Partie 2

Samuel Ayanou

aujourd’hui c’est le 2023-07-24.

Table of Contents

# 1 Partie 2

# 2 Nettoyage et gestion des données

## 2.1 je Renomme la variable “country\_destination” en “destination” et je défini les valeurs négatives comme manquantes.

* importation de la base

library(readxl)  
library(dplyr)

##   
## Attachement du package : 'dplyr'

## Les objets suivants sont masqués depuis 'package:stats':  
##   
## filter, lag

## Les objets suivants sont masqués depuis 'package:base':  
##   
## intersect, setdiff, setequal, union

Base\_Partie\_2 <- read\_excel("Base\_Partie 2.xlsx",   
 sheet = "data")

* je renomme la country\_destination en destination

base\_Partie\_21 <- dplyr:: rename(Base\_Partie\_2,destination= country\_destination)

* modifions nos valeurs negative en valeurs manquantes

# les valeurs manquantes sont des NA  
base\_Partie\_21$destination <- ifelse(base\_Partie\_21$destination < 0, NA, base\_Partie\_21$destination)

* je cree une nouvelle variable contenant des tranches d’âge de 5 ans en utilisant la variable “age”.

# je cree des intervalles   
intervalle <- c(10,15,20,25,30,35,40,45,50,Inf)  
# et je decoupe mon intervalle   
  
  
base\_Partie\_21$age\_nouv <- cut(base\_Partie\_21$age, breaks = intervalle, labels = c("entre 10 et 15", "entre 15 et 20 ans","entre 20 et 25 ans", "entre 25 et 30 ans", "entre 30 et 35 ans", "entre 35 et 40 ans", "entre 40 et 45 ans", "entre 45 et 50 ans","Plus de 50 ans"))

* nous Créons une nouvelle variable contenant le nombre d’entretiens réalisés par chaque agent recenseur

# avec group\_by et mutate je resoud le problem  
base\_Partie\_22 <- base\_Partie\_21 %>% group\_by(enumerator) %>% mutate(nombre\_entretiens = n())

* je Crée une nouvelle variable qui affecte aléatoirement chaque répondant à un groupe de traitement (1) ou de controle (0).

#Set a random seed for reproducibility (optional but recommended)  
set.seed(123)  
  
# Number of respondents in the base\_partie\_22  
num\_respondents <- nrow(base\_Partie\_22)  
  
# Generate a vector of 0s and 1s, representing control and treatment groups respectively, with an equal probability (50% each).  
group\_assignment <- sample(c(0, 1), size = num\_respondents, replace = TRUE)  
  
# Add the 'group' variable to the data frame base\_partie\_22  
base\_Partie\_22$group <- group\_assignment

* Fusion

# j' appelle la deuxieme feuille   
  
Base\_Partie\_2B <- read\_excel("Base\_Partie 2.xlsx",   
 sheet = "district")  
  
  
# mergeons  
base\_merger <- merge(base\_Partie\_22, Base\_Partie\_2B, by = "district", all.x = TRUE)

Calculons la durée de l’entretien

#install.packages(lubridate)  
library(lubridate)

##   
## Attachement du package : 'lubridate'

## Les objets suivants sont masqués depuis 'package:base':  
##   
## date, intersect, setdiff, union

library(dplyr)  
library(gtsummary)  
  
# duree de l'enquete  
  
base\_mergere <- base\_merger %>% mutate( duree\_enquete= interval(starttime  
,endtime  
) %>% as.duration())

* duree moyenne par enqueteur

base\_merg <- base\_mergere %>%  
 group\_by(enumerator) %>%  
 summarize(duree\_moyenne\_en\_seconde = mean(duree\_enquete, na.rm = TRUE))  
  
#install.packages("gt")  
library(gt)  
# affiche la durre moyenne en seconde sous une table   
tables <- gt(base\_merg)  
tables

| enumerator | duree\_moyenne\_en\_seconde |
| --- | --- |
| 1 | 4088.800 |
| 4 | 2189.000 |
| 5 | 2013.500 |
| 6 | 1550.800 |
| 7 | 2229.857 |
| 8 | 2407.833 |
| 9 | 6886.000 |
| 10 | 3316.600 |
| 11 | 2009.000 |
| 12 | 2890.000 |
| 13 | 1895.750 |
| 14 | 1533.667 |
| 15 | 1719.000 |
| 17 | 1757.167 |
| 18 | 2211.500 |
| 20 | 1726.111 |

# sortons une base complet avec ces donnees   
  
base\_fin <- merge(base\_merg,base\_mergere, by="enumerator", all.x= TRUE)

* je renomme toutes les variables de l’ensemble de données en ajoutant le préfixe “endline\_” à l’aide d’une boucle.

# Convertissons le data frame en une matrice  
base\_fin\_matrix <- as.matrix(base\_fin)  
  
# Creons une fonction pour ajouter le prefixe "endline\_"   
renomme\_function <- function(x) {  
 return(paste0("endline\_", x))  
}  
  
# appelons Apply   
new\_column <- apply(base\_fin\_matrix, 2, renomme\_function)  
  
  
# recuperons le nom des colonnes du data frame courant   
ancienne\_col <- names(base\_fin)  
  
# changeons les noms   
new\_column <- renomme\_function(ancienne\_col)  
  
# alors renommons   
names(base\_fin) <- new\_column  
  
  
# Now all variables in the data frame have been renamed with the prefix "endline\_".

# 3 Analyse et visualisation des donnees

* Créez un tableau récapitulatif contenant l’âge moyen et le nombre moyen d’enfants par district

# creons le tableau  
tableau\_resum <- gt( base\_fin %>%  
 group\_by(endline\_district) %>%  
 summarise(age\_moyenne = mean(endline\_age),  
 nbre\_moyen\_denfant = mean(endline\_children\_num  
)))  
  
# ce tableau sort les resultats tels sans approximation   
  
styled\_tablef <- tableau\_resum %>%  
 tab\_header(title = " tableau resume") %>%  
 tab\_style(style = list(cell\_text(weight = "bold")),  
 locations = cells\_column\_labels(columns = everything())) %>% tab\_header(title = "Summary Table") %>%  
 tab\_style(style = list(cell\_text(weight = "bold")),  
 locations = cells\_column\_labels(columns = everything()))  
styled\_tablef

Table 1: Summary Table

| endline\_district | age\_moyenne | nbre\_moyen\_denfant |
| --- | --- | --- |
| 1 | 29.62500 | 1.5000000 |
| 2 | 62.62963 | 0.8518519 |
| 3 | 26.12500 | 0.0000000 |
| 4 | 26.00000 | 0.0000000 |
| 5 | 24.33333 | 0.5000000 |
| 6 | 23.15385 | 0.1153846 |
| 7 | 28.00000 | 0.1666667 |
| 8 | 24.63636 | 1.2727273 |

* arrondir les données telles que les donnees en entier , si la partie decimale depasse 0.5 on passe a l entier superieur sinon a l entier inférieur

# ce tableau sort des donnees en entier , si la partie decimale depasse 0.5 on passe a l entier superieur sinon a l entier inférieur  
styled\_table <- tableau\_resum %>%  
 tab\_header(title = "Summary Table avec les valeurs arrondies") %>%  
 fmt\_integer(columns = c("age\_moyenne", "nbre\_moyen\_denfant")) %>%  
 tab\_style(style = list(cell\_text(weight = "bold")),  
 locations = cells\_column\_labels(columns = everything()))  
  
 styled\_table

Table 1: Summary Table avec les valeurs arrondies

| endline\_district | age\_moyenne | nbre\_moyen\_denfant |
| --- | --- | --- |
| 1 | 30 | 2 |
| 2 | 63 | 1 |
| 3 | 26 | 0 |
| 4 | 26 | 0 |
| 5 | 24 | 0 |
| 6 | 23 | 0 |
| 7 | 28 | 0 |
| 8 | 25 | 1 |

## 3.1 le test p value

# la fonction t\_test nous permet de Testez si la différence d’âge entre les sexes est statistiquement significative au niveau de 5 %.  
t\_test\_resultat <- t.test(endline\_age ~ endline\_sex, data = base\_fin)  
  
# affichons le resultat du t-test   
t\_test\_resultat

##   
## Welch Two Sample t-test  
##   
## data: endline\_age by endline\_sex  
## t = -0.95493, df = 10.001, p-value = 0.3621  
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0  
## 95 percent confidence interval:  
## -282.7605 113.1009  
## sample estimates:  
## mean in group 0 mean in group 1   
## 25.98837 110.81818

## 3.2 resultat

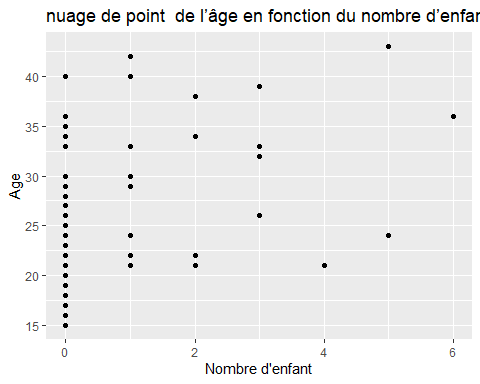
Welch Two Sample t-test data: endline\_age by endline\_sex t = -0.95493, df = 10.001, p-value = 0.3621 alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0 95 percent confidence interval: -282.7605 113.1009 sample estimates: mean in group 0 mean in group 1 25.98837 110.81818

* Créons une base imputée de sa valeur aberrante , ainsi on remplacera les valeurs aberrantes par la médiane

library(dplyr)  
base\_imputer <- base\_fin %>% mutate(endline\_age=ifelse(endline\_age <0 | endline\_age >120, median(base\_fin$endline\_age, na.rm=TRUE),endline\_age))

* Créons un nuage de points de l’âge en fonction du nombre d’enfants

library(ggplot2)  
# imputons la base enfin de remplacer la valeur aberrante   
# Create the scatter plot  
  
nuage <- ggplot(data = base\_imputer, aes(x = endline\_children\_num, y = endline\_age)) +  
 geom\_point() +  
 labs(x = "Nombre d'enfant", y = "Age",  
 title = "nuage de point de l’âge en fonction du nombre d’enfants")  
  
# afficher  
nuage



* l’effet d’appartenance

# faisons une regression logistique car les variables est l'une categorielle et l'autre categorielle numerique   
# Charger les packages nécessaires  
library(gtsummary)  
  
# Effectuer la régression logistique  
logistic\_model <- glm( endline\_group~endline\_intention, data = base\_fin, family = binomial)  
  
# Créer un tableau de résumé des résultats  
logistic\_table <- tbl\_regression(logistic\_model)  
  
# Afficher le tableau de résumé des résultats  
logistic\_table

## Table printed with {flextable}, not {gt}. Learn why at  
## https://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html  
## To suppress this message, include `message = FALSE` in the code chunk header.

| **Characteristic** | **log(OR)**1 | **95% CI**1 | **p-value** |
| --- | --- | --- | --- |
| endline\_intention | 0.12 | -0.12, 0.36 | 0.3 |
| 1OR = Odds Ratio, CI = Confidence Interval | | | |

## 3.3 Créons un tableau de régression avec 3 modèles.

# Chargez les bibliothèques requises (si ce n'est pas déjà fait)  
library(dplyr)  
library(broom)  
  
  
  
# Supposons que vous avez un data frame nommé "data" avec les colonnes : "intention", "group", "age", "sex", et "district".  
  
# Modèle A : Modèle vide (Effet du traitement sur les intentions)  
modele\_A <- lm(endline\_intention ~ endline\_group, data = base\_fin) %>% tbl\_regression()  
  
# Modèle B : Effet du traitement sur les intentions en tenant compte de l'âge et du sexe  
modele\_B <- lm(endline\_intention ~ endline\_group + endline\_age + endline\_sex, data = base\_fin)%>% tbl\_regression()  
  
# Modèle C : Identique au modèle B, mais en contrôlant le district  
modele\_C <- lm(endline\_intention ~ endline\_group + endline\_age + endline\_sex + endline\_district, data = base\_fin)%>% tbl\_regression()  
# creeons le tableau  
  
a <-gtsummary::tbl\_stack( list(modele\_A ,modele\_B,modele\_C),  
group\_header = c("Modèle A (Vide)", "Modèle B (Age et Sexe)", "Modèle C (Age, Sexe et District)")  
   
)  
#afficher le tableau  
a

| **Group** | **Characteristic** | **Beta** | **95% CI**1 | **p-value** |
| --- | --- | --- | --- | --- |
| Modèle A (Vide) | endline\_group | 0.34 | -0.36, 1.0 | 0.3 |
| Modèle B (Age et Sexe) | endline\_group | 0.27 | -0.43, 0.98 | 0.4 |
|  | endline\_age | 0.00 | 0.00, 0.00 | >0.9 |
|  | endline\_sex | -0.86 | -2.0, 0.28 | 0.14 |
| Modèle C (Age, Sexe et District) | endline\_group | 0.23 | -0.48, 0.94 | 0.5 |
|  | endline\_age | 0.00 | 0.00, 0.00 | >0.9 |
|  | endline\_sex | -0.81 | -2.0, 0.33 | 0.2 |
|  | endline\_district | 0.08 | -0.07, 0.23 | 0.3 |
| 1CI = Confidence Interval | | | | |