Projet final - Logiciel statistique R

DIAGNE Pathé ISEP 3 ENSAE

2023-07-24

Table of Contents

[Avant Propos 2](#_Toc141129430)

[INTRODUCTION 3](#_Toc141129431)

[Partie 1 4](#_Toc141129432)

[I- Préparation des données 4](#_Toc141129433)

[I.1 Description 4](#_Toc141129434)

[I.1.2 Importation et mise en forme 4](#_Toc141129435)

[I.1.3 un tableau qui resume les valeurs manquantes par variable 5](#_Toc141129436)

[pour ameliorer la visibilite de nos valeurs manquantes, colorions les en bleu 7](#_Toc141129437)

[I.2 Création de variables 10](#_Toc141129438)

[I.2.1 Rénommer la variable q1 en region , q2 en departement , q23 en sexe. 10](#_Toc141129439)

[I.2.2 Créer la variable sexe\_2 qui vaut 1 si sexe égale à Femme et 0 sinon. 10](#_Toc141129440)

[1.2.3 Créons un data.frame nommé langues qui prend les variables key et les variables correspondantes décrites plus haut. 11](#_Toc141129441)

[I.3 Analyses descriptives 12](#_Toc141129442)

[I.3.1 Creons un tableau descriptif sommaire 12](#_Toc141129443)

[I.3.2 les statistiques descriptives de notre choix sur les autres variables 14](#_Toc141129444)

[I.4 Un peu de cartographie 14](#_Toc141129445)

[I.4.1 Preparation des donnees 14](#_Toc141129446)

[I.4.2 realisons une réprésentation spatiale des PME suivant le sexe 15](#_Toc141129447)

[I.4.3 Faisons une réprésentation spatiale des PME suivant le niveau d’instruction 16](#_Toc141129448)

[I.4.4 Realisons une analyse spatiale de votre choix 16](#_Toc141129449)

[Conclusion 18](#_Toc141129450)

# Avant Propos

Selon Emmanuel KANT, la pratique sans la théorie est aveugle et que la théorie sans la pratique est absurde. Ainsi, les connaissances théoriques apprises ne seront d’aucune valeur sans la mise en application effective. C’est dans cette optique que l’Ecole Nationale de la Statistique et de l’Analyse Économique Pierre Ndiaye et monsieur **Hema Aboubacar**, accordent une importance et une considération particulière à la pratique au cours de la formation des élèves. Dans ce sens, l’élaboration de ce document est une bonne opportunité pour nous élèves Ingénieurs statisticiens économistes d’affermir un peu plus les connaissances théoriques et pratiques acquises lors de notre cours de Pojet Statistique avec R.

# INTRODUCTION

L’objectif de ce projet est de mettre en pratique les compétences acquises dans le cadre du cours sur lelogiciel statistique R en utilisant une étude de cas réelle. L’utilisation de R permet d’appliquer des techniques statistiques et de manipulation de données pour analyser les problèmes et prendre des décisions éclairées.

En utilisant R, nous pouvons explorer les données, effectuer des analyses descriptives, appliquer des modèles statistiques, effectuer des tests d’hypothèses et produire des visualisations graphiques pour communiquer les résultats de manière claire et concise.

Ce projet vise à intégrer les différentes étapes de l’analyse statistique, depuis la préparation des données jusqu’à l’interprétation des résultats. Il permet de se familiariser avec les différentes fonctions, packages et techniques disponibles dans R, ainsi que d’acquérir une expérience pratique dans l’utilisation de ces outils pour résoudre des problèmes réels.

L’étude de cas réelle utilisée dans ce projet offre l’occasion d’appliquer les concepts et les méthodes statistiques appris, d’identifier les variables pertinentes, de formuler des hypothèses, d’effectuer des analyses appropriées et de tirer des conclusions basées sur les résultats.

# Partie 1

## I- Préparation des données

### I.1 Description

Le fichier Base\_Partie1.xlsx contient 250 observations et 33 variables. La première colonne key correspond à l’identifiant de la PME. Les variables sont les suivantes :

### I.1.2 Importation et mise en forme

library(readxl)  
  
projet <- read\_excel("Base\_Partie 1.xlsx")  
  
# Vérification des valeurs manquantes pour la variable "key"  
valeurs\_manquantes <- projet[is.na(projet$key), ]  
  
# Vérification des valeurs manquantes pour la variable "key"  
valeurs\_manquantes <- projet[is.na(projet$key), ]  
  
if (nrow(valeurs\_manquantes) > 0) {  
 cat("Il y a des valeurs manquantes dans la variable 'key'.\n")  
 cat("Nombre de valeurs manquantes : ", nrow(valeurs\_manquantes), "\n")  
} else {  
 cat("Il n'y a pas de valeurs manquantes dans la variable 'key'.\n")  
}

## Il n'y a pas de valeurs manquantes dans la variable 'key'.

library(dplyr)  
  
## 1.2.2 Selection les variables mentionnees dans la section description.  
  
projet\_selected <- select(projet, q1, q2, q23, q24, q24a\_1, q24a\_2, q24a\_3, q24a\_4, q24a\_5, q24a\_6, q24a\_7, q24a\_9, q24a\_10, q25, q26, q12, q14b, q16, q17, q19, q20, filiere\_1, filiere\_2, filiere\_3, filiere\_4, q8, q81, gps\_menlatitude,gps\_menlongitude, submissiondate,start,today)

### I.1.3 un tableau qui resume les valeurs manquantes par variable

Dans ce code, nous avons ajouté une étape de filtrage après le calcul des valeurs manquantes pour ne conserver que les lignes où Valeurs\_manquantes > 0. Le tableau résultant ne contiendra donc que les variables qui ont des valeurs manquantes.

library(tidyr)  
library(flextable)  
  
# Compter les valeurs manquantes dans chaque variable  
missing\_counts <- projet\_selected %>%  
 summarise\_all(~ sum(is.na(.))) %>%  
 gather(key = "Variable", value = "Valeurs\_manquantes")  
  
# Filtrer les variables avec des valeurs manquantes  
missing\_counts\_filtered <- missing\_counts %>%  
 filter(Valeurs\_manquantes > 0)  
  
# Calculer le pourcentage de chaque valeur manquante  
missing\_counts\_filtered <- missing\_counts\_filtered %>%  
 mutate(Pourcentage = round(Valeurs\_manquantes / sum(Valeurs\_manquantes) \* 100, 3))  
  
# Créer une flextable avec les en-têtes de colonne spécifiques  
table\_flex <- flextable(missing\_counts\_filtered)  
  
# Personnaliser le tableau  
table\_flex <- set\_header\_labels(table\_flex, Variable = "Variable", Valeurs\_manquantes = "Valeurs manquantes", Pourcentage = "Pourcentage en %")  
table\_flex <- align(table\_flex, align = "center", part = "all")  
  
# Ajouter un titre au tableau  
title <- "Variables contenant des valeurs manquantes dans ma base"  
table\_flex <- add\_header\_lines(table\_flex, title)  
  
# Formater la colonne Pourcentage avec trois chiffres après la virgule  
table\_flex <- colformat\_num(table\_flex, col\_keys = "Pourcentage", digits = 3)

| Variables contenant des valeurs manquantes dans ma base | | |
| --- | --- | --- |
| Variable | Valeurs manquantes | Pourcentage en % |
| q14b | 1 | 0.395 |
| q16 | 1 | 0.395 |
| q17 | 131 | 51.779 |
| q19 | 120 | 47.431 |

* supposons que nous voulions afficher tous le tableau et avec les valeurs manquantes. Il s’agira pour nous de mettre en un tableau descriptif sommaire pour incluant les valeurs manquantes, en effet nous souhaitons voir les statistiques de base concernant les valeurs manquantes.
* Cela peut être utile pour avoir une vue d’ensemble de nos données et pour comprendre la proportion de valeurs manquantes pour chaque variable. En incluant les valeurs manquantes dans le tableau descriptif sommaire, nous pouvons ainsi voir combien de valeurs manquantes il y a pour chaque variable et prendre des décisions sur la façon de gérer ces valeurs manquantes dans nos analyses ultérieures

## pour ameliorer la visibilite de nos valeurs manquantes, colorions les en bleu

# Compter les valeurs manquantes dans chaque variable  
missing\_counts <- projet\_selected %>%  
 summarise\_all(~ sum(is.na(.))) %>%  
 gather(key = "Variable", value = "Valeurs\_manquantes")  
  
# Calculer le pourcentage de chaque valeur manquante et le formater en pourcentage avec 3 chiffres après la virgule  
missing\_counts <- missing\_counts %>%  
 mutate(Pourcentage = sprintf("%.3f%%", Valeurs\_manquantes / sum(Valeurs\_manquantes) \* 100))  
  
# Créer une flextable avec les en-têtes de colonne spécifiques  
table\_flex <- flextable(missing\_counts)  
  
# Personnaliser le tableau  
table\_flex <- set\_header\_labels(table\_flex, Variable = "Variable", Valeurs\_manquantes = "Valeurs manquantes", Pourcentage = "Pourcentage")  
table\_flex <- align(table\_flex, align = "center", part = "all")  
  
# Ajouter un titre au tableau  
title <- "Valeurs manquantes par variable"  
table\_flex <- add\_header\_lines(table\_flex, title)  
  
# Colorer les valeurs manquantes en bleu  
table\_flex <- color(table\_flex, i = ~ Valeurs\_manquantes > 0, j = ~ Valeurs\_manquantes, color = "blue")

| Valeurs manquantes par variable | | |
| --- | --- | --- |
| Variable | Valeurs manquantes | Pourcentage |
| q1 | 0 | 0.000% |
| q2 | 0 | 0.000% |
| q23 | 0 | 0.000% |
| q24 | 0 | 0.000% |
| q24a\_1 | 0 | 0.000% |
| q24a\_2 | 0 | 0.000% |
| q24a\_3 | 0 | 0.000% |
| q24a\_4 | 0 | 0.000% |
| q24a\_5 | 0 | 0.000% |
| q24a\_6 | 0 | 0.000% |
| q24a\_7 | 0 | 0.000% |
| q24a\_9 | 0 | 0.000% |
| q24a\_10 | 0 | 0.000% |
| q25 | 0 | 0.000% |
| q26 | 0 | 0.000% |
| q12 | 0 | 0.000% |
| q14b | 1 | 0.395% |
| q16 | 1 | 0.395% |
| q17 | 131 | 51.779% |
| q19 | 120 | 47.431% |
| q20 | 0 | 0.000% |
| filiere\_1 | 0 | 0.000% |
| filiere\_2 | 0 | 0.000% |
| filiere\_3 | 0 | 0.000% |
| filiere\_4 | 0 | 0.000% |
| q8 | 0 | 0.000% |
| q81 | 0 | 0.000% |
| gps\_menlatitude | 0 | 0.000% |
| gps\_menlongitude | 0 | 0.000% |
| submissiondate | 0 | 0.000% |
| start | 0 | 0.000% |
| today | 0 | 0.000% |

## I.2 Création de variables

La création de variables est une étape importante dans l’analyse des données. Elle permet de stocker des données, des résultats d’analyses ou des valeurs calculées dans des objets qui peuvent être utilisés ultérieurement dans le code.

Les objectifs de la création de variables dans R peuvent varier en fonction de l’analyse en cours. La création de variables peut également faciliter l’organisation et la lisibilité du code en permettant de nommer et de référencer les objets de manière claire et concise.

### I.2.1 Rénommer la variable q1 en region , q2 en departement , q23 en sexe.

# ici on Rénomme la variable q1 en region  
  
projet <- rename(projet, region = q1)  
  
# ici on Rénomme q2 en departement  
projet <- rename(projet, departement = q2)  
  
# ici on Rénomme q23 en sexe  
projet <- rename(projet, sexe = q23)

### I.2.2 Créer la variable sexe\_2 qui vaut 1 si sexe égale à Femme et 0 sinon.

#variable sexe\_2 qui vaut 1 si sexe égale à Femme et 0 sinon  
  
projet <- mutate(projet, sexe\_2 = ifelse(sexe == "Femme", 1, 0))  
  
#Créer un data.frame nommé langues qui prend les variables key et les variables correspondantes décrites plus haut  
  
langues <- projet %>%  
 select(key, starts\_with("q24a\_"))

### 1.2.3 Créons un data.frame nommé langues qui prend les variables key et les variables correspondantes décrites plus haut.

# Créer une variable parle qui est égale au nombre de langue parlée par le dirigeant de la PME  
langues <- langues %>%  
 mutate(parle = rowSums(select(., -key)))  
  
# Sélectionner uniquement les variables key et parle, l'objet de retour sera langues  
langues <- langues %>%  
 select(key, parle)

# Créer une variable parle qui est égale au nombre de langue parlée par le dirigeant de la PME.  
  
langues <- select(projet, key, q24a\_1, q24a\_2, q24a\_3, q24a\_4, q24a\_5, q24a\_6, q24a\_7, q24a\_9, q24a\_10)  
projet <- mutate(projet, parle = rowSums(select(projet, q24a\_1:q24a\_10)))  
  
# Sélectionnez uniquement les variables key et parle, l’objet de retour sera langues.  
  
langues <- select(projet, key, parle)  
projet\_langues <- left\_join(projet, langues, by = "key")

## I.3 Analyses descriptives

L’analyse descriptive est une méthode pour résumer et décrire les caractéristiques principales d’un ensemble de données. Elle peut inclure des mesures de tendance centrale (moyenne, médiane, mode), des mesures de dispersion (écart-type, variance, étendue), des fréquences et des proportions.

### I.3.1 Creons un tableau descriptif sommaire

Si vous souhaitez effectuer une analyse par filière, cela signifie que vous souhaitez examiner les caractéristiques de vos données pour chaque filière séparément. Vous pouvez le faire en filtrant vos données pour chaque filière, puis en effectuant des analyses descriptives sur les données filtrées.

|  | | **Filière1: Arachide** | | **Filière2: Anacarde** | | **Filière3: Mangue** | | **Filière4: Riz** | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Group** | **Characteristic** | **Femme**, N = 931 | **Homme**, N = 151 | **Femme**, N = 401 | **Homme**, N = 211 | **Femme**, N = 681 | **Homme**, N = 211 | **Femme**, N = 771 | **Homme**, N = 151 |
| \*\*Niveau d'étude\*\* | q25 |  |  |  |  |  |  |  |  |
|  | Aucun niveau | 38 (41%) | 5 (33%) | 12 (30%) | 1 (4.8%) | 22 (32%) | 4 (19%) | 10 (13%) | 1 (6.7%) |
|  | Niveau primaire | 20 (22%) | 3 (20%) | 15 (38%) | 2 (9.5%) | 20 (29%) | 4 (19%) | 26 (34%) | 0 (0%) |
|  | Niveau secondaire | 30 (32%) | 4 (27%) | 9 (23%) | 6 (29%) | 21 (31%) | 4 (19%) | 28 (36%) | 4 (27%) |
|  | Niveau Superieur | 5 (5.4%) | 3 (20%) | 4 (10%) | 12 (57%) | 5 (7.4%) | 9 (43%) | 13 (17%) | 10 (67%) |
| \*\*Statut juridique\*\* | q12 |  |  |  |  |  |  |  |  |
|  | Association | 2 (2.2%) | 0 (0%) | 1 (2.5%) | 2 (9.5%) |  |  | 0 (0%) | 2 (13%) |
|  | GIE | 70 (75%) | 9 (60%) | 27 (68%) | 8 (38%) | 62 (91%) | 11 (52%) | 73 (95%) | 4 (27%) |
|  | Informel | 20 (22%) | 3 (20%) | 10 (25%) | 2 (9.5%) | 3 (4.4%) | 2 (9.5%) | 1 (1.3%) | 2 (13%) |
|  | SA | 0 (0%) | 2 (13%) | 0 (0%) | 2 (9.5%) | 1 (1.5%) | 2 (9.5%) | 0 (0%) | 3 (20%) |
|  | SARL | 0 (0%) | 1 (6.7%) | 1 (2.5%) | 5 (24%) | 1 (1.5%) | 5 (24%) | 1 (1.3%) | 4 (27%) |
|  | SUARL | 1 (1.1%) | 0 (0%) | 1 (2.5%) | 2 (9.5%) | 1 (1.5%) | 1 (4.8%) | 2 (2.6%) | 0 (0%) |
| \*\*Propriétaire/Locataire\*\* | q81 |  |  |  |  |  |  |  |  |
|  | Locataire | 9 (9.7%) | 3 (20%) | 3 (7.5%) | 4 (19%) | 8 (12%) | 3 (14%) | 8 (10%) | 1 (6.7%) |
|  | Propriétaire | 84 (90%) | 12 (80%) | 37 (93%) | 17 (81%) | 60 (88%) | 18 (86%) | 69 (90%) | 14 (93%) |
| 1n (%) | | | | | | | | | |

### I.3.2 les statistiques descriptives de notre choix sur les autres variables

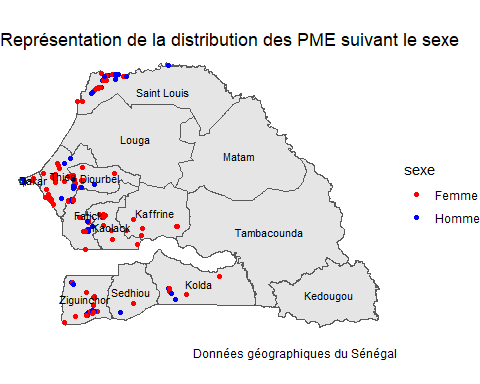
## I.4 Un peu de cartographie

### I.4.1 Preparation des donnees

# Cette ligne de code nous permet d'xtraire les noms de région du fichier de formes  
region\_names <- senegal$ADM1\_FR  
  
# Créer un bloc de données avec les noms de région  
regions <- data.frame(region = region\_names)  
  
# Calculons les centroïdes des régions  
senegal\_centroids <- st\_centroid(senegal)  
  
# Convertir l’objet senegal\_centroids en trame de données  
senegal\_centroids\_df <- as.data.frame(st\_coordinates(senegal\_centroids))  
senegal\_centroids\_df$region <- senegal$ADM1\_FR

### I.4.2 realisons une réprésentation spatiale des PME suivant le sexe

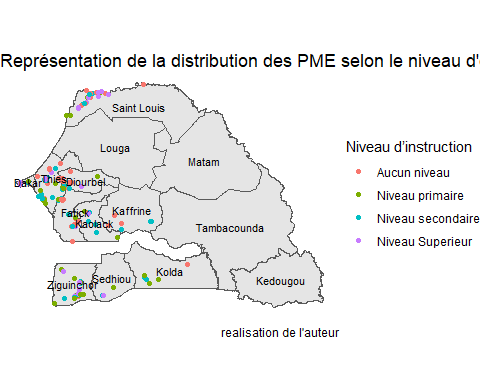
library(ggplot2)  
 # ceci nous permet d'jouter les noms des régions à la carte  
ggplot() +  
 geom\_sf(data = senegal) +  
 geom\_sf(data = projet\_map, aes(color = sexe)) +  
 geom\_text(data = senegal\_centroids\_df, aes(x = X, y = Y, label = region), size = 3) +  
 scale\_color\_manual(values = c("Femme" = "red", "Homme" = "blue")) +  
 theme\_void() +  
 labs(title = "Représentation de la distribution des PME suivant le sexe",  
   
 caption = "Données géographiques du Sénégal")



nous avons créé une représentation spatiale des PME par sexe en utilisant les packages sf et ggplot2. La carte montre la distribution spatiale des PME par sexe au Sénégal. Les points rouges représentent les PME détenues par des femmes, tandis que les points bleus représentent les PME détenues par des hommes.

### I.4.3 Faisons une réprésentation spatiale des PME suivant le niveau d’instruction.

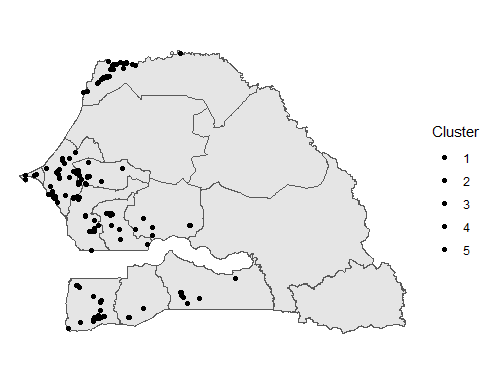
# Tracons les données avec le fichier de formes Sénégal  
ggplot() +  
 geom\_sf(data = senegal) +  
 geom\_sf(data = projet\_map, aes(color = q25)) +  
 geom\_text(data = senegal\_centroids\_df, aes(x = X, y = Y, label = region), size = 3) +  
 scale\_color\_discrete(name = "Niveau d’instruction") +  
 theme\_void()+  
 labs(title = "Représentation de la distribution des PME selon le niveau d'étude",  
   
 caption = "realisation de l'auteur")



### I.4.4 Realisons une analyse spatiale de votre choix

Nous proposons une analyse spatiale qui utilise **l’algorithme k-means** pour effectuer une analyse des clusters spatiaux sur un ensemble de données. L’algorithme k-means est une méthode de partitionnement de données qui vise à diviser les observations en k clusters en minimisant la somme des distances au carré entre chaque observation et le centroïde de son cluster. L’analyse des clusters spatiaux peut vous aider à identifier les régions où les PME sont regroupées. En regroupant les points en fonction de leur proximité spatiale.

# notre base"projet\_map" contient des variables"geometry" pour les coordonnées spatiales  
  
coords <- st\_coordinates(projet\_map)  
  
# Effectuer l'analyse des clusters spatiaux avec k-means  
clusters <- kmeans(coords, centers = 5)$cluster  
  
# Ajouter les résultats de l'analyse des clusters au projet\_map  
projet\_map$cluster <- clusters  
  
# Tracer les résultats sur la carte du Sénégal  
ggplot() +  
 geom\_sf(data = senegal) +  
 geom\_sf(data = projet\_map, aes(fill = factor(cluster))) +  
 scale\_fill\_discrete(name = "Cluster") +  
 theme\_void()



* **interpretations**

Nous pouvons voir qu’il existe des groupes de PME détenues par des femmes dans certaines zones, tandis que d’autres zones ont une plus grande concentration de PME détenues par des hommes. Ces informations peuvent être utiles pour comprendre la dynamique de l’entrepreneuriat au Sénégal et pour identifier les domaines dans lesquels des interventions ciblées peuvent être nécessaires pour soutenir les femmes entrepreneurs.

# Conclusion

Cette étude nous a permis d’obtenir des informations importantes sur les bioénergies durables pour les PME agroalimentaires d’Afrique de l’Ouest. Les tableaux récapitulatifs nous ont aidés à identifier les tendances et les caractéristiques importantes pour chaque filière étudiée. Ces résultats constituent une base solide pour des actions futures visant à promouvoir l’utilisation de bioénergies durables dans cette région