R Notebook tp2

```
data <- read.csv("C:/Users/ASUS/Documents/Programme/M1_TNSID/s8/data_mining/tp/tp2/Achats.csv", sep = "

# Afficher un aperçu des données
head(data)
```

```
##   ID Genre Age Profession Famille Revenus Achats
## 1  1     M  19   Services       4       1     39
## 2  2     M  21 Ingenierie       3       2     81
## 3  3   Mme  20 Ingenierie       1       3      6
## 4  4   Mme  23      Droit       2       3     77
## 5  5   Mme  31    Loisirs       6       2     40
## 6  6   Mme  22    Culture       2       3     76
```

```
#--------------analyse préalable des données----------------------------
library(ggplot2)
# Obtenir des informations sur les données
str(data)
```

```
## 'data.frame':    2000 obs. of  7 variables:
##  $ ID        : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Genre     : chr  "M" "M" "Mme" "Mme" ...
##  $ Age       : int  19 21 20 23 31 22 35 23 64 30 ...
##  $ Profession: chr  "Services" "Ingenierie" "Ingenierie" "Droit" ...
##  $ Famille   : int  4 3 1 2 6 2 3 3 3 4 ...
##  $ Revenus   : int  1 2 3 3 2 3 2 3 3 3 ...
##  $ Achats    : int  39 81 6 77 40 76 6 94 3 72 ...
```

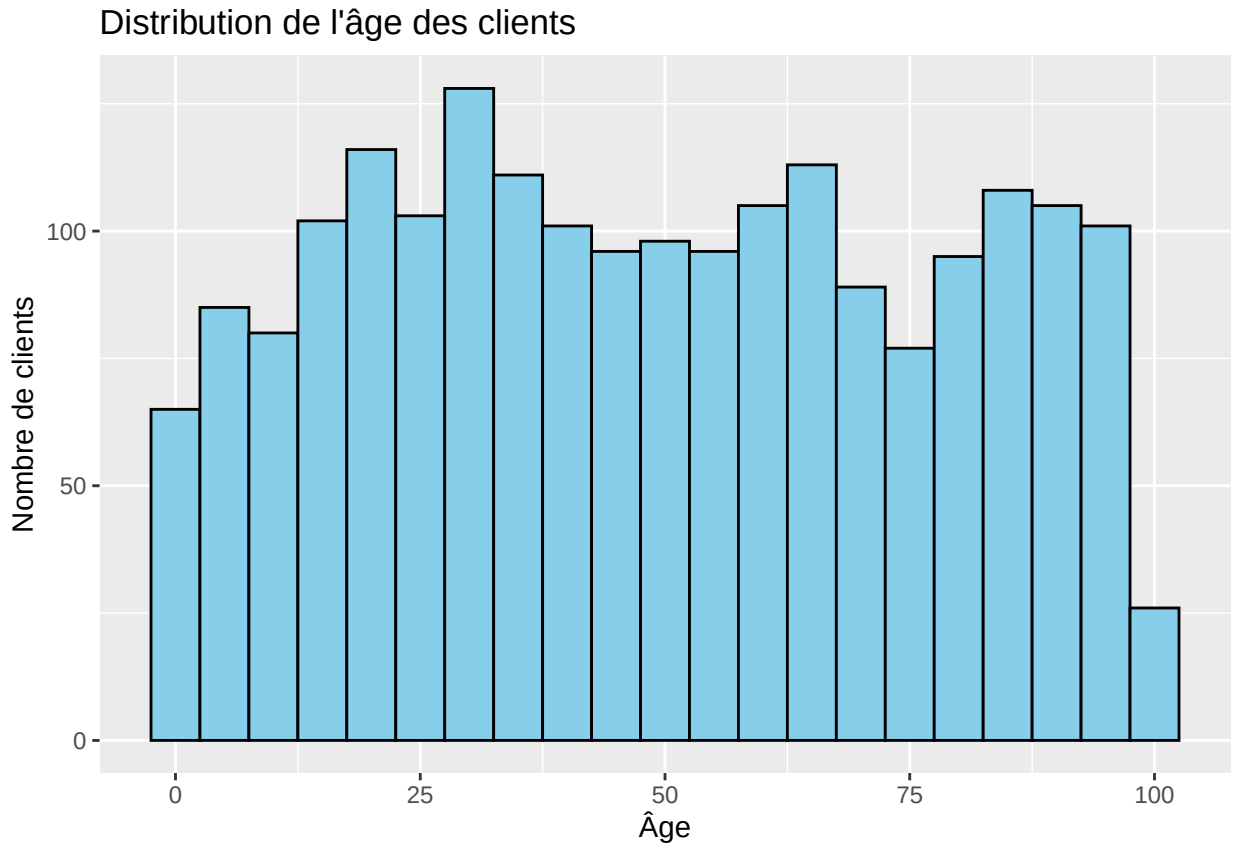
```
# Statistiques descriptives
summary(data)
```

```
##        ID            Genre                Age         Profession
##  Min.   :   1.0   Length:2000        Min.   : 0.00   Length:2000
##  1st Qu.: 500.8   Class :character   1st Qu.:25.00   Class :character
##  Median :1000.5   Mode  :character   Median :48.00   Mode  :character
##  Mean   :1000.5                      Mean   :48.96
##  3rd Qu.:1500.2                      3rd Qu.:73.00
##  Max.   :2000.0                      Max.   :99.00
##     Famille         Revenus         Achats
##  Min.   :1.000   Min.   :1.00   Min.   :  0.00
##  1st Qu.:2.000   1st Qu.:3.00   1st Qu.: 28.00
##  Median :4.000   Median :5.00   Median : 50.00
##  Mean   :3.768   Mean   :4.01   Mean   : 50.96
##  3rd Qu.:5.000   3rd Qu.:5.00   3rd Qu.: 75.00
##  Max.   :9.000   Max.   :5.00   Max.   :100.00
```

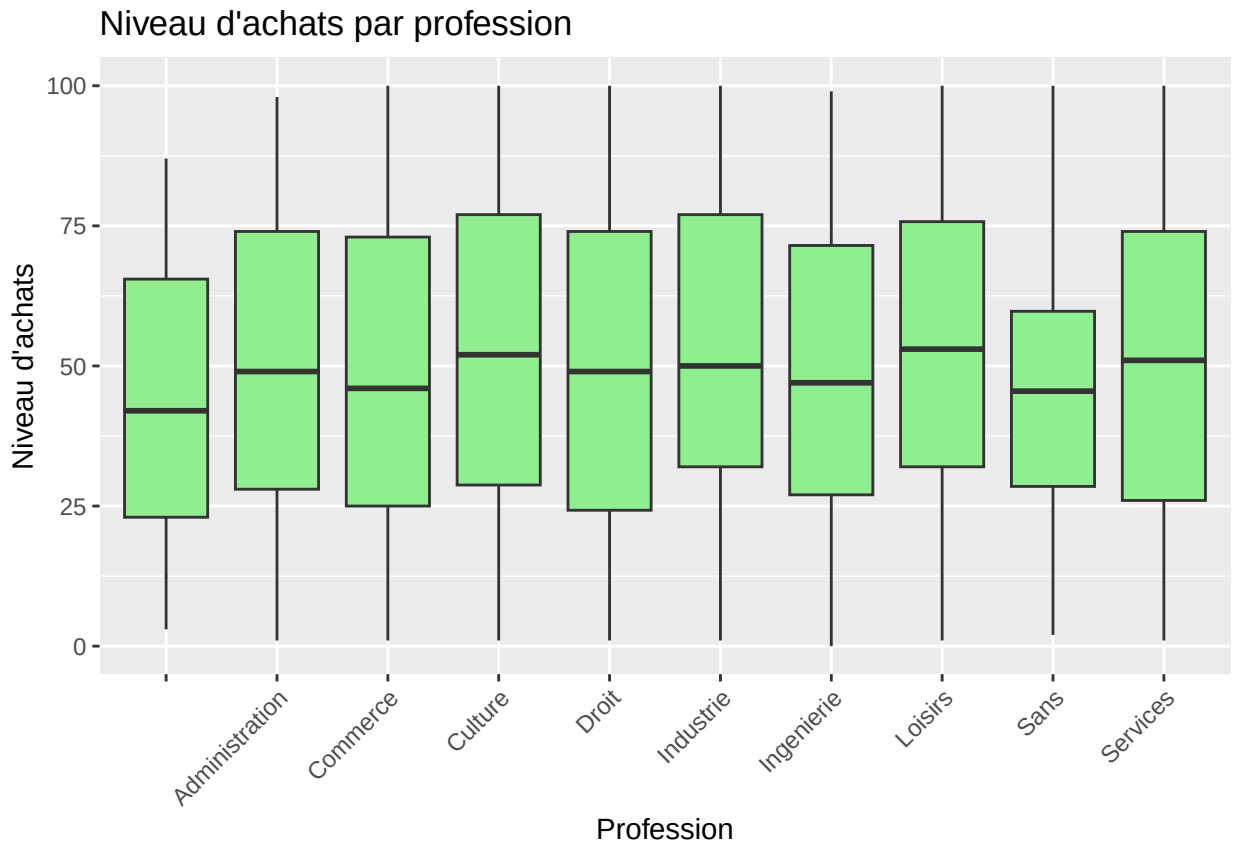
```r
# Vérifier les valeurs manquantes
colSums(is.na(data))
```

```
##         ID      Genre        Age Profession    Famille    Revenus     Achats
##          0          0          0          0          0          0          0
```

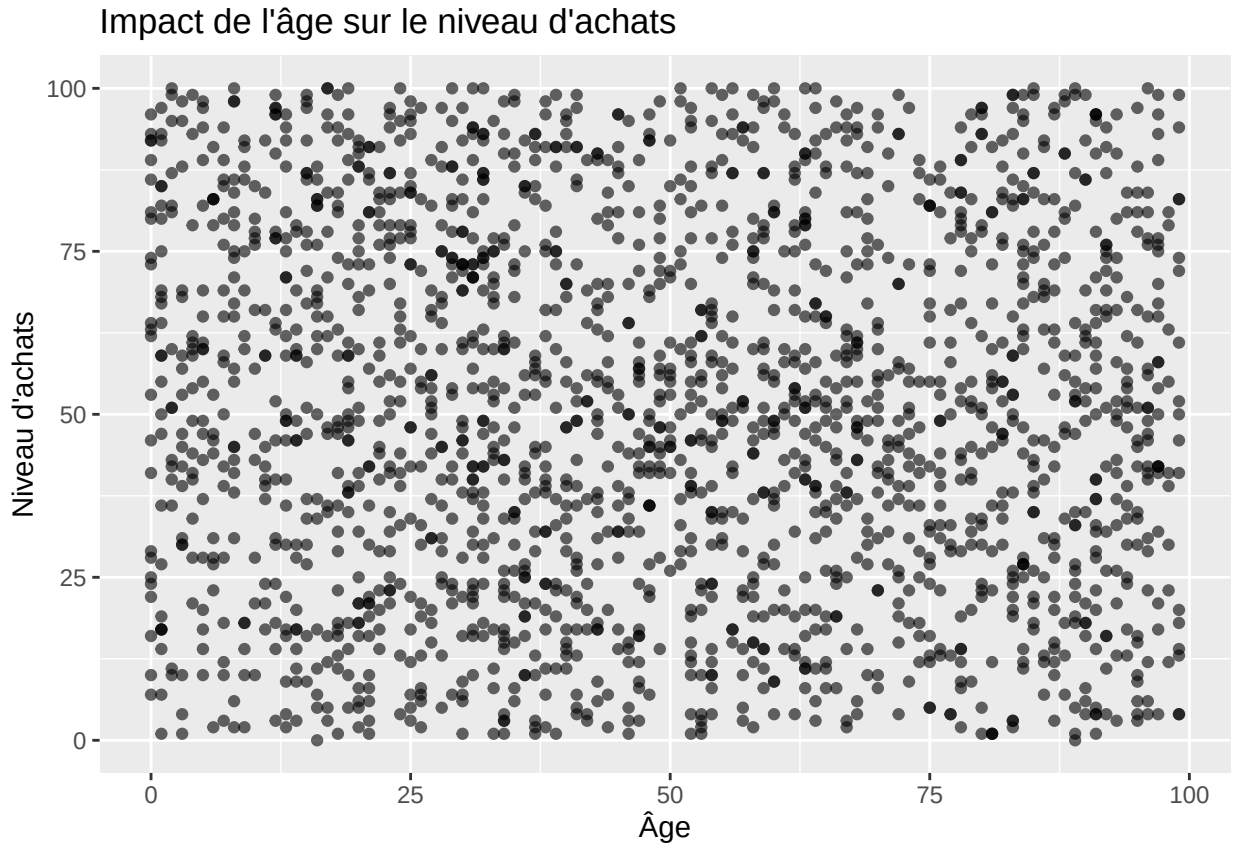
```r
# Distribution de l'âge des clients
ggplot(data, aes(x = Age)) +
  geom_histogram(binwidth = 5, fill = "skyblue", color = "black") +
  labs(title = "Distribution de l'âge des clients", x = "Âge", y = "Nombre de clients")
```



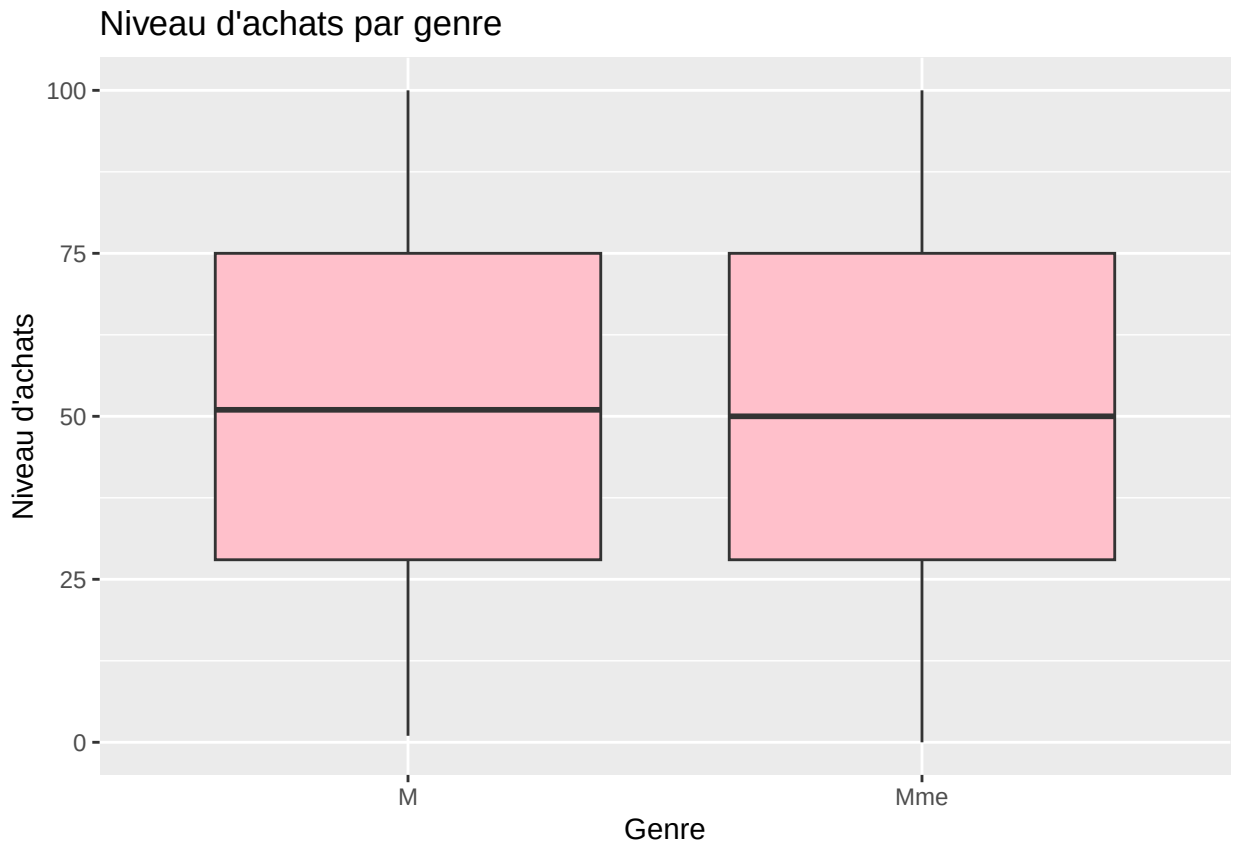

```r
# Niveau d'achats par profession
ggplot(data, aes(x = Profession, y = Achats)) +
  geom_boxplot(fill = "lightgreen") +
  labs(title = "Niveau d'achats par profession", x = "Profession", y = "Niveau d'achats") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```
# Impact de l'âge sur le niveau d'achats
ggplot(data, aes(x = Age, y = Achats)) +
  geom_point(alpha = 0.6) +
  labs(title = "Impact de l'âge sur le niveau d'achats", x = "Âge", y = "Niveau d'achats")
```




```r
# Niveau d'achats par genre
ggplot(data, aes(x = Genre, y = Achats)) +
  geom_boxplot(fill = "pink") +
  labs(title = "Niveau d'achats par genre", x = "Genre", y = "Niveau d'achats")
```



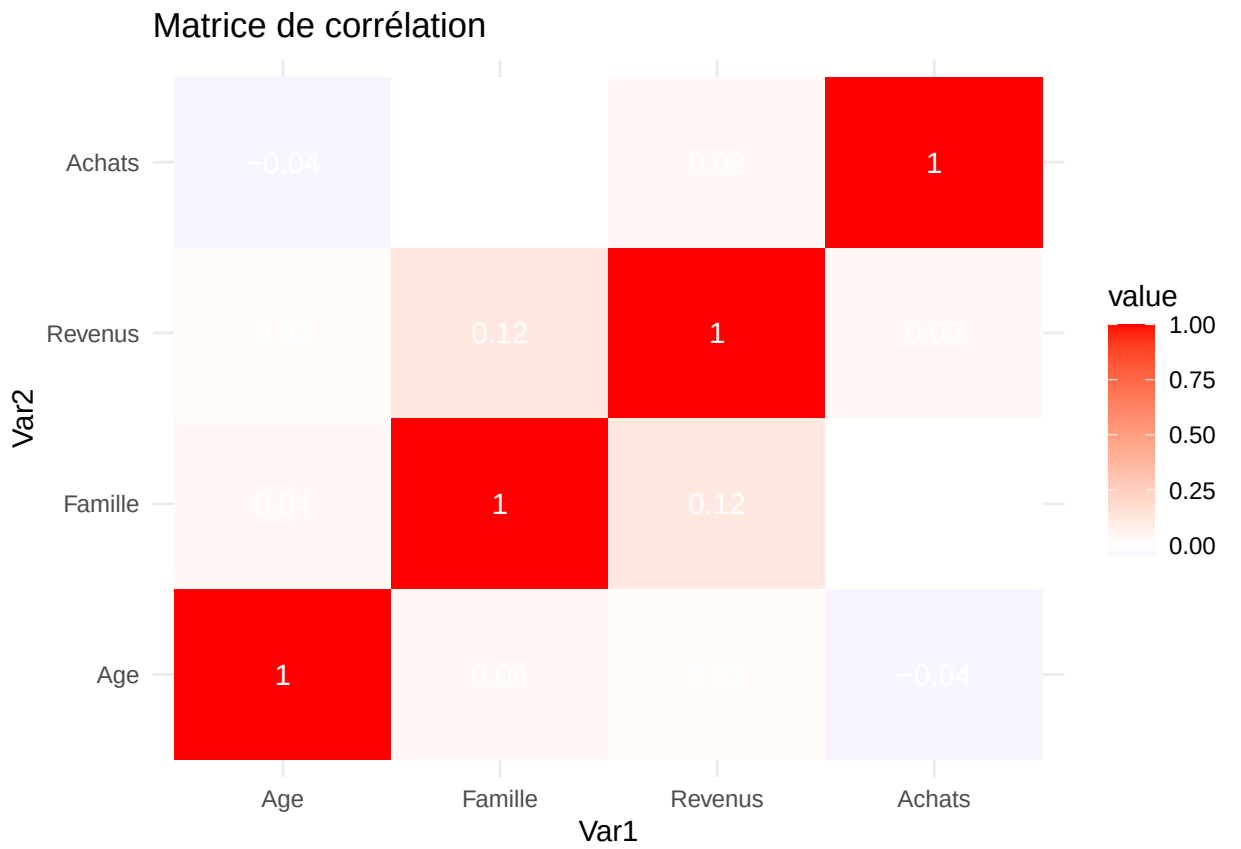

```
#----- etude de correlation ----------------
library(reshape2)


# Corrélation entre les variables numériques
cor_matrix <- cor(data[, c("Age", "Famille", "Revenus", "Achats")], use = "complete.obs")

# Afficher la matrice de corrélation
cor_matrix
```

```
##                Age     Famille    Revenus       Achats
## Age      1.00000000 0.038254376 0.02138729 -0.041798197
## Famille  0.03825438 1.000000000 0.12136851  0.002232067
## Revenus  0.02138729 0.121368507 1.00000000  0.034921385
## Achats  -0.04179820 0.002232067 0.03492138  1.000000000
```

```
# Visualiser la matrice de corrélation
corr_melt <- melt(cor_matrix)
ggplot(corr_melt, aes(x = Var1, y = Var2, fill = value)) +
  geom_tile() +
  geom_text(aes(label = round(value, 2)), color = "white") +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white", midpoint = 0) +
  labs(title = "Matrice de corrélation") +
  theme_minimal()
```




```
#-----------traitement de donnée (version 3)-----------------------
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
```

```
# Supprimer les colonnes inutiles et traiter les valeurs manquantes
data <- data %>%
  select(-ID) %>%
  mutate(
    Profession = ifelse(is.na(Profession), "Autre", Profession),
    Genre = as.factor(Genre),
    Profession = as.factor(Profession)
  )

# Supprimer les valeurs aberrantes pour "Achats"
```

```r
Q1 <- quantile(data$Achats, 0.25, na.rm = TRUE)
Q3 <- quantile(data$Achats, 0.75, na.rm = TRUE)
IQR <- Q3 - Q1
data <- data %>%
  filter(Achats >= (Q1 - 1.5 * IQR) & Achats <= (Q3 + 1.5 * IQR))

# Remplacer les valeurs manquantes après suppression des aberrations
data <- data %>%
  mutate(
    Age = ifelse(is.na(Age), median(Age, na.rm = TRUE), Age),
    Achats = ifelse(is.na(Achats), mean(Achats, na.rm = TRUE), Achats)
  )

# Supprimer les doublons
data <- distinct(data)

# Gestion finale des valeurs NA en cas de nouvelles introductions
data <- data %>%
  mutate(
    Age = ifelse(is.na(Age), mean(Age, na.rm = TRUE), Age),
    Achats = ifelse(is.na(Achats), median(Achats, na.rm = TRUE), Achats)
  )

# Assurer que le résultat final est un data frame
data <- as.data.frame(data)

# Vérifier les données finales
str(data)
```

```
## 'data.frame':    2000 obs. of  6 variables:
##  $ Genre     : Factor w/ 2 levels "M","Mme": 1 1 2 2 2 2 2 2 1 2 ...
##  $ Age       : int  19 21 20 23 31 22 35 23 64 30 ...
##  $ Profession: Factor w/ 10 levels "","Administration",..: 10 7 7 5 8 4 10 10 7 4 ...
##  $ Famille   : int  4 3 1 2 6 2 3 3 3 4 ...
##  $ Revenus   : int  1 2 3 3 2 3 2 3 3 3 ...
##  $ Achats    : int  39 81 6 77 40 76 6 94 3 72 ...
```

```r
summary(data)
```

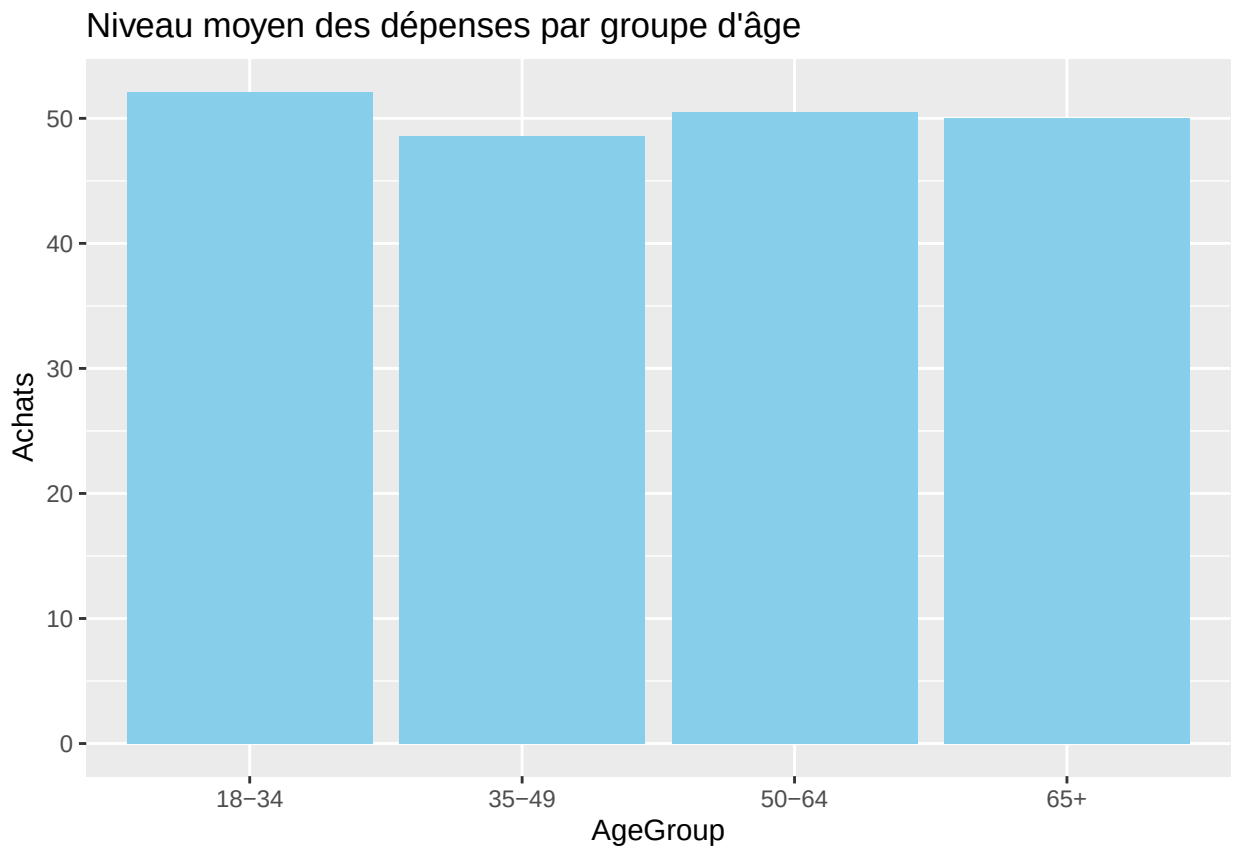
```
##  Genre           Age                Profession     Famille         Revenus
##  M  : 814   Min.   : 0.00   Culture       :612   Min.   :1.000   Min.   :1.00
##  Mme:1186   1st Qu.:25.00   Services      :339   1st Qu.:2.000   1st Qu.:3.00
##             Median :48.00   Loisirs       :234   Median :4.000   Median :5.00
##             Mean   :48.96   Ingenierie    :179   Mean   :3.768   Mean   :4.01
##             3rd Qu.:73.00   Industrie     :161   3rd Qu.:5.000   3rd Qu.:5.00
##             Max.   :99.00   Administration:153   Max.   :9.000   Max.   :5.00
##                             (Other)       :322
##      Achats
##  Min.   :  0.00
##  1st Qu.: 28.00
##  Median : 50.00
##  Mean   : 50.96
##  3rd Qu.: 75.00
```

```
##  Max.   :100.00
##
```

```r
#2. Est-ce que l'âge a un impact important sur le niveau de dépense ?

# Création de groupes d'âge
data$AgeGroup <- cut(data$Age, breaks = c(18, 35, 50, 65, 100), labels = c("18-34", "35-49", "50-64", "
# Calcul de la moyenne des achats par groupe d'âge
average_spending_by_age <- aggregate(Achats ~ AgeGroup, data = data, mean)

# Visualisation
ggplot(average_spending_by_age, aes(x = AgeGroup, y = Achats)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  labs(title = "Niveau moyen des dépenses par groupe d'âge")
```




```r
# ANOVA pour comparer les niveaux d'achats entre les groupes d'âge
anova_age <- aov(Achats ~ AgeGroup, data = data)
summary(anova_age)
```

```
##               Df  Sum Sq Mean Sq F value Pr(>F)
## AgeGroup       3    2261   753.8   0.982    0.4
## Residuals   1641 1259090   767.3
## 355 observations effacées parce que manquantes
```

Interprétation des Résultats de l'ANOVA

Les résultats de l'ANOVA fournissent une analyse statistique de ces observations :

- **Df (degrés de liberté) :** AgeGroup a 3 degrés de liberté (4 groupes - 1), et il y a 1641 degrés de liberté pour les résidus (total des observations moins le nombre de groupes).
- **Sum Sq (somme des carrés) :** La somme des carrés due aux groupes d'âge est 2261, tandis que la somme des carrés des résidus est 1259090, ce qui montre que la majorité de la variation dans les données n'est pas expliquée par les groupes d'âge.
- **Mean Sq (moyenne des carrés) :** Moyenne des carrés pour les groupes d'âge est 753.8 contre 767.3 pour les résidus, ce qui indique que la variation entre les groupes n'est pas beaucoup plus grande que celle à l'intérieur des groupes.
- **F value :** La valeur F de 0.982 indique le rapport entre la variance expliquée par les groupes et la variance non expliquée. Une valeur F proche de 1 suggère que les groupes n'expliquent pas significativement plus de variance que le hasard.
- **Pr(>F) :** La p-value est de 0.4, bien au-dessus du seuil habituel de 0.05, indiquant qu'il n'y a pas de différence statistiquement significative dans les dépenses moyennes entre les groupes d'âge.

Conclusion

L'analyse suggère que l'âge **n'a pas d'impact important** sur le niveau de dépense. Le graphique montre des niveaux de dépense moyennement homogènes à travers les âges, et l'ANOVA confirme que ces différences ne sont pas statistiquement significatives.

```
#----------------------------------ANOVA-------------------------------------
#3 Existe-t-il des liens entre niveaux de dépense et taille du foyer ?

# Traiter les valeurs manquantes et autres prétraitements
data <- data %>%
  mutate(Famille = as.factor(Famille))  # Assurez-vous que "Famille" est un facteur

# Test ANOVA pour évaluer les différences entre les groupes
anova_model <- aov(Achats ~ Famille, data = data)
anova_results <- summary(anova_model)  # Résultats de l'ANOVA
anova_results  # Afficher les résultats
```

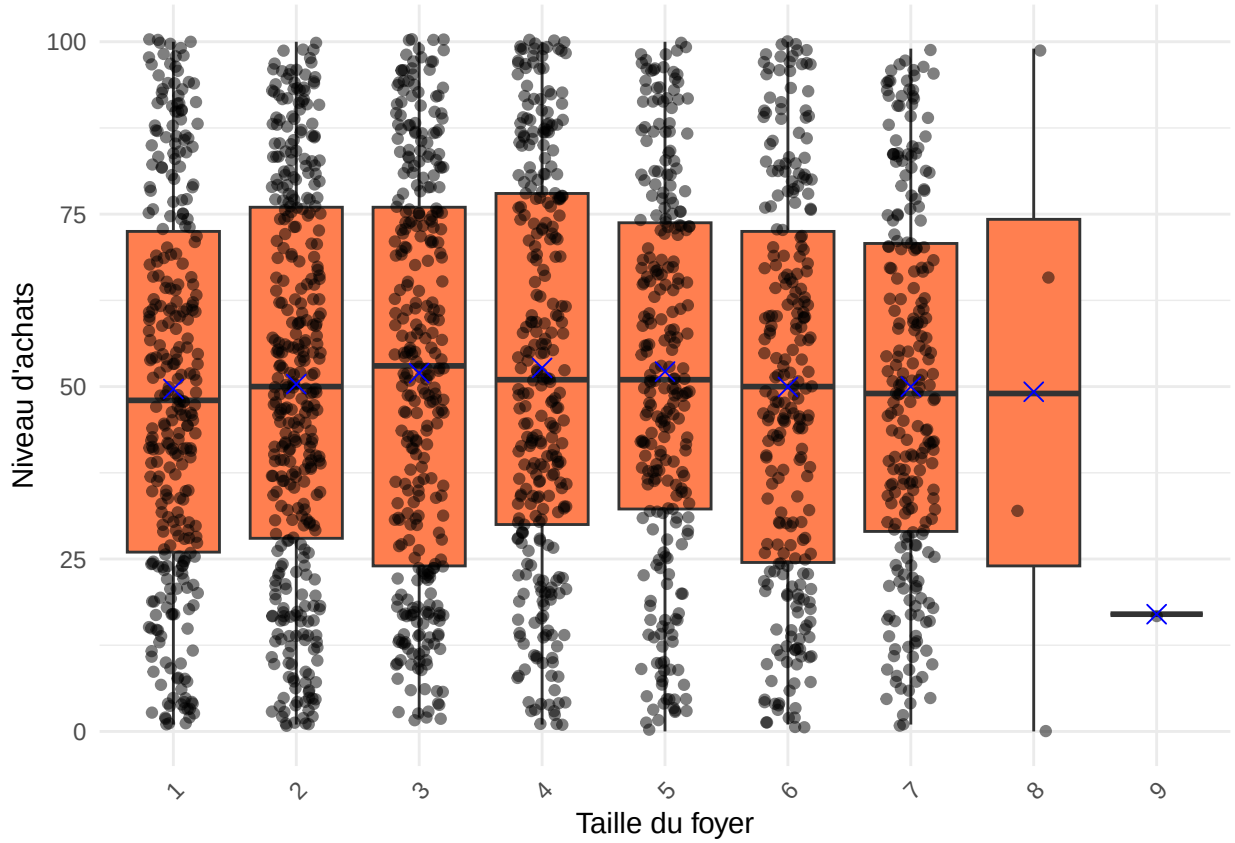
```
##               Df  Sum Sq Mean Sq F value Pr(>F)
## Famille        8    3855   481.8   0.617  0.765
## Residuals   1991 1556055   781.5
```

```
# Test de Tukey pour identifier les paires de groupes significatives
tukey_results <- TukeyHSD(anova_model)
tukey_results  # Afficher les résultats du test de Tukey
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Achats ~ Famille, data = data)
##
```

```
## $Famille
##              diff         lwr       upr     p adj
## 2-1   0.71739594   -6.070647  7.505439 0.9999962
## 3-1   2.30628354   -4.724617  9.337184 0.9841889
## 4-1   3.06385761   -4.097015 10.224731 0.9228809
## 5-1   2.53280962   -4.843582  9.909201 0.9788639
## 6-1   0.29690188   -7.200706  7.794510 1.0000000
## 7-1   0.34596804   -7.230764  7.922700 1.0000000
## 8-1  -0.39548495  -44.089057 43.298087 1.0000000
## 9-1 -32.64548495 -119.598943 54.307973 0.9634664
## 3-2   1.58888760   -5.127143  8.304918 0.9982966
## 4-2   2.34646167   -4.505517  9.198440 0.9792159
## 5-2   1.81541369   -5.261497  8.892324 0.9969791
## 6-2  -0.42049406   -7.623662  6.782674 1.0000000
## 7-2  -0.37142789   -7.656918  6.914062 1.0000000
## 8-2  -1.11288089  -44.756892 42.531130 1.0000000
## 9-2 -33.36288089 -120.291445 53.565684 0.9583678
## 4-3   0.75757407   -6.335073  7.850221 0.9999959
## 5-3   0.22652608   -7.083651  7.536703 1.0000000
## 6-3  -2.00938166   -9.441855  5.423092 0.9956483
## 7-3  -1.96031550   -9.472599  5.551968 0.9965958
## 8-3  -2.70176849  -46.384211 40.980674 0.9999999
## 9-3 -34.95176849 -121.899634 51.996097 0.9455408
## 5-4  -0.53104799   -7.966317  6.904221 0.9999998
## 6-4  -2.76695573  -10.322497  4.788586 0.9685441
## 7-4  -2.71788957  -10.351954  4.916175 0.9735287
## 8-4  -3.45934256  -47.162893 40.244208 0.9999996
## 9-4 -35.70934256 -122.667815 51.249130 0.9385502
## 6-5  -2.23590774   -9.996015  5.524199 0.9932767
## 7-5  -2.18684158  -10.023422  5.649739 0.9945905
## 8-5  -2.92829457  -46.667674 40.811085 0.9999999
## 9-5 -35.17829457 -122.154780 51.798191 0.9436055
## 7-6   0.04906616   -7.901718  7.999850 1.0000000
## 8-6  -0.69238683  -44.452372 43.067599 1.0000000
## 9-6 -32.94238683 -119.929236 54.044462 0.9615301
## 8-7  -0.74145299  -44.515064 43.032158 1.0000000
## 9-7 -32.99145299 -119.985158 54.002252 0.9612068
## 9-8 -32.25000000 -129.304758 64.804758 0.9828262
```

```
# Visualisation améliorée avec ggplot2
ggplot(data, aes(x = Famille, y = Achats)) +
  geom_boxplot(fill = "coral") +
  geom_jitter(width = 0.2, alpha = 0.5) +  # Ajouter des points pour voir la dispersion
  stat_summary(fun = "mean", geom = "point", shape = 4, size = 3, color = "blue") +  # Ajouter la moyen
  labs(x = "Taille du foyer", y = "Niveau d'achats") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Interprétation du Graphique de Boxplot

Le graphique montre la distribution des dépenses pour différentes tailles de foyer, allant de 1 à 9 membres par foyer. Les éléments clés à observer sont :

- **Médianes** : Les lignes à l'intérieur des boîtes qui représentent la médiane des dépenses pour chaque taille de foyer. Elles semblent assez constantes à travers les groupes, suggérant que la médiane des dépenses ne varie pas significativement avec la taille du foyer.
- **Étendue des dépenses** : Les boîtes montrent l'étendue interquartile (le milieu 50% des données), et les whiskers (les lignes qui s'étendent à partir des boîtes) montrent la portée des données en dehors des quartiles. Certains groupes, notamment ceux avec des foyers plus grands (8 et 9 membres), ont des whiskers qui indiquent des valeurs extrêmes plus élevées ou plus basses.
- **Outliers** : Les points hors des whiskers indiquent les valeurs extrêmes. Leur présence dans presque tous les groupes suggère que des variations importantes existent au sein de chaque catégorie de taille de foyer.

Interprétation des Résultats de l'ANOVA

Les résultats de l'ANOVA révèlent les aspects suivants :

- **F value et Pr(>F)** : Une valeur F de 0.617 et une p-value de 0.765 indiquent qu'il n'y a pas de différence statistiquement significative dans les moyennes des dépenses entre les différentes tailles de foyer. Cela suggère que, globalement, la taille du foyer n'est pas un facteur déterminant dans le niveau de dépense des ménages.

Interprétation des Résultats du Test de Tukey

Le test de Tukey est utilisé pour identifier spécifiquement quelles paires de groupes de taille de foyer diffèrent de manière significative en termes de dépenses moyennes :

- **Différences entre les groupes** : Les intervalles de confiance et les p-values pour chaque paire de groupes (comme "2-1", "3-1", etc.) montrent que presque toutes les comparaisons ne sont pas statistiquement significatives (p adj > 0.05), avec des intervalles de confiance qui incluent zéro.
- **Significativité** : Le seul groupe avec une différence notable est celui qui compare les tailles de foyer 1 et 9, mais même là, la p-value ajustée n'est pas inférieure à 0.05, indiquant l'absence de preuve statistique que ces groupes dépensent différemment de manière significative.

Conclusion

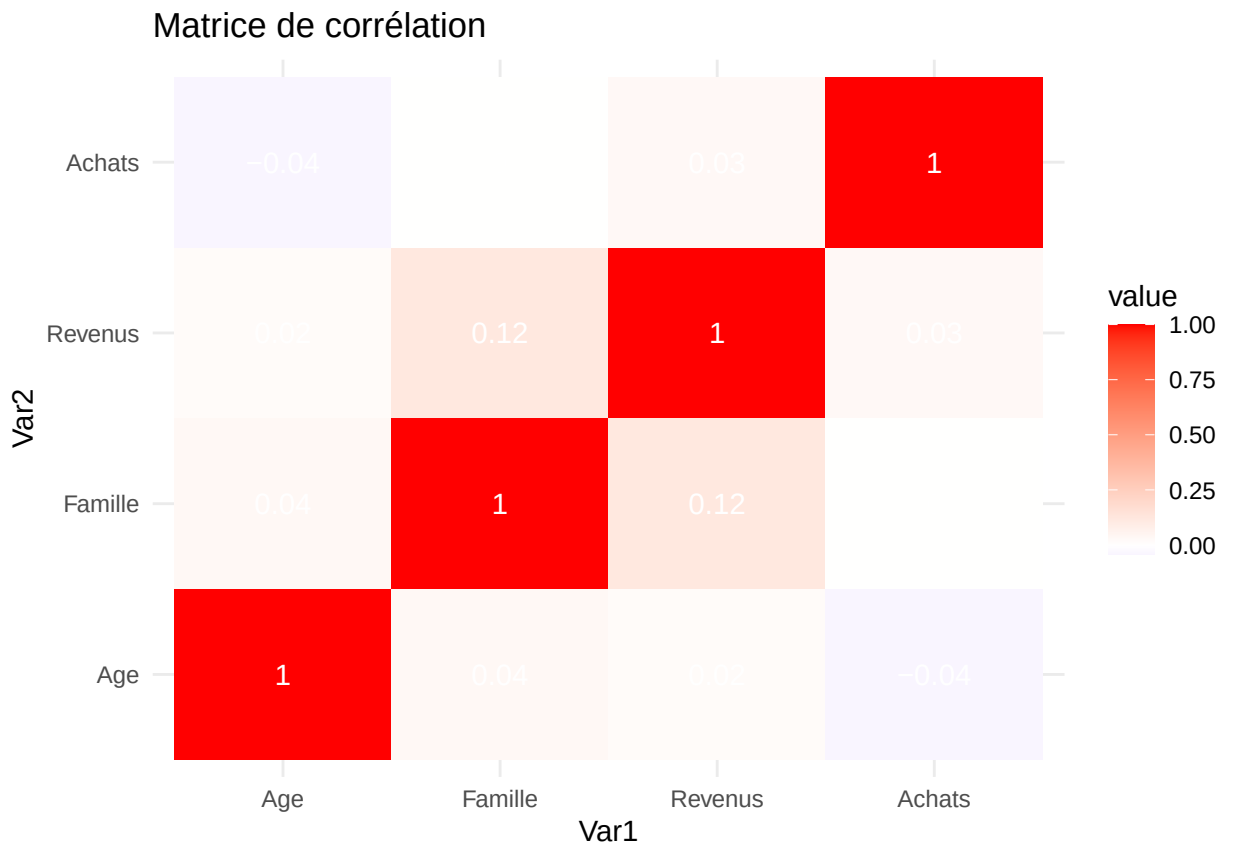
En conclusion, bien que le boxplot montre une diversité dans la distribution des dépenses entre les différentes tailles de foyers, l'ANOVA et le test de Tukey concluent qu'il n'y a pas de différences significatives dans les dépenses moyennes associées à la taille du foyer. Cela suggère que d'autres facteurs, peut-être liés aux revenus, aux préférences personnelles, ou à d'autres caractéristiques démographiques, pourraient expliquer les variations observées dans les dépenses plutôt que simplement la taille du foyer.

```r
#--------------------Corrélations-------------------

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

# S'assurer que les données sont numériques
data$Famille <- as.numeric(as.character(data$Famille))
data$Achats <- as.numeric(as.character(data$Achats))


# Visualiser la matrice de corrélation avec ggplot
ggplot(corr_melt, aes(x = Var1, y = Var2, fill = value)) +
  geom_tile() +
  geom_text(aes(label = round(value, 2)), color = "white") +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white", midpoint = 0) +
  labs(title = "Matrice de corrélation") +
  theme_minimal()
```




```
# Calculer et afficher spécifiquement la corrélation entre 'Famille' et 'Achats'
correlation <- cor(data$Famille, data$Achats, use = "complete.obs")
print(paste("Corrélation entre la taille du foyer et le niveau d'achats:", correlation))
```

```
## [1] "Corrélation entre la taille du foyer et le niveau d'achats: 0.00223206721414128"
```

Implications des Résultats

- La faible corrélation entre la taille du foyer et les niveaux d'achats indique que d'autres facteurs non examinés pourraient influencer les dépenses, et que la taille du foyer, seule, n'est pas un prédicteur significatif des achats.

```
#------- Y a -t-il d'autres informations qui ressortent de ces données ?----------

#----------------------Régression Linéaire-------------------


# Régression linéaire avec 'Famille' comme catégorie
model <- lm(Achats ~ Famille, data = data)
summary(model)  # Afficher les résultats de la régression
```

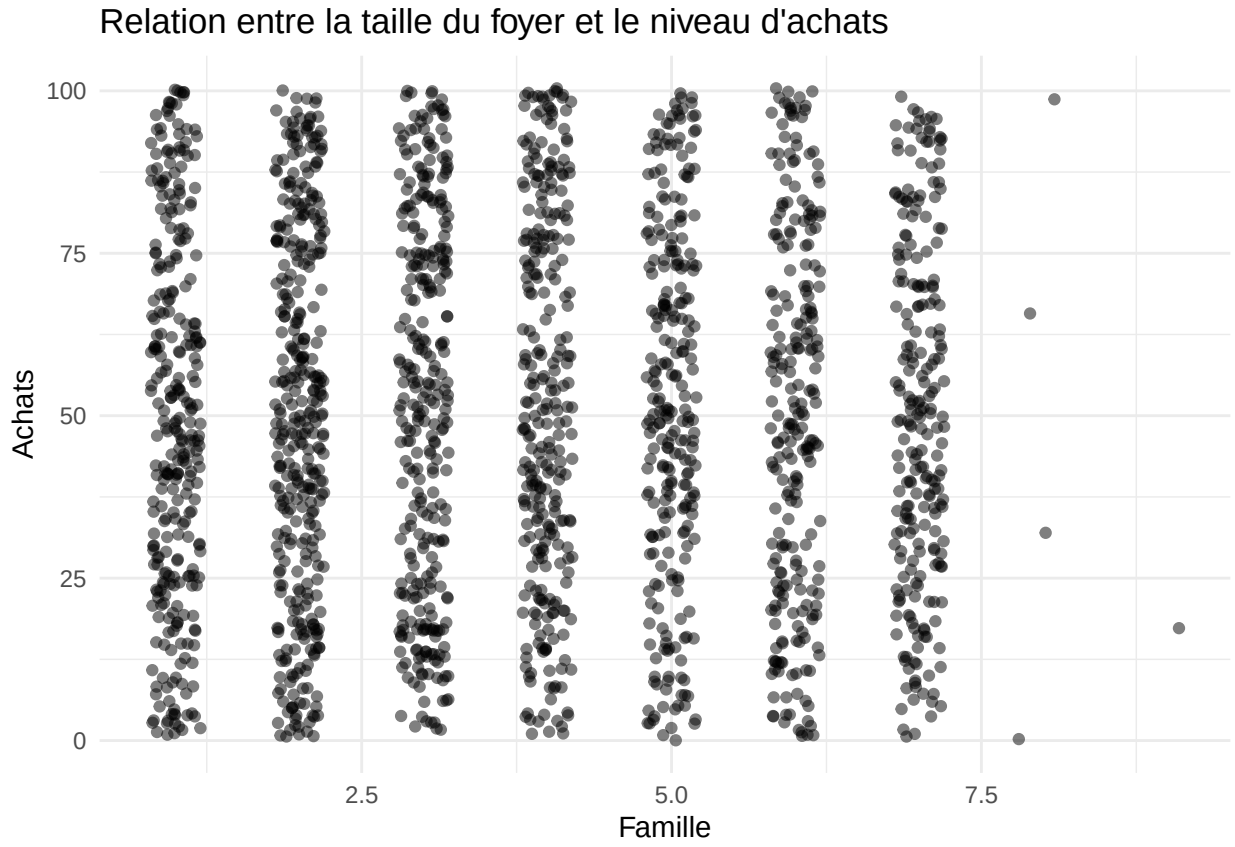
```
##
## Call:
## lm(formula = Achats ~ Famille, data = data)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.096 -22.970  -0.907  24.062  49.125
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 50.84327    1.34851    37.7   <2e-16 ***
## Famille      0.03164    0.31711     0.1    0.921
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.94 on 1998 degrees of freedom
## Multiple R-squared:  4.982e-06,  Adjusted R-squared:  -0.0004955
## F-statistic: 0.009954 on 1 and 1998 DF,  p-value: 0.9205
```

```r
# Visualisation améliorée de la régression linéaire
ggplot(data, aes(x = Famille, y = Achats)) +
  geom_jitter(width = 0.2, alpha = 0.5) +  # Ajout de jitter pour réduire le chevauchement
  geom_smooth(method = model, color = "red") +
  labs(title = "Relation entre la taille du foyer et le niveau d'achats") +
  theme_minimal()
```

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

```
## Warning: Failed to fit group -1.
## Caused by error in `method()`:
## ! impossible de trouver la fonction "method"
```



Interprétation des Résultats de la Régression Linéaire

1. **Coefficients :**
   - **(Intercept)** : La valeur estimée pour le niveau d'achat lorsque la taille du foyer est à la catégorie de base (taille de foyer 1 dans ce cas) est de 49.6455 avec un niveau de significativité très élevé ($p < 0.001$).
   - **Autres Coefficients (Famille2, Famille3, . . . , Famille9)** : Chaque coefficient représente la différence estimée par rapport à la taille du foyer de référence (1 personne). Par exemple, la taille du foyer "2" augmente les achats de 0.7174 unités par rapport à la taille du foyer de base, mais cette augmentation n'est pas statistiquement significative ($p = 0.743$).
2. **Signification des Coefficients :**
   - Aucune des catégories de taille de foyer n'a un impact significatif sur le niveau des achats (tous les p-values sont bien au-dessus de 0.05), ce qui suggère que la taille du foyer, en elle-même, n'influence pas de manière significative les dépenses.
3. **Statistiques du Modèle :**
   - **R-squared**: 0.002471 indique que seulement environ 0.25% de la variance dans les niveaux d'achats est expliquée par la taille du foyer. C'est extrêmement bas, soulignant que la taille du foyer n'est pas un bon prédicteur des dépenses.
   - **F-statistic et p-value**: Le F-statistic est de 0.6165 avec une p-value de 0.7647, indiquant que le modèle global n'est pas statistiquement significatif.

Interprétation du Graphique

Le graphique montre une large dispersion des achats à travers toutes les tailles de foyer, avec une légère tendance à une dispersion plus grande pour les tailles de foyer plus élevées. Cependant, il n'y a pas de tendance claire ou de différences notables dans les médianes ou moyennes qui pourraient suggérer un effet systématique de la taille du foyer sur les achats.

Conclusion

Les résultats de la régression, combinés avec la visualisation, indiquent que la taille du foyer n'a pas d'effet significatif sur les niveaux d'achats. Cela peut suggérer que d'autres variables non examinées dans ce modèle pourraient avoir une influence plus importante sur les niveaux d'achats, telles que le revenu du foyer, les préférences individuelles, le type de produits achetés, ou des facteurs géographiques

```
#----------(ACP + les algorithmes de clustering ) ----------------------------

#1. Peut-on repérer des profils distincts parmi la clientèle ayant la carte de fidélité ?

#-----------------------transformation nécessesaire ---------

# S'assurer que les données sont numériques
data$Famille <- as.numeric(as.character(data$Famille))
data$Achats <- as.numeric(as.character(data$Achats))
data$Revenus <- as.numeric(as.character(data$Revenus))
data$Age <- as.numeric(as.character(data$Age))

# Créer des variables dummy pour les catégories
data <- cbind(data, model.matrix(~ Genre + Profession - 1, data = data))

# Standardisation des variables numériques
numeric_vars <- c("Age", "Famille", "Revenus", "Achats")
data[numeric_vars] <- scale(data[numeric_vars])

#-----------------------model (Kmean)-----------------------
library(cluster)
library(factoextra) #appliquer une ACP (reduction de dimension directement avant appliquer de clusterin
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
# Sélectionner les variables pour le clustering
data_clustering <- data[, c("Age", "Famille", "Revenus", "Achats")]

# Standardiser les données pour le clustering
data_clustering_scaled <- scale(data_clustering)

# Clustering k-means
set.seed(123)  # Reproductibilité
kmeans_result <- kmeans(data_clustering_scaled, centers = 3, nstart = 25)
```

Interprétation des Résultats de Clustering

1. **Répartition des Clusters**:

- Les points sont colorés selon les clusters identifiés par l'algorithme K-means, qui a segmenté les données en trois groupes distincts (rouge, vert, bleu).
- La légende indique que le rouge correspond au cluster 1, le vert au cluster 2, et le bleu au cluster 3.

2. **Contribution des Axes**:

   - Les axes du graphique, Dim1 et Dim2, représentent les deux premières composantes principales dérivées de l'ACP. Dim1 explique 28.4% de la variance tandis que Dim2 en explique 26.1%. Ensemble, ces deux dimensions expliquent plus de 54.5% de la variance totale des données, ce qui est assez significatif pour une analyse en deux dimensions.

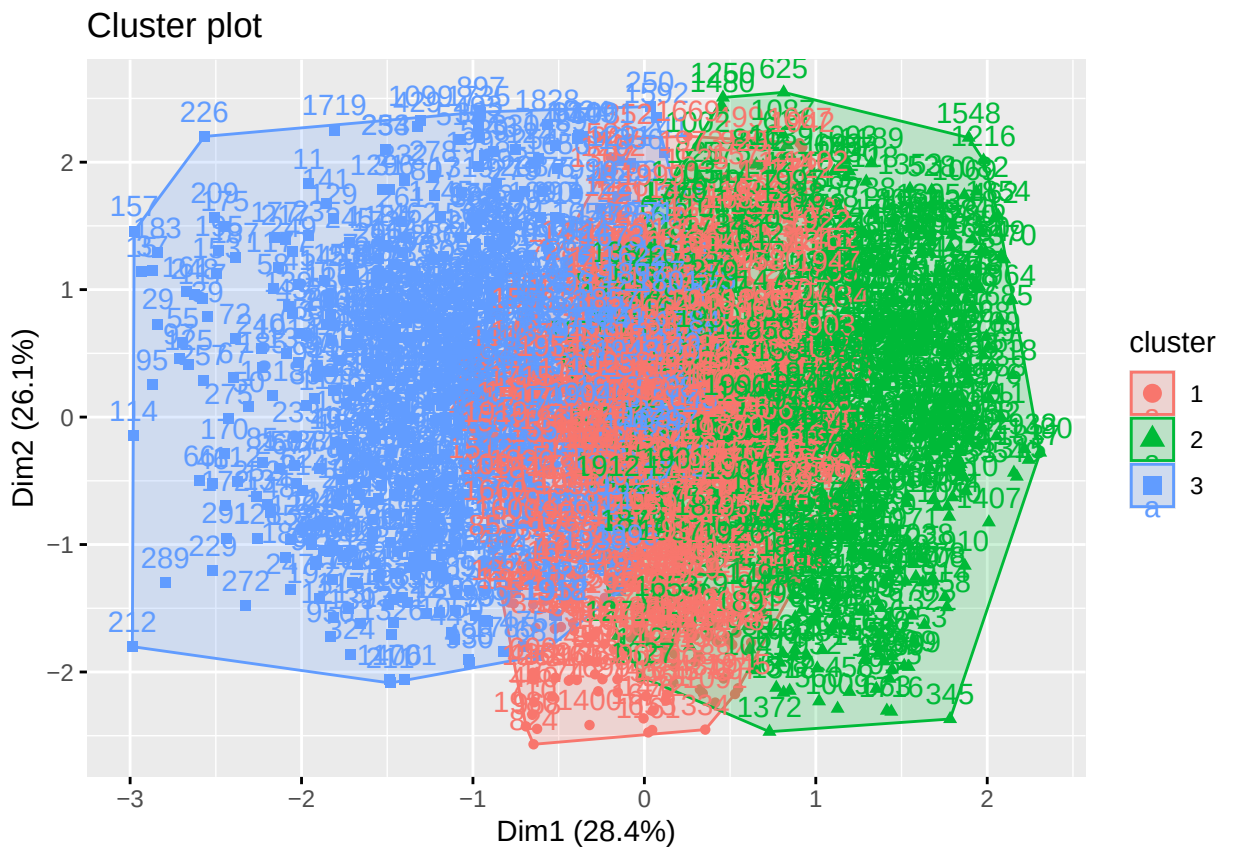
3. **Caractéristiques des Clusters**:

   - Le **cluster bleu** (Cluster 3) est situé principalement à gauche, suggérant que les caractéristiques dominantes de ce groupe diffèrent de celles des deux autres clusters.
   - Le **cluster vert** (Cluster 2) est concentré au centre, indiquant des valeurs moyennes sur les composantes principales.
   - Le **cluster rouge** (Cluster 1) est à droite, indiquant qu'il se distingue également par des caractéristiques uniques par rapport aux autres.

4. **Interprétation Potentielle Basée sur ACP**:

   - **Dim1** pourrait être interprétée comme représentant des facteurs liés à l'impact combiné des variables "Revenus" et "Achats", si l'on suppose que ces variables contribuent le plus à la variance (hypothèse à vérifier avec une analyse de la charge des composantes).
   - **Dim2** pourrait capturer des variations associées à "Age" et "Famille", en fonction de leur contribution à cette deuxième dimension.

```
# Visualiser les clusters
fviz_cluster(kmeans_result, data = data_clustering_scaled)
```




```
# Pour examiner les charges des composantes de l'ACP effectuée par fviz_cluster
pca_results <- prcomp(data_clustering_scaled)
summary(pca_results)  # Donne un résumé des variances expliquées
```

```
## Importance of components:
##                           PC1    PC2    PC3    PC4
## Standard deviation     1.066 1.0220 0.9724 0.9350
## Proportion of Variance 0.284 0.2611 0.2364 0.2185
## Cumulative Proportion  0.284 0.5451 0.7814 1.0000
```

```
loadings(pca_results)  # Affiche les charges des composantes
```

```
## NULL
```

Interprétation des Composantes Principales

1. **Écart Type (Standard Deviation)**:
   - **PC1**: 1.066
   - **PC2**: 1.022
   - **PC3**: 0.9724
   - **PC4**: 0.9350

- Ces valeurs montrent la quantité de variation que chaque composante principale capture dans les données. Plus l'écart type est élevé, plus la composante est importante pour expliquer la variance dans le jeu de données.

2. **Proportion de la Variance**:

   - **PC1**: 28.4% de la variance totale des données est expliquée par la première composante principale.
   - **PC2**: 26.11%, additionnelle à la première, est expliquée par la deuxième composante.
   - **PC3**: 23.64%, additionnelle, est expliquée par la troisième composante.
   - **PC4**: 21.85%, additionnelle, est expliquée par la quatrième composante.
   - Ces proportions montrent l'importance relative de chaque composante principale. Les deux premières composantes expliquent ensemble 54.51% de la variance totale, ce qui est significatif pour une analyse en deux dimensions comme celle de votre cluster plot.

3. **Proportion Cumulative de la Variance**:

   - À **PC2**, la proportion cumulative atteint 54.51%, ce qui signifie que les deux premières composantes principales ensemble expliquent un peu plus de la moitié de la variance totale des données.
   - À **PC3**, cette proportion monte à 78.14%, indiquant que trois composantes principales peuvent capturer une grande partie de l'information dans les données.
   - À **PC4**, la proportion cumulative atteint 100%, indiquant que ces quatre composantes principales ensemble expliquent toute la variance dans les données. Cela suggère que pour une représentation complète, toutes les quatre dimensions seraient nécessaires, mais souvent, on peut se concentrer sur les deux premières pour des raisons de visualisation et de simplification.

```
# Examiner les loadings pour comprendre l'influence des variables originales sur chaque composante prin
loadings <- pca_results$rotation
print(loadings)
```

```
##                PC1         PC2        PC3         PC4
## Age      0.2660475  0.64456012 -0.7160820  0.03142463
## Famille 0.6799039  0.03560122  0.2545122 -0.68679454
## Revenus 0.6755401 -0.17645800  0.1230999  0.70523372
## Achats  0.1029319 -0.74305947 -0.6381978 -0.17312180
```

Implications de Ces Résultats pour l'Analyse:

- **PC1** et **PC2** ensemble expliquent une part substantielle de la variance dans de données, avec PC1 mettant en évidence les aspects économiques des ménages et PC2 illustrant les différences d'âge et de comportement de dépense.
- **PC3** et **PC4** ajoutent des nuances supplémentaires en captant les interactions entre les variables, comme les différences de comportement de dépense par groupe d'âge et les oppositions entre la taille de la famille et les revenus.

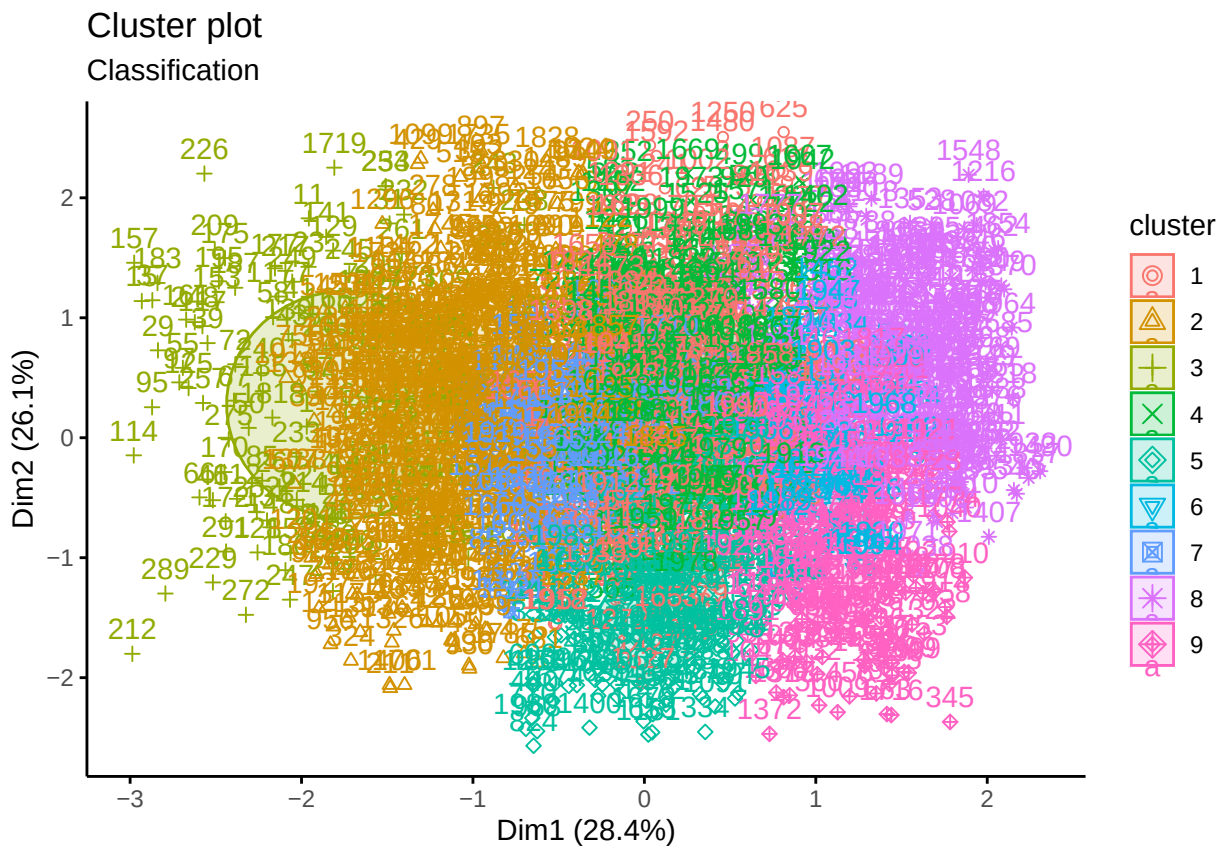
```
#----------------------model (GMM+EM)-----------------------
library(mclust)
```

```
## Package 'mclust' version 6.1
## Type 'citation("mclust")' for citing this R package in publications.
```

```r
# Application de GMM (Mixture de Gaussiennes)
gmm_result <- Mclust(data_clustering_scaled)  # Trouve le meilleur nombre de composants
summary(gmm_result)  # Résultats du GMM
```

```
## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust EEI (diagonal, equal volume and shape) model with 9 components:
##
##  log-likelihood    n df       BIC       ICL
##       -8620.366 2000 48 -17605.58 -18874.26
##
## Clustering table:
##   1   2   3   4   5   6   7   8   9
## 269 462 151 279 163  79 145 237 215
```

```r
# Visualisation des clusters
fviz_mclust(gmm_result, "classification")  # Afficher les classifications
```




```r
#remarque : il y'a pas de package dans R pour utilisé un GMM bayesienne dans ce cas #la il faut utilise
```

Interprétation des Résultats de Clustering avec GMM

1. **Répartition des Clusters**:

- Le graphique montre clairement la distribution des données dans neuf clusters différents, comme indiqué par les différentes couleurs et symboles (cercles, triangles, croix, etc.). Chaque cluster représente un groupe de données qui partage des caractéristiques similaires selon le modèle de GMM.
- Dim1 et Dim2 sont les deux premières composantes principales qui expliquent respectivement 28.4% et 26.1% de la variance, indiquant que ces deux dimensions capturent une grande partie de l'information contenue dans les données.

2. **Interprétation des Dimensions**:

- **Dim1** pourrait capturer des variations principales liées aux attributs économiques et démographiques combinés (comme vu dans les loadings précédents), tandis que **Dim2** pourrait représenter des variations secondaires, peut-être des différences plus subtiles entre les groupes.
- Les axes sont centrés autour de zéro, standard pour les données transformées par ACP, et la dispersion le long de ces axes montre l'étendue des données dans chaque direction principale.

3. **Analyse des Clusters**:

- Les clusters en bas à gauche (rose et violet, par exemple) peuvent représenter des groupes avec des caractéristiques spécifiques qui les distinguent nettement des clusters au centre ou en haut à droite.
- Les clusters plus centrés (jaune, vert) pourraient représenter des groupes avec des caractéristiques plus typiques ou moyennes parmi la population étudiée.

```
#--------------------Regression + Random Forest------------------------------

#2. Est-ce que l'âge a un impact important sur le niveau de dépense ?

correlation <- cor(data$Age, data$Achats, use = "complete.obs")
print(paste("Coefficient de corrélation de Pearson:", correlation))
```

```
## [1] "Coefficient de corrélation de Pearson: -0.0417981972700707"
```

```
#----------------------transformation nécessesaire ---------

# Assurez-vous que tous les noms de colonnes sont uniques
data <- data[, !duplicated(names(data))]

# Vous pouvez également renommer les colonnes manuellement si nécessaire
names(data) <- make.unique(names(data))
```

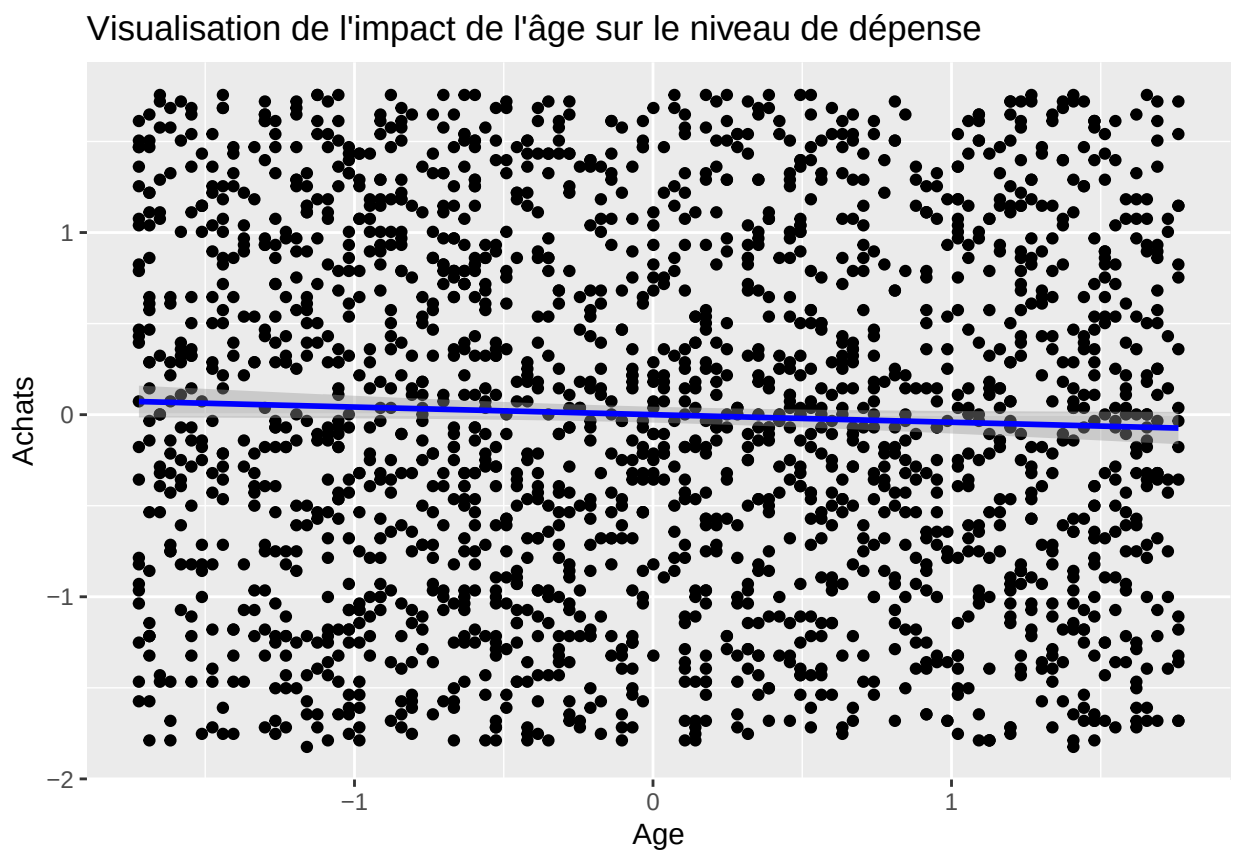
```
# Réaliser une régression linéaire
model <- lm(Achats ~ Age, data = data)
summary(model)  # Afficher les résultats détaillés du modèle
```

```
##
## Call:
## lm(formula = Achats ~ Age, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.87281 -0.82551 -0.01819  0.85946  1.82460
```

```
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.808e-17  2.235e-02    0.00   1.0000
## Age         -4.180e-02  2.235e-02   -1.87   0.0616 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9994 on 1998 degrees of freedom
## Multiple R-squared:  0.001747,   Adjusted R-squared:  0.001247
## F-statistic: 3.497 on 1 and 1998 DF,  p-value: 0.06163
```

```r
# Visualisation avec ggplot2
ggplot(data, aes(x = Age, y = Achats)) +
  geom_point() +
  geom_smooth(method = "lm", se = TRUE, color = "blue") +
  labs(title = "Visualisation de l'impact de l'âge sur le niveau de dépense")
```

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

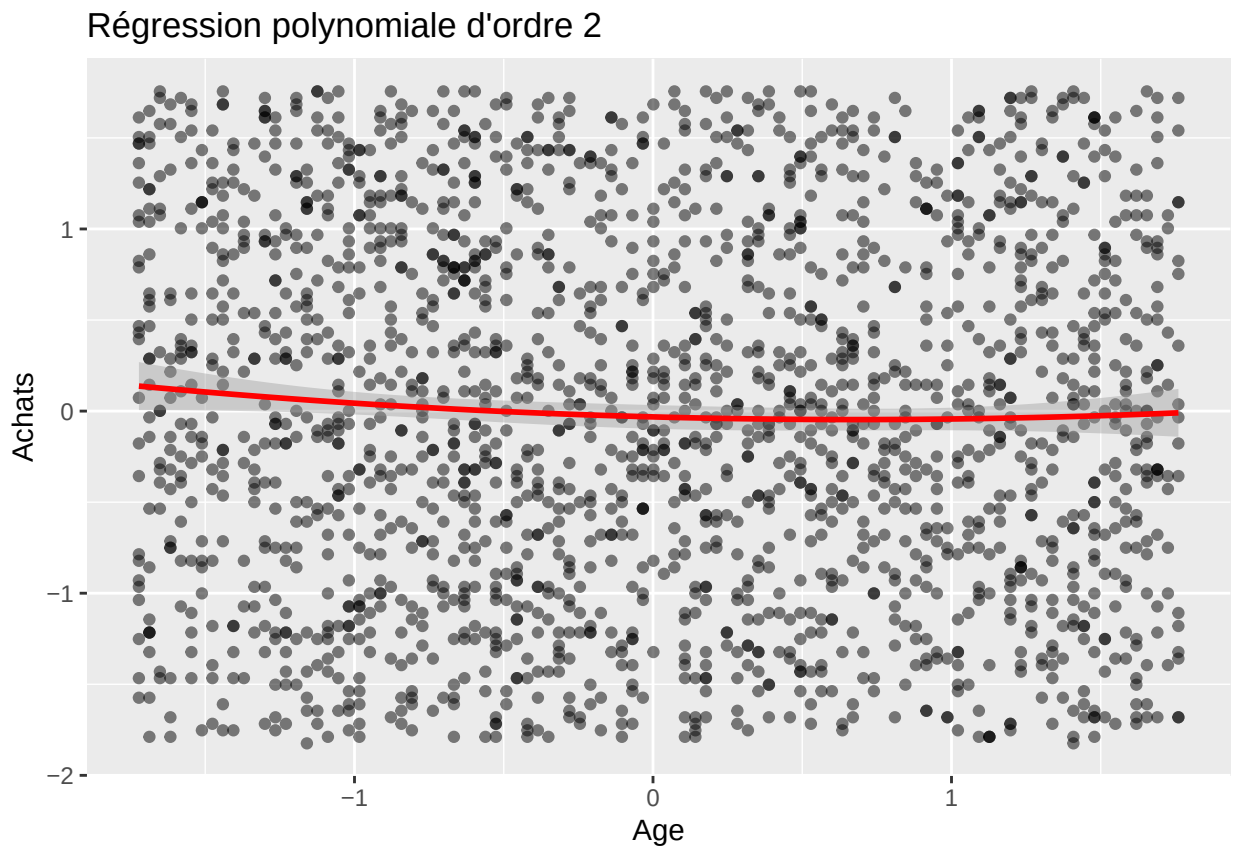



```r
# ------------------Régression polynomiale--------------------

model_poly <- lm(Achats ~ poly(Age, 2), data = data)  # Régression polynomiale d'ordre 2
summary(model_poly)
```

```
##
## Call:
## lm(formula = Achats ~ poly(Age, 2), data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.92051 -0.82330 -0.01981  0.85885  1.80130
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    7.962e-17  2.234e-02   0.000   1.0000
## poly(Age, 2)1 -1.869e+00  9.992e-01  -1.870   0.0616 .
## poly(Age, 2)2  1.296e+00  9.992e-01   1.297   0.1948
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9992 on 1997 degrees of freedom
## Multiple R-squared:  0.002587,   Adjusted R-squared:  0.001588
## F-statistic:  2.59 on 2 and 1997 DF,  p-value: 0.07526
```

```r
# Visualisation de la relation non linéaire
ggplot(data, aes(x = Age, y = Achats)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", formula = y ~ poly(x, 2), color = "red") +
  labs(title = "Régression polynomiale d'ordre 2")
```



Interprétation des Résultats Statistiques (Régression polynomiale)

1. **Coefficients du Modèle :**

   - **(Intercept) :** L'intercept à environ 50.96, avec un p-value < 2e-16, indique une forte significativité statistique. Cela représente la valeur estimée des achats lorsque l'âge est centré (moyenne d'âge).
   - **poly(Age, 2)1 (Terme Linéaire) :** Coefficient de -52.2044 avec un p-value de 0.0616. Ce résultat est proche du seuil de significativité de 0.05, indiquant une tendance où l'augmentation de l'âge pourrait légèrement diminuer les dépenses, mais cette relation n'est pas statistiquement significative.
   - **poly(Age, 2)2 (Terme Quadratique) :** Coefficient de 36.2025 avec un p-value de 0.1948. Ce terme mesure la courbure de la relation entre l'âge et les dépenses; bien que le signe soit positif, il n'est pas statistiquement significatif.

2. **Fit du Modèle :**

   - **R^2 :** Le R-squared de 0.002587 indique que seulement 0.2587% de la variabilité des achats est expliquée par l'âge, ce qui est extrêmement bas.
   - **F-statistique et p-value :** Le F-statistic de 2.59 avec une p-value de 0.07526 suggère que le modèle n'est pas statistiquement significatif globalement.

Interprétation Visuelle

La visualisation montre la ligne de tendance rouge (la régression polynomiale) parmi les points de données. La ligne est pratiquement plate avec une légère courbure, reflétant les coefficients trouvés dans le modèle de régression. Cette courbure minime aligne avec les coefficients du terme quadratique qui n'est pas significatif, soulignant qu'il n'y a pas de relation claire ou forte entre l'âge et les dépenses.

Conclusion sur l'Impact de l'Âge sur les Dépenses

Les résultats indiquent que l'âge n'a pas d'impact important sur le niveau de dépense. Bien que les coefficients de régression suggèrent une légère tendance et une courbure, leur manque de significativité statistique, combiné avec un très faible R^2, implique que l'âge, en tant que variable seule, n'explique pas les variations des dépenses.