offre à elles présentement est celle offerte par les nouvelles technologies de l'information et de la communication. A travers certains procédés, ces nouvelles technologies permettent à celles-ci d'enregistrer toutes les informations en lien avec leurs activités d'approvisionnements, de productions et de commercialisations dans une base de données unique. En effet, la disponibilité de cette base de données riche en informations devient de plus en plus incontournable pour permettre à ces dernières de mettre en place des plans stratégiques pour optimiser leurs coûts de production ou minimiser l'ensemble des contraintes pouvant ralentir la bonne marche de l'entreprise et par ricochet augmenter leurs performances.

## 3 Étape 3 : Décrivons nos données (type de variables, statistiques descriptives, données aberrantes, présence de doublons).

### 3.1 Description type de variables

```
setwd("C:/Users/lenovo/Documents/R pour scientifique sémaine1/formation en R Awa Diop ULaval/Apprendre par la pratique Statistics&Coding/data analysis LSTP session 1")
```

```
library(tidyverse)
library(haven)
library(readr)
library(readxl)
library(openxlsx)
library(labelled)
library(ggplot2)
```

```
data_student_perfmnce <- read.csv(file = "StudentsPerformance.csv", header = TRUE, sep=",")
data_student_perfmnce$Sum_Score <- rowSums(data_student_perfmnce[,c("math.score","reading.score","writing.score")], na.rm=TRUE)
data_student_perfmnce$Mean_Score <- rowMeans(data_student_perfmnce[,c("math.score","reading.score","writing.score")], na.rm=TRUE)
str(data_student_perfmnce)
```

```
## 'data.frame': 1000 obs. of 10 variables:
## $ gender : chr "female" "female" "female" "male" ...
## $ race.ethnicity : chr "group B" "group C" "group B" "group A" ...
## $ parental.level.of.education: chr "bachelor's degree" "some college" "master's degree" "associate's degree" ...
## $ lunch : chr "standard" "standard" "standard" "free/reduced" ...
## $ test.preparation.course : chr "none" "completed" "none" "none" ...
## $ math.score : int 72 69 90 47 76 71 88 40 64 38 ...
## $ reading.score : int 72 90 95 57 78 83 95 43 64 60 ...
## $ writing.score : int 74 88 93 44 75 78 92 39 67 50 ...
## $ Sum_Score : num 218 247 278 148 229 232 275 122 195 148 ...
## $ Mean_Score : num 72.7 82.3 92.7 49.3 76.3 ...
```

```
genre<- as.factor(data_student_perfmnce$gender)
ethnicity<-as.factor(data_student_perfmnce$race.ethnicity)
parental_levels_education <-as.factor(data_student_perfmnce$parental.level.of.education)
luncheon <-as.factor(data_student_perfmnce$lunch)
test_preparation <-as.factor(data_student_perfmnce$test.preparation.course)
```

```
mean(data_student_perfmnce$Mean_Score)
```

```
## [1] 67.77067
```

```
data_student_perfmnce$Mean_Score_Bin <- factor(ifelse(data_student_perfmnce$Mean_Score >= 67.77067, "Performant", "No_performant"))
Mean_Score_Binair<- as.factor(data_student_perfmnce$Mean_Score_Bin)
Mean_Scor <- data_student_perfmnce$Mean_Score
```

```
library(questionr)
freq(Mean_Score_Binair)
```

```
##          n    % val%
## No_performant 474 47.4 47.4
## Performant     526 52.6 52.6
```

Cette base contenant l'ensemble des informations susceptibles d'expliquer la performance des étudiants en mathématique est constituée d'un échantillon de 1000 observations et 8 variables, dont cinq de types chaînes de caractère et trois de type entier et ne révèle l'existence pas l'existence de données manquantes .

## 3.2 Statistiques descriptives

```
library(summarytools)
st_options(plain.ascii = FALSE,
           style = "rmarkdown",
           dfSummary.varnumbers = FALSE,
           dfSummary.valid.col = FALSE,
           lang ="fr")
```

```
dfSummary(data_student_perfmnce, plain.ascii = FALSE, style = "grid",
          graph.magnif = 0.75, valid.col = FALSE, tmp.img.dir = "/tmp")
```

```
## temporary images written to 'C:\tmp'
```

### 3.2.1 Tableau-synthèse

#### 3.2.1.1 data\_student\_perfmnce

**Dimensions:** 1000 x 11

**Doublons:** 0

Variable	Stats / valeurs	Fréq. (% de valide)	Diagramme	Manquant
gender [character]	1. female 2. male	518 (51.8%) 482 (48.2%)		0 (0.0%)
race.ethnicity [character]	1. group A 2. group B 3. group C 4. group D 5. group E	89 (8.9%) 190 (19.0%) 319 (31.9%) 262 (26.2%) 140 (14.0%)		0 (0.0%)
parental.level.of.education [character]	1. associate's degree 2. bachelor's degree 3. high school 4. master's degree 5. some college 6. some high school	222 (22.2%) 118 (11.8%) 196 (19.6%) 59 (5.9%) 226 (22.6%) 179 (17.9%)		0 (0.0%)
lunch [character]	1. free/reduced 2. standard	355 (35.5%) 645 (64.5%)		0 (0.0%)
test.preparation.course [character]	1. completed 2. none	358 (35.8%) 642 (64.2%)		0 (0.0%)
math.score [integer]	Moy (é-t) : 66.1 (15.2) min < med < max: 0 < 66 < 100 ÉIQ (CV) : 20 (0.2)	81 valeurs uniques		0 (0.0%)
reading.score [integer]	Moy (é-t) : 69.2 (14.6) min < med < max: 17 < 70 < 100 ÉIQ (CV) : 20 (0.2)	72 valeurs uniques		0 (0.0%)
writing.score [integer]	Moy (é-t) : 68.1 (15.2) min < med < max: 10 < 69 < 100 ÉIQ (CV) : 21.2 (0.2)	77 valeurs uniques		0 (0.0%)
Sum_Score [numeric]	Moy (é-t) : 203.3 (42.8) min < med < max: 27 < 205 < 300 ÉIQ (CV) : 58 (0.2)	194 valeurs uniques		0 (0.0%)
Mean_Score [numeric]	Moy (é-t) : 67.8 (14.3) min < med < max: 9 < 68.3 < 100 ÉIQ (CV) : 19.3 (0.2)	194 valeurs uniques		0 (0.0%)
Mean_Score_Bin [factor]	1. No_performant 2. Performant	474 (47.4%) 526 (52.6%)		0 (0.0%)

### 3.3 Détection de données aberrantes

Le test statistique de Grubbs.

Le **test statistique de Grubbs** permet de **tester si la valeur la plus faible est un outlier**. Son hypothèse nulle spécifie que la valeur la plus élevée n'est pas un outlier, alors que son hypothèse alternative spécifie que la valeur la plus élevée est un outlier. Si la p-value du test est inférieure au seuil de significativité choisi (en général 0.05) alors on concluera que la valeur la plus élevée est outlier. Pour réaliser ce test avec R, on utilise la fonction **grubbs.test()** du package “outliers” :

```
library(outliers)
grubbs.test(Mean_Scor)
```

```
Grubbs test for one outlier
```

```
data: Mean_Scor G = 4.12214, U = 0.98297, p-value = 0.01746 alternative hypothesis: lowest value 9 is an outlier
```

Ici, on constate que la p-value est inférieure au seuil de significativité donc, on rejette l'hypothèse nulle, ainsi on peut conclure que la valeur la plus élevée de la moyenne des scores de performance est une donnée aberrante

Il est possible de tester le caractère outlier de la valeur la plus élevée en utilisant l'argument **opposite=TRUE** :

```
grubbs.test(Mean_Scor, opposite = TRUE)
```

```
Grubbs test for one outlier
```

```
data: Mean_Scor G = 2.26055, U = 0.99488, p-value = 1 alternative hypothesis: highest value 100 is an outlier
```

Etant donné que la p-value est supérieure au seuil de significativité donc, on peut dire que la valeur la plus élevée de la moyenne des scores de performance n'est pas une donnée aberrante.

extension du test de grubbs. En fait ce test est itératif. par exemple si la plus grande valeur est aberrante, il faut continuer à test si la valeur qui vient avant cette valeur est aussi aberrante.

### 3.4 Teste de présence de doublons

```
sum(duplicated(data_student_perfmnce))
```

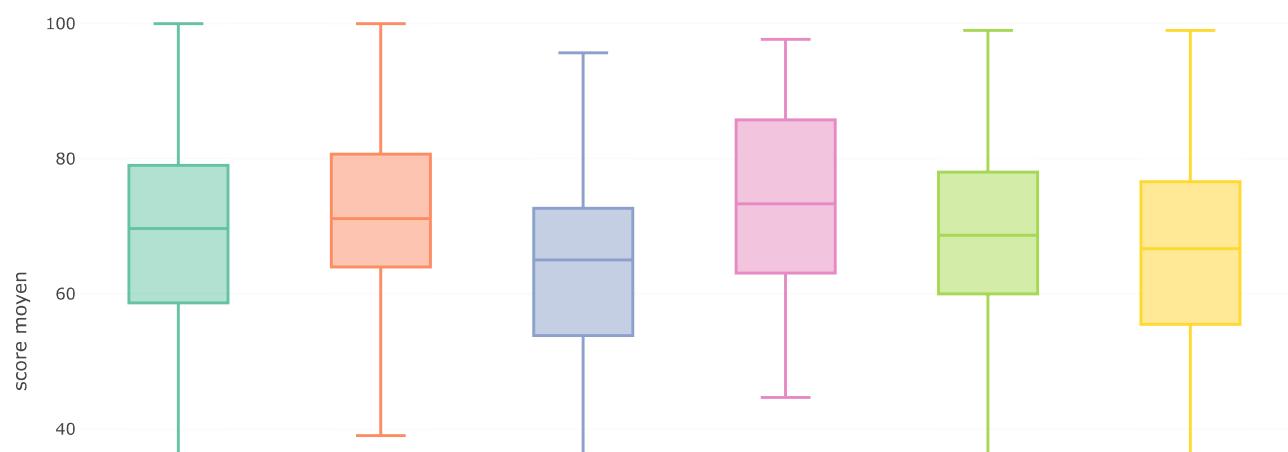
```
[1] 0
```

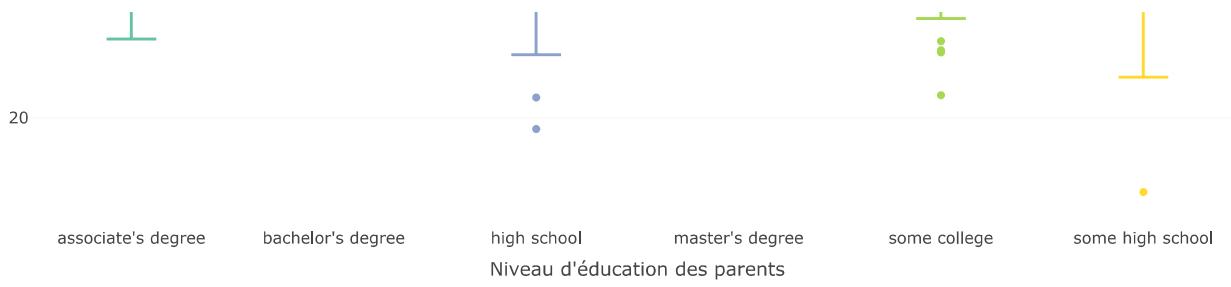
On remarque qu'aucune des variables ne présente de doublons mais également le tableau de synthèse des statistiques descriptives confirme la même chose.

## 4 Étape 4 : Réalisons quelques visualisations simples de nos données: diagramme en boîte, diagramme en barre, nuage de points, nuage de mots.

```
library(plotlyGeoAssets)
library(stats)
library(MASS)
library(car)
library(plotly)
p <- plot_ly(ggplot2::diamonds, y = ~Mean_Scor,
             color = ~ parental_levels_education,
             type = "box")%>%layout( title= "Figure N°1: Répartition de la moyenne des scores de l'étudiant sur les trois matières selon le niveau d'éducation des parents",
             xaxis=list(title="Niveau d'éducation des parents"),
             yaxis=list(title="score moyen"))
p
```

Figure N°1: Répartition de la moyenne des scores de l'étudiant sur les trois matières selon le niveau d'éducatio

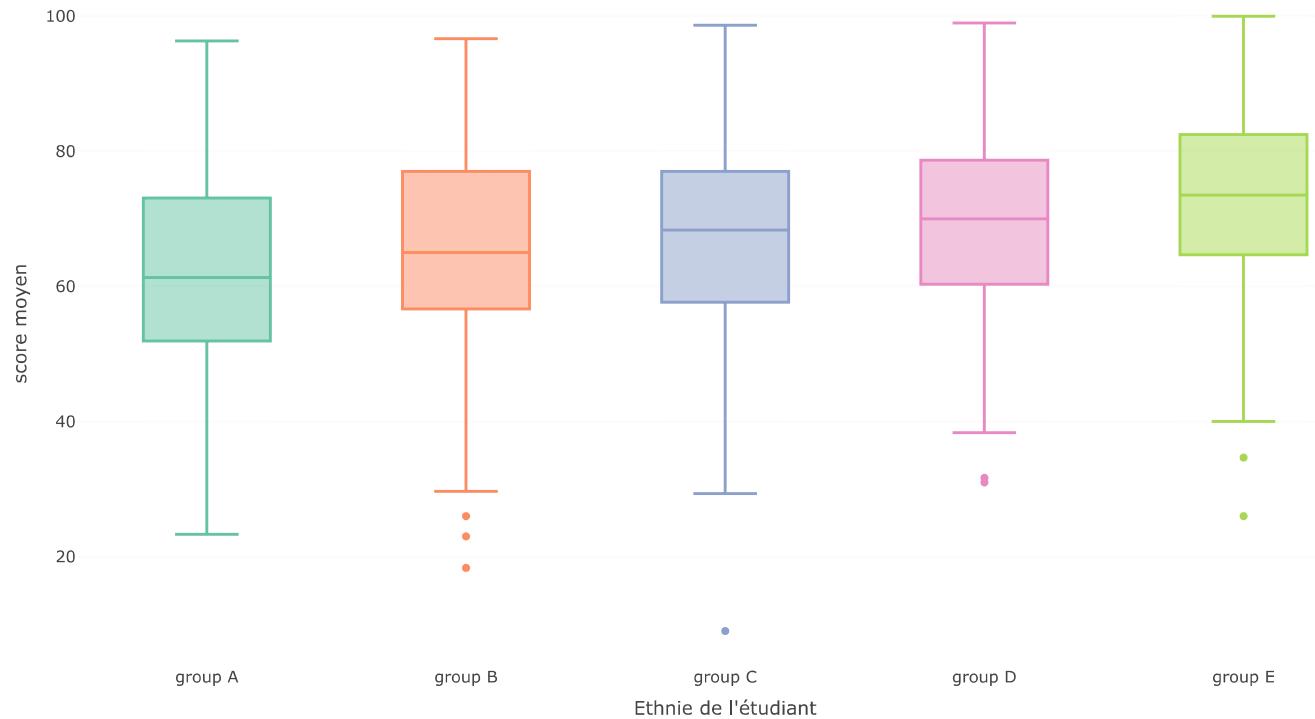




On remarque que les étudiants qui ont au moins le niveau Bachelor enregistrent en moyenne des scores de performance relativement plus élevés que ceux qui ont des parents des classes inférieures, cet écart pourrait être expliqué sans doute par le fait que les parents avec au moins un niveau bachelor ont plus de chance d'avoir des emplois avec des revenus qui leurs permettent payer des cours particuliers à leurs enfants .

```
p <- plot_ly(ggplot2::diamonds, y = ~Mean_Scor,
             color = ~ ethnicity,
             type = "box")%>%layout( title= "Figure N°2: Répartition de la moyenne des scores de l'étudiant sur les trois matières selon l'appartenance ethnique de l'étudiant",
             xaxis=list(title="Ethnie de l'étudiant"),
             yaxis=list(title="score moyen"))
p
```

Figure N°2: Répartition de la moyenne des scores de l'étudiant sur les trois matières selon l'appartenance ethnique de l'étudiant



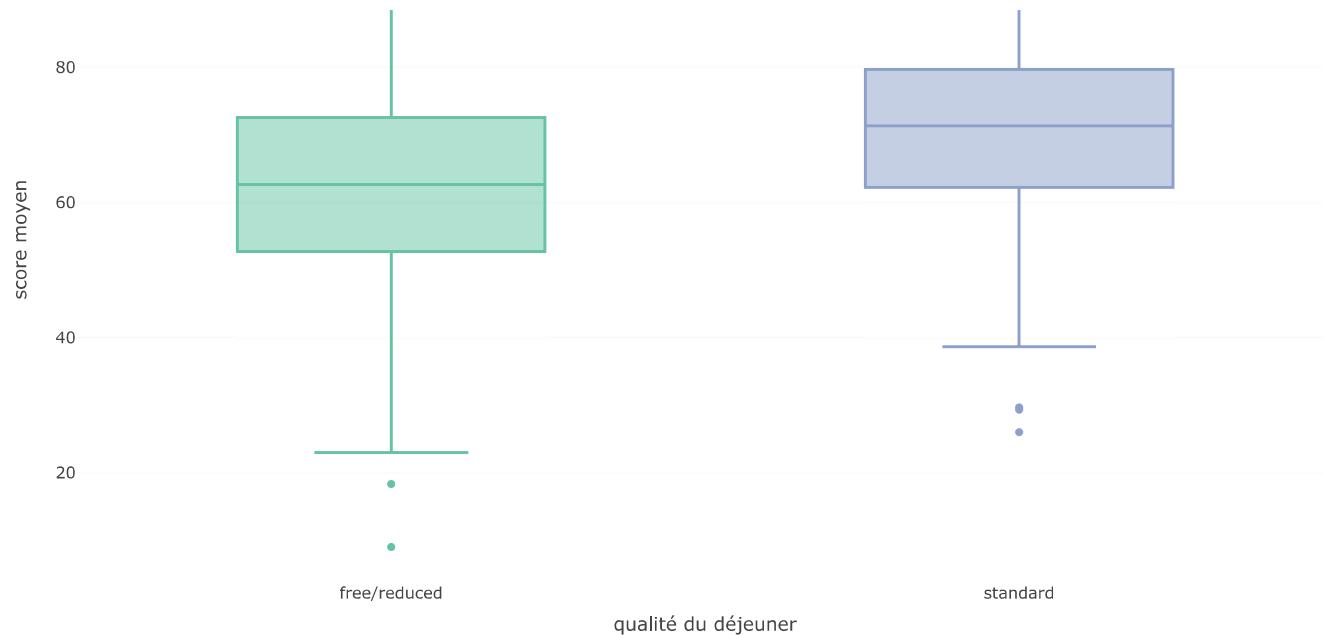
Ici, on note une inégalité croissante en termes de score de performance obtenu par les étudiants entre ceux qui sont membres de l'ethnie A jusqu'à ceux de l'ethnie E. Etant donné que nous ne disposons pas de l'ensemble des facteurs pouvant expliquer les particularités de chaque groupe d'ethnie, on est donc dans l'incapacité de formuler des hypothèses sur des éventuelles causes les disparités observées entre ces derniers .

```
library(plotly)

p <- plot_ly(ggplot2::diamonds, y= ~Mean_Scor ,
             color = ~ luncheon,
             type = "box")%>%layout(title= "Figure N°3: Répartition de la moyenne des scores sur les trois matières en fonction de la catégorie des qualités du déjeuner dont l'étudiant a pu s'offrir",
             xaxis=list(title="qualité du déjeuner"),
             yaxis=list(title="score moyen"))
p
```

Figure N°3: Répartition de la moyenne des scores sur les trois matières en fonction de la catégorie des qualités du déjeuner



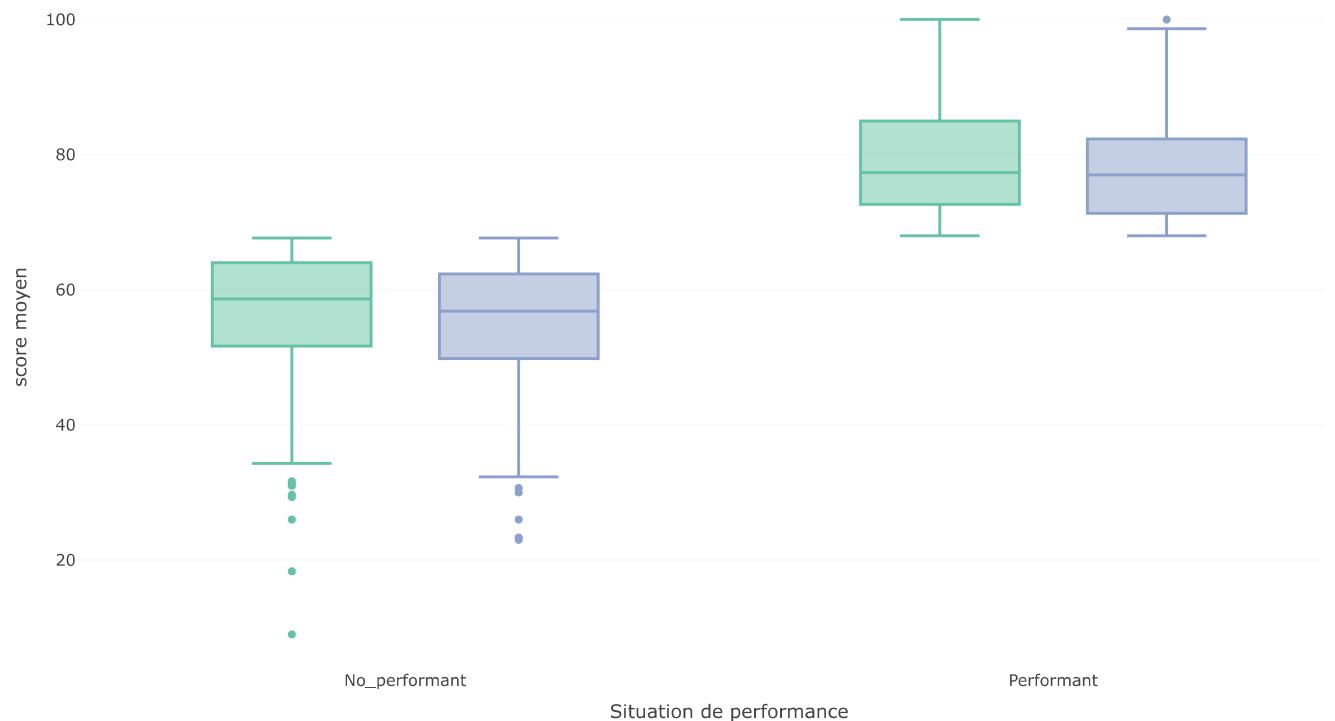


On note que les étudiants qui bénéficient d'un déjeuner de qualité standard sont plus performant que les autres, cette situation pourrait être liée par les différences d'apport calorifiques entre ces déjeuners de qualité différentes. En effet, il est assez trivial que les étudiants qui ont des déjeuners avec des pris standards ont plus de force pour bien suivre leurs cours que ceux qui ont des déjeuners de qualités inférieures car l'apport énergétique est moins important par rapport à celui procuré par les déjeuners de qualité standard .

```
fig <- plot_ly(ggplot2::diamonds, x= ~Mean_Score_Binair, y = ~Mean_Scor, color = ~genre, type = "box")
fig <- fig %>% layout(boxmode = "group", title= "Figure N°4: Répartition de la moyenne des scores sur les trois matières selon le genre et la nature de performance de l'étudiant",
                      xaxis=list(title="Situation de performance"),
                      yaxis=list(title="score moyen"))

fig
```

Figure N°4: Répartition de la moyenne des scores sur les trois matières selon le genre et la nature de performance

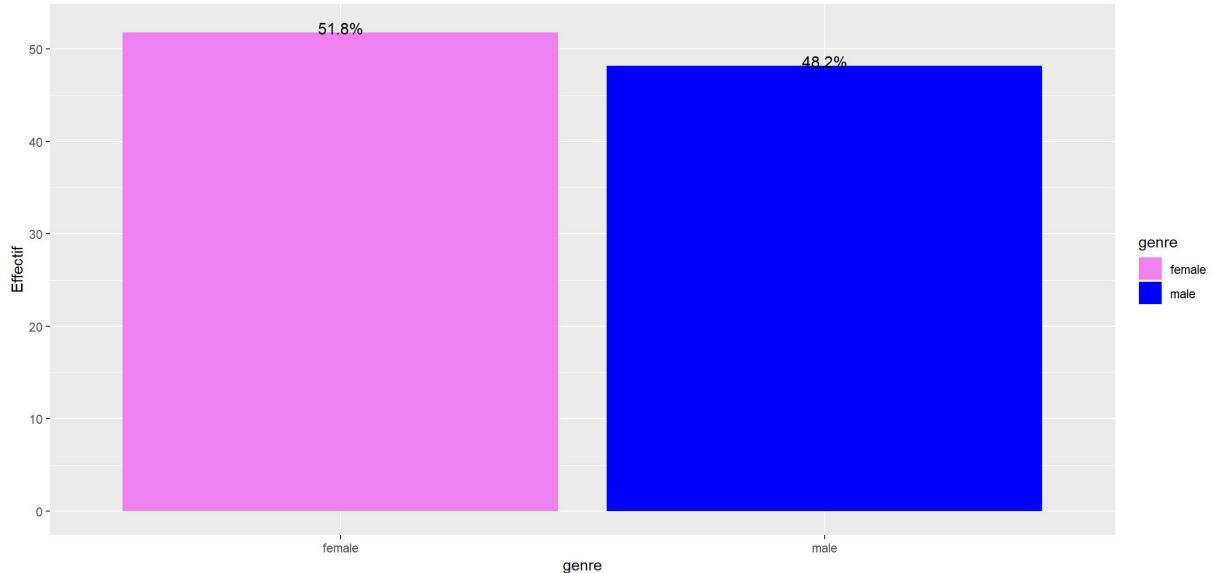


```

ggplot(data = data_student_perfmnce, aes(x = genre, y= Mean_Scor, fill = genre)) +
  geom_bar(aes(y = prop.table(..count..) * 100),
           position = "dodge") +
  geom_text(aes(y = prop.table(..count..) * 100 + 0.5,
                label = paste0(round(prop.table(..count..),3 )* 100, '%')),
            stat = 'count',
            position = position_dodge(.9),
            size = 4) + ggtitle("Figure N°5: Répartition de la moyenne des scores en fonction du genre de l'étudiant ") +
  labs(x= 'genre', y= 'Effectif') + scale_fill_manual(values = c("violet", "blue"))

```

Figure N°5: Répartition de la moyenne des scores en fonction du genre de l'étudiant



```

library(questionr)
freq(genre)

```

## 4.0.1 Tableau de fréquences

### 4.0.1.1 genre

Type: Facteur

	Fréq.	% Valide	% Valide cum.	% Total	% Total cum.
female	518	51.80	51.80	51.80	51.80
male	482	48.20	100.00	48.20	100.00
<NA>	0			0.00	100.00
Total	1000	100.00	100.00	100.00	100.00

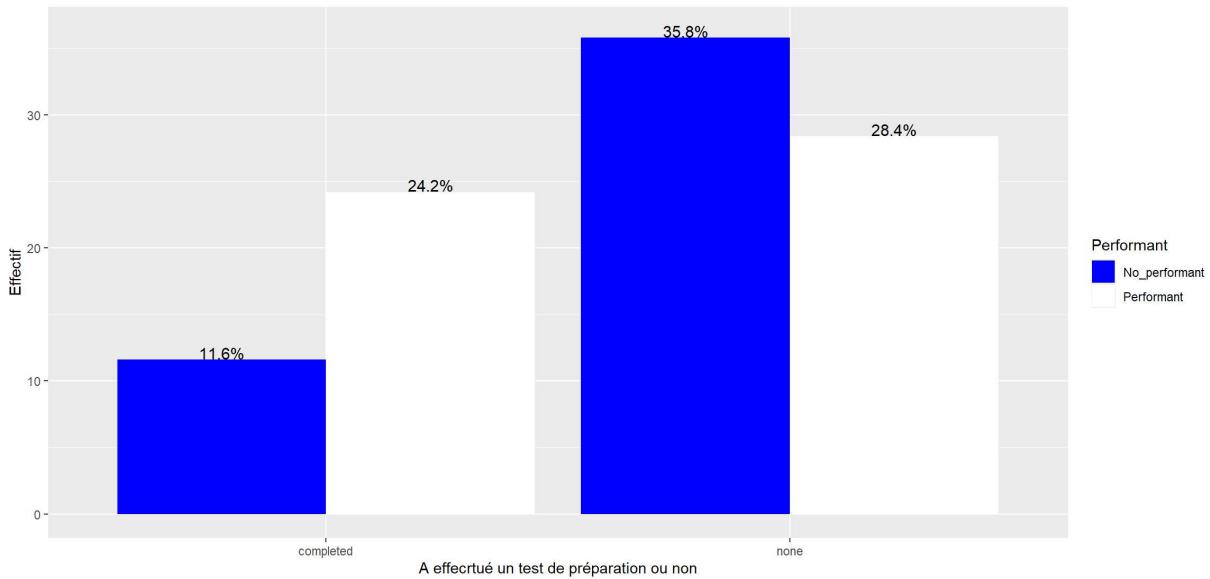
Le graphique N°6, nous montre que les étudiants de sexe féminin ont des moyennes plus élevées que ceux du sexe opposé, cependant si l'on analyse le graphique N°5, on note une faible disparité vis-à-vis du genre parmi les apprenants considérés comme plus performants, c'est-à-dire ce qui ont un score supérieur à la moyenne (67.77). Donc, on peut supposer que la supériorité de la moyenne des scores des étudiantes par rapport à celles des étudiants pourraient être expliquée par la prédominance du genre féminin dans l'échantillon. D'ailleurs, si l'on prend les étudiants les plus performants, on remarque que le score médian chez les filles (77.3333) est presque égal à celui des garçons (77), donc une supériorité de 0.3333, qui à notre égard n'est pas assez pour dire que c'est la gent féminine qui domine.

```

ggplot(data = data_student_perfmnce, aes(x = test_preparation, fill = data_student_perfmnce$Mean_Score_Bin)) +
  geom_bar(aes(y = prop.table(..count..) * 100),
           position = "dodge") +
  geom_text(aes(y = prop.table(..count..) * 100 + 0.5,
                label = paste0(round(prop.table(..count..), 3 )* 100, '%')),
            stat = 'count',
            position = position_dodge(.9),
            size = 4) + ggtitle("Figure N°6: Répartition de la situation de performance de l'étudiant selon que celui-ci ait subi un test de préparation ou non ") +
  labs(x = "A effectué un test de préparation ou non", y = 'Effectif', fill = 'Performant')+ scale_fill_manual(values = c("blue1", "white"))

```

Figure N°6: Répartition de la situation de performance de l'étudiant selon que celui-ci ait subi un test de préparation ou non



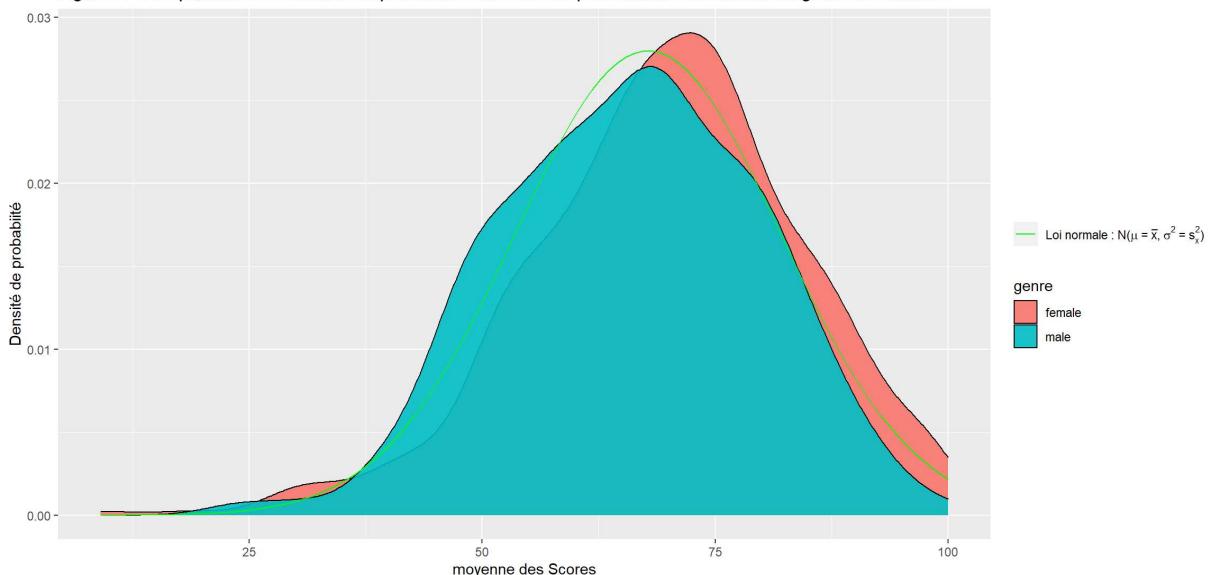
Le graphique ainsi obtenu nous montre que les étudiants qui ont subi des tests de préparations sont beaucoup plus performants que ceux qui ne l'ont pas fait. Toutefois, on constate, une supériorité de 4% en termes de performance des étudiants qui font partie de la modalité inverse. Ces deux observables nous renseignent dans une certaine mesure que le fait d'effectuer des tests de préparations augmente certes les chances de l'apprenants d'avoir de bonnes notes mais cela n'exclut en rien à ce qui n'ont pas faits de test de préparation d'avoir aussi de bonne de note. Donc, comme hypothèse, on peut supposer qu'il n'y a point une forte relation de cause à effet entre le fait d'avoir subi des tests de préparation et le fait d'avoir une bonne note mais néanmoins on peut dire que qu'il y a une forte corrélation positive entre ces deux variables .

```

moy <- mean(Mean_Scor)
eqrt <- sd(Mean_Scor)
ggplot(data = data_student_perfmnce) +
  geom_density(mapping = aes(x = Mean_Scor, fill = genre), alpha = 0.9) +
  stat_function(aes(colour = "c2"), fun = dnorm, args = list(mean = moy,
    sd = eqrt), alpha = 2) +
  scale_colour_manual(name = "",
    values = c(c2 = "green"),
    labels = c(c2 = expression(paste("Loi normale : N(", mu, " = ", bar(x), ", ",
      sigma^2, " = ", s[x]^2, ")")))) +
  ggtitle("Figure N°7: Répartition des densités de probabilités du score de performance en fonction du genre de l'étudiant ")
+ 
  labs(x= ' moyenne des Scores ', y= 'Densité de probabilité')

```

Figure N°7: Répartition des densités de probabilités du score de performance en fonction du genre de l'étudiant



La densité de probabilité de la moyenne des scores des étudiants est presque comparable à celle de la loi de Gauss, de ce fait, on peut en déduire que les données sur la moyenne des scores chez les garçons sont normalement distribuées. Cependant, la densité de probabilité obtenue chez les filles n'est pas comparable à celle de la loi normale nonobstant du fait qu'elle soit unimodal mais par ailleurs, la distribution des données est asymétrique vers la gauche et centrée à environ 68.75 et la plupart des données sont à peu près entre 44 et 100 alors que chez les garçons, elles sont comprises entre 37.5 et 94. Ces informations laissent entendre qu'il est plus probable pour le genre féminin d'avoir de meilleurs scores que le genre masculin .

## 5 Modèle de régression logit-binomial.

### 5.1 Estimation des déterminants de performance de l'étudiant

```
library(MASS)
library(car)
library(stats)
library(ordinal)
library(gtsummary)

reg_bin <- glm(Mean_Score_Binaire ~ genre + parental_levels_educ + test_prepar +
luncheon + ethnicity, data = data_student_perfmnce, family = binomial(link= "logit"))

tbl_regression(reg_bin)
```

Characteristic	log(OR) <sup>1</sup>	95% CI <sup>1</sup>	p-value
Sexe de l'élève			
female	—	—	
male	-0.53	-0.80, -0.25	<0.001
Niveau d'éducation des parents			
high school	—	—	
associate's degree	0.60	0.18, 1.0	0.005
bachelor's degree	0.84	0.34, 1.4	0.001
master's degree	1.0	0.34, 1.7	0.003
some college	0.45	0.03, 0.87	0.035
some high school	0.13	-0.31, 0.57	0.6
Test de préparation complété ou pas			
none	—	—	
completed	1.1	0.81, 1.4	<0.001
qualité du déjeuner			
free/reduced	—	—	
standard	1.1	0.79, 1.4	<0.001
l'éthnie auquelle appartient l'élève			
group A	—	—	
group B	0.27	-0.29, 0.84	0.4
group C	0.62	0.09, 1.2	0.022
group D	1.0	0.45, 1.5	<0.001
group E	1.3	0.68, 1.9	<0.001

<sup>1</sup> OR = Odds Ratio, CI = Confidence Interval

En considérant la variable genre, on remarque que le fait d'être un garçon a un effet négatif sur la chance de l'apprenant d'être plus performant par rapport au fait d'être une fille et cette estimation est hautement significative car la p-value correspondante est strictement inférieure au seuil de 1%. On observe les mêmes tendances en termes d'influence et de niveau de significativité de la modalité : "avoir complété un test de préparation" mais également de la modalité : "avoir un déjeuner de qualité standard" sur la performance des étudiants. Pour ce qui est du niveau d'éducation des parents, on note que, le fait pour l'étudiant que ses parents aient au moins un niveau d'éducation de niveau collégial influence positivement ses chances d'être plus performant mais sauf que cette prédiction n'est pas significative même au seuil de 10% pour la modalité "some high school". Quand il s'agit de l'appartenance ethnique de l'étudiant, on observe une influence positive sur le niveau de performance de l'apprenant et cela quel que soit le groupe d'ethnie considéré juste qu'on ne note aucune significativité pour ceux qui sont du groupe B .

## 5.2 Estimation des risques relatifs ou rapport des chances (Odds ratios)

```
library(gtsummary)
theme_gtsummary_language("fr", decimal.mark = ",", big.mark = " ")
tbl_regression(reg_bin, exponentiate = TRUE) %>% add_global_p(keep = TRUE)
```

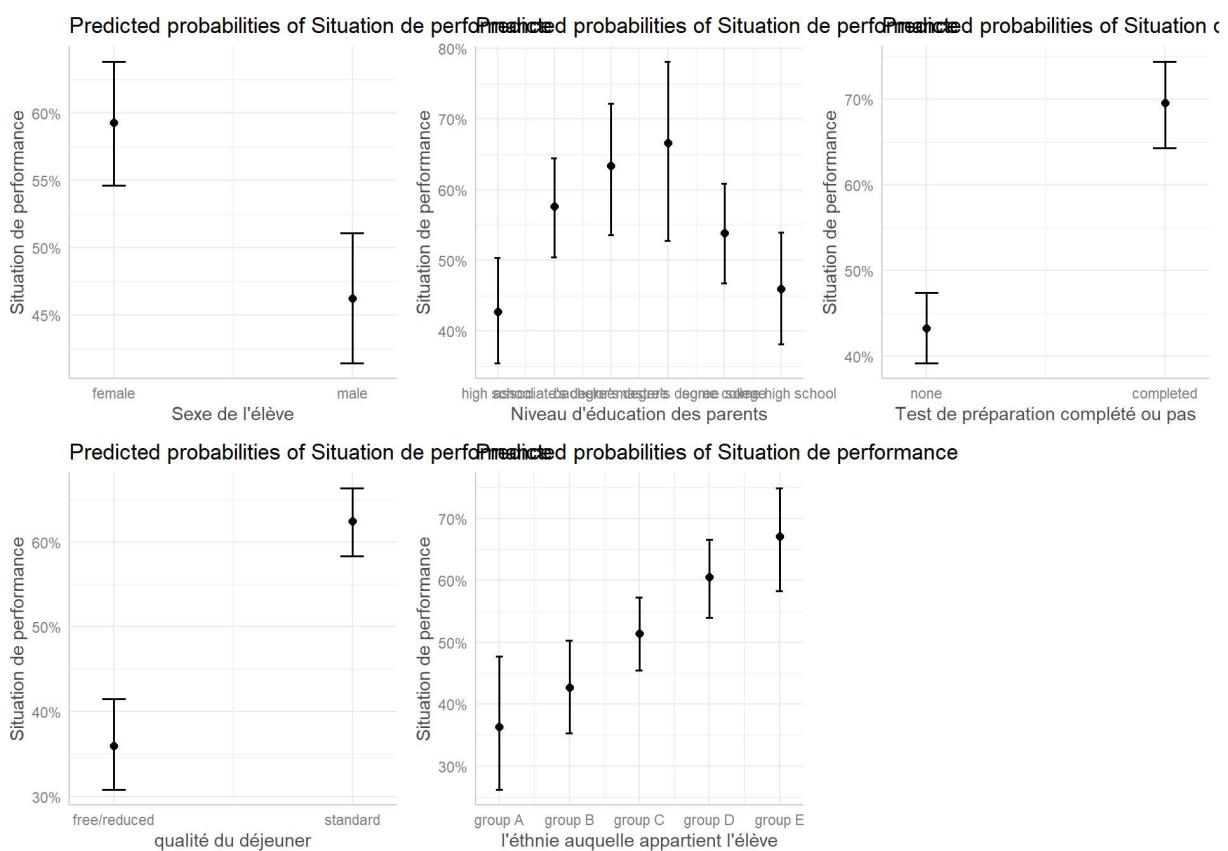
Caractéristique	OR <sup>1</sup>	95% IC <sup>1</sup>	p-valeur
Sexe de l'élève			<0,001
female	—	—	
male	0,59	0,45 – 0,78	<0,001
Niveau d'éducation des parents			0,001
high school	—	—	
associate's degree	1,82	1,20 – 2,78	0,005
bachelor's degree	2,32	1,40 – 3,87	0,001
master's degree	2,68	1,40 – 5,24	0,003
some college	1,57	1,03 – 2,38	0,035
some high school	1,14	0,73 – 1,78	0,6
Test de préparation complété ou pas			<0,001
none	—	—	
completed	3,00	2,24 – 4,03	<0,001
qualité du déjeuner			<0,001
free/reduced	—	—	
standard	2,95	2,21 – 3,96	<0,001
l'éthnie auquelle appartient l'élève			<0,001
group A	—	—	
group B	1,31	0,75 – 2,31	0,4
group C	1,86	1,10 – 3,18	0,022
group D	2,69	1,57 – 4,67	<0,001
group E	3,58	1,97 – 6,63	<0,001

<sup>1</sup> OR = rapport de cotes, IC = intervalle de confiance

L'ensemble des odds ratios obtenus confirme les signes (nature de l'influence) du signe des paramètres des modalités de chaque variable explicative par rapport à la variable endogène (performance de l'étudiant). En effet, on remarque que les chances de performance sont plus élevées chez les filles que chez les garçons et il y a deux fois plus de chance de performance lorsque les parents d'élève ont un niveau d'éducation de types licence et maîtrise que par rapport à la modalité de référence (high school). La même tendance est notée pour l'exogène "Test de préparation complété ou pas" avec un rapport des chances 3 fois plus élevé pour la catégorie "completed". Il en est de même pour la variable en lien avec la qualité du déjeuner (ici modalité standard) bénéficiée par l'apprenant. Et pour fermer la boucle, on observe que le rapport des chances augmente au fur et à mesure qu'on part du groupe ethnique A qui est la modalité de référence vers le groupe ethnique E avec une valeur trois fois plus importante pour ceux qui sont de cette catégorie par rapport au premier groupe .

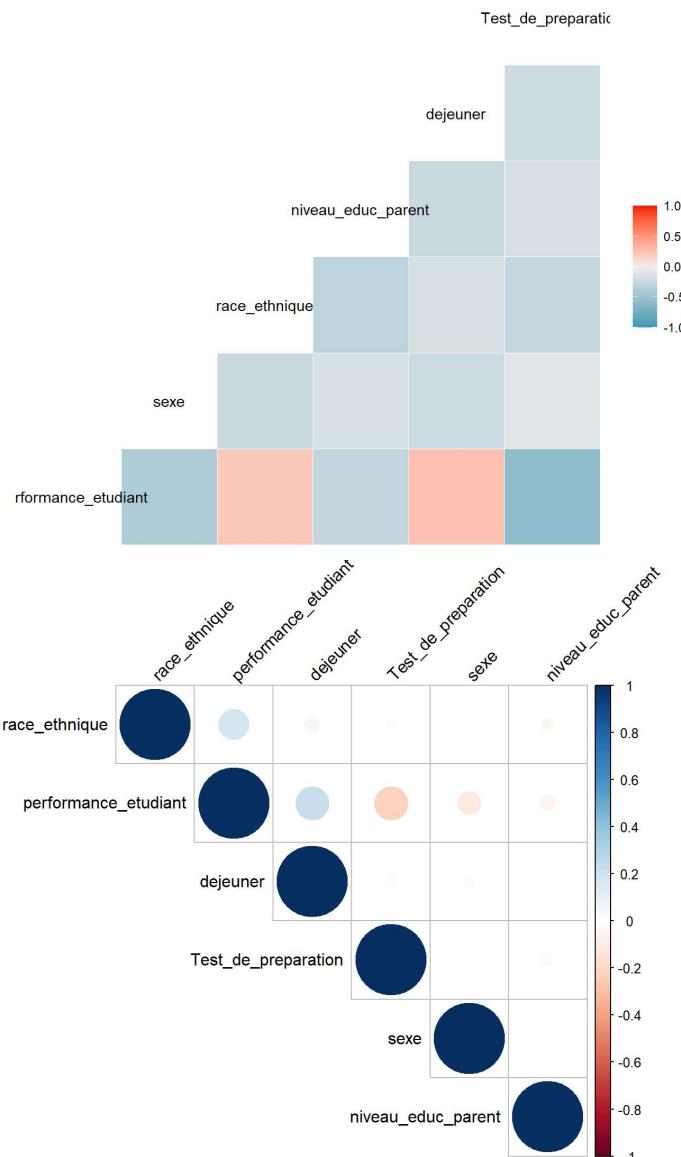
## 5.3 Estimation des effets marginaux

```
library(ggeffects)
cowplot::plot_grid(plotlist = plot(ggeffect(reg_bin)))
```



Ici, les effets marginaux (probabilité de prédiction) viennent en appuis ou confirmer les valeurs des risques relatifs enregistrés pour chaque variable indépendante.

## 5.4 Test de corrélation : Corrélogramme



## 5.5 Tests de significativité et de la qualité du modèle

### 5.5.1 Test de significativité

#### Test de Wald

```
waldtest(reg_bin)
```

Wald test

Model 1: Mean\_Score\_Binair ~ genre + parental\_levels\_educ + test\_prep + luncheon + ethnicity Model 2: Mean\_Score\_Binair ~ 1 Res.Df Df F Pr(>F)  
1 987  
2 999 -12 11.594 < 2.2e-16 \*\*\* — Signif. codes: 0 ‘‘ **0.001** ’’ 0.01 ’ 0.05 ‘ 0.1 ’’ 1

#### Test LR (Log\_Likelihood) Ratio de Vraisemblance

```
library(lmtest)
lrtest(reg_bin)
```

Likelihood ratio test

Model 1: Mean\_Score\_Binair ~ genre + parental\_levels\_educ + test\_prep + luncheon + ethnicity Model 2: Mean\_Score\_Binair ~ 1 #Df LogLik Df Chisq Pr(>Chisq)  
1 13 -604.00  
2 1 -691.79 -12 175.59 < 2.2e-16 \*\*\* — Signif. codes: 0 ‘‘ **0.001** ’’ 0.01 ’ 0.05 ‘ 0.1 ’’ 1

Si c'est seulement la valeur du log de vraisemblance qui nous intéresse, on peut utiliser le code suivant.

```
logLik(reg_bin)
```

```
'log Lik.' -604.0011 (df=13)
```

La valeur du test de Wald montre que le modèle est globalement significatif et celle du ratio de vraisemblance nous informe qu'au moins une des variables explicatives a une influence hautement significative sur la variable dépendante.

### 5.5.2 Test de la Qualité de l'estimation

Pour mesurer la qualité d'ajustement du modèle estimé, nous utilisons deux indicateurs classiques qui sont le pseudo R2 de McFadden et le taux d'erreur de prédiction.

#### calcul R2 McFadden pseudo R-squared

```
library(pscl)
pR2(reg_bin)
```

```
fitting null model for pseudo-r2 llh llhNull G2 McFadden r2ML r2CU -604.0010642 -691.7945706 175.5870128 0.1269069 0.1610356 0.2149082
```

On peut utiliser aussi la formule suivante Si c'est seulement la valeur du R2 de McFadden qu'on veut avoir.

```
reg_bin_0 <- glm(Mean_Score_Binair ~ 1,
                  data = data_student_perfmnce, family = binomial(link= "logit"))
R2<- (reg_bin_0$deviance - reg_bin$deviance)/reg_bin_0$deviance
R2
```

```
[1] 0.1269069
```

```
R2McF<- 1 - logLik(reg_bin)/logLik(reg_bin_0)
R2McF
```

```
'log Lik.' 0.1269069 (df=13)
```

Le Pseudo R2 de McFadden trouvé vaut 0.13. Ainsi, le pouvoir explicatif du modèle est de 13% ; ce qui signifie que, 13% de la probabilité qu'un étudiant fasse partie des plus performants est expliquée par les variables exogènes retenues.

#### Test du Taux d'erreur de prédiction

On convient que la qualité prédictive du modèle est mauvaise lorsque  $t > 0.5$ .

#### Matrice de confusion

```
mc <- table(Mean_Score_Binair, pred.mod)
print(mc)
```

```
pred.mod
```

```
Mean_Score_Binair 0 1 No_performant 301 173 Performant 146 380
```

```
t0 = (mc[1, 2] + mc[2, 1]) / sum(mc)
print(t0)
```

```
[1] 0.319
```

#### Autre méthode pour calculer la matrice de confusion

1	2	3	4	5	6
---	---	---	---	---	---

```
0.5928704 0.8072944 0.6270814 0.1489853 0.4518887 0.5337631 Mean_Score_Binair No_performant Performant FALSE 301 146 TRUE 173
380
```

```
mc1 <- table(Mean_Score_Binair$pred > 0.5, Mean_Score_Binair)
print(mc1)
```

```
Mean_Score_Binair
No_performant Performant
```

FALSE 301 146 TRUE 173 380

```
t1 = (mc1[1, 2] + mc1[2, 1]) / sum(mc1)
t1
```

[1] 0.319

Le taux d'erreur de prédiction est la proportion des modalités prédictives qui diffèrent des modalités observées (c'est aussi la somme des 2 valeurs non-diagonales de la matrice de confusion divisée par n). Plus le taux est proche de 0, meilleur est la qualité prédictive du modèle. Il est convenu que la qualité prédictive du modèle est mauvaise lorsque  $t > 0.5$ . En effet, d'après le résultat obtenu, on constate que ce taux est inférieur à 0.5, ce qui nous permet de dire que la qualité prédictive du modèle est très bonne. Ce taux qui est de 32%, veut dire en réalité, si l'on considère la probabilité d'être plus performant pour ces apprenants, elle a été mal prédictée pour 32% d'entre eux.

# Apprendre la manipulation de données avec R par la pratique - Session 1\_Partie\_II

Ibrahima Diallo

2022-05-18

- Fusion des données du dossier archive.
- Question 1 : Les pays exclus à la suite de cette fusion? .
- Question 2 : La quantité totale de pesticide utilisée en Amérique, en Afrique et en Australie? .
- Question 3 : La moyenne et médiane de rendement de céréales (hectogramme par hectare (Hg/Ha)) pour l'Amérique, l'Afrique et l'Australie? .
- Question 4 : Exploitation de la base de données portant sur les commandes Amazon.
  - Ville où le nombre d'envoie (shipping) est le plus élevé? .
  - Conversion de la variable order\_date en années (Recommandation : utilisez le package lubridate).
  - Calcul de la moyenne de frais d'envois par année (Recommandation : utilisez le package stringr).

## Semaine 3 à 4

```
setwd("C:/Users/lenovo/Documents/R pour scientifique sémaine1/formation en R Awa Diop ULaval/  
Apprendre par la pratique Statistics&Coding/data analysis LSTP session 1")  
  
library(tidyverse)  
library(haven)  
library(readr)  
library(readxl)  
library(openxlsx)  
library(labelled)
```

## Fusion des données du dossier archive.

```
data_cereal <- read.csv("CerealCropYield_1961-2018.csv", sep = ",", header = TRUE)  
data_ferti <- read.csv("FertilizerConsumption_1961-2018.csv", sep = ",", header = TRUE)  
data_pest <- read.csv("PesticideUsage_1990-2017.csv", sep = ",", header = TRUE)
```

```
library(dplyr)  
  
data_merg<- merge(data_cereal, data_ferti, by = "Country" )  
  
data_merge<- merge(data_merg, data_pest, by = "Country")  
  
glimpse(data_merge)
```

```

## Rows: 147
## Columns: 5
## $ Country           <chr> "Albania", "Algeria", "A...
## $ Yield..hg.ha.     <int> 1651049, 534659, 362959, ...
## $ Area.harvested..ha. <dbl> 14972582, 158067578, 646...
## $ FertilizerQuantity <dbl> 4316765, 8578144, 203069...
## $ Total.Pesticides.use.per.area.of.land..kg.ha. <dbl> 14.25, 12.88, 0.43, 98.1...

```

## Question 1 : Les pays exclus à la suite de cette fusion? .

```

## Rows: 202
## Columns: 1
## $ data_cereal.Country..in..data_merge.Country <lgl> FALSE, FALSE, TRUE, TRUE, ...

```

```

## Rows: 202
## Columns: 3
## $ Country           <chr> "Afghanistan", "Africa", "Albania", "Algeria", "Am...
## $ Yield..hg.ha.     <int> 808952, 692014, 1651049, 534659, 2135761, 362959, ...
## $ Area.harvested..ha. <dbl> 167842553, 4829674858, 14972582, 158067578, 750292...

```

```

library(dplyr)
country_deleted <- data.frame(anti_join(data_cereal, data_ferti, data_pest, by="Country"))
glimpse(country_deleted)

```

```

## Rows: 24
## Columns: 3
## $ Country           <chr> "Africa", "Americas", "Antigua and Barbuda", "Barb...
## $ Yield..hg.ha.     <int> 692014, 2135761, 959479, 1497173, 2032873, 1325172...
## $ Area.harvested..ha. <dbl> 4829674858, 7502921817, 1646, 25747, 16692803, 694...

```

Les pays qui sont exclus suite à la fusion des trois bases données sont au nombre de 24, dont : Africa, Americas, Antigua and Barbuda, Barbados, Belgium-Luxembourg, Central America, Cote d'Ivoire, Dominica, Ethiopia PDR, Grenada, Guadeloupe, Guam, Hong Kong, Maldives, Melanesia, Micronesia (country), Micronesia (region), Middle Africa, Montserrat, Reunion, Saint Lucia, Saint Vincent and the Grenadines, Sudan (former) et Western Sahara.

## Question 2 : La quantité totale de pesticide utilisée en Amérique, en Afrique et en Australie? .

Quantité totale en Australie

```
tot_pept<- data_pest$Total.Pesticides.use.per.area.of.land..kg.ha.  
county_pept<- data_pest$Country  
tot_perti_Australia <- sum(tot_pept[county_pept == "Australia"])  
tot_perti_Australia
```

```
## [1] 41.11
```

La quantité de pesticide utilisée en Australie est de 41.11 kg.ha, cependant cette information n'est pas disponible pour l'Afrique et l'Amérique.

## Question 3 : La moyenne et médiane de rendement de céréales (hectogramme par hectare (Hg/Ha)) pour l'Amérique, l'Afrique et l'Australie? .

```
##Moyenne et médiane en Afrique
```

```
county_cerea <- data_cereal$Country  
rend_afric <- data_cereal$Yield..hg.ha.[county_cerea== "Africa"]  
summary(rend_afric)
```

```
##      Min. 1st Qu. Median   Mean 3rd Qu.   Max.  
## 692014 692014 692014 692014 692014 692014
```

la moyenne et la médiane de rendement de céréales (hectogramme par hectare (Hg/Ha)) pour l'Afrique sont pareilles avec une valeur estimée à 692014.

```
##Moyenne et médiane en Amérique
```

```
rend_americ <- data_cereal$Yield..hg.ha.[county_cerea== "Americas"]  
summary(rend_americ)
```

```
##      Min. 1st Qu. Median   Mean 3rd Qu.   Max.  
## 2135761 2135761 2135761 2135761 2135761 2135761
```

la moyenne et la médiane de rendement de céréales (hectogramme par hectare (Hg/Ha)) pour l'Amérique sont identiques corrondant à 213576 .

```
##Moyenne et médiane en Australe
```

```
rend_austra <- data_cereal$Yield..hg.ha.[county_cerea== "Australia"]  
summary(rend_austra)
```

```
##      Min. 1st Qu. Median   Mean 3rd Qu.   Max.  
## 924103 924103 924103 924103 924103 924103
```

la moyenne et la médiane de rendement de céréales (hectogramme par hectare (Hg/Ha)) pour l'Australie aussi sont similaires avec une valeur estimée à 924103 .

## Question 4 : Exploitation de la base de données portant sur les commandes Amazon.

Importons tout d'abord la base de données

```
library(dplyr)  
library("readxl")  
  
data_amazon <- read_excel( "orders_data.xlsx")  
  
glimpse(data_amazon)
```

```
## Rows: 171  
## Columns: 12  
## $ order_no      <chr> "405-9763961-5211537", "404-3964908-7850720", "171-810318...  
## $ order_date    <chr> "Sun, 18 Jul, 2021, 10:38 pm IST", "Tue, 19 Oct, 2021, 6:...  
## $ buyer         <chr> "Mr.", "Minam", "yatipertin", "aciya", "Susmita", "Subini...  
## $ ship_city     <chr> "CHANDIGARH,", "PASIGHAT,", "PASIGHAT,", "DEVARAKONDA,", ...  
## $ ship_state    <chr> "CHANDIGARH", "ARUNACHAL PRADESH", "ARUNACHAL PRADESH", "...  
## $ sku           <chr> "SKU: 2X-3C0F-KNJE", "SKU: DN-0wdx-vyot", "SKU: DN-0wd...  
## $ description   <chr> "100% Leather Elephant Shaped Piggy Coin Bank | Block Pri...  
## $ quantity       <chr> "1", "1", "1", "1", "1", "1", "1", "1", "1", "1", "1", "1...  
## $ item_total    <chr> "₹449.00", "₹449.00", "₹449.00", NA, "₹1,099.00", "₹200.0...  
## $ shipping_fee  <chr> NA, "₹60.18", "₹60.18", NA, "₹84.96", NA, NA, "₹84.96", "...  
## $ cod            <chr> NA, NA, NA, "Cash On Delivery", NA, NA, "Cash On Delivery...  
## $ order_status   <chr> "Delivered to buyer", "Delivered to buyer", "Delivered to...
```

La base de données ainsi importée est constituée de 171 observations et 12 variables.

## Ville où le nombre d'envoie (shipping) est le plus élevé? .

```
names(which.max(rowSums(table(data_amazon$ship_city,data_amazon$quantity))))  
  
## [1] "MUMBAI,"
```

On constate que c'est à MUMBAI que le nombre d'envoie (shipping) est le plus élevé .

## Conversion de la variable order\_date en années (Recommandation : utilisez le package lubridate).

```
library("lubridate")
data_amazon$order_date_reo <- dmy_hm(data_amazon$order_date)
data_amazon$order_date_year <- year(data_amazon$order_date_reo)
View(data_amazon$order_date_year)
```

```
tail(data_amazon$order_date)
```

```
## [1] "Sat, 25 Dec, 2021, 4:03 pm IST"  "Mon, 13 Dec, 2021, 11:30 am IST"
## [3] "Wed, 1 Dec, 2021, 12:18 pm IST"   "Thu, 9 Dec, 2021, 6:55 pm IST"
## [5] "Wed, 23 Feb, 2022, 12:43 am IST"  "Sun, 26 Dec, 2021, 6:21 pm IST"
```

La commande **\*\*tail\*\*** permet de voir la façon dont la variable date a été ordonnée

## Calcul de la moyenne de frais d'envois par année (Recommandation : utilisez le package stringr). . .

Supprimons d'abord le caractère spécial (ici dévise) qui accompagne la valeur des frais d'envois

```
library(stringr)
data_amazon$item_total_su<- str_sub(data_amazon$item_total, 2, 6)
```

```
years<- factor(data_amazon$order_date_year, labels = c("2021" , "2022"))
```

```
data_amazon$item_total_su_num<- as.numeric(data_amazon$item_total_su)
```

```
moyennes3 <- by(data_amazon$item_total_su_num, years, FUN = mean , na.rm=TRUE)
moyennes3
```

```
## years: 2021
## [1] 428.1226
## -----
## years: 2022
## [1] 451.0769
```

la moyenne des frais d'envois pour l'année 2021 est de 428.1226 et 451.0769 pour l'année 2022 .

# Apprendre par la pratique - Session 1

Armel TINDO

07 mai 2022

## Table of Contents

Définition .....	1
Explication du contexte des données .....	1
Description des données (type de variables, statistiques descriptives, données aberrantes, présence de doublons) .....	1
Performance des étudiants .....	1
Tweets.....	4
Visualisation des données.....	5
Performance des étudiants .....	5
Sentiments des tweets .....	6

## PARTIE 1

### Définition

La **manipulation des données** est une série de processus qui consiste à extraire les données, les transformer (recodage, traitement des données manquantes...)

### Explication du contexte des données

Plusieurs jeux de données ont été utilisées dans cette session.

- **Commentaires sur Twitter** : il s'agit ici des commentaires de certains utilisateurs de Twitter qui ont permis de définir leurs sentiments.
- **Performance des étudiants** : ce jeu de données présente les notes des étudiants

### Description des données (type de variables, statistiques descriptives, données aberrantes, présence de doublons)

#### Performance des étudiants

```
str(data_perform)
```

```
## 'data.frame': 1000 obs. of 9 variables:  
## $ gender : Factor w/ 2 levels "female","male": 1 1 1  
2 2 1 1 2 2 1 ...  
## $ race.ethnicity : Factor w/ 5 levels "group A","group B",...  
2 3 2 1 3 2 2 2 4 2 ...  
## $ parental.level.of.education: Factor w/ 6 levels "associate's  
degree",...: 2 5 4 1 5 1 5 5 3 3 ...  
## $ lunch : Factor w/ 2 levels "free/reduced",...: 2 2  
2 1 2 2 2 1 1 1 ...  
## $ test.preparation.course : Factor w/ 2 levels "completed","none": 2 1  
2 2 2 2 1 2 1 2 ...  
## $ math.score : int 72 69 90 47 76 71 88 40 64 38 ...  
## $ reading.score : int 72 90 95 57 78 83 95 43 64 60 ...  
## $ writing.score : int 74 88 93 44 75 78 92 39 67 50 ...  
## $ moyenne : num 72.7 82.3 92.7 49.3 76.3 ...
```

- **Variables qualitatives**

```
data_perform %>%tbl_summary(include = c("gender", "race.ethnicity",  
"parental.level.of.education", "lunch", "test.preparation.course"), )
```

Characteristic	N = 1,000 <sup>1</sup>
gender	
female	518 (52%)
male	482 (48%)
race.ethnicity	
group A	89 (8.9%)
group B	190 (19%)
group C	319 (32%)
group D	262 (26%)
group E	140 (14%)
parental.level.of.education	
associate's degree	222 (22%)
bachelor's degree	118 (12%)
high school	196 (20%)
master's degree	59 (5.9%)
some college	226 (23%)
some high school	179 (18%)
lunch	
free/reduced	355 (36%)
standard	645 (64%)
test.preparation.course	
completed	358 (36%)
none	642 (64%)

<sup>1</sup>n (%)

- **Notes en maths, lecture et rédaction**

```
summary(data_perform[6:8])
```

```
##      math.score    reading.score    writing.score
##  Min.   :  0.00   Min.   :17.00   Min.   : 10.00
##  1st Qu.: 57.00   1st Qu.:59.00   1st Qu.: 57.75
##  Median : 66.00   Median :70.00   Median : 69.00
```

```

##  Mean    : 66.09   Mean    : 69.17   Mean    : 68.05
##  3rd Qu.: 77.00   3rd Qu.: 79.00   3rd Qu.: 79.00
##  Max.   :100.00   Max.   :100.00   Max.   :100.00

```

- **Moyenne générale**

```

summary(data_perform$moyenne)

##      Min. 1st Qu. Median     Mean 3rd Qu.     Max.
##      9.00  58.33  68.33  67.77  77.67 100.00

```

Le jeu de données sur les performances des étudiants contient les données de 1000 étudiants dont 52% de femmes contre 48% d'hommes. Ils sont répartis suivant l'ethnie, le niveau d'éducation de leurs parents, le type de déjeuner, le niveau de préparation de l'examen. Les notes sont celles obtenues en mathématiques, en lecture et en rédaction.

On note que plus de 50% des parents des étudiants ont au plus le niveau "high school". 64% des étudiants prennent un déjeuner standard contre 36% qui obtiennent un déjeuner gratuit ou subventionné. 64% des étudiants n'ont pas pu faire le test de préparation à l'examen.

Le premier de la classe a obtenu une moyenne de score 100 contre 9 pour le dernier.

```

summary(data_perform$moyenne)

##      Min. 1st Qu. Median     Mean 3rd Qu.     Max.
##      9.00  58.33  68.33  67.77  77.67 100.00

```

## Tweets

```

str(data_tweet)

## 'data.frame': 27481 obs. of 4 variables:
## $ textID       : chr "cb774db0d1" "549e992a42" "088c60f138" "9642c003ef"
...
## $ text         : chr "I`d have responded, if I were going" "Sooo SAD I will miss you here in San Diego!!!" "my boss is bullying me..." "what interview! leave me alone" ...
## $ selected_text: chr "I`d have responded, if I were going" "Sooo SAD" "bullying me" "leave me alone" ...
## $ sentiment    : Factor w/ 3 levels "negative","neutral",...: 2 1 1 1 1 2 3 2 2 3 ...

```

- **Sentiments des utilisateurs**

```

freq(data_tweet[4])

##          n   % val%
## negative 7781 28.3 28.3
## neutral 11118 40.5 40.5
## positive 8582 31.2 31.2

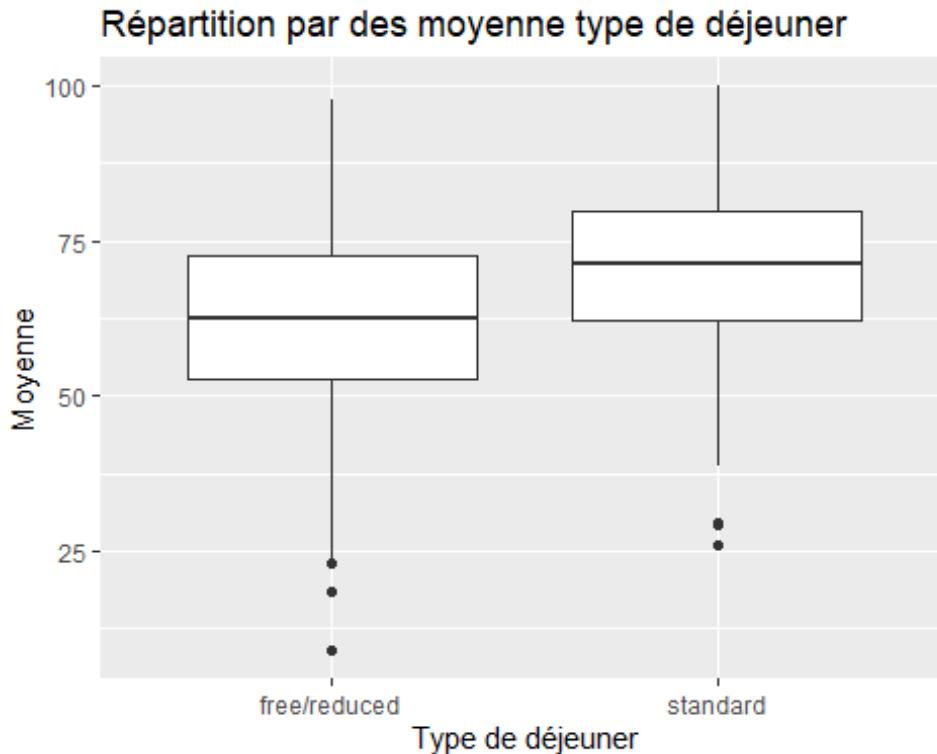
```

Les tweets sont relatifs à 27 481 utilisateurs de Twitter. Il ressort que 28,3% des commentaires sont négatifs, 31,2% sont positifs et 40,5% étaient des avis neutres.

## Visualisation des données

### Performance des étudiants

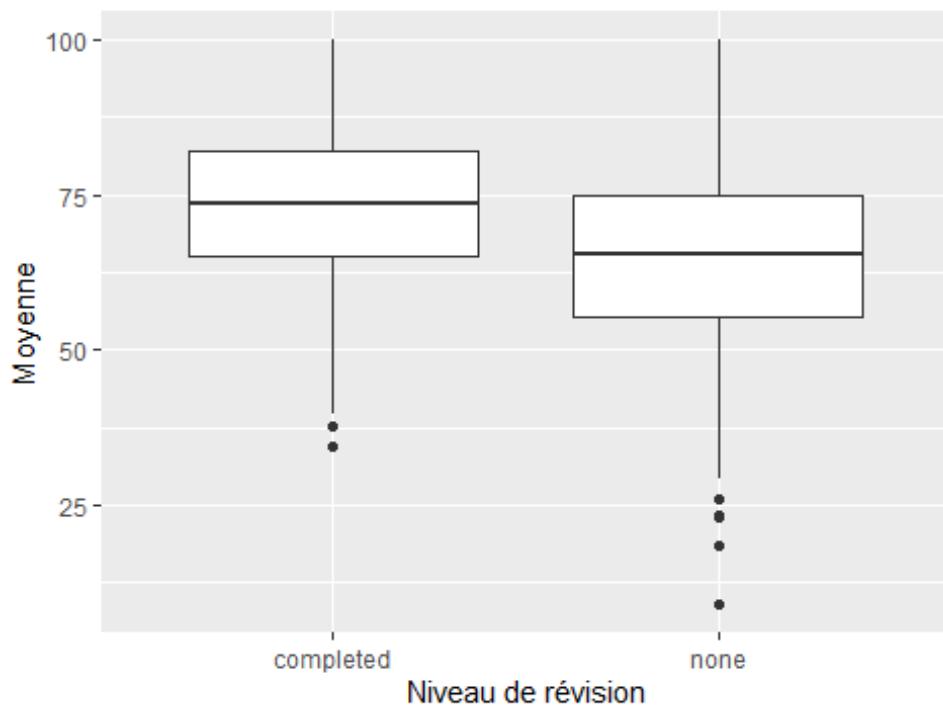
```
ggplot(data_perform) +  
  aes(x = lunch, y = moyenne) +  
  geom_boxplot() +  
  xlab("Type de déjeuner") +  
  ylab("Moyenne") +  
  ggtitle("Répartition par des moyenne type de déjeuner")
```



Les étudiants qui prennent un déjeuner standard ont une moyenne supérieure à ceux qui ne prennent pas de déjeuner ou ont un déjeuner réduit.

```
ggplot(data_perform) +  
  aes(x = test.preparation.course, y = moyenne) +  
  geom_boxplot() +  
  xlab("Niveau de révision") +  
  ylab("Moyenne") +  
  ggtitle("Répartition par niveau de révision des moyenne")
```

### Répartition par niveau de révision des moyenne

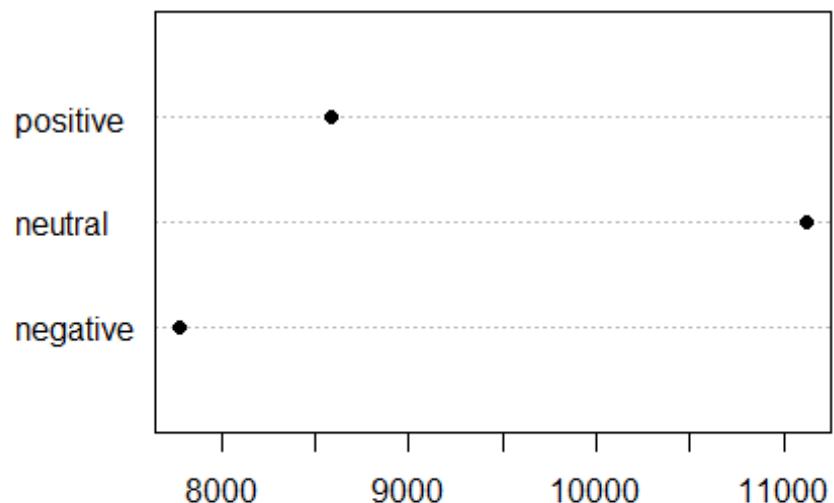


Les étudiants qui ont fait une révision complète en vue de la préparation de leur examen travaillent mieux que ceux qui n'ont fait aucune préparation.

### Sentiments des tweets

```
dotchart(table(data_tweet$sentiment), main = "Sentiment des utilisateurs de Twitter", pch = 19)
```

## Sentiment des utilisateurs de Twitter



# Rapport Phase 2, Session Apprendre R par la pratique

Armel TINDO

2022-05-24

Installation des packages

Importation et fusion des bases de données

## Données sur l'Agriculture

### Les pays exclus à la suite de la fusion

Au total **9 pays ont été exclus** après la fusion.

```
pays_exclus <- pesticide %>%
  anti_join(cereal, by = "Country") %>%
  anti_join(fertilizer, by = "Country")

subset(pays_exclus, select = c(Country))
```

```
## # A tibble: 9 x 1
##   Country
##   <chr>
## 1 Australia & New Zealand
## 2 Cook Islands
## 3 French Polynesia
## 4 Saint Kitts and Nevis
## 5 Samoa
## 6 Seychelles
## 7 Southern Europe
## 8 Tonga
## 9 Western Asia
```

### Quantité totale de pesticide utilisée en Amérique, en Australie et en Afrique

```

amerique <- c("Argentina", "Armenia", "Australia", "Belize", "Bolivia", "Brazil", "Canada", "Chile", "Ecuador", "Guatemala", "Guyana", "Haiti", "Honduras", "Jamaica", "Mexico", "Nicaragua", "Panama", "Paraguay", "Peru", "Suriname", "United States", "Uruguay", "Venezuela")

afrique <- c("Algeria", "Angola", "Botswana", "Burkina Faso", "Burundi", "Cameroon", "Cape Verde", "Central African Republic", "Chad", "Comoros", "Congo", "Egypt", "Eritrea", "Ethiopia", "Gambia", "Ghana", "Guinea", "Guinea-Bissau", "Kenya", "Lesotho", "Libya", "Madagascar", "Malawi", "Malaysia", "Mali", "Mauritania", "Morocco", "Mozambique", "Namibia", "Niger", "Rwanda", "Senegal", "South Africa", "Sudan", "Tanzania", "Togo", "Tunisia", "Uganda", "Zambia", "Zimbabwe")

pays_select <- c("Australia", amerique, afrique)

data2 <- subset(data_complet, Country %in% pays_select)

```

## Visualisation de la quantité totale de pesticide utilisée en Australie et dans les pays de l'Afrique et de l'Amérique

```

colnames(data2) <- c("Country", "Pesticides")

pesticide_map <- joinCountryData2Map(data2, joinCode = "NAME", nameJoinColumn = "Country")

```

```

## 63 codes from your data successfully matched countries in the map
## 0 codes from your data failed to match with a country code in the map
## 180 codes from the map weren't represented in your data

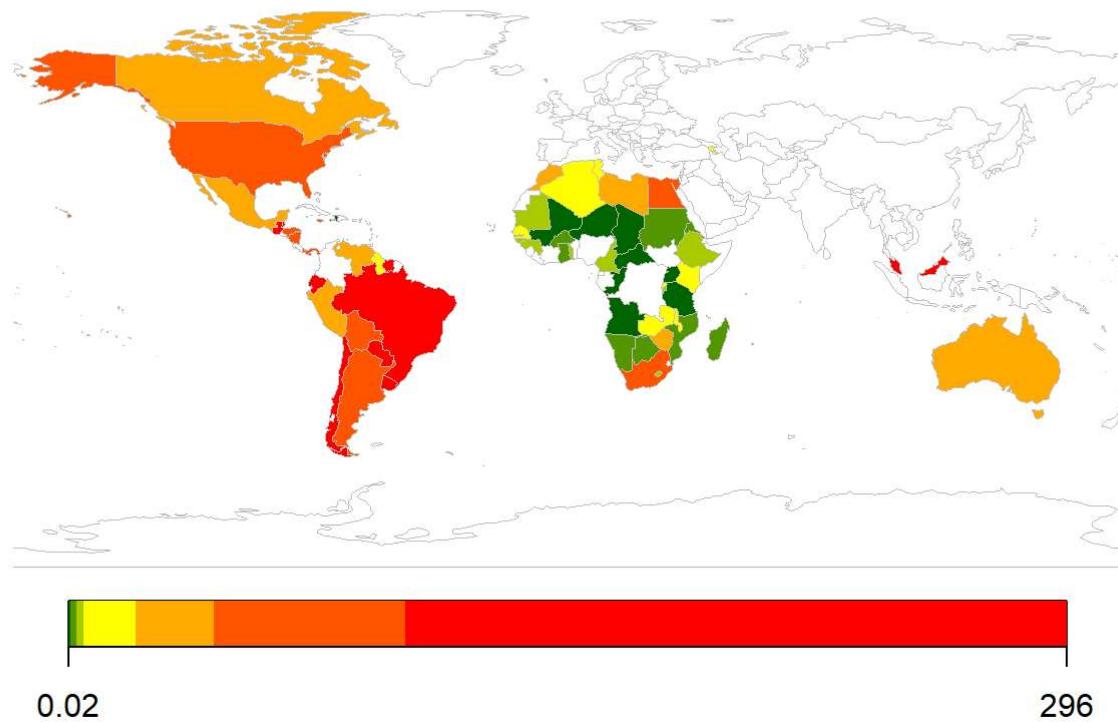
```

```

mapCountryData(pesticide_map, nameColumnToPlot = "Pesticides",
               colourPalette = c("darkgreen", "yellow", "red"), mapTitle = "Pesticides",
               mapRegion = "america")

```

## Pesticides



Cette carte montre que les pays africains utilisent moins de pesticides contrairement aux pays de l'Amérique du Sud et aux Etats Unis d'Amérique.

```
# Amérique
sum(subset(data_complet, Country %in% amerique)$Pesticides)
```

```
## [1] 2080.93
```

```
#Afrique
sum(subset(data_complet, Country %in% afrique)$Pesticides)
```

```
## [1] 522.22
```

```
# Australie
sum(subset(data_complet, Country %in% c("Australia"))$Pesticides)
```

```
## [1] 41.11
```

La quantité totale de pesticide utilisée :

- en Amérique est de 2080.93 kg/ha
- en Afrique est de 522.22 kg/ha

- en Australie est de 41.11 kg/ha

# Moyenne et médiane de rendement de céréales (hectogramme par hectare (Hg/Ha)) pour l'Amérique, l'Afrique et l'Australie

## En Amérique

- Rendement moyen de céréales (hg/ha)

```
mean(subset(data_complet, Country %in% amerique)$Yield)
```

```
## [1] 1329047
```

- Rendement médian de céréales (hg/ha)

```
median(subset(data_complet, Country %in% amerique)$Yield)
```

```
## [1] 1216282
```

## En Afrique

- Rendement moyen de céréales (hg/ha)

```
mean(subset(data_complet, Country %in% afrique)$Yield)
```

```
## [1] 657683.6
```

- Rendement médian de céréales (hg/ha)

```
median(subset(data_complet, Country %in% afrique)$Yield)
```

```
## [1] 589721
```

## En Australie

- Rendement moyen de céréales (hg/ha)

```
mean(subset(data_complet, Country %in% c("Australia"))$Yield)
```

```
## [1] 924103
```

- Rendement médian de céréales (hg/ha)

```
median(subset(data_complet, Country %in% c("Australia"))$Yield)
```

```
## [1] 924103
```

## Commandes Amazon

### Ville où le nombre d'envoie (shipping) est le plus élevé

Il s'agit de la ville de **Numbai** avec **21 commandes**.

```
amazon_data %>% count(str_to_upper(str_extract_all(ship_city, pattern = "[a-zA-Z]+")), sort = TRUE, name = "Nombre maximum de commande")
```

ship_city	Nombre maximum de commande
"MUMBAI"	21
"BENGALURU"	15
"KOLKATA"	14
"CHENNAI"	9
"HYDERABAD"	9
"C(\"NEW\", \"DELHI\")"	8
"GURUGRAM"	7
"CHANDIGARH"	4
"PUNE"	4
"VISAKHAPATNAM"	4
# ... with 63 more rows	

## Conversion de la colonne order en année

```
annee <- str_sub(amazon_data$order_date, 6, -5)
amazon_data$order_date <- year(parse_date_time(annee, '%d%m%y, %I:%M %p'))
```

```
str(amazon_data$order_date)
```

```
## num [1:171] 2021 2021 2021 2021 2021 ...
```

## Clacul de la moyenne des frais d'envoi

En moyenne, les frais d'envoин s'élèvent à 85.79 yuan en 2021 et 81.05 yuan en 2022.

```
tapply(as.double(substr.amazon_data$shipping_fee, 2, nchar.amazon_data$shipping_fee)), amazon_data$order_date, mean, na.rm = TRUE)
```

```
##      2021     2022
## 85.79540 81.05125
```

# Apprendre par la pratique s1

Mariel A.-F.

25/04/2022

## Etape 1

Une manipulation de données consiste à modifier la structure des données pour répondre à des besoins d'analyse.

## Étape 2 : Le contexte des données

Notre base de données contient des données de différents TED talks. On y trouve le titre, l'auteur, le mois de parution, le nombre de vues, le nombre de likes, et le lien du TED talk.

```
#Répertoire de travail
setwd("D:/Apprendre par la pratique")
#TED Talks
data_ted <- read.csv("data.csv", sep = ",", header = TRUE)
str(data_ted)

## 'data.frame': 5440 obs. of 6 variables:
## $ title : chr "Climate action needs new frontline leadership" "The dark history of the overthrow o...
## $ author: chr "Ozawa Bineshi Albert" "Sydney Iaukea" "Martin Reeves" "James K. Thornton" ...
## $ date  : chr "December 2021" "February 2022" "September 2021" "October 2021" ...
## $ views : int 404000 214000 412000 427000 2400 422000 412000 455000 66000 584000 ...
## $ likes : int 12000 6400 12000 12000 72 12000 12000 13000 1900 17000 ...
## $ link  : chr "https://ted.com/talks/ozawa_bineshi_albert_climate_action_needs_new_frontline_leade...
```

## Étape 3 : Description des données (type de variables, statistiques descriptives, données aberrantes, présence de doublons)

Les deux variables views et like sont des variables quantitatives alors que les autres variables sont des variables qualitatives. Le nombre moyen de vue est de 2 061 576. Une vidéo compte en moyenne 62608 likes. Il y a des vidéos comportant un nombre extrêmement élevé de vues allant jusqu'à 72 millions de vues. De même pour les likes qui peuvent atteindre les 2 millions. La transformation log a été appliquée pour réduire l'impact des valeurs aberrantes et de rendre la distribution plus symétrique dans les représentations en box plot. Après vérification, on constate qu'il n'y a aucun doublon dans la base de données.

### Statistiques descriptives du nombre de vues

```
summary(data_ted$views)

##      Min.   1st Qu.    Median     Mean   3rd Qu.   Max.
##      532    670750  1300000  2061576  2100000 72000000
```

### Statistiques descriptives du nombre de likes

```
summary(data_ted$like)

##      Min.   1st Qu.    Median     Mean   3rd Qu.   Max.
##      15     20000    40500    62608    65000  2100000
```

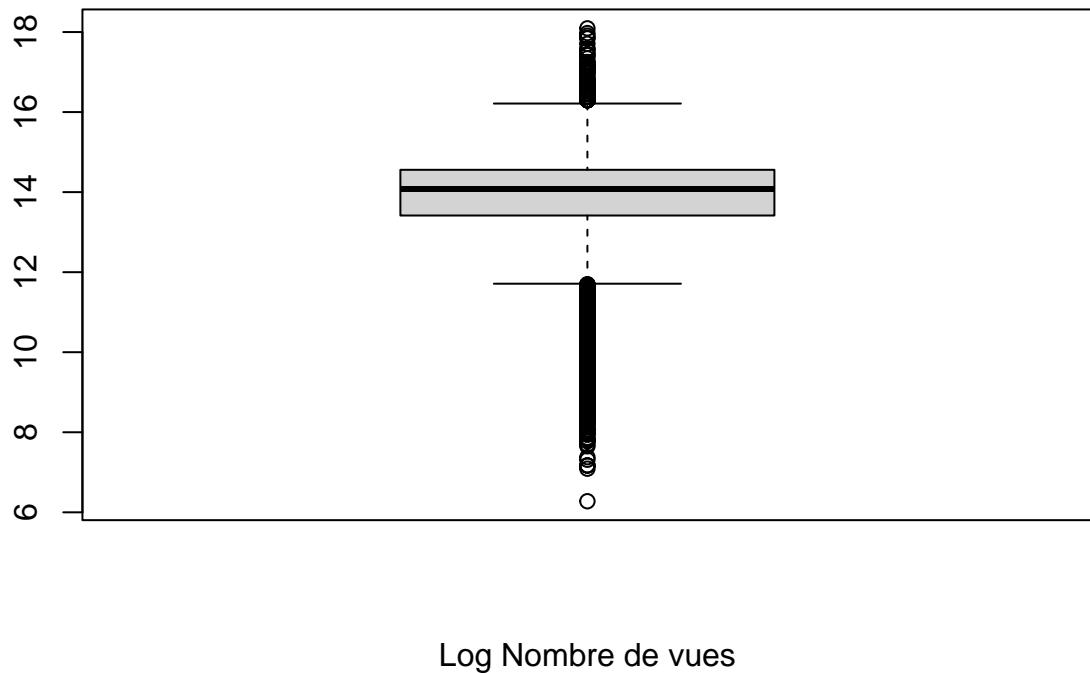
### Nombre de doublons

```
sum(duplicated(data_ted))

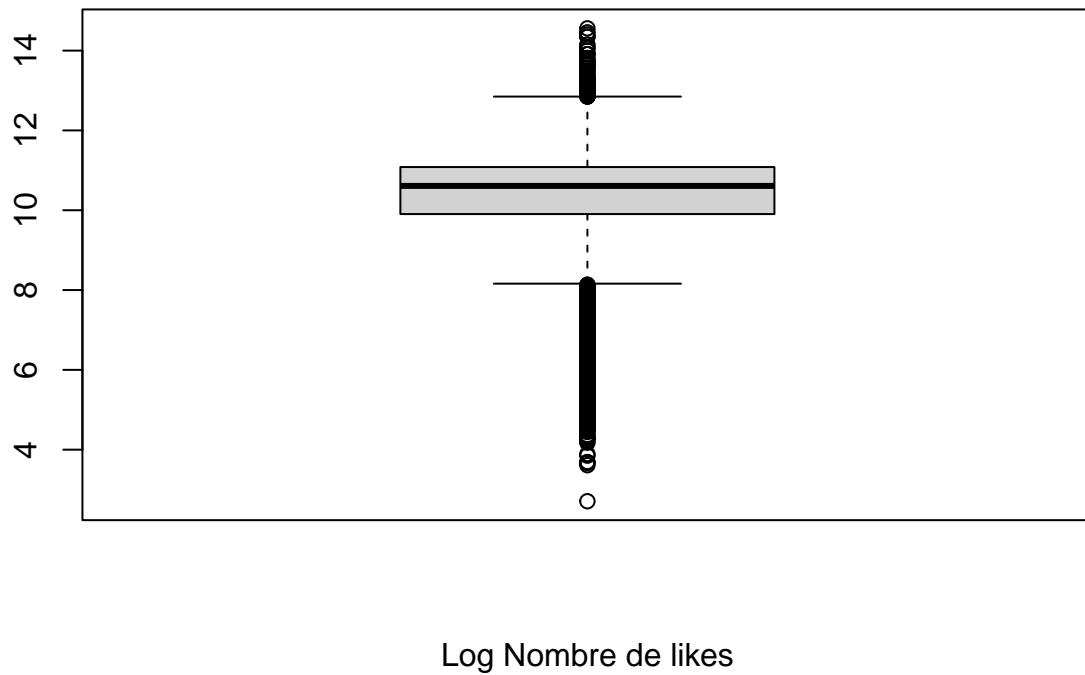
## [1] 0
```

### Diagramme en boite

```
log_views <- log(data_ted$views)
boxplot(log_views,data=data_ted,xlab="Log Nombre de vues")
```



```
log_like <- log(data_ted$like)
boxplot(log_like,data=data_ted,xlab="Log Nombre de likes")
```



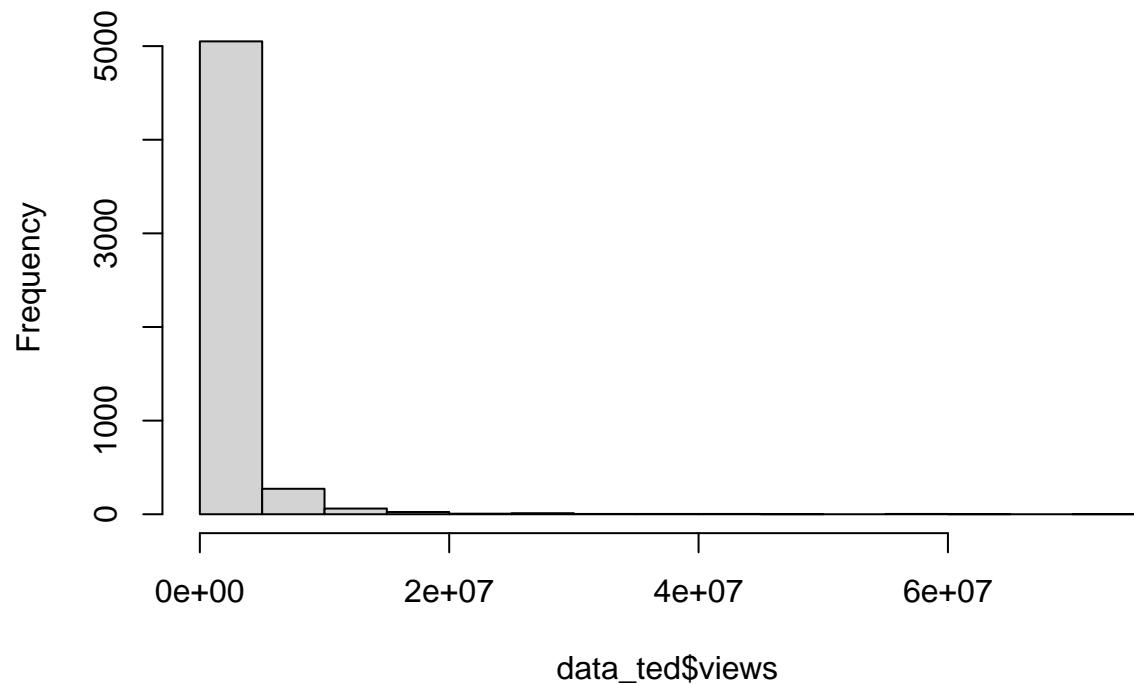
#### Étape 4 : Quelques visualisations simples

Le nuage de points nous informe que ce sont les vidéos les plus vues qui sont les plus likées. Le nuage de mots des noms des auteurs permet d'identifier les prénoms ou les noms les plus fréquents. A savoir "David", "Dan", "Michael", "Gandler", etc...

#### Diagrammes en barre du nombre de vues

```
hist(data_ted$views)
```

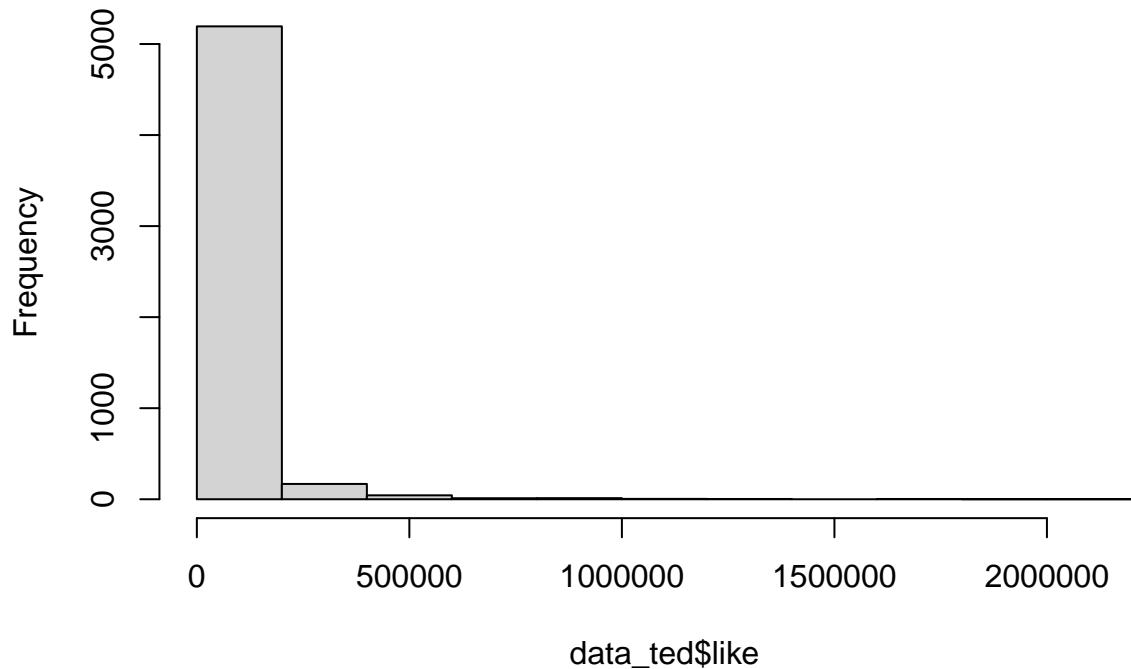
**Histogram of data\_ted\$views**



Diagrammes en barre du nombre de likes

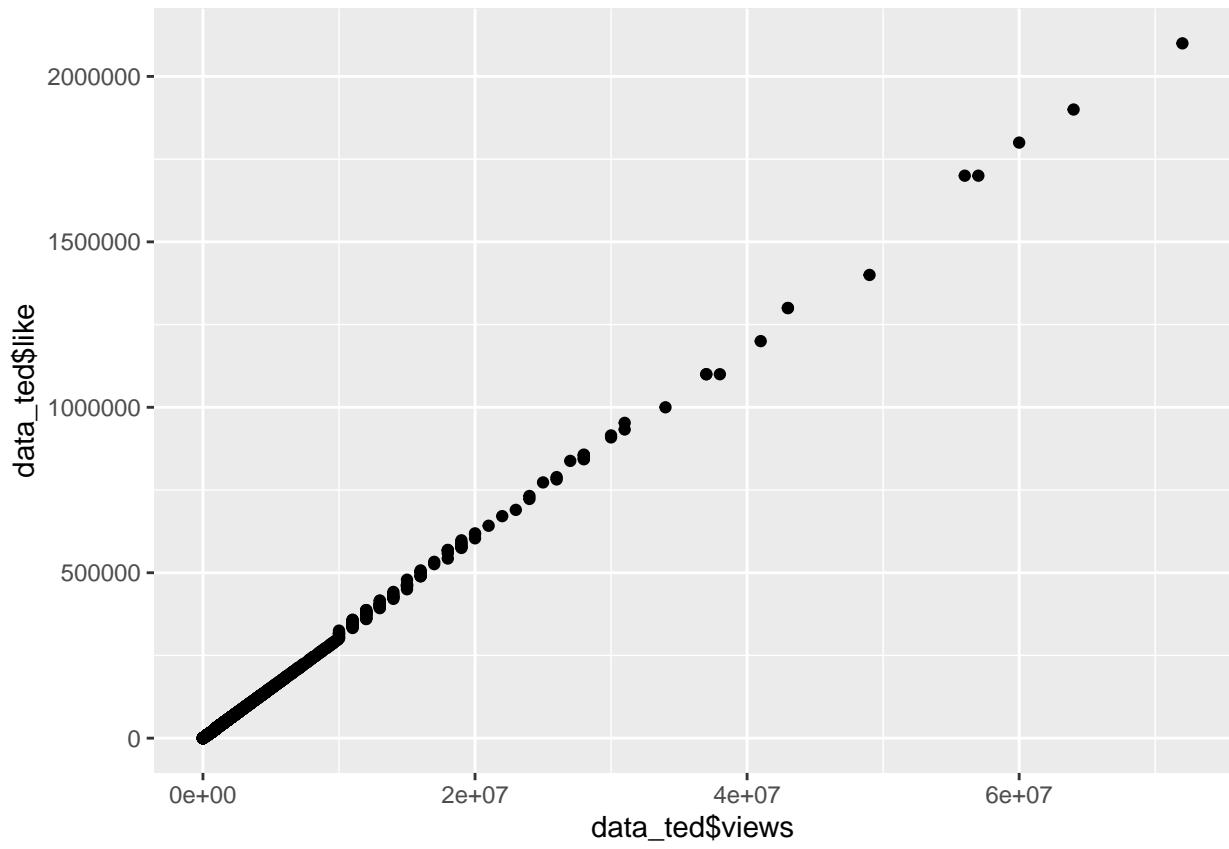
```
hist(data_ted$like)
```

**Histogram of data\_ted\$like**



NUAGE DE POINTS DU NOMBRE DE VUES ET DU NOMBRE DE LIKES

```
library(ggplot2)
ggplot(data_ted, aes(x=data_ted$views, y=data_ted$like)) + geom_point()
```



### NUAGE DE MOTS DES AUTEURS

```
library(tm) # ce package propose un ensemble de fonctions facilitant le traitement de donnees textuelles
library(wordcloud) # ce package permet la creation de wordcloud
Texte <- data_ted$author
text_corpus <- Corpus(VectorSource(Texte))
text_corpus <- tm_map(text_corpus, content_transformer(tolower))
text_corpus <- tm_map(text_corpus, removePunctuation)
text_corpus <- tm_map(text_corpus, function(x)removeWords(x,stopwords(kind = "fr")))
set.seed(123456) # permet de "fixer un graine" pour l'alea, afin de pouvoir regenerer plusieurs fois le nuage de mots
wordcloud(text_corpus, max.words = 200, colors = brewer.pal(8, "Dark2"), rot.per=0)
```

thomas  
iseult brown  
finkel will  
naomi green  
charles christopher  
rosenthal  
emma anderson gillespie megan philip  
sophie larry christina  
megan marie  
anne julia adam paul george  
laura hans adam caroline sebastian  
william sam alex robinson schwartz  
robin karen william sam alex joy teded jacknoah kim  
davis jen kelly jennifer dan gunter kingjill chang  
greg brooks steve zaidanadams natalie mary  
monica chen wangalexanderian zaidanadams sharma  
jeremy alan barry rose johnson sarah marco wilson amanda  
grant wright bryce scott tim anthony joe jay carl  
jackson lewis dennis rosa esther clay matthew rebecca  
sara liusimon jason nina francis  
mike danielle matt walker nathan katie jessica mona gore  
ellen nicholas lee nick aaronlinenriquez rachel julian jim jeff  
douglas rosling gage graham emily gil stone liz ryan jane seth amy  
rob ted joseph nancy margaret susan marc jones jacobs  
helen gary brian michael todd ray jon tom smith  
angela van eric kate susanne andrew the bill lisa cox  
joshua steven stephen peter kristen  
martin richard bell james  
daniel elizabeth williams howard john  
mark ben wendy jonathan christian  
benjamin chalabi harris gandler and

# Apprendre par la pratique s3 et s4

Mariel A.-F.

28/04/2022

## Question 1

Les pays exclus de cette fusion sont les pays qui ne sont pas présents à la fois dans les quatre base de données.

```
setwd("D:/Apprendre par la pratique")

Cereal <- read.csv("CerealCropYield_1961-2018.csv", sep = ",", header = TRUE)
Fertilizer <- read.csv("FertilizerConsumption_1961-2018.csv", sep = ",", header = TRUE)
Pesticide <- read.csv("PesticideUsage_1990-2017.csv", sep = ",", header = TRUE)
data_binded_1 <- merge(Fertilizer, Cereal, by = intersect(names(Fertilizer), names(Cereal)), all = FALSE)
data_binded_2 <- merge(data_binded_1, Pesticide, by = intersect(names(data_binded_1), names(Pesticide)), all = FALSE)
names(data_binded_2)[names(data_binded_2) == "FertilizerQuantity"] <- "Fertilizer"
names(data_binded_2)[names(data_binded_2) == "Yield..hg.ha."] <- "Yield"
names(data_binded_2)[names(data_binded_2) == "Area.harvested..ha."] <- "Area"
names(data_binded_2)[names(data_binded_2) == "Total.Pesticides.use.per.area.of.land..kg.ha."] <- "Pesticide"
library("readxl")
continent <- read_excel("continent.xlsx")
data_binded_3 <- merge(data_binded_2, continent, by = intersect(names(data_binded_2), names(continent)), all = FALSE)
```

## Question 2 : La quantité totale de pesticide utilisée en Amérique, en Afrique et en Australie

La quantité totale de pesticide utilisée est de 560,16 kg/Ha pour l'Afrique; 2110,65 kg/Ha pour l'Amérique; et 41,11 kg/Ha pour l'Australie.

```
aggregate(data_binded_3$Pesticide,
          by = list(Continent = data_binded_3$Continent),
          sum)

##   Continent      x
## 1    Africa  560.16
## 2    America 2110.65
## 3     Asia  2330.43
```

```

## 4    Europe 2135.65
## 5    Oceania 351.82

library(dplyr)
data_binded_3 %>%
  group_by(Country) %>%
  summarise(
    pesticide = sum(Pesticide, na.rm = TRUE)
  ) %>%
  filter(
    Country=="Australia"
  )

## # A tibble: 1 x 2
##   Country  pesticide
##   <chr>      <dbl>
## 1 Australia     41.1

```

### Question 3 : La moyenne et médiane de rendement de céréales (hectogramme par hectare (Hg/Ha)) pour l'Amérique, l'Afrique et l'Australie?

En moyenne, le rendement de céréales est de 684876,6 Hg/Ha pour l'Afrique; 1349710,9 Hg/Ha pour l'Amérique; et 924103 Hg/Ha pour l'Australie. Le rendement médian de céréales est de 589721 Hg/Ha pour l'Afrique; 1306386 Hg/Ha pour l'Amérique; et 924103 Hg/Ha pour l'Australie.

```

aggregate(data_binded_3$Yield,
          by = list(Continent = data_binded_3$Continent),
          mean)

##   Continent      x
## 1 Africa  684876.6
## 2 America 1349710.9
## 3 Asia   1334872.3
## 4 Europe  1751073.6
## 5 Oceania 1503567.5

aggregate(data_binded_3$Yield,
          by = list(Continent = data_binded_3$Continent),
          median)

##   Continent      x
## 1 Africa  589721
## 2 America 1306386
## 3 Asia   1150580
## 4 Europe  1651049
## 5 Oceania 1501646

library(dplyr)
data_binded_3 %>%
  group_by(Country) %>%

```

```

summarise(
  mean_yield = mean(Yield, na.rm = TRUE)
) %>%
filter(
  Country=="Australia")

## # A tibble: 1 x 2
##   Country   mean_yield
##   <chr>        <dbl>
## 1 Australia     924103

data_binded_3 %>%
  group_by(Country) %>%
  summarise(
    median_yield = median(Yield, na.rm = TRUE)
) %>%
filter(
  Country=="Australia")

## # A tibble: 1 x 2
##   Country   median_yield
##   <chr>        <int>
## 1 Australia     924103

```

La ville où le nombre d'envoie (shipping) est le plus élevé est Mumbai avec 17 envoi au total, ce qui vaut 9,9% des envois.

```

library("readxl")
amazon <- read_excel("orders_data.xlsx")
max_number_ship <- amazon %>%
  group_by(ship_city) %>%
  summarise(nb=n(), prop = round(nb/nrow(amazon)*100,digits = 2)) %>%
  slice_max(nb)

```

La moyenne de frais d'envois par année est de 69,74 Yuan pour l'année 2021 et 81,05 Yuan pour l'année 2022.

```

library("stringr")
amazon$shipping_fee_num <- as.numeric(str_replace_all(amazon$shipping_fee,
"₹", ""))
amazon[c('day of the week', 'day of the year','year','time')] <-
str_split_fixed(amazon$order_date, ',', 4)
amazon[["shipping_fee_num"]][is.na(amazon[["shipping_fee_num"]])] <- 0
aggregate(amazon$shipping_fee_num,
  by = list(year = amazon$year),
  mean)

##      year x
## 1 2021 0
## 2 2022 0

```