Projet final - Logiciel statistique R

Dieynaba Ka

2023-07-20

## 1 partie 1

L’objectif de ce projet est que nous appliquons les outils que nous avons étudiés dans le cours du logiciel statistique R, dans le cas d’une étude de cas réelle.

#les packages essentiels pour le projet  
library("haven")  
library("readxl")  
library("janitor")  
library("gtsummary")  
library("dplyr")  
library("knitr")  
library(ggplot2)  
library(sf)  
library(st)

## 1.2 Importation et mise en forme

Dans cette question,il sagit d’abord d’importer la base, de creer le dataframe projet puis de sélectionner les variables mentionnées.

Base\_partie1<- read\_excel("Base\_Partie 1.xlsx") #on importe la base  
projet<-data.frame(Base\_partie1)#création du dataframe  
base1<-projet|> dplyr :: select(-c("key"))#selection de toutes les variables de la base sauf la variable key

Créons le tableau qui résume les valeurs manquantes par variables. Pour celà, on utilise la fonction kable() qui est dans le package knitr.

val\_manquantes<-data.frame(nbr\_valmanquantes=colSums(is.na(projet)),frequency=(colSums(is.na(projet)/nrow(projet))\*100))  
kable(val\_manquantes)# création du tableau qui résume les valeurs manquantes par variables

|  | nbr\_valmanquantes | frequency |
| --- | --- | --- |
| key | 0 | 0.0 |
| q1 | 0 | 0.0 |
| q2 | 0 | 0.0 |
| q23 | 0 | 0.0 |
| q24 | 0 | 0.0 |
| q24a\_1 | 0 | 0.0 |
| q24a\_2 | 0 | 0.0 |
| q24a\_3 | 0 | 0.0 |
| q24a\_4 | 0 | 0.0 |
| q24a\_5 | 0 | 0.0 |
| q24a\_6 | 0 | 0.0 |
| q24a\_7 | 0 | 0.0 |
| q24a\_9 | 0 | 0.0 |
| q24a\_10 | 0 | 0.0 |
| q25 | 0 | 0.0 |
| q26 | 0 | 0.0 |
| q12 | 0 | 0.0 |
| q14b | 1 | 0.4 |
| q16 | 1 | 0.4 |
| q17 | 131 | 52.4 |
| q19 | 120 | 48.0 |
| q20 | 0 | 0.0 |
| filiere\_1 | 0 | 0.0 |
| filiere\_2 | 0 | 0.0 |
| filiere\_3 | 0 | 0.0 |
| filiere\_4 | 0 | 0.0 |
| q8 | 0 | 0.0 |
| q81 | 0 | 0.0 |
| gps\_menlatitude | 0 | 0.0 |
| gps\_menlongitude | 0 | 0.0 |
| submissiondate | 0 | 0.0 |
| start | 0 | 0.0 |
| today | 0 | 0.0 |

Au regard du tableau ci\_dessus on voit que la variable key ne contient aucune valeur manquante.

## 1.3 Création des variables

Avec la fonction rename(), on renomme nos trois variables que l’on met dans un nouveau dataframe, puis on utilise la fonction mutate() pour créer la variable sexe. La création du dataframe langues se fait facilement en appliquant la fonction select() à notre base.

newdata<-projet%>% #création d'un nouveau dataframe où les 3variables seront renommés  
 rename(region=q1,departement=q2,sexe=q23)  
newdata<-newdata%>% # création de la variable sexe2 en conservant notre base initiale  
 mutate(sexe\_2=ifelse(sexe=="Femme",1,0))  
langues<-select(newdata,'key',starts\_with('q24a'))#création du dataframe Langues

• Créons une variable parle qui est égale au nombre de langue parlée par le dirigeant de la PME.

langues<-langues%>%  
 mutate(parle=rowSums(langues[,2:10]))#création de la variable parle en considérant uniquement les colonnes de 2 à 10 car c'est elles qui contiennent les informations relatives à chaque langue parlée par le dirigeant de la PME.

• Sélectionnons uniquement les variables key et parle, l’objet de retour sera langues• Mergons les data.frame projet et langues:

langues<-langues%>%dplyr::select('key','parle')# selection des variables key et parle  
proj\_langue<-data.frame(newdata%>%merge(langues,by="key"))

## 2 Analyses descriptives

Dans cette partie, pour répondre aux différentes questions posées, on résume tout dans un tableau en utilisant la fonction tbl\_summary() qui est dans le package gtsummary. En effet, ce tableau donne la répartion des PME suivant les quatre variables : sexe, le niveau d’instruction,le statut juridique et la varable propriètaire/locataire. De plus, en utilisant la variable sexe comme variable de groupement, on otient le croisement entre la variable sexe et les trois autres variables. Ce qui donne le résultat suivannt :

gtsummary::tbl\_summary(projet,include=c(q23,q25,q12,q81),  
 by=q23, #variable de groupement  
 label=list(q12~ "Statut juridique",q81~"propriétaire/locataire",q25~"Niveau d'instuction"),#on labélise nos variables  
 statistic=list( all\_categorical()~"{n}/{N}({p}%)"),#on donne les statistics souhaités  
 missing="always",# on affiche les statistiques sur les valeurs manquantes  
 missing\_text="Missing",# mettre le texte 'Missing pour les valeurs maquantes  
 percent="column"# opter pour le profil colonne  
 )%>%  
 add\_overall()#pour ajouter les totaux

| **Characteristic** | **Overall**, N = 250 | **Femme**, N = 191 | **Homme**, N = 59 |
| --- | --- | --- | --- |
| Niveau d’instuction |  |  |  |
| Aucun niveau | 79/250(32%) | 70/191(37%) | 9/59(15%) |
| Niveau primaire | 56/250(22%) | 48/191(25%) | 8/59(14%) |
| Niveau secondaire | 74/250(30%) | 56/191(29%) | 18/59(31%) |
| Niveau Superieur | 41/250(16%) | 17/191(8.9%) | 24/59(41%) |
| Missing | 0 | 0 | 0 |
| Statut juridique |  |  |  |
| Association | 6/250(2.4%) | 3/191(1.6%) | 3/59(5.1%) |
| GIE | 179/250(72%) | 149/191(78%) | 30/59(51%) |
| Informel | 38/250(15%) | 32/191(17%) | 6/59(10%) |
| SA | 7/250(2.8%) | 1/191(0.5%) | 6/59(10%) |
| SARL | 13/250(5.2%) | 2/191(1.0%) | 11/59(19%) |
| SUARL | 7/250(2.8%) | 4/191(2.1%) | 3/59(5.1%) |
| Missing | 0 | 0 | 0 |
| propriétaire/locataire |  |  |  |
| Locataire | 24/250(9.6%) | 16/191(8.4%) | 8/59(14%) |
| Propriétaire | 226/250(90%) | 175/191(92%) | 51/59(86%) |
| Missing | 0 | 0 | 0 |

##Analyse descriptive Faisons une statistique descriptive sur les variables de notre choix. Dans cette partie, il s’agira de voir la répartion des PME suivant la region et le sexe du dirigeant de la PME. Le tableau ci\_dessous montre cette répartition.

tab<-table(projet$q1,projet$q23)  
knitr::kable(tab,caption="Répartion des PME suivant le sexe et la région",escape=F)

Répartion des PME suivant le sexe et la région

|  | Femme | Homme |
| --- | --- | --- |
| Dakar | 0 | 1 |
| Diourbel | 28 | 6 |
| Fatick | 23 | 7 |
| Kaffrine | 8 | 0 |
| Kaolack | 20 | 1 |
| Kolda | 4 | 5 |
| Saint-Louis | 22 | 20 |
| Sédhiou | 1 | 3 |
| Thiès | 48 | 3 |
| Ziguinchor | 37 | 13 |

Dans cette partie, nous allons voir la répartiotion des PME suivant l’age et le sexe du responsable de la PME. Mais avant celà, il est necessaire de regrouper la variable age en tranche d’age pour faciliter l’analyse.

Age<-projet$q24# on récupere la varible qui contient l'age des dirigeants des PME  
points\_de\_coupe<-c(0,30,40,50,60,Inf)# on définit les points de coupure   
tranche\_age<-cut(Age,breaks=points\_de\_coupe,labels=c("0-30","31-40","41-50","51-60","61+"))#on met la variable age sous forme de tranche d'age  
tab<-table(projet$q23,tranche\_age)# puis on fait le croisement  
knitr::kable(tab,caption="Répartion des PME suivant le sexe et l'age ",escape=F)

Répartion des PME suivant le sexe et l’age

|  | 0-30 | 31-40 | 41-50 | 51-60 | 61+ |
| --- | --- | --- | --- | --- | --- |
| Femme | 3 | 20 | 42 | 59 | 67 |
| Homme | 3 | 16 | 12 | 16 | 12 |

Dans cette section , il s’agira tout simplement de voir la répartition des PME suivant le nombre d’annees d’expérience et le niveau d’instruction.

Nbr\_annee<-projet$q26  
points\_de\_coupe<-c(0,5,15,25,35,Inf)  
tranche<-cut(Nbr\_annee,breaks=points\_de\_coupe,labels=c("0-5","6-15","16-25","26-35","36+"))  
tab<-table(projet$q25,tranche)  
knitr::kable(tab,caption="Répartion des PME suivant le niveau d'instruction et le nombre d'années d'expérience ",escape=F)

Répartion des PME suivant le niveau d’instruction et le nombre d’années d’expérience

|  | 0-5 | 6-15 | 16-25 | 26-35 | 36+ |
| --- | --- | --- | --- | --- | --- |
| Aucun niveau | 11 | 32 | 25 | 9 | 2 |
| Niveau primaire | 16 | 20 | 10 | 8 | 2 |
| Niveau secondaire | 15 | 33 | 17 | 6 | 3 |
| Niveau Superieur | 20 | 13 | 7 | 0 | 1 |

## Partie 2

## Netoyage et gestion des données

On importe d’abord la base avant de procéder au netoyage des données.

Base\_partie2<- read\_excel("Base\_Partie 2.xlsx")# on importe d'abord la base  
projet2<-data.frame(Base\_partie2)# création du dataframe projet2

• Renommons la variable “country\_destination” en “destination”et définissons les valeurs négatives comme manquantes.

projet2<-projet2%>%   
 rename(destination="country\_destination")# on renomme la country\_destination en destination  
projet2$destination<-ifelse(projet2$destination<0,NA,projet2$destination)# on définit les valeurs négatives commes manquante

• Créer une nouvelle variable contenant des tranches d’âge de 5 ans en utilisant la variable “age”.

Var<-projet2$age# on récupere la variable age dans la base projet2  
points\_de\_coupe<-c(5,10,15,20,25,30,35,40,Inf)# definition des extremités de nos tranches d'age  
Ag<-cut(Var,breaks=points\_de\_coupe,labels=c("5-10","11-15","16-20","21-25","26-30","31-35","36-40","41+"))# creation de la variable qui recupere les tranches d'age  
projet2<-projet2%>% # creation de la variable Age au niveau de projet2 en utilisant la variable Ag  
 mutate(Age=Ag)

• Créons une nouvelle variable contenant le nombre d’entretiens réalisés par chaque agent recenseur.

projet2<-projet2%>%  
 dplyr::group\_by(enumerator)%>%# on regroupe d'abord les agents recenseurs selon leurs code car un agent peu enqueter plusieurs personnes  
 mutate(Nbr\_entretien=n())%>%distinct()#on crée la variable 'Nbr\_entretien' au niveau de la base projet2

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

projet2<-projet2%>%dplyr::mutate(Groupe=sample(c(0,1),1))# pour l'alea on utilise la fontion sample puis on crée la variable groupe au niveau de la base

• Fusionnons la taille de la population de chaque district (feuille 2) avec l’ensemble de données (feuille 1) afin que toutes les personnes interrogées aient une valeur correspondante représentant la taille de la population du district dans lequel elles vivent.

feuill2<-read\_excel("Base\_Partie 2.xlsx",sheet="district")#on importe la feuille2  
projet2<-projet2%>%merge(feuill2,by="district")# on fusionne cette feuilleavec notre base en utilisantt la fonction merge() et avec comme clé la variable district

• Calculons la durée de l’entretien et indiquons la durée moyenne de l’entretien par enquêteur.

projet2<-projet2%>%dplyr::mutate(Durée=endtime-starttime)#on crée d'abord la variable Durée qui contient la durée de chaque enquête   
projet2<-projet2%>%dplyr::group\_by(enumerator)%>%# on regroupe d'abord les enqueteurs selon leurs code  
 mutate(Durée\_moyenne=mean(Durée,na.rm = T ))#puis on crée la variable durée moyenne en faisant la moyenne de la variable Durée mais sans les valeurs manquantes

• Renommons toutes les variables de l’ensemble de données en ajoutant le préfixe “endline\_” en utilisant une boucle

nb<-ncol(projet2)# le nomdre de colonnes de notre base  
prefixe<-"endline\_"#on définit le prefixe  
for (i in 1:nb)# on utilise la boucle pour parcourir toutes les colonnes de la base  
 {newname<-paste0(prefixe,names(projet2)[i])# on ajoute devant chaque variable le prefixe  
 names(projet2)[i]<-newname# on remet les nouveaux noms dans notre base  
}

## Analyse et visualisation des données

• Créons un tableau récapitulatif contenant l’âge moyen et le nombre moyen d’enfants par district.

projet2

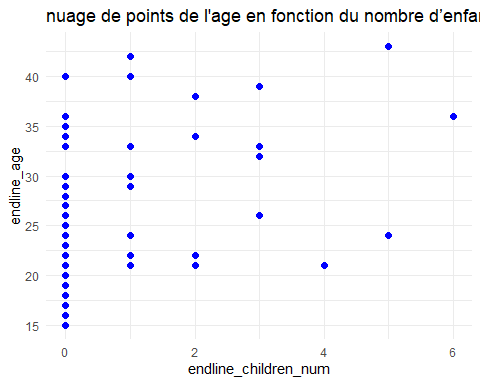
## # A tibble: 97 × 16  
## # Groups: endline\_enumerator [16]  
## endline\_district endline\_id endline\_starttime endline\_endtime   
## <dbl> <dbl> <dttm> <dttm>   
## 1 1 2 2019-01-14 14:56:37 2019-01-14 15:11:10  
## 2 1 3 2019-01-14 16:12:22 2019-01-14 16:45:52  
## 3 1 4 2019-01-14 17:15:47 2019-01-14 17:45:47  
## 4 1 1 2019-01-14 13:44:43 2019-01-14 14:16:09  
## 5 1 5 2019-01-14 14:16:35 2019-01-14 14:36:20  
## 6 1 30 2019-01-14 14:20:17 2019-01-14 14:52:30  
## 7 1 32 2019-01-14 15:31:49 2019-01-14 15:54:43  
## 8 1 31 2019-01-14 14:56:39 2019-01-14 15:23:24  
## 9 2 92 2019-01-14 14:36:32 2019-01-14 15:06:00  
## 10 2 69 2019-01-14 16:02:28 2019-01-14 16:25:49  
## # ℹ 87 more rows  
## # ℹ 12 more variables: endline\_enumerator <dbl>, endline\_age <dbl>,  
## # endline\_sex <dbl>, endline\_children\_num <dbl>, endline\_intention <dbl>,  
## # endline\_destination <dbl>, endline\_Age <fct>, endline\_Nbr\_entretien <int>,  
## # endline\_Groupe <dbl>, endline\_population <dbl>, endline\_Durée <drtn>,  
## # endline\_Durée\_moyenne <drtn>

gtsummary::tbl\_summary(projet2,include=c(endline\_district,endline\_age,endline\_children\_num),  
 by=endline\_district, #variable de groupement  
 label=list(endline\_age~ "Age\_Moyen",endline\_children\_num~"Nombre\_Moyen\_d'enfants"),#on labélise nos variables  
 statistic =list(all\_continuous() ~"{mean}",all\_categorical()~ "{p}"),#statistiques souhaités  
 missing="always",# on affiche les statistiques sur les valeurs manquantes  
 missing\_text="Missing",# mettre le texte 'Missing pour les valeurs maquantes  
 percent="column"# opter pour le profil colonne  
 )%>%  
 add\_overall()

| **Characteristic** | **Overall**, N = 97 | **1**, N = 8 | **2**, N = 27 | **3**, N = 8 | **4**, N = 5 | **5**, N = 6 | **6**, N = 26 | **7**, N = 6 | **8**, N = 11 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Age\_Moyen | 36 | 30 | 63 | 26 | 26 | 24 | 23 | 28 | 25 |
| Missing | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Nombre\_Moyen\_d’enfants |  |  |  |  |  |  |  |  |  |
| 0 | 75 | 50 | 63 | 100 | 100 | 67 | 88 | 83 | 64 |
| 1 | 9.3 | 13 | 11 | 0 | 0 | 17 | 12 | 17 | 0 |
| 2 | 6.2 | 13 | 11 | 0 | 0 | 17 | 0 | 0 | 9.1 |
| 3 | 5.2 | 0 | 11 | 0 | 0 | 0 | 0 | 0 | 18 |
| 4 | 1.0 | 13 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 2.1 | 13 | 3.7 | 0 | 0 | 0 | 0 | 0 | 0 |
| 6 | 1.0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 9.1 |
| Missing | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

• Créeons un nuage de points de l’âge en fonction du nombre d’enfants

imputer<-projet2%>%dplyr:: mutate(endline\_age=if\_else(endline\_age>100,median(projet2$endline\_age,na.rm=T),endline\_age))  
ggplot(imputer)+  
aes(x=endline\_children\_num,y=endline\_age)+  
 labs(title = "nuage de points de l'age en fonction du nombre d’enfants"  
,x = "endline\_children\_num",y="endline\_age")+  
 geom\_point(color="blue",size=2)+ theme\_minimal()



## 3 Un peu de cartographie

##3.1Transformation du dataframe en données géographiques Avant de tranformer notre data frame en données geographiques, on télécharge d’abord les données du sénégal sur le site gadm avec le format shapefile qui contient les données spatiales.

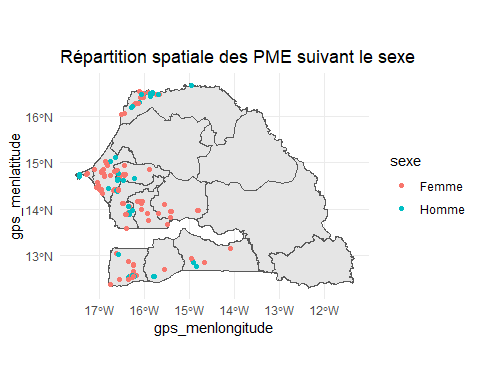
senegal <- st\_read("gadm41\_SEN\_shp/gadm41\_SEN\_1.shp")# pour lire les fichiers spatiales

## Reading layer `gadm41\_SEN\_1' from data source   
## `C:\Users\user\Documents\Projet\_R\_2023\_KA\_Dieynaba\gadm41\_SEN\_shp\gadm41\_SEN\_1.shp'   
## using driver `ESRI Shapefile'  
## Simple feature collection with 14 features and 11 fields  
## Geometry type: MULTIPOLYGON  
## Dimension: XY  
## Bounding box: xmin: -17.54319 ymin: 12.30786 xmax: -11.34247 ymax: 16.69207  
## Geodetic CRS: WGS 84

projet\_map<-st\_as\_sf(projet,coords=c("gps\_menlongitude","gps\_menlatitude"), crs =st\_crs(senegal))# tranformation du dataframe en données géographiques  
projet\_map<- st\_join(projet\_map, senegal)# on joint les deux

##3.2 Représentation spatiales des PME suivant le sexe Pour faire la représentation spatiale, on utise la fonction ggplot() qui est dans le package ggplot2.

ggplot(newdata) +# on vas chercher les variables dans la base newdata,puis  
 geom\_sf(data=senegal) +# faire la représentation à partir de la carte du senegal  
 geom\_point(aes(x = gps\_menlongitude,y=gps\_menlatitude, color=sexe))+#on fait la représentation suivant la latitude et la longitude et on colorie suivant le sexe  
 labs(title = "Répartition spatiale des PME suivant le sexe",x = "gps\_menlongitude",y="gps\_menlatitude")+# titre de notre carte  
 theme\_minimal()

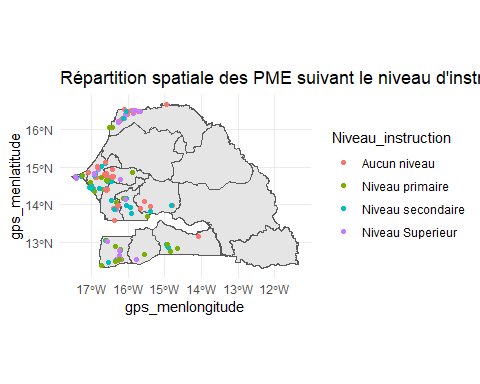


##3.3 Représentation spatiales des PME suivant le niveau d’instruction Renommons d’abord la variable q25 en Niveau d’instruction.

newdata<-newdata%>%  
 rename(Niveau\_instruction=q25)

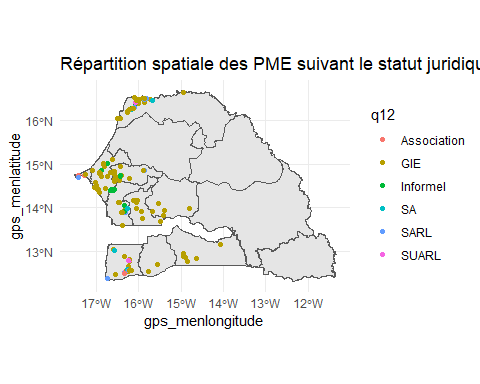
On applique le même procédé que précédemment sauf que à la place de la variable sexe on a la variable Niveau\_instruction.

ggplot(newdata)+  
 geom\_sf(data=senegal) +  
 geom\_point(aes(x = gps\_menlongitude,y=gps\_menlatitude, color=Niveau\_instruction))+  
 labs(title = "Répartition spatiale des PME suivant le niveau d'instruction",x = "gps\_menlongitude",y="gps\_menlatitude")+  
 theme\_minimal()



##3.4 Faisons une analyse spatiale de notre choix Il serait aussi interessant de voir la répartion des PME suivant le statut juridique afin d’avoir une idée sur la nature des PME ( informel,association, GIE etc.). La représentation des PME suivant le statut juridique, donne la représentation suivante.

ggplot(newdata)+  
 geom\_sf(data=senegal) +  
 geom\_point(aes(x = gps\_menlongitude,y=gps\_menlatitude, color=q12))+  
 labs(title = "Répartition spatiale des PME suivant le statut juridique",x = "gps\_menlongitude",y="gps\_menlatitude")+  
 theme\_minimal()



Il serait également interressant de voir la répartition des PME suivant l’activité principale. La représentation des PME suivant le statut juridique, donne la représentation suivante.

ggplot(newdata)+  
 geom\_sf(data=senegal) +  
 geom\_point(aes(x = gps\_menlongitude,y=gps\_menlatitude, color=q8))+  
 labs(title = "Répartition spatiale des PME suivant le statut juridique",x = "gps\_menlongitude",y="gps\_menlatitude")+  
 theme\_minimal()

