

```
# Chargement des packages
```

```
library(stringi)
```

```
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
```

```
library(knitr)
```

```
library(kableExtra)
```

```
##
```

```
## Attachement du package : 'kableExtra'
```

```
## L'objet suivant est masqué depuis 'package:dplyr':
```

```
##
```

```
##     group_rows
```

```
library(tidyr)
```

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
```

```
## v forcats   1.0.0      v readr     2.1.5
```

```
## v ggplot2   3.5.1      v stringr  1.5.1
```

```
## v lubridate 1.9.4      v tibble   3.2.1
```

```
## v purrr     1.0.2
```

```
## -- Conflicts ----- tidyverse_conflicts() --
```

```
## x dplyr::filter()          masks stats::filter()
```

```
## x kableExtra::group_rows() masks dplyr::group_rows()
```

```
## x dplyr::lag()             masks stats::lag()
```

```
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(summarytools)
```

```
##
```

```
## Attachement du package : 'summarytools'
```

```
##
```

```
## L'objet suivant est masqué depuis 'package:tibble':
```

```
##
```

```
##     view
```

```
library(gridExtra)
```

```
##  
## Attachement du package : 'gridExtra'  
##  
## L'objet suivant est masqué depuis 'package:dplyr':  
##  
##      combine
```

```
library(purrr)  
library(skimr)  
library(spdep)
```

```
## Le chargement a nécessité le package : spData  
## To access larger datasets in this package, install the spDataLarge  
## package with: 'install.packages('spDataLarge',  
## repos='https://nowosad.github.io/drat/', type='source')'  
## Le chargement a nécessité le package : sf  
## Linking to GEOS 3.10.2, GDAL 3.4.1, PROJ 8.2.1; sf_use_s2() is TRUE
```

```
library(geosphere)
```

```
# Importation des données de démonstration
```

```
demo <- read.csv("../data/demof2.csv", sep = ";", dec=",")  
#View(demo)  
#str(demo)  
names(demo)[names(demo) == "Libellé"] <- "libelle_maj"
```

```
## Fonction prenant en entrée un base et nettoie les noms des colonnes
```

```
nettoyer_noms_colonnes <- function(data){  
  names(data) <- names(data) %>%  
    stri_trans_general("Latin-ASCII") %>% # Suppression des accents  
    gsub("\\s+", "_", .) %>% # Remplacement des espaces par des underscores  
    gsub("\\.+", "_", .) %>% # Remplacement des points par des underscores  
    tolower() # Conversion en minuscules  
  return (data)  
}
```

```
## Nettoyage des colonnes de la base demo  
demo <- nettoyer_noms_colonnes(demo)  
#names(demo)
```

```
# Fusion des bases et création des variables
```

```
## Importation de la base generalise  
generalise <- read.csv("../data/generalise.csv", sep=";")  
#str(generalise)
```

```
## Importation de la base pour les lon et lat manquantes
```

```

donnees_manquantes <- read.csv(
  "../data/communes_manquantes_latitudes_longitudes.csv", sep=";", dec=".")
#str(donnees_manquantes)

donnees_manquantes$longitude <- donnees_manquantes$longitude %>%
  str_replace_all(",", "") %>% # Supprime les virgules
  as.numeric()

## Nettoyage dans les noms des colonnes
generalise <- nettoyer_noms_colonnes(generalise)
donnees_manquantes <- nettoyer_noms_colonnes(donnees_manquantes)

## Fusion des bases
data <- demo %>%
  inner_join(generalise, by = "code") %>%
  left_join(donnees_manquantes, by = "code") %>%
  mutate(
    longitude = ifelse(is.na(longitude.x), longitude.y, longitude.x),
    latitude = ifelse(is.na(latitude.x), latitude.y, latitude.x)
  ) %>%
  select(-longitude.x, -longitude.y, -latitude.x, -latitude.y)

#nrow(demo)
#nrow(generalise)
#nrow(data)

## Filtrons les communes n'appartenant pas au département 97
data <- data %>% filter(departement != 97)

## Création de la variable taux de visites
data <- data %>%
  mutate(taux_visites = nb_visite/population_municipale_2021_x)

## Création de la variable taux de visites pour les plus de 19 ans
data <- data %>%
  mutate(pop_19_ans_ou_plus = pop_15_ans_ou_plus - pop_15_19_ans,
    taux_visites_19_ans_ou_plus = nb_visite / pop_19_ans_ou_plus)

#summary(data$taux_visites)
#summary(data$taux_visites_19_ans_ou_plus)

#skim(data)

## Exportation de la base finale
write.csv(data, "../data/data.csv", row.names = FALSE)

```

Analyse descriptive

1. Description de la population d'étude

Notre population d'étude est une population assez homogène en matière d'âge. Cependant plus on dépasse les 75 ans et moins on rencontre de personnes. D'autres part notre population est fortement masculine avec une forte proportion des hommes quelle que soit la tranche d'âge à l'exception des tranches du troisième âge.

```
library(ggplot2)
library(dplyr)
library(tidyr)
df = data
colnames(df) <- gsub("homme_", "hommes_", colnames(df))
colnames(df) <- gsub("hommes_70_47$", "hommes_70_74", colnames(df))

# Sélectionner les colonnes de la pyramide des âges
age_groups <- c("0_4", "5_9", "10_14", "15_19", "20_24", "25_29", "30_34", "35_39",
               "40_44", "45_49", "50_54", "55_59", "60_64", "65_69", "70_74",
               "75_79", "80_84", "85_89", "90_94", "95_plus")

# Restructurer les données pour la visualisation
hommes_vars <- intersect(colnames(df), paste0("hommes_", age_groups))
femmes_vars <- intersect(colnames(df), paste0("femmes_", age_groups))

# Créer la pyramide des âges avec les hommes d'abord, puis les femmes
pyramide <- data.frame(
  Age = rep(age_groups, 2), # Liste tous les âges d'abord pour les hommes, puis pour les femmes
  Sexe = c(rep("Homme", length(age_groups)), rep("Femme", length(age_groups))),
  Population = c(-colSums(df[paste0("hommes_", age_groups)], na.rm=TRUE),
                +colSums(df[paste0("femmes_", age_groups)], na.rm=TRUE)) # Hommes en négatif
)

# Convertir Age en facteur ordonné pour garantir le bon ordre
pyramide$Age <- factor(pyramide$Age, levels = age_groups, ordered = TRUE)

ggplot(pyramide, aes(x=Age, y=Population, fill=Sexe)) +
  geom_bar(stat="identity", width=0.8) +
  coord_flip() + # Pour afficher en pyramide
  scale_y_continuous(labels = abs) + # Afficher les valeurs absolues
  labs(
    x="Tranche d'âge",
    y="Population",
    fill="Sexe") +
  theme_minimal() +
  scale_fill_manual(values=c("pink", "blue")) # Couleurs pour Femme/Homme
```

Dans cette partie, nous allons réaliser quelques statistiques descriptives sur nos données.

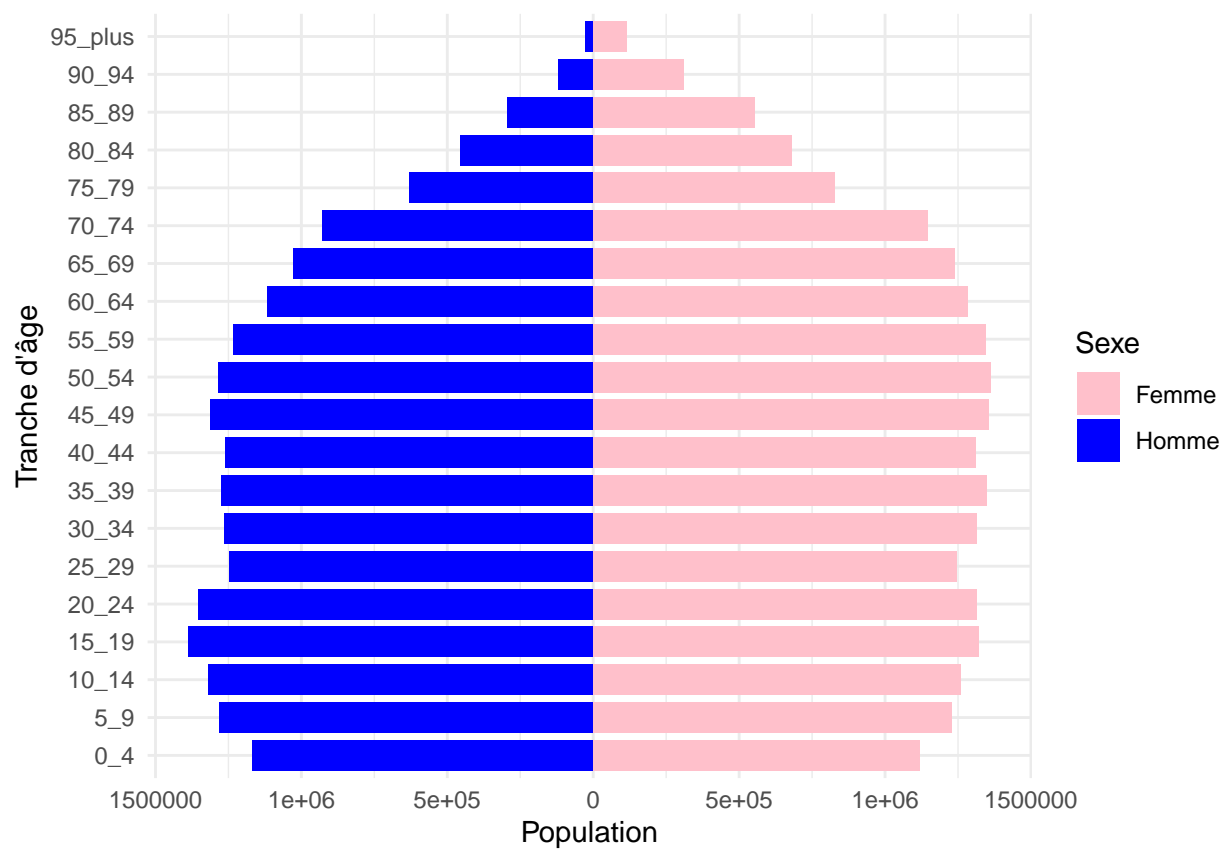


Figure 1: Pyramide des âges

Analyse univariée

```
# Stats desc sur le nombre de visites
stats <- summary(data$nb_visite)

# Sous forme de data frame
summary_df <- data.frame(
  Statistique = names(stats),
  Valeur = as.numeric(stats)
) %>%
  pivot_wider(names_from = Statistique, values_from = Valeur)

# Génération du tableau en LaTeX
summary_df %>% kable(format = "latex",
  booktabs = TRUE,
  caption = "Résumé statistique du nombre de visites") %>%
  kable_styling(latex_options = c("striped",
    "HOLD_position"))
```

Table 1: Résumé statistique du nombre de visites

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1037	5993	9127	19129.63	17290	765833

1. Taux et Nombre de visites

L'analyse des statistiques descriptives sur le nombre de consultations annuelles de médecin généraliste entre 2018 et 2022 révèle une distribution fortement asymétrique à droite, avec une grande dispersion des données. La moyenne de 19130 consultations, nettement supérieure à la médiane de 9127, indique la présence de valeurs extrêmes tirant la distribution vers le haut. Cette asymétrie est confirmée par l'écart considérable entre le minimum de 1037 et le maximum de 765833 consultations par an.

La moitié des médecins généralistes effectuent entre 5993 et 17290 consultations annuellement, ce qui suggère une variabilité importante dans la charge de travail. La médiane de 9127 consultations par an, équivalant à environ 25 consultations par jour ouvrable, semble plus représentative de l'activité typique d'un médecin généraliste que la moyenne influencée par les valeurs extrêmes. Ces statistiques mettent en lumière la diversité des pratiques et des charges de travail parmi les médecins généralistes, avec potentiellement quelques cas atypiques présentant un volume de consultations exceptionnellement élevé.

Le nombre de visites pouvant potentiellement être influencé par la taille de la commune et donc par sa population, nous avons éliminé cet effet en calculant le taux de consultations qui n'est autre que le nombre de consultations moyennes par personnes.

2. Taux de mortalité et de Natalité

Dans les communes étudiées, le taux de natalité et de mortalité sont un peu élevées avec la plupart des taux variant entre 5 et 15 pour 1000 en ce qui concerne la natalité et 0 et 20 pour 1000 pour la mortalité. On remarque une corrélation négative entre ces deux taux. Néanmoins cette corrélation n'a à priori aucun sens. Par ailleurs, l'observation des distribution permet de constater que la natalité est de façon générale élevée par rapport à la mortalité dans les communes étudiées. En vue de mieux de mieux voir peut être l'effet de la mortalité sur la natalité, nous allons nous intéresser alors à une analyse de la corrélation entre les deux

taux par groupe d'âge. Nous avons considéré les groupes d'âge suivants : 0-24, 25-44, 45-60, 60 et plus en fonction des variables disponibles et aussi à partir de l'information sur l'âge des femmes en âge de procréer qui est de l'ordre de 25-45 et des personnes âgées dont l'âge est de plus 60 ans. Ne disposant pas du taux de mortalité dans chaque groupe, alors nous avons dans notre analyse opté plutôt pour le pourcentage des femmes de chaque groupe en partant du principe que la natalité est très souvent liée aux femmes et du fait que nous pouvons analyser une diminution du pourcentage comme étant dû à une mortalité. Ainsi sur la base de cette nouvelle hypothèse, voici nos nouveaux résultats

```
# Charger les bibliothèques nécessaires
library(ggplot2)
library(readr)
library(reshape2)
```

```
##
## Attachement du package : 'reshape2'

## L'objet suivant est masqué depuis 'package:tidyr':
##
## smiths
```

```
library(gridExtra)
```

```
# Créer les groupes d'âge en sommant les colonnes correspondantes
```

```
data$pop_femmes <- data$femmes_0_4 + data$femmes_5_9 + data$femmes_10_14 + data$femmes_15_19 + data$femmes_20_24 +
  data$femmes_25_29 + data$femmes_30_34 + data$femmes_35_39 + data$femmes_40_44 + data$femmes_45_49 +
  data$femmes_50_54 + data$femmes_55_59 + data$femmes_60_64 + data$femmes_65_69 + data$femmes_70_74 +
  data$femmes_75_79 + data$femmes_80_84 + data$femmes_85_89 + data$femmes_90_94 + data$femmes_95_plus
```

```
data$femmes_0_24 <- (data$femmes_0_4 + data$femmes_5_9 + data$femmes_10_14 + data$femmes_15_19 + data$femmes_20_24) /
data$femmes_25_45 <- (data$femmes_25_29 + data$femmes_30_34 + data$femmes_35_39 + data$femmes_40_44) /
data$femmes_45_60 <- (data$femmes_45_49 + data$femmes_50_54 + data$femmes_55_59) / data$pop_femmes
data$femmes_60_plus <- (data$femmes_60_64 + data$femmes_65_69 + data$femmes_70_74 + data$femmes_75_79 +
  data$femmes_80_84 + data$femmes_85_89 + data$femmes_90_94 + data$femmes_95_plus) / data$pop_femmes
```

```
# Extraire le taux de natalité
```

```
data$taux_natalite <- data$taux_de_natalite_annuel_moyen_2015_2021
```

```
# ---- Création des graphiques ----
```

```
g1 <- ggplot(data, aes(x = femmes_0_24, y = taux_natalite)) +
  geom_point(color = "blue", alpha = 0.6) +
  geom_smooth(method = "lm", se = TRUE, color = "blue") +
  labs(title = "0-24 ans",
       x = "Proportion de Femmes 0-24 ans",
       y = "Taux de Natalité") +
  theme_minimal()
```

```
g2 <- ggplot(data, aes(x = femmes_25_45, y = taux_natalite)) +
  geom_point(color = "red", alpha = 0.6) +
  geom_smooth(method = "lm", se = TRUE, color = "red") +
  labs(title = "25-45 ans",
       x = "Proportion de Femmes",
       y = "Taux de Natalité") +
  theme_minimal()
```

```

theme_minimal()

g3 <- ggplot(data, aes(x = femmes_45_60, y = taux_natalite)) +
  geom_point(color = "green", alpha = 0.6) +
  geom_smooth(method = "lm", se = TRUE, color = "green") +
  labs(title = "45-60 ans",
       x = "Proportion de Femmes",
       y = "Taux de Natalité") +
  theme_minimal()

g4 <- ggplot(data, aes(x = femmes_60_plus, y = taux_natalite)) +
  geom_point(color = "purple", alpha = 0.6) +
  geom_smooth(method = "lm", se = TRUE, color = "purple") +
  labs(title = "60+ ans",
       x = "Proportion de Femmes",
       y = "Taux de Natalité") +
  theme_minimal()

# ---- Affichage en grille (2x2) ----
grid.arrange(g1, g2, g3, g4, ncol = 2, nrow = 2)

```

```
## 'geom_smooth()' using formula = 'y ~ x'
```

```
## 'geom_smooth()' using formula = 'y ~ x'
```

```
## 'geom_smooth()' using formula = 'y ~ x'
```

```
## 'geom_smooth()' using formula = 'y ~ x'
```

Les résultats nous montrent un lien croissant pour les tranches d'âge 0-24 et 25-45 ans montrant ainsi que dans ces tranches d'âge si le pourcentage des femmes diminuent (quer l'on pourrait assimiler à une mort des femmes) alors le taux de natalité diminue. Par ailleurs ceux de la tranche 45-60 semble n'avoir aucun lien sur le taux de natalité. Enfin il a été constaté un lien négatif pour la tranche d'âge 60 ans et plus.

```

# Vérification et conversion des variables avant de tracer
# Vérification et conversion du taux de mortalité
if (class(df$taux_de_mortalite_annuel_moyen_2015_2021) != "numeric") {
  df$taux_de_mortalite_annuel_moyen_2015_2021 <- as.numeric(as.character(df$taux_de_mortalite_annuel_moyen_2015_2021))
}

# Vérification et conversion du taux de natalité
if (class(df$taux_de_natalite_annuel_moyen_2015_2021) != "numeric") {
  df$taux_de_natalite_annuel_moyen_2015_2021 <- as.numeric(as.character(df$taux_de_natalite_annuel_moyen_2015_2021))
}

# Histogramme du taux de mortalité
p1 <- ggplot(df, aes(x = taux_de_mortalite_annuel_moyen_2015_2021)) +
  geom_histogram(bins = 30, fill = "red", alpha = 0.7, color = "black") +
  labs(
    x = "Taux de mortalité",
    y = "Nombre de communes") +
  theme_minimal()

# Histogramme du taux de natalité

```

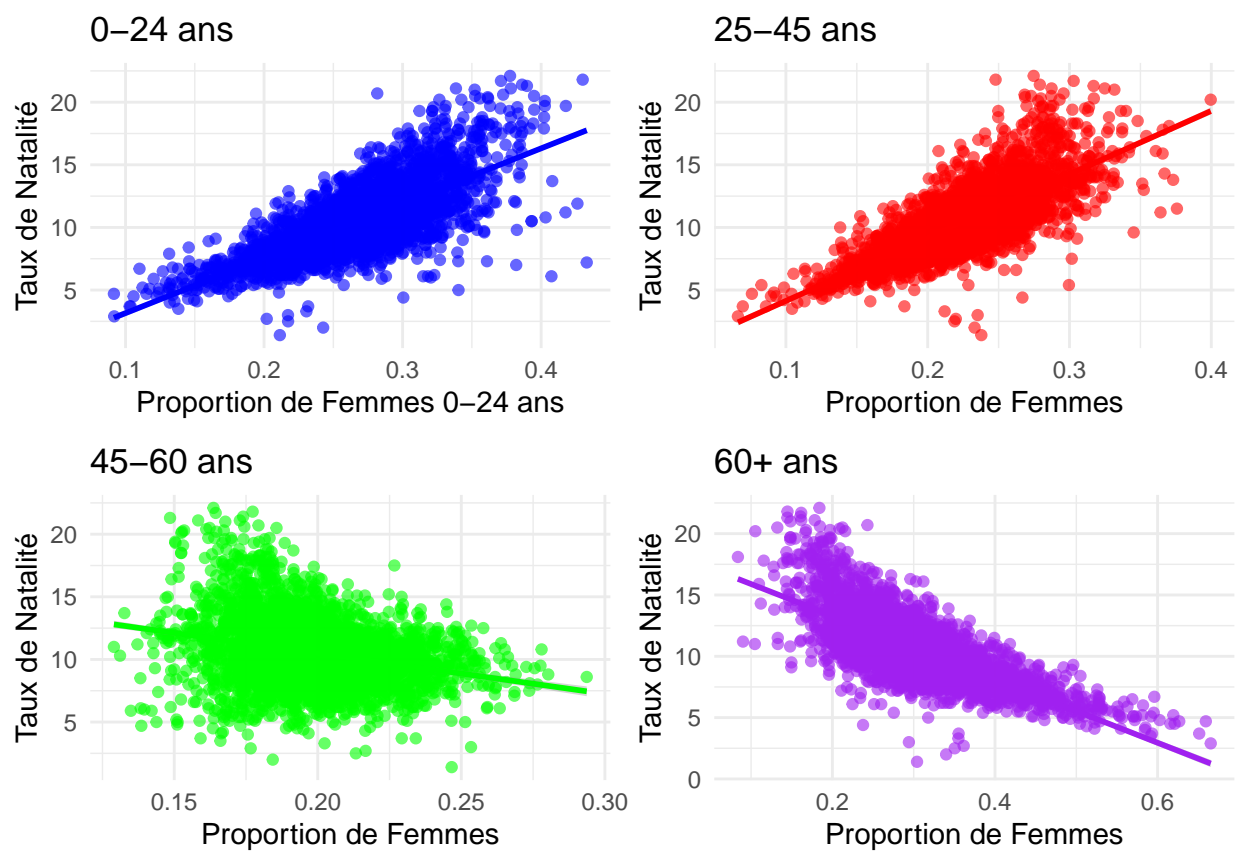



Figure 2: Taux de Natalité et Pourcentage des femmes dans chaque groupe

```

p2 <- ggplot(df, aes(x = taux_de_natalite_annuel_moyen_2015_2021)) +
  geom_histogram(bins = 30, fill = "blue", alpha = 0.7, color = "black") +
  labs(
    x = "Taux de natalité",
    y = "Nombre de communes") +
  theme_minimal()

# Nuage de points pour voir la relation entre mortalité et natalité
p3 <- ggplot(df, aes(x = taux_de_mortalite_annuel_moyen_2015_2021,
                     y = taux_de_natalite_annuel_moyen_2015_2021)) +
  geom_point(alpha = 0.7, color = "purple") +
  geom_smooth(method = "lm", color = "black", linetype = "dashed") + # Ajout d'une tendance linéaire
  labs(
    x = "Taux de mortalité",
    y = "Taux de natalité") +
  theme_minimal()

# Courbes de densité pour mieux voir la distribution
p4 <- ggplot(df) +
  geom_density(aes(x = taux_de_mortalite_annuel_moyen_2015_2021, fill = "Mortalité"), alpha = 0.5, color = "black") +
  geom_density(aes(x = taux_de_natalite_annuel_moyen_2015_2021, fill = "Natalité"), alpha = 0.5, color = "blue") +
  labs(
    x = "Taux",
    y = "Densité") +
  scale_fill_manual(values = c("Mortalité" = "red", "Natalité" = "blue")) +
  theme_minimal()

# Affichage de tous les graphiques ensemble
library(gridExtra)
grid.arrange(p1, p2, p3, p4, ncol = 2)

## 'geom_smooth()' using formula = 'y ~ x'

```

```

library(ggplot2)
library(patchwork)

# Premier graphique pour nb_visite
plot1 <- ggplot(data) +
  aes(x = nb_visite) +
  geom_histogram(bins = 30L, fill = "gray") +
  theme_minimal() +
  ylab("Nombre de consultations") +
  xlab("")

# Deuxième graphique pour taux_visites_19_ans_ou_plus
plot2 <- ggplot(data) +
  aes(x = taux_visites_19_ans_ou_plus) +
  geom_histogram(bins = 30L, fill = "gray") +
  theme_minimal() +
  ylab("Taux de consultations") +
  xlab("")

```

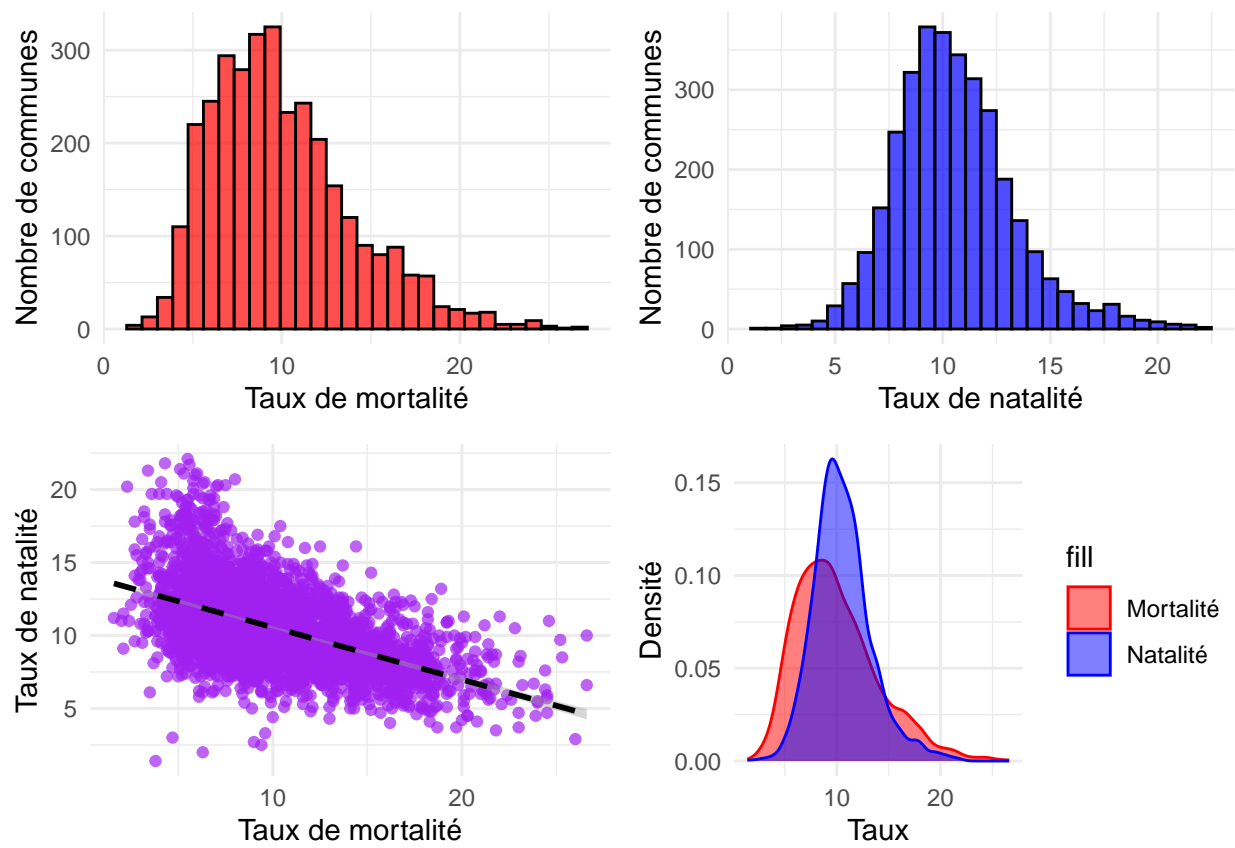


Figure 3: Taux de Natalité et Taux de Mortalité

```
# Combinaison des deux graphiques
plot1 + plot2
```

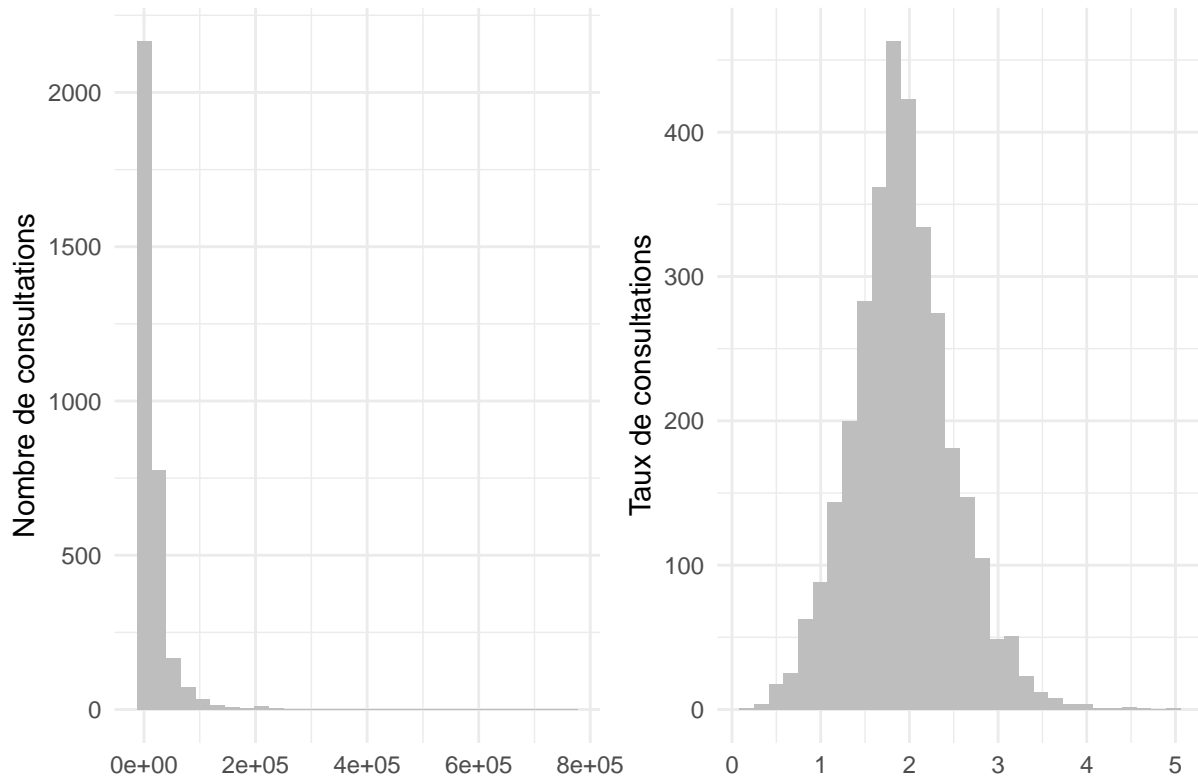


Figure 4: Répartition du nombre et du taux de consultations

Analyse bivariée

Nous allons ici, voir s'il y a un lien à priori entre le taux de consultation et certaines de nos variables explicatives. Ainsi, nous avons d'abord réalisé une analyse descriptive bivariée puis nous avons calculé la corrélation de Pearson pour évaluer le lien linéaire entre le taux de consultation et des variables telles que la population totale, la part des personnes âgées (75 ans et plus), la part de quelques CSP (ouvriers et retraités).

Taux de consultation et population totale

```
# Calcul des quantiles
quantiles <- quantile(data$population_municipale_2021_x, probs = c(1/3, 2/3), na.rm = TRUE)

# Création des classes avec les bornes des intervalles
data_pop <- data %>%
```

```

mutate(taille_commune = case_when(
  population_municipale_2021_x <= quantiles[1] ~ paste0("Petite (<= ", round(quantiles[1]), ")",
  population_municipale_2021_x <= quantiles[2] ~ paste0("Moyenne (", round(quantiles[1] + 1), " - ",
  TRUE ~ paste0("Grande (> ", round(quantiles[2]), ")",
)) %>%
group_by(taille_commune) %>%
summarise("Taux de consultations" = mean(taux_visites, na.rm = TRUE))

# Génération du tableau en LaTeX
data_pop %>% kable(format = "latex",
  booktabs = TRUE,
  caption = "Taux de consultations selon la taille de la commune") %>%
kable_styling(latex_options = c("striped",
  "HOLD_position"))

```

Table 2: Taux de consultations selon la taille de la commune

taille_commune	Taux de consultations
Grande (> 8974)	1.526810
Moyenne (4849 - 8974)	1.456356
Petite (<= 4848)	1.383861

En divisant les communes en trois groupes égaux (ou presque égaux) en fonction de la population totale, il ressort qu'en moyenne, plus la taille de la commune est importante plus le taux de consultations est élevé.

Taux de consultation et population âgée

```

# Calcul de la médiane
mediane <- median(data$nb_de_pers_agees_de_75_ans_ou_plus_2021, na.rm = TRUE)

# Création des classes avec les bornes des intervalles
data_age <- data %>%
  mutate(population_agee_importante = case_when(
    nb_de_pers_agees_de_75_ans_ou_plus_2021 <= mediane ~ paste0("Non (<= ", round(mediane), ")",
    TRUE ~ paste0("Oui (> ", round(mediane), ")",
  )) %>%
  group_by(population_agee_importante) %>%
  summarise(consultations_moyennes = mean(taux_visites, na.rm = TRUE))

# Génération du tableau en LaTeX
data_age %>% kable(format = "latex",
  booktabs = TRUE,
  caption = "Taux de consultations selon la population âgée") %>%
kable_styling(latex_options = c("striped",
  "HOLD_position"))

```

Table 3: Taux de consultations selon la population âgée

population_agee_importante	consultations_moyennes
Non (≤ 670)	1.501111
Oui (> 670)	1.410213

```
ggplot(data = data, aes(x = part_des_pers_agees_de_75_ans_ou_2021 , y = taux_visites_19_ans_ou_plus)) +
  geom_point(color = "blue", size = 3) + # Points bleus
  geom_smooth(method = "lm", se = TRUE, color = "red") + # Droite de tendance (modèle linéaire)
  labs(
    x = "Part des plus de 75 ans",
    y = "Taux de consultation"
  ) +
  theme_minimal()
```

```
## 'geom_smooth()' using formula = 'y ~ x'
```

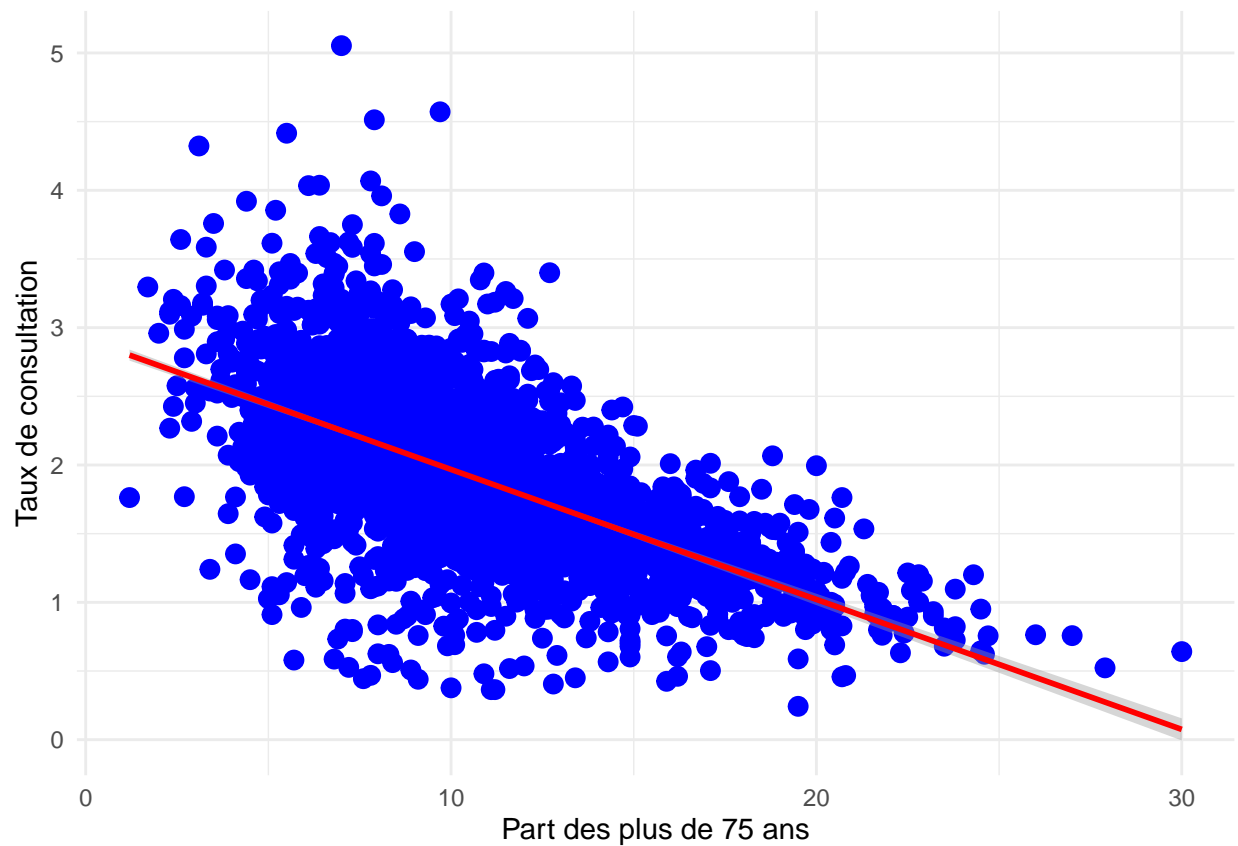


Figure 5: Relation entre taux de consultations et part des plus de 75 ans

Les communes avec une population âgée importante (communes dont la population âgée de 75 ans ou plus est supérieure à la médiane) ont en moyenne un taux de consultations plus faible.

Taux de consultation et CSP

Aucune catégorie ne semble montrer une relation linéaire évidente avec le taux de visite. Par ailleurs, pour toutes les catégories socio-professionnelles, la majorité des communes se situent dans une plage de proportions faibles, ce qui limite la variabilité observable dans les relations. Une analyse statistique supplémentaire, comme le calcul de corrélations, serait nécessaire pour confirmer ou infirmer les relations observées visuellement.

Analyse de corrélation

Les résultats de la corrélation de Pearson sont consignées dans le tableau suivant :

```
# Fonction de conversion
conversion_en_numeric <- function(data, columns) {
  resultat <- data %>%
    mutate(across(all_of(columns), as.numeric))
  return (resultat)
}

# Liste des variables à tester avec taux_de_consultation
variables <- c("population_municipale_2021_x", "part_des_pers_agees_de_75_ans_ou_2021",
              "population_de_15_ans_ou_selon_la_csp_2021_retraites", "population_de_15_ans_ou_selon_la_csp_2021_actifs")

data <- conversion_en_numeric(data, variables)

# Initialisation du tableau pour stocker les résultats
resultats <- data.frame(Variable = character(), Correlation = numeric(), P_value = numeric())

# Calcul de la corrélation pour chaque variable et tester la significativité
for (var in variables) {
  test <- cor.test(data$taux_visites_19_ans_ou_plus, data[[var]], method = "pearson")
  resultats <- rbind(resultats, data.frame(
    Variable = var,
    Correlation = test$estimate,
    P_value = test$p.value
  ))
}

# Format du tableau avec la significativité
resultats$Significatif <- ifelse(resultats$P_value < 0.05, "Oui", "Non")

# Génération du tableau en LaTeX
resultats %>% kable(format = "latex",
                  booktabs = TRUE,
                  caption = "Corrélations de Pearson entre le taux de consultation et les autres variables",
                  kable_styling(latex_options = c("striped",
                                                    "HOLD_position")))
```

Table 4: Corrélations de Pearson entre le taux de consultation et les autres variables

	Variable	Correlation	P_value	Significatif
cor	population municipale 2021_x	0.0765022	0.0000118	Oui
cor1	part_des_pers_agees_de_75_ans_ou_2021	-0.6258560	0.0000000	Oui
cor2	population_de_15_ans_ou_selon_la_csp_2021_retraites	-0.0285517	0.1024362	Non
cor3	population_de_15_ans_ou_selon_la_csp_2021_ouvriers	0.1077559	0.0000000	Oui

Les résultats nous montrent que le taux de consultation est positivement corrélé à la population ainsi qu'à celle de plus de 15 ans. Cependant la corrélation est faible. Par ailleurs, la corrélation est négative avec la part des personnes âgées de plus de 75 ans. Cela dit, plus la part des plus de 75 ans augmente moins est le taux de consultations dans une commune. Cela peut vouloir dire que les personnes de plus de 75 ans sont ceux qui ne se consultent pas assez.

```

# Sélectionner uniquement les variables d'intérêt
# Définition des noms lisibles pour les variables
nom_variables <- c(
  "taux_de_mortalite_annuel_moyen_2015_2021" = "Mortalité",
  "taux_de_natalite_annuel_moyen_2015_2021" = "Natalité",
  "part_des_familles_sans_enf_de_de_25_ans_2021" = "Sans enfants",
  "part_des_familles_avec_1_enf_de_de_25_ans_2021" = "Un enfant",
  "part_des_familles_avec_3_enf_ou_plus_de_de_25_ans_2021" = "Trois enfants",
  "nb_visite" = "Nombre de visites"
)

# Sélectionner les variables d'analyse
variables_analyse <- names(nom_variables)

df_analyse <- df[variables_analyse]

# Convertir toutes les colonnes en numérique
df_analyse <- df_analyse %>% mutate(across(everything(), as.numeric))

# Supprimer les valeurs manquantes
df_analyse <- na.omit(df_analyse)

# Calculer les corrélations
cor_matrix <- cor(df_analyse, use="complete.obs")

# Trier les corrélations par ordre décroissant
cor_target <- sort(cor_matrix["nb_visite", ], decreasing=TRUE)

# Remplacer les noms de variables par des noms plus lisibles
cor_data <- data.frame(
  Variable = names(cor_target),
  Correlation = cor_target
)

# Appliquer les nouveaux noms
cor_data$Variable <- nom_variables[cor_data$Variable]

# Exclure "Nombre de visites" du graphique

```



```
cor_data <- cor_data[cor_data$Variable != "Nombre de visites", ]

# Afficher le barplot des corrélations
ggplot(cor_data, aes(x = reorder(Variable, Correlation), y = Correlation, fill = Correlation)) +
  geom_bar(stat="identity") +
  coord_flip() +
  scale_fill_gradient2(low="blue", mid="white", high="red", midpoint=0) +
  labs(
    x="Variables",
    y="Coefficient de corrélation") +
  theme_minimal()
```

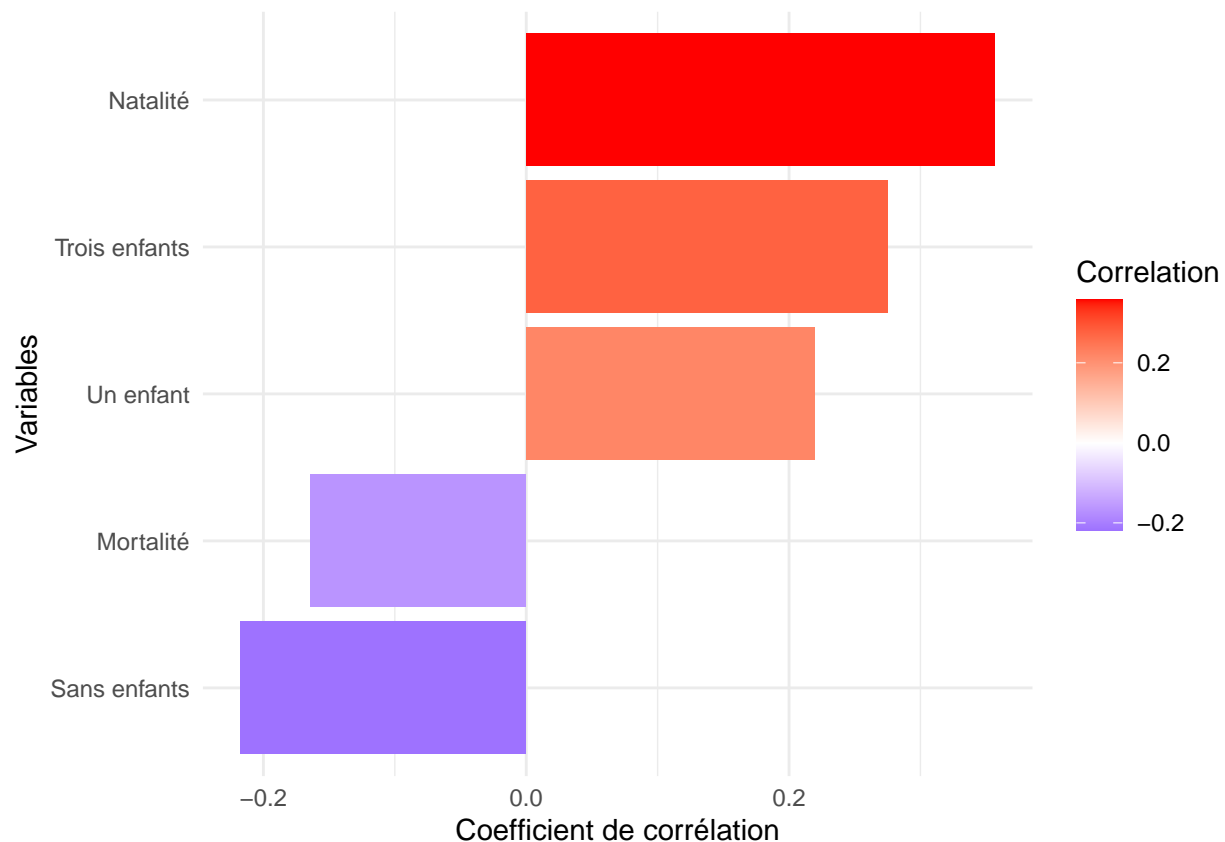


Figure 6: Corrélations entre le nombre de visite et quelques variables

Autocorrélation

L'autocorrélation spatiale est une mesure essentielle pour analyser la dépendance entre des observations géographiques. Dans notre étude nos données sont des données portant sur des communes. Ainsi il peut exister une dépendance entre nos taux de consultations du fait de la proximité des communes ou de l'appartenance à un même département ou région. Ainsi nous allons mesurer cette dépendance en évaluant l'autocorrélation spatiale. Dans ce contexte, l'**indice de Moran** est largement utilisé pour quantifier cette dépendance en fournissant une mesure globale de l'autocorrélation spatiale.

Définition de l'indice de Moran

L'indice de Moran (I) évalue la similitude des valeurs d'une variable entre différentes entités géographiques (par exemple, des communes) en fonction de leur proximité spatiale. Il se base sur la matrice de poids spatiale (W), qui définit les relations entre ces entités.

Formule de l'indice de Moran

La formule mathématique de l'indice de Moran est la suivante :

$$I = \frac{n}{\sum_{i=1}^n \sum_{j=1}^n w_{ij}} \cdot \frac{\sum_{i=1}^n \sum_{j=1}^n w_{ij} (x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Où :

- n : Nombre total d'entités spatiales (Ici, le nombre de communes).
- x_i, x_j : Valeurs observées de la variable pour les entités i et j (Ici le taux de consultations)
- \bar{x} : Moyenne de la variable x .
- w_{ij} : Poids spatial définissant la relation entre i et j .

La matrice de W peut être construite sur la base du voisinage entre les deux communes ou soit de la distance entre les deux communes. Dans le premier cas alors $w_{ij} = 1$ si i et j sont voisins et $w_{ij} = 0$ sinon. Dans le second cas $w_{ij} = d_{ij}$. Nous allons dans notre cas utiliser une matrice de poids basée sur la distance, notamment celle d'Haversine.

Matrice de poids basée sur la distance de Haversine

La distance de Haversine est une mesure de la distance entre deux points sur une sphère, basée sur leurs coordonnées géographiques (*latitude* et *longitude*). Elle est particulièrement utile pour les données géographiques projetées sur une surface sphérique, comme la Terre.

Formule de la distance de Haversine

Si l'on considère deux points (i) et (j), la distance (d_{ij}) entre ces deux points sur la surface d'une sphère de rayon (r) est donnée par :

$$d_{ij} = 2r \cdot \arcsin \left(\sqrt{\sin^2 \left(\frac{\phi_j - \phi_i}{2} \right) + \cos(\phi_i) \cos(\phi_j) \sin^2 \left(\frac{\lambda_j - \lambda_i}{2} \right)} \right)$$

Où :

- r : Rayon de la Terre (environ 6371 km).
- ϕ_i, ϕ_j : Latitudes des points i et j (en radians).
- λ_i, λ_j : Longitudes des points i et j (en radians). Après calcul nous avons ces statistiques sur nos distances.

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]


```
ggplot(dist_df, aes(x = Distance)) +
  geom_density(fill = "blue", alpha = 0.4) +
  theme_minimal() +
  labs(x = "Distance", y = "Densité")
```

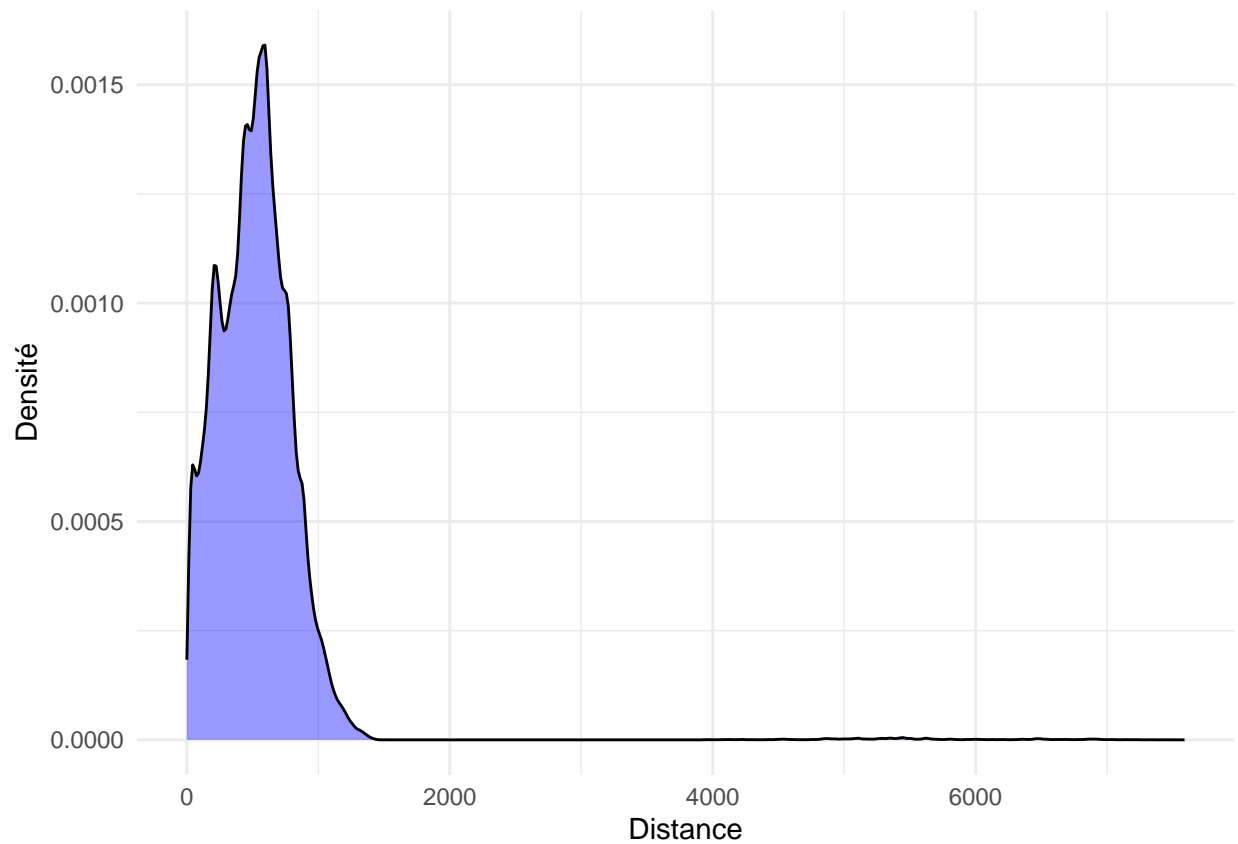


Figure 7: Densité des distances

Construction de la matrice de poids

Pour construire la matrice de poids, nous avons alors suivi ces étapes.

1. Calculer les distances de Haversine entre chaque paire d'entités.
2. Définir un seuil de distance maximale (d_{max}) :
 - Si $d_{ij} < d_{max}$, $w_{ij} = \frac{1}{d_{ij}}$.
 - Sinon, $w_{ij} = 0$.
3. Normaliser les poids pour que chaque ligne de la matrice ait une somme égale à 1 :

$$w_{ij}^{norm} = \frac{w_{ij}}{\sum_j w_{ij}}.$$

```

# Créer la matrice de poids (inverse des distances)
weight_matrix <- 1 / dist_matrix
diag(weight_matrix) <- 0 # Aucun poids pour soi-même

# Gérer les cas où les distances sont nulles ou infinies
weight_matrix[is.infinite(weight_matrix)] <- 0

# Créer l'objet spatial de pondération
W <- mat2listw(weight_matrix, style = "W")

# Vérifier que la variable à analyser existe et ne contient pas de NA
if (!("taux_visites_19_ans_ou_plus" %in% names(data))) {
  stop("La colonne 'taux_visites_19_ans_ou_plus' doit exister dans la base de données.")
}
values <- data$taux_visites_19_ans_ou_plus

if (anyNA(values)) {
  stop("La colonne 'taux_visites_19_ans_ou_plus' ne doit pas contenir de valeurs manquantes.")
}

# Calcul de l'indice de Moran
moran_result <- moran.test(values, W, zero.policy = TRUE)

# Génération du tableau en LaTeX
moran_result[["estimate"]] %>% kable(format = "latex",
                                     booktabs = TRUE,
                                     caption = "Résultats du test de Moran") %>%
  kable_styling(latex_options = c("striped",
                                   "HOLD_position"))

```

Table 5: Résultats du test de Moran

	x
Moran I statistic	0.1275832
Expectation	-0.0003056
Variance	0.0000029

Ainsi dans notre étude, nous avons trouvé un indice de Moran égale à 0.1275832. Le test nous a permis d'obtenir une p-value de 0. Ce qui permet de conclure qu'il y a effectivement une autocorrélation positive et significative entre les communes selon leur taux de consultations.