Projet Clustering

Natacha Babalola.

2024-05-18

Table of Contents

Contexte :

Dans le cadre de l'amélioration de la compréhension des dynamiques de propriété des logements, notre étude se concentre sur l'analyse du taux de propriété des logements par communes et régions en France sur une période de plus de 60 ans, de 1960 à 2022. Cette analyse approfondie vise à identifier les facteurs influençant l'acquisition de logements et à fournir des recommandations pour renforcer l'attractivité des différentes régions.

Pour ce faire, nous avons intégré des données démographiques, notamment l'âge moyen des populations, afin d'examiner l'hypothèse selon laquelle l'âge des habitants pourrait jouer un rôle dans le taux de propriété des logements. De plus, nous avons incorporé des données socio-politiques, incluant les taux de criminalité et de divers délits, pour évaluer l'impact de la sécurité sur l'acquisition de logements.

En analysant ces différents aspects, notre objectif est de mettre en lumière les interactions complexes entre la démographie, la sécurité et la propriété des logements. Nous espérons que cette étude fournira des insights précieux aux décideurs politiques, les aidant à mettre en œuvre des stratégies efficaces pour renforcer l'attractivité et la sécurité des régions, et ainsi promouvoir une croissance durable du taux de propriété des logements.

Source : Julia Cagé et Thomas Piketty (2023) : Une histoire du conflit politique. Élections et inégalités sociales en France, 1789-2022, Paris, Le Seuil.

1. Exploration des données

1.1. Téléchargement des bibliothèques

```
#library(readr)
#proprietairescommunes <- read_csv("~/Documents/Module 3/Apprentissage Non-supervisé - Madalina/Apprentissage non-supervisé-20240405/Projet/proprietairescommunes.csv")
#View(proprietairescommunes)

# Chargement des bibliothèques nécessaires
library(tidyverse)
library(readr)
library(ggplot2)
```

```r
library(FactoMineR)
library(cluster)
library(factoextra)
library(dbscan)
#library(Rlof)
#library(ruptures)

library(readr)
# Chargement des données
data <- read_csv("proprietairescommunes.csv")
```

```
## Rows: 37937 Columns: 426
## ── Column specification
──────────────────────────────────────────────────────
## Delimiter: ","
## chr   (4): dep, nomdep, codecommune, nomcommune
## dbl (422): ppropri1960, ppropri1961, ppropri1962, ppropri1963,
ppropri1964, ...
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this
message.
```

```r
data <- na.omit(data)
View(data)
```

1.2. Exploration et Nettoyage des Données

```r
# Affichage des premières lignes des données
head(data_f)
```

```
## # A tibble: 6 × 67
##   dep   nomdep codecommune nomcommune        ppropri1960 ppropri1961
ppropri1962
##   <chr> <chr>  <chr>       <chr>                   <dbl>       <dbl>
<dbl>
## 1 01    AIN    01001       ABERGEMENT-CLEME…       0.358       0.378
0.402
## 2 01    AIN    01002       ABERGEMENT-DE-VA…       0.756       0.762
0.786
## 3 01    AIN    01004       AMBERIEU-EN-BUGEY       0.386       0.382
0.379
## 4 01    AIN    01005       AMBERIEUX-EN-DOM…       0.300       0.328
0.358
## 5 01    AIN    01006       AMBLEON                 0.794       0.771
0.771
## 6 01    AIN    01007       AMBRONAY                0.637       0.629
0.621
## # ℹ 60 more variables: ppropri1963 <dbl>, ppropri1964 <dbl>, ppropri1965
<dbl>,
## #   ppropri1966 <dbl>, ppropri1967 <dbl>, ppropri1968 <dbl>, ppropri1969
<dbl>,
```

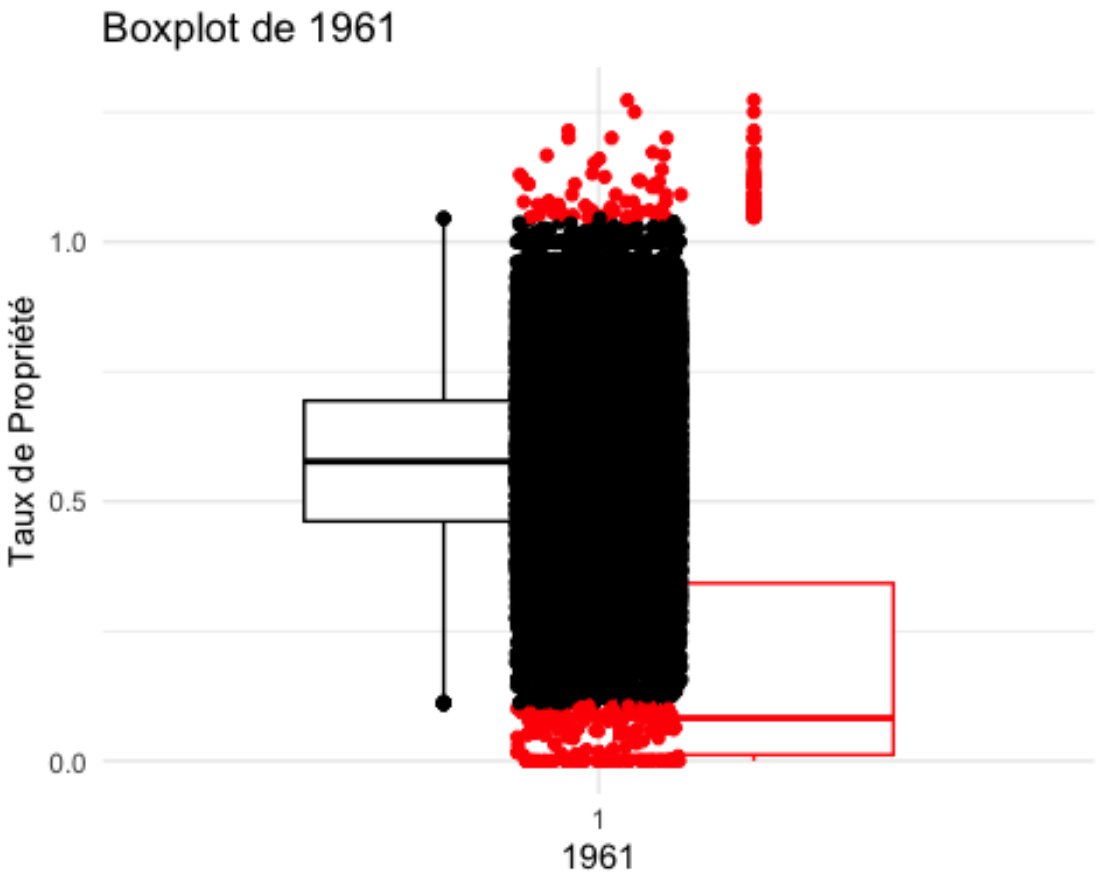
```
## #   ppropri1970 <dbl>, ppropri1971 <dbl>, ppropri1972 <dbl>, ppropri1973
<dbl>,
## #   ppropri1974 <dbl>, ppropri1975 <dbl>, ppropri1976 <dbl>, ppropri1977
<dbl>,
## #   ppropri1978 <dbl>, ppropri1979 <dbl>, ppropri1980 <dbl>, ppropri1981
<dbl>,
## #   ppropri1982 <dbl>, ppropri1983 <dbl>, ppropri1984 <dbl>, ppropri1985
<dbl>,
## #   ppropri1986 <dbl>, ppropri1987 <dbl>, ppropri1988 <dbl>, …
```

```r
# Récupération des noms de colonnes existants
noms_colonnes <- names(data_f)

# Application d'une expression régulière pour extraire les années
annees <- gsub("ppropri", "", noms_colonnes)  # Supprimer "ppropri" des noms
de colonnes

# Renommage des colonnes avec les années extraites
names(data_f) <- annees

# Affichage des premières lignes du dataframe pour vérifier les changements
head(data_f)
```

```
## # A tibble: 6 × 67
##   dep   nomdep codecommune nomcommune  `1960` `1961` `1962` `1963` `1964`
`1965`
##   <chr> <chr>  <chr>       <chr>        <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
<dbl>
## 1 01    AIN    01001       ABERGEMENT…  0.358  0.378  0.402  0.426  0.449
0.484
## 2 01    AIN    01002       ABERGEMENT…  0.756  0.762  0.786  0.810  0.814
0.837
## 3 01    AIN    01004       AMBERIEU-E…  0.386  0.382  0.379  0.377  0.374
0.372
## 4 01    AIN    01005       AMBERIEUX-…  0.300  0.328  0.358  0.387  0.413
0.442
## 5 01    AIN    01006       AMBLEON      0.794  0.771  0.771  0.75   0.75
0.730
## 6 01    AIN    01007       AMBRONAY     0.637  0.629  0.621  0.614  0.607
0.598
## # ℹ 57 more variables: `1966` <dbl>, `1967` <dbl>, `1968` <dbl>, `1969`
<dbl>,
## #   `1970` <dbl>, `1971` <dbl>, `1972` <dbl>, `1973` <dbl>, `1974` <dbl>,
## #   `1975` <dbl>, `1976` <dbl>, `1977` <dbl>, `1978` <dbl>, `1979` <dbl>,
## #   `1980` <dbl>, `1981` <dbl>, `1982` <dbl>, `1983` <dbl>, `1984` <dbl>,
## #   `1985` <dbl>, `1986` <dbl>, `1987` <dbl>, `1988` <dbl>, `1989` <dbl>,
## #   `1990` <dbl>, `1991` <dbl>, `1992` <dbl>, `1993` <dbl>, `1994` <dbl>,
## #   `1995` <dbl>, `1996` <dbl>, `1997` <dbl>, `1998` <dbl>, `1999` <dbl>,
…
```

1.3. Détection valeurs abbérantes

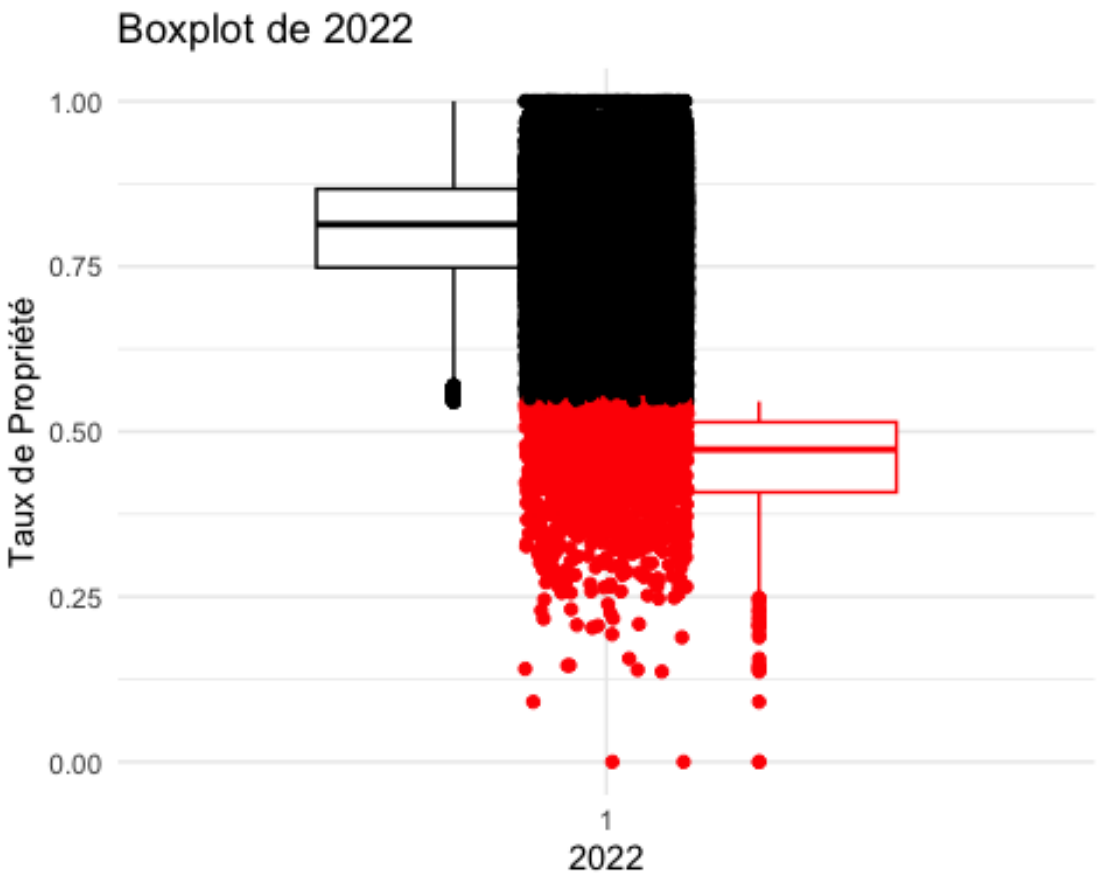
1.3.1. Cas de l'année 1961



1.3.2. Interpretation :

La ligne noire à l'intérieur de la boîte représente la médiane des taux de propriété en 1961. Elle semble être autour de 0.65 La boîte s'étend du premier quartile (Q1) au troisième quartile (Q3). Cela montre que la majorité des taux de propriété se situe entre environ 0.5 et 0.95. Les points rouges en dehors des moustaches sont des valeurs aberrantes. Pour 1961, il semble y avoir un nombre significatif de valeurs aberrantes au-dessus de la moustache supérieure, indiquant des taux de propriété exceptionnellement élevés pour certaines communes.

1.3.3 Cas de l'année 2022



####

1.3.4. Interpretation :

La ligne noire à l'intérieur de la boîte représente la médiane des taux de propriété en 2022. Elle semble être autour de 0.85. Les points rouges en dehors des moustaches sont des valeurs aberrantes. Pour 2022, il y a également un nombre significatif de valeurs aberrantes, mais cette fois, on observe des taux de propriété exceptionnellement bas pour certaines communes, indiquant une plus grande dispersion et variabilité dans les taux de propriété

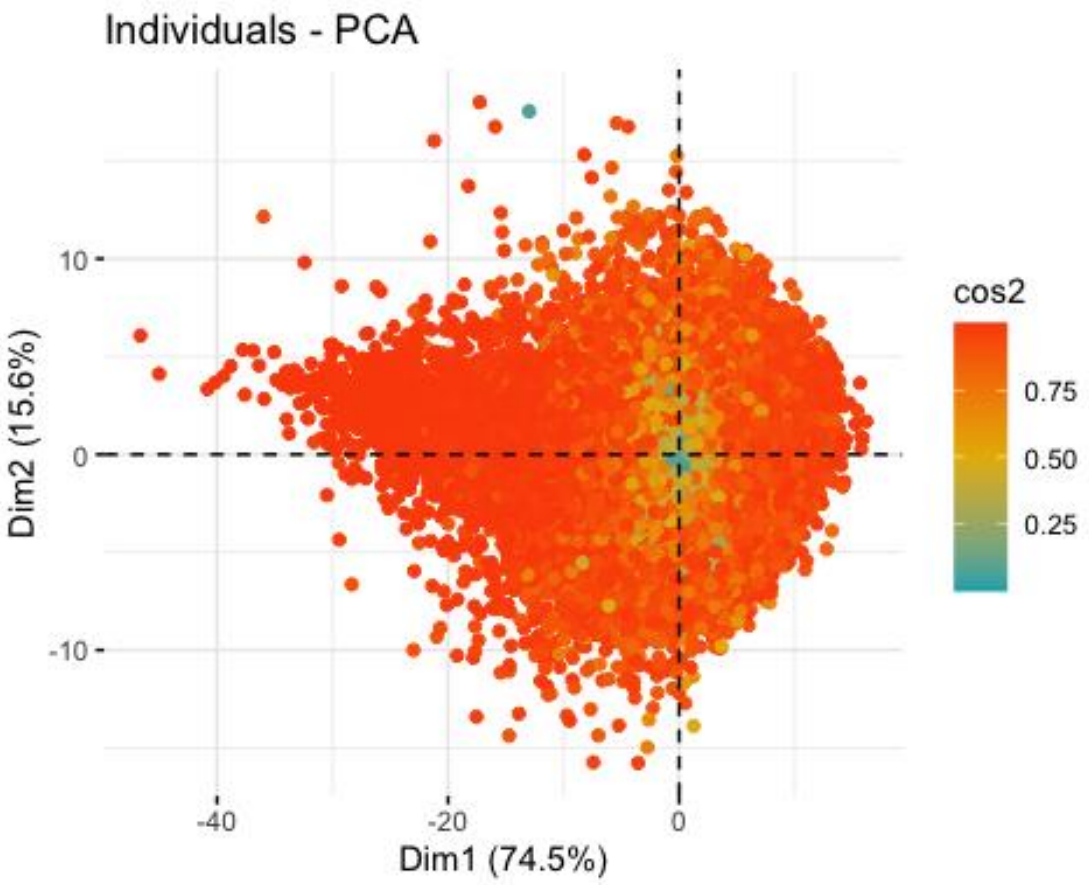
Comparaison entre 1961 et 2022 Médiane : La médiane a légèrement augmenté de 1961 à 2022. Dispersion : La dispersion des taux de propriété (IQR) semble être légèrement plus grande en 2022 qu'en 1961, indiquant une variabilité accrue. Valeurs Aberrantes : En 1961, les valeurs aberrantes étaient principalement des taux de propriété élevés, tandis qu'en 2022, on observe des valeurs aberrantes aussi bien élevées que basses, indiquant une plus grande variabilité dans les taux de propriété.

Conclusion Évolution des Taux de Propriété : Les taux de propriété semblent être plus dispersés en 2022 qu'en 1961, avec une plus grande variabilité et des valeurs aberrantes significatives des deux côtés de la distribution. Variabilité Géographique : La présence de nombreuses valeurs aberrantes indique des différences significatives dans les taux de

propriété entre différentes communes, tant en 1961 qu'en 2022. Cette variabilité pourrait être due à des facteurs économiques, politiques, ou sociaux différents entre les périodes.

2. Méthode de Réduction de Dimension (Application de l'ACP)

```
# Application de l'ACP
res.pca <- PCA(data_f[, 5:ncol(data_f)], graph = FALSE)

# Visualisation des résultats de l'ACP
fviz_pca_ind(res.pca, geom.ind = "point", col.ind = "cos2",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"), repel =
TRUE)
```



2.1. Interpretation :

Axe des composantes principales :

Dim1 (74.5%) : L'axe horizontal (Dim1) explique 74.5% de la variance totale des données. C'est la composante principale qui capture la plus grande proportion de la variance dans les données. Dim2 (15.6%) : L'axe vertical (Dim2) explique 15.6% de la variance totale des données. C'est la deuxième composante principale qui capture une partie importante de la variance restante. Couleur des points (cos2) :

cos2 : La couleur des points représente le cos2 (cosinus carré) des individus, qui mesure la qualité de la représentation des individus sur le plan factoriel. Les valeurs plus élevées de cos2 (plus proches de 1) indiquent que l'individu est bien représenté par les composantes principales choisies. La palette de couleurs allant du bleu (#00AFBB) à l'orange (#FC4E07) montre que les points en bleu sont moins bien représentés par ces deux composantes, tandis que les points en orange sont mieux représentés.

Distribution des points :

Les points plus proches du centre (0, 0) sont des individus qui ne sont pas bien expliqués par les deux premières composantes principales. Ils sont plus proches de la moyenne de toutes les variables. Les points éloignés du centre représentent des individus qui sont plus distincts et bien expliqués par les deux premières composantes principales. Ils contribuent davantage à la variance expliquée par ces composantes.

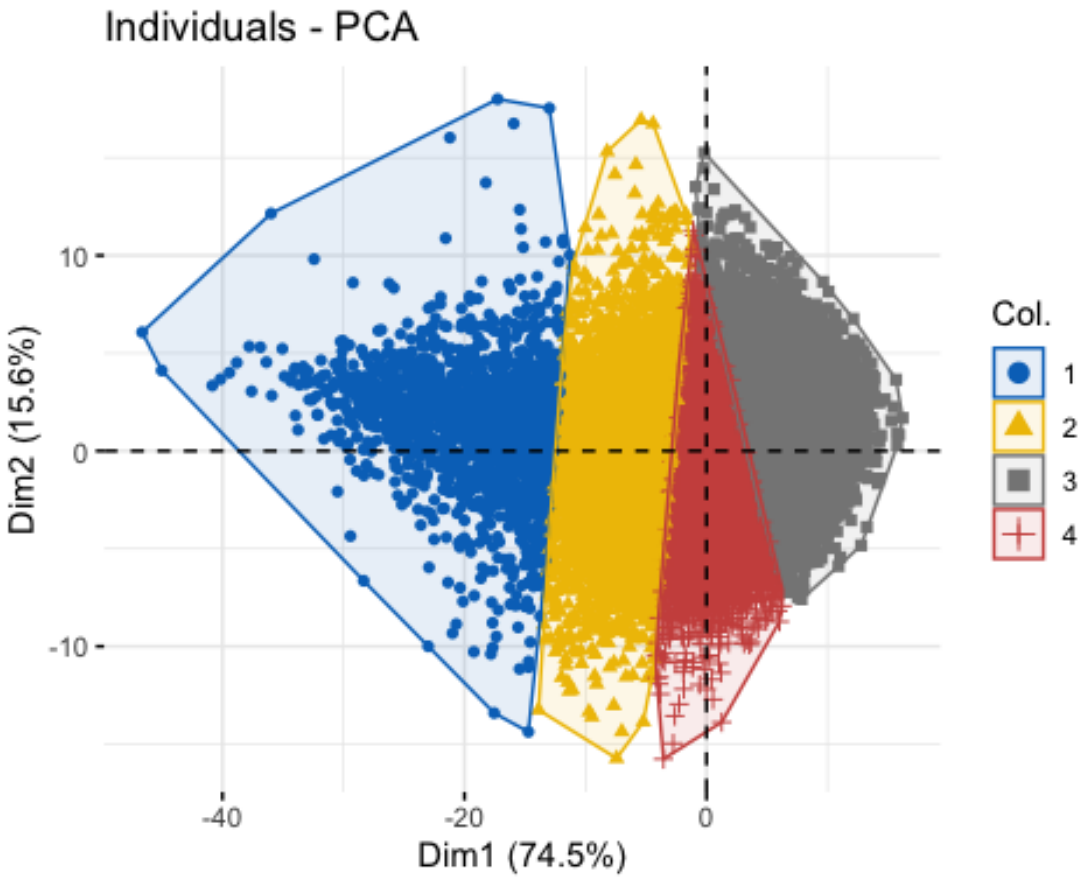
Clusters potentiels :

La répartition et la densité des points peuvent suggérer des clusters ou des regroupements naturels dans les données. Si nous observons des groupes de points distincts, cela peut indiquer des similarités ou des différences significatives entre les individus de ces groupes.

3. K-Means

3.1. Application d'un cluster à 4 niveaux

```r
# Exécution de k-means clustering sur les données projetées
set.seed(123)
clustering <- kmeans(res.pca$ind$coord, centers = 4)  # Choisissez un nombre de clusters approprié

# Ajout des clusters aux données projetées et convertir les clusters en facteur
data_clusters <- data.frame(res.pca$ind$coord, cluster = as.factor(clustering$cluster))

# Visualisation des clusters
library(factoextra)
fviz_pca_ind(res.pca, geom.ind = "point", col.ind = data_clusters$cluster, palette = "jco", addEllipses = TRUE, ellipse.type = "convex", repel = TRUE)
```



####

3.1.1 Interpretation :

- **Axes des composantes principales :**

Dim1 (74.5%) : L'axe horizontal (Dim1) explique 74.5% de la variance totale des données. Cette composante principale capture la majorité de la variance. Dim2 (15.6%) : L'axe vertical (Dim2) explique 15.6% de la variance totale des données. Cette composante principale capture une partie importante de la variance restante.

- **Points colorés par cluster :**

Les points sont colorés et symbolisés en fonction des clusters auxquels ils appartiennent. Chaque couleur et symbole représente un cluster différent. Dans ce graphique, nous avons quatre clusters : Cluster 1 (bleu, cercle) Cluster 2 (jaune, triangle) Cluster 3 (gris, carré) Cluster 4 (rouge, croix) Ellipses ou contours des clusters :

Les contours autour des clusters montrent les zones de densité des points pour chaque cluster. Cela nous donne une idée de la répartition et de l'étendue des clusters dans l'espace des composantes principales.

- **Interprétation des clusters :**

Cluster 1 (bleu) :

Les points du cluster bleu sont principalement situés sur la gauche du graphique. Ce cluster représente une partie significative de la variance capturée par Dim1.

Cluster 2 (jaune) :

Les points du cluster jaune sont principalement situés au centre du graphique, indiquant une variance modérée expliquée par Dim1 et Dim2.

Cluster 3 (gris) :

Les points du cluster gris sont dispersés principalement à droite du graphique. Ce cluster capture une portion différente de la variance par rapport aux autres clusters.

Cluster 4 (rouge) :

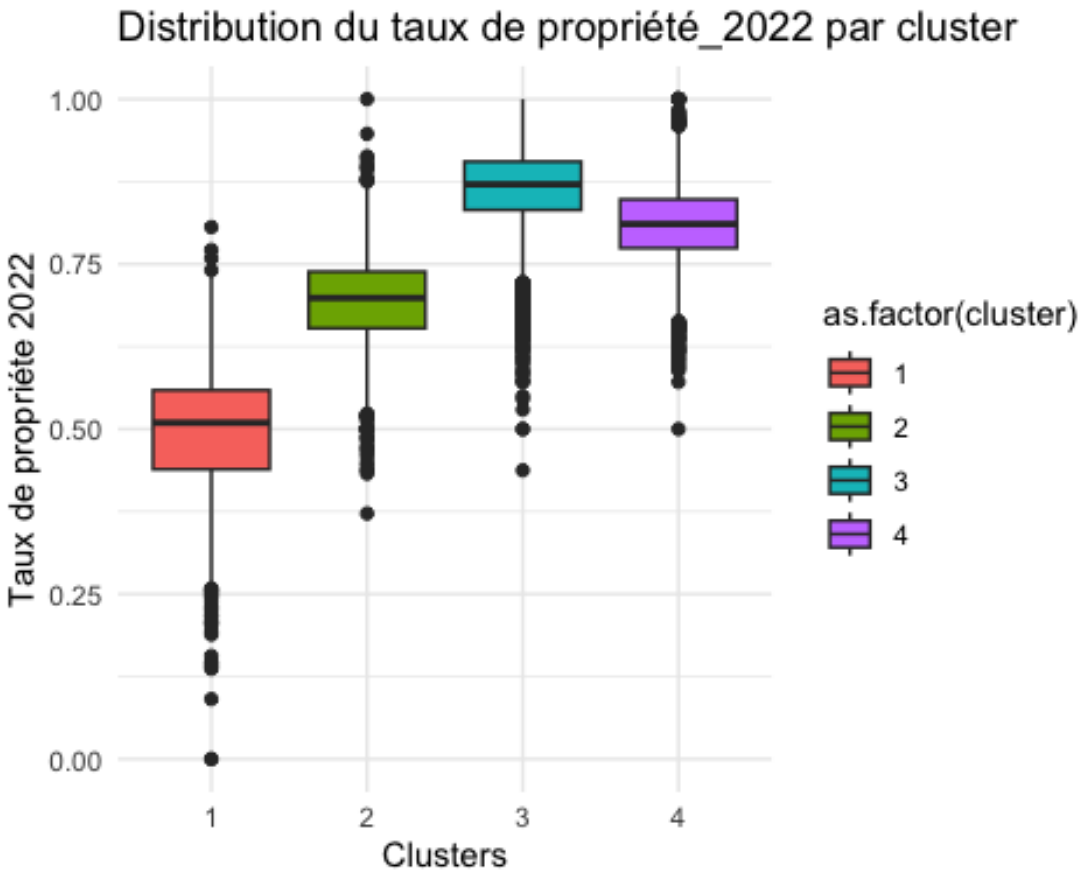
Les points du cluster rouge sont situés principalement à droite et en bas du graphique. Ce cluster a des caractéristiques distinctes des autres clusters, comme l'indiquent les positions des points dans l'espace des composantes principales.

3.2. Analyse des Clusters :

3.2.1. Caractéristiques de chaque clusters

3.2.2.Cas du Taux de propriété en 2022 en box

```
# Visualiser la distribution de ppropri2022 par cluster
ggplot(data_with_clusters, aes(x = as.factor(cluster), y = `2022`, fill =
as.factor(cluster))) +
  geom_boxplot() +
  labs(title = "Distribution du taux de propriété_2022 par cluster",
       x = "Clusters",
       y = "Taux de propriéte 2022") +
  theme_minimal()
```



####

3.2.3. Interpretation :

Le boxplot généré montre la distribution du taux de propriété pour l'année 2022 pour chaque cluster.

- Les whiskers s'étendent jusqu'à 1,5 fois l'IQR au-dessus et en dessous de Q1 et Q3, respectivement.
- Les points en dehors des whiskers sont des valeurs aberrantes, ce qui signifie qu'ils sont significativement plus élevés ou plus bas que la majorité des données.

3.2.4. Analyse des Clusters

1. **Cluster 1** :
   - La médiane du taux de propriété est d'environ 0.5.
   - Il y a une large plage de valeurs, avec des valeurs aberrantes significativement plus basses et quelques-unes plus élevées.
2. **Cluster 2** :
   - La médiane du taux de propriété est également d'environ 0.7.
   - Ce cluster a une plage de valeurs plus concentrée avec quelques valeurs aberrantes basses.
3. **Cluster 3** :
   - La médiane du taux de propriété est d'environ 0.9.

- Il y a plusieurs valeurs aberrantes basses.

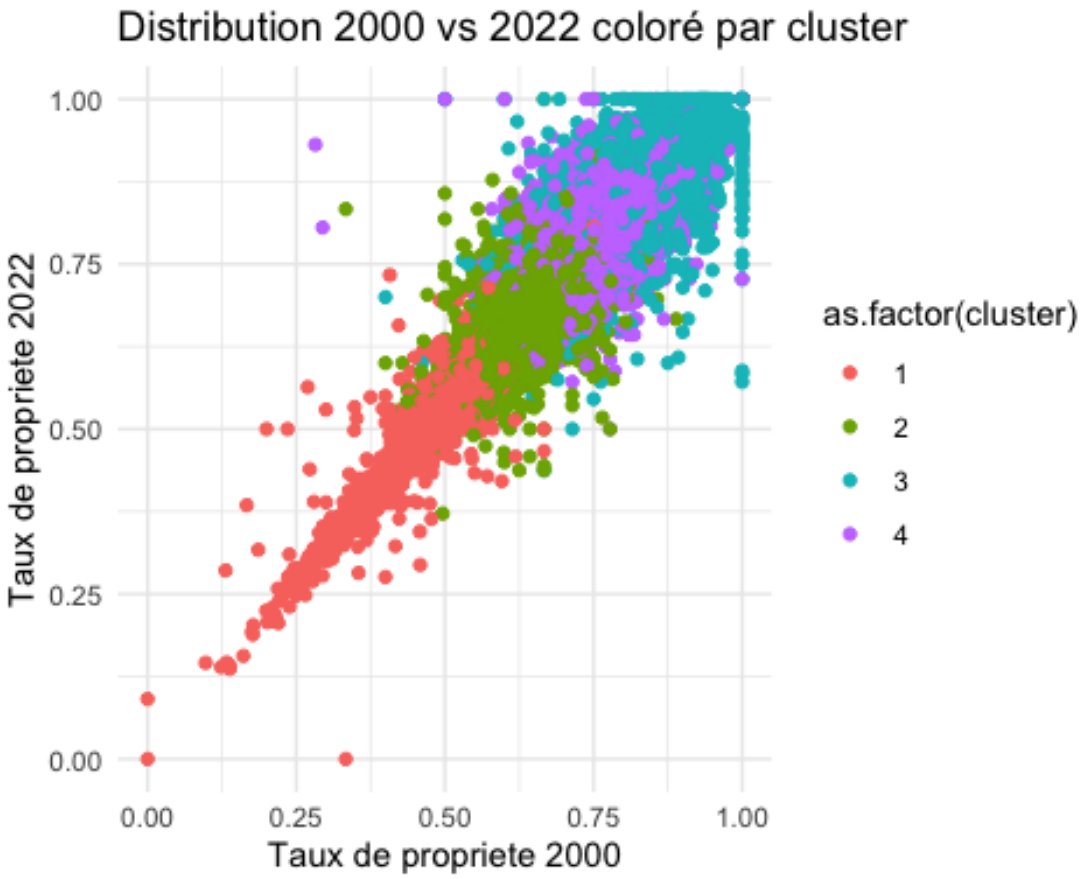
4. **Cluster 4** :
   - La médiane du taux de propriété est d'environ 0.8.
   - Il y a une distribution plus concentrée, mais avec quelques valeurs aberrantes basses et hautes.

3.2.5. Conclusions Possibles

- **Comparaison des Clusters** :
  - Les clusters 3 et 4 semblent avoir des taux de propriété plus élevés en général, avec des médianes autour de 0.9 et 0.8, respectivement.
  - Les clusters 1 et 2 ont des taux de propriété plus bas, avec des médianes autour de 0.5 et 0.7, respectivement.
- **Homogénéité des Clusters** :
  - Les clusters 3 et 4 sont plus homogènes en termes de taux de propriété, car ils ont des plages de valeurs plus concentrées et moins de dispersion.
  - Les clusters 1 et 2 montrent plus de variabilité avec des plages plus larges et plus de valeurs aberrantes.
- **Valeurs Aberrantes** :
  - La présence de nombreuses valeurs aberrantes peut indiquer des départements avec des caractéristiques spécifiques ou des anomalies dans les données.

3.3. Visualisation des Clusters avec d'autres Variables :

```
# Visualiser les clusters avec d'autres variables
ggplot(data_with_clusters, aes(x = `2000`, y = `2022`, color =
as.factor(cluster))) +
  geom_point() +
  labs(title = "Distribution 2000 vs 2022 coloré par cluster",
       x = "Taux de propriete 2000",
       y = "Taux de propriete 2022") +
  theme_minimal()
```



3.3.1. Interpretation :

Le graphique de dispersion que nous avons généré montre la relation entre les taux de propriété de 2000 et ceux de 2022, coloré par cluster.

1. **Points Colorés par Cluster** :
   - Les points sont colorés en fonction des clusters auxquels ils appartiennent, ce qui permet de visualiser la distribution des clusters en termes de taux de propriété sur ces deux années.

3.3.2. Analyse des Clusters

1. **Cluster 1 (Rouge)** :
   - Les points du cluster 1 sont principalement situés dans la partie inférieure gauche du graphique.
   - Ces départements ont tendance à avoir des taux de propriété relativement bas en 2000 et en 2022.
   - La majorité des points sont proches de la diagonale, indiquant une relation linéaire entre les taux de propriété de 2012 et 2022 pour ce cluster.
2. **Cluster 2 (Vert)** :
   - Les points du cluster 2 sont répartis dans la partie centrale du graphique.
   - Ces départements ont des taux de propriété modérés en 2000 et en 2022.

- Il y a une certaine variabilité dans ce cluster, avec des points plus dispersés autour de la diagonale.

3. **Cluster 3 (Bleu)** :
   - Les points du cluster 3 sont principalement situés dans la partie supérieure droite du graphique.
   - Ces départements ont tendance à avoir des taux de propriété relativement élevés en 2000 et en 2022.
   - La majorité des points sont proches de la diagonale, indiquant une forte relation linéaire entre les taux de propriété de 2000 et 2022 pour ce cluster.
4. **Cluster 4 (Violet)** :
   - Les points du cluster 4 sont répartis dans diverses parties du graphique, mais beaucoup sont dans la partie supérieure droite.
   - Ces départements montrent une grande variabilité avec certains ayant des taux de propriété très élevés en 2022.
   - La dispersion des points indique une variabilité plus élevée dans ce cluster comparé aux autres.

3.3.3. Observations Clés

1. **Relation Linéaire** :
   - La forte concentration de points le long de la diagonale (de 0 à 1) indique une relation linéaire positive entre les taux de propriété de 2000 et 2022. Cela signifie que les départements ayant un taux de propriété élevé en 2000 tendent à avoir également un taux de propriété élevé en 2022.
2. **Variabilité entre Clusters** :
   - Les clusters 1 et 3 montrent une relation plus forte et linéaire entre les années comparé aux clusters 2 et 4, qui présentent plus de variabilité.
   - Les clusters avec des couleurs différentes indiquent que les clusters capturent des différences significatives dans les tendances de propriété au fil des ans.

3.3.4. Conclusions

1. **Stabilité des Taux de Propriété** :
   - La majorité des départements maintiennent des taux de propriété similaires entre 2000 et 2022.
2. **Identité des Clusters** :
   - Les clusters permettent de distinguer les départements avec des taux de propriété différents et de capturer des variations importantes au fil du temps.
   - Les clusters les plus hauts (3 et 4) indiquent des départements avec des taux de propriété généralement plus élevés.

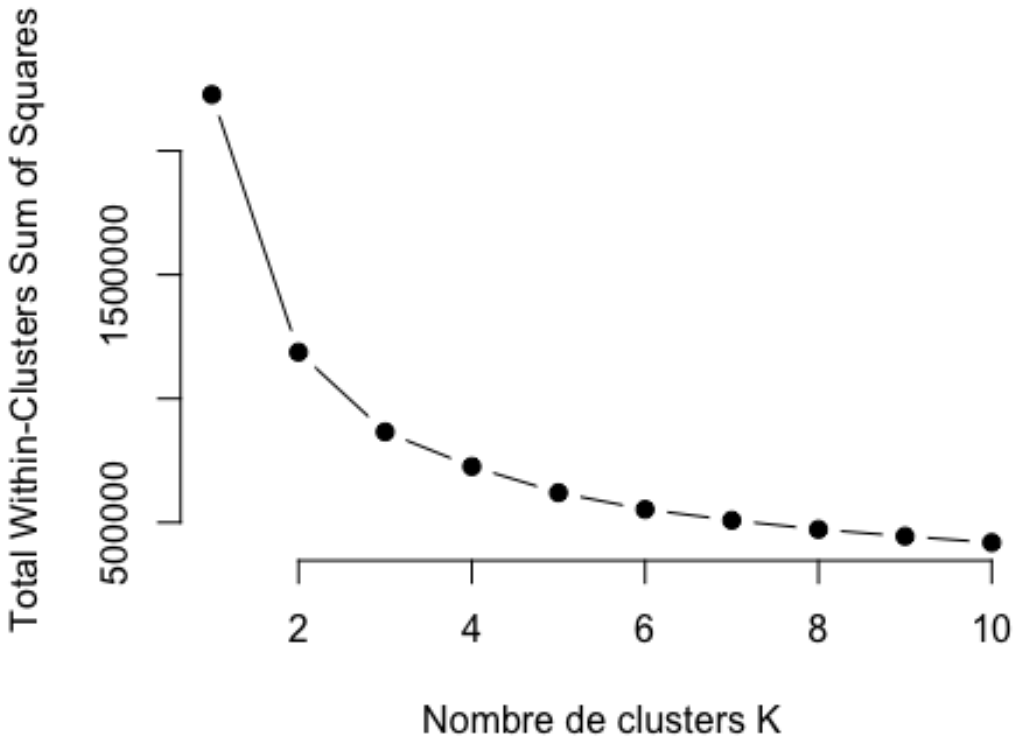
3.4. Identification du meilleur nombre de clusters à utiliser

3.4.1. Méthode du Coude

La méthode du coude implique de tracer la somme des distances intra-cluster (Within-cluster Sum of Squares, WCSS) pour différents nombres de clusters et de choisir le nombre de clusters où une “courbure” ou un “coude” apparaît.

Regardons le graphique et identifions le point où la réduction du WCSS commence à diminuer. Ce point est souvent appelé le “coude”. Le nombre de clusters correspondant à ce point est généralement considéré comme optimal.

```r
# Charger les packages nécessaires
#library(factoextra)
#library(cluster)

# Préparer les données (utiliser données transformées avec PCA ou autres)
data_n_clusters <- res.pca$ind$coord # Si PCA, sinon utiliser un dataframe approprié

# Calculer le WCSS pour différents nombres de clusters
wss <- function(k) {
  kmeans(data_n_clusters, k, nstart = 10)$tot.withinss
}

k.values <- 1:10

# Calculer le WCSS pour chaque valeur de k
wss_values <- map_dbl(k.values, wss)

# Tracer la méthode du coude
plot(k.values, wss_values,
     type = "b", pch = 19, frame = FALSE,
     title = "Methode du coude",
     xlab = "Nombre de clusters K",
     ylab = "Total Within-Clusters Sum of Squares")
```



3.4.2. Interpretation :

Le principe est de choisir le nombre de clusters où l'ajout d'un cluster supplémentaire n'améliore plus de manière significative la performance du modèle. Cela se traduit par un changement moins prononcé dans la somme des carrés des distances intra-cluster.

Pour interpréter ce graphique :

En regardant notre graphique, le coude semble se situer autour de **4 clusters**. À partir de ce point, l'amélioration de la somme des carrés intra-cluster devient beaucoup moins marquée.

Donc, selon ce graphe, **4 clusters** serait un bon choix pour le nombre de clusters dans notre analyse.

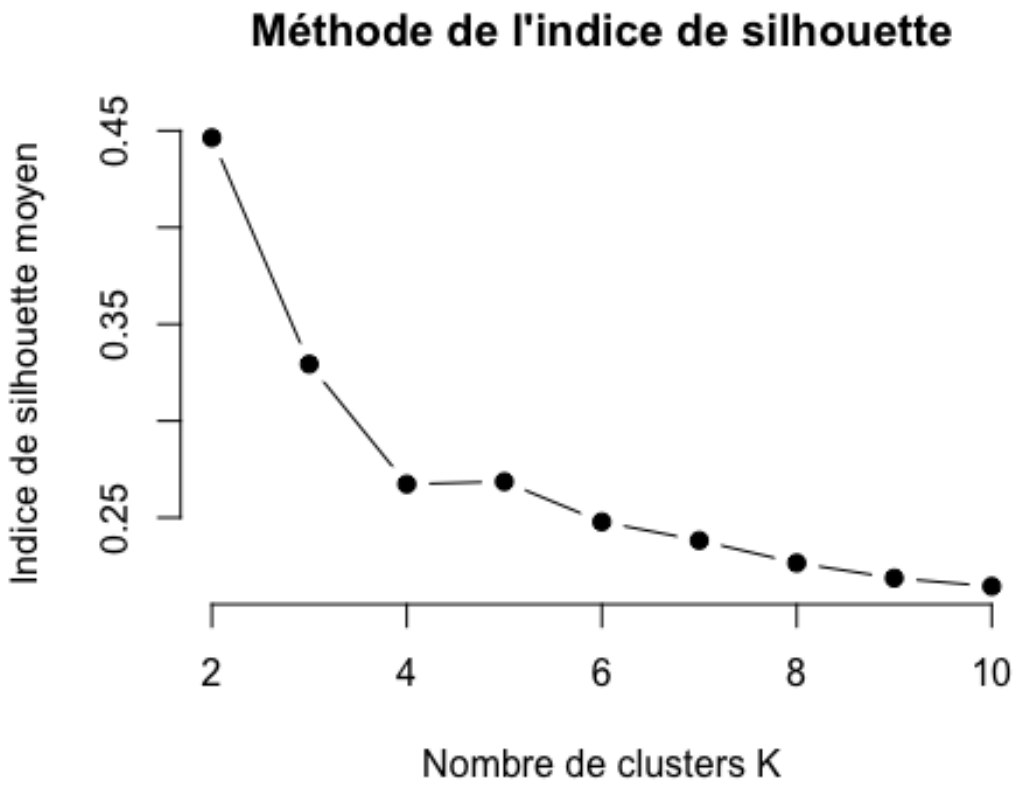
Pour confirmer ce choix, il est également recommandé de vérifier d'autres indicateurs tels que l'indice de silhouette que nous avons discuté précédemment. Une bonne pratique consiste à combiner plusieurs méthodes pour obtenir une vue plus complète et robuste sur le nombre optimal de clusters.

3.4.3. Méthode de l'Indice de Silhouette

La méthode de l'indice de silhouette évalue la qualité du clustering en calculant l'indice de silhouette pour différents nombres de clusters. Un indice de silhouette plus élevé indique des clusters plus cohérents.

Regardons le graphique et identifions le nombre de clusters avec l'indice de silhouette moyen le plus élevé. Ce nombre est considéré comme le plus optimal.

```
##      Dim.1              Dim.2               Dim.3              Dim.4        
##  Min.   :-46.632   Min.   :-15.76339   Min.   :-10.25267   Min.   :-8.2375  
##  1st Qu.: -3.198   1st Qu.: -1.99335   1st Qu.: -1.01562   1st Qu.:-0.5914  
##  Median :  1.246   Median :  0.02966   Median : -0.07699   Median :-0.0364  
##  Mean   :  0.000   Mean   :  0.00000   Mean   :  0.00000   Mean   : 0.0000  
##  3rd Qu.:  4.692   3rd Qu.:  2.01094   3rd Qu.:  0.93937   3rd Qu.: 0.5625  
##  Max.   : 16.210   Max.   : 17.99881   Max.   : 16.82065   Max.   :10.4597  
##      Dim.5        
##  Min.   :-9.93018  
##  1st Qu.:-0.48547  
##  Median : 0.00242  
##  Mean   : 0.00000  
##  3rd Qu.: 0.48557  
##  Max.   :39.87276
```



3.4.4. Interpretation :

L'indice de silhouette mesure à quel point chaque point de données est bien assigné à son cluster par rapport à d'autres clusters.

En observant le graphique, l'indice de silhouette moyen est le plus élevé pour **2 clusters**. Cependant, après 2 clusters, il y a encore une valeur relativement élevée pour **3 clusters**, puis une diminution significative par la suite. Cela suggère que 2 ou 3 clusters pourraient être des choix raisonnables.

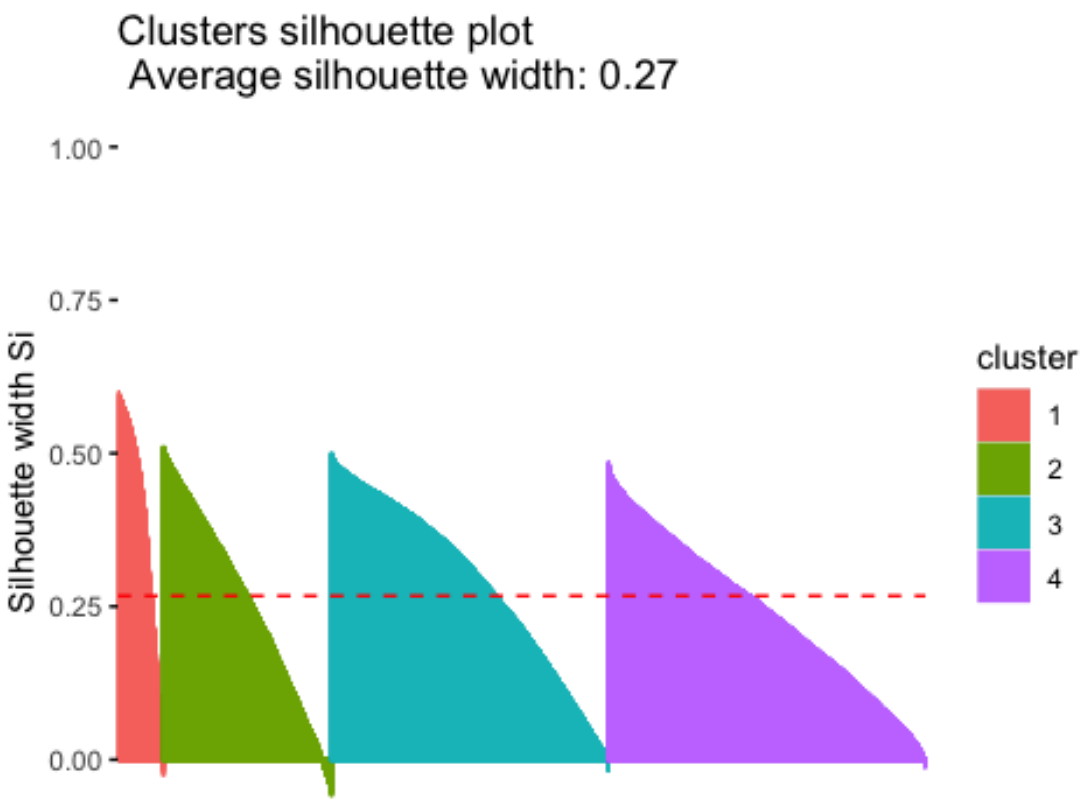
Pour déterminer plus précisément le meilleur nombre de clusters, nous pouvons considérer à la fois les résultats du "Elbow Plot" et de l'indice de silhouette : - Le "Elbow Plot" suggère 4 clusters. - L'indice de silhouette est le plus élevé pour 2 clusters, avec une bonne valeur pour 3 clusters.

Compte tenu de ces informations, **3 clusters** pourrait être un bon compromis, car il a un indice de silhouette encore élevé et est également raisonnablement bien positionné sur le "Elbow Plot". Toutefois, en fonction de notre contexte spécifique et des exigences de notre analyse, nous pourrions choisir 2, 3, ou 4 clusters.

3.5. Évaluer la Qualité du Clustering :

3.5.1. Visualisons les scores des silouhette et comparaison de chaque groupe

```
##   cluster  size ave.sil.width
## 1       1  2030          0.39
## 2       2  7481          0.25
## 3       3 12337          0.29
## 4       4 14125          0.24
```



3.5.2. Interprétation du Graphique de Silhouette

Le graphique de silhouette que nous avons obtenu permet d'évaluer la qualité du clustering.

1. **Largeur Moyenne de la Silhouette** :
   - La largeur moyenne de la silhouette est de 0.27.
   - La valeur de la largeur de silhouette varie de -1 à 1.
     - Une valeur proche de 1 indique que les points sont bien regroupés et séparés des autres clusters.
     - Une valeur proche de 0 indique que les points sont proches de la limite des clusters voisins.
     - Une valeur négative indique que les points sont probablement mal classifiés.
2. **Clusters Individuels** :
   - Chaque couleur représente un cluster différent.
   - Les segments colorés montrent la silhouette de chaque point au sein du cluster.
   - La ligne rouge en pointillés représente la largeur moyenne de la silhouette globale.

3.5.3. Analyse Détailée

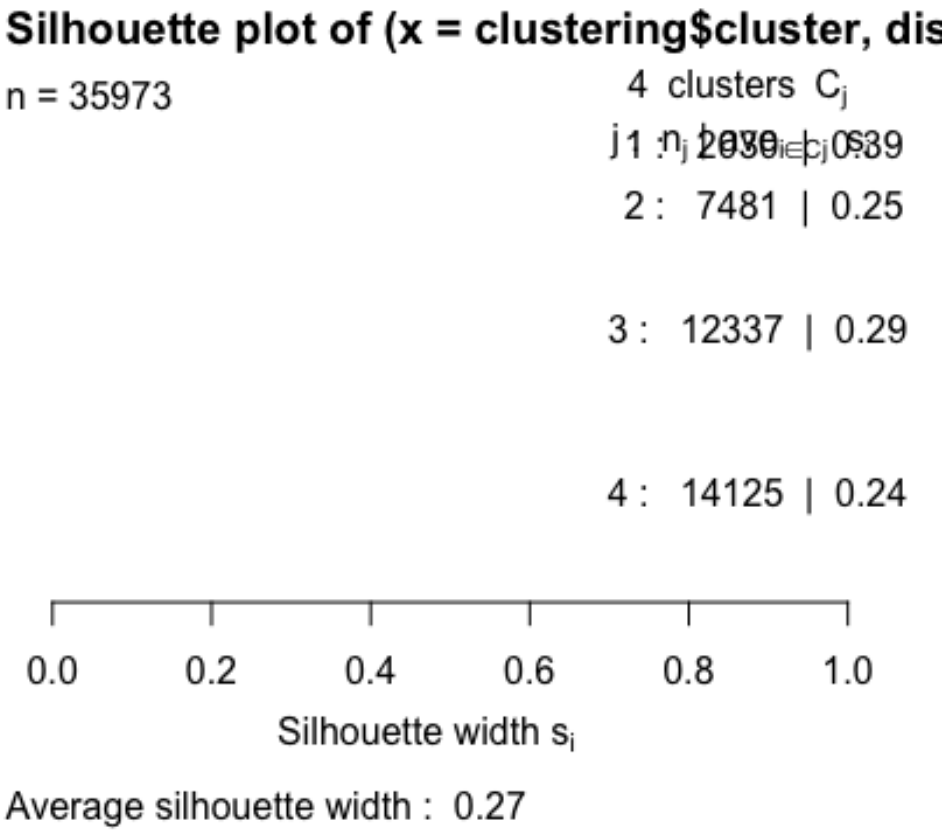
1. **Cluster 1 (Rouge)** :
   - Les points du Cluster 1 ont une silhouette variée, certains sont bien regroupés avec une silhouette proche de 0.5, tandis que d'autres sont proches de la limite de leurs voisins.
2. **Cluster 2 (Vert)** :
   - Le Cluster 2 montre une variation de silhouette similaire à celle du Cluster 1, mais la largeur moyenne est légèrement inférieure.
3. **Cluster 3 (Bleu)** :
   - Le Cluster 3 semble avoir une silhouette plus stable, avec la majorité des points ayant une valeur autour de 0.3 à 0.4.
4. **Cluster 4 (Violet)** :
   - Le Cluster 4 a une largeur de silhouette moyenne comparable aux autres clusters, avec une distribution relativement large des valeurs.

3.5.4. Conclusion

- **Qualité du Clustering** :
  - Une largeur moyenne de silhouette de 0.27 indique une qualité de clustering modérée. Ce n'est pas excellent mais pas non plus trop mauvais.
  - La largeur de silhouette montre que certains groupes sont mieux définis que d'autres.
- **Amélioration Possible** :
  - Nous pouvons essayer de varier le nombre de clusters pour voir si cela améliore la largeur moyenne de la silhouette.
  - L'analyse plus approfondie de chaque cluster pourrait révéler des points mal classifiés qui peuvent nécessiter une révision du modèle de clustering ou des données.

En résumé, ce graphique indique que le clustering est correct mais pourrait être amélioré pour obtenir des clusters mieux définis. Nous pourrons essayer différentes approches de clustering ou ajuster les paramètres pour améliorer les résultats.

3.5.5. Évaluation de 4 clusters



3.5.6. Interpretation :

- **Largeur de Silhouette par Cluster :**
  - Cluster 1 : Taille = 2030, Largeur moyenne de silhouette = 0.39
  - Cluster 2 : Taille = 7481, Largeur moyenne de silhouette = 0.25
  - Cluster 3 : Taille = 12337, Largeur moyenne de silhouette = 0.29
  - Cluster 4 : Taille = 14125, Largeur moyenne de silhouette = 0.24

Analyse

- **Cluster 1 :**

A la meilleure qualité de clustering avec une largeur moyenne de silhouette de 0.39. Les points dans ce cluster sont bien regroupés et distincts des autres clusters.

- **Cluster 3 :**

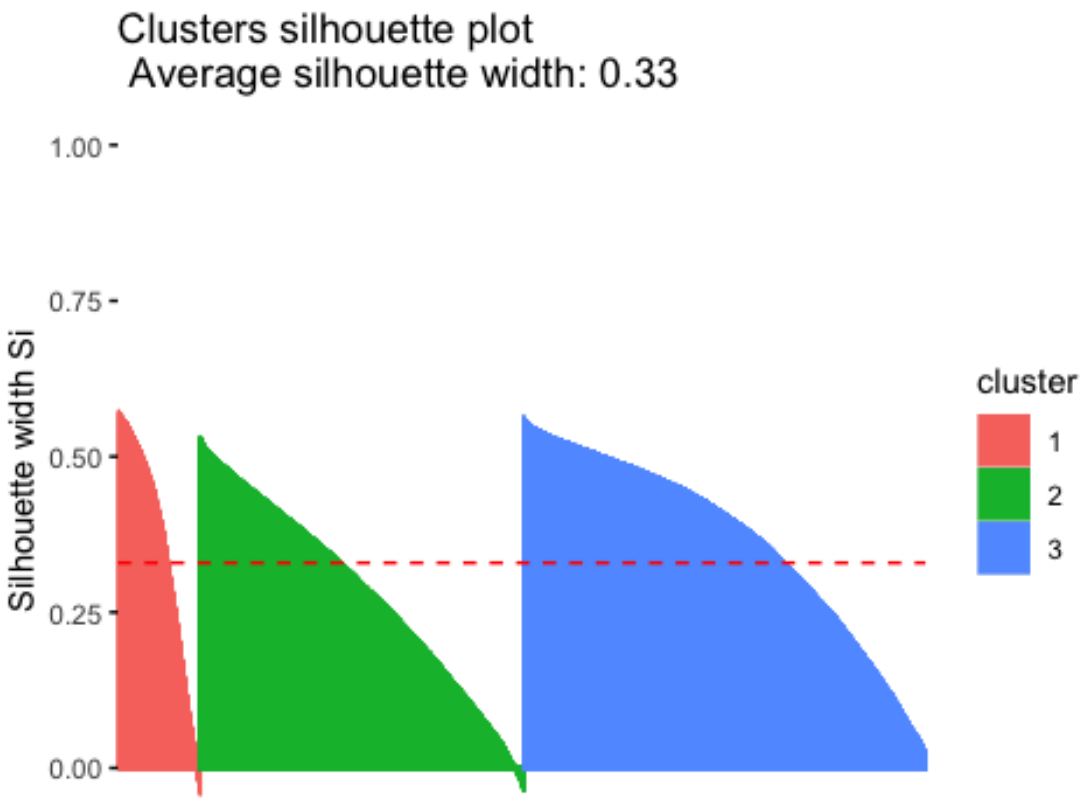
Relativement bon clustering avec une largeur moyenne de silhouette de 0.29.

- **Cluster 2 et 4 :**

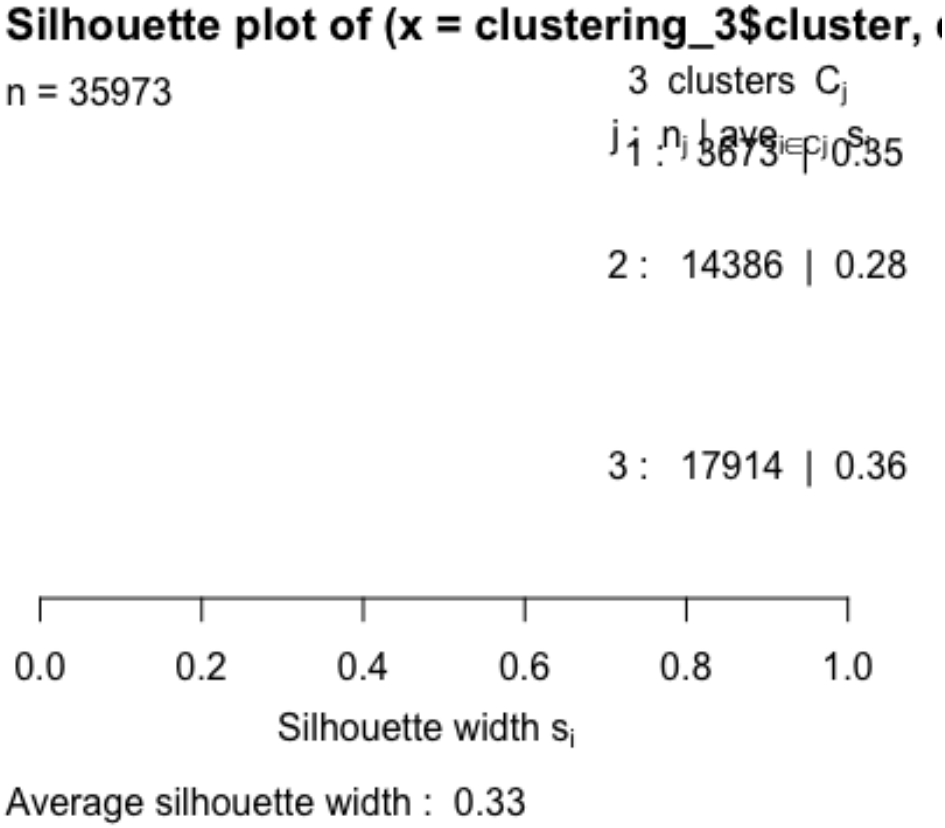
Ont des largeurs de silhouette plus basses (0.25 et 0.24 respectivement), indiquant que les points dans ces clusters sont moins distincts et peuvent être plus proches des clusters voisins.

3.6. Cas de 3 clusters

```
##   cluster  size ave.sil.width
## 1       1  3673          0.35
## 2       2 14386          0.28
## 3       3 17914          0.36
```

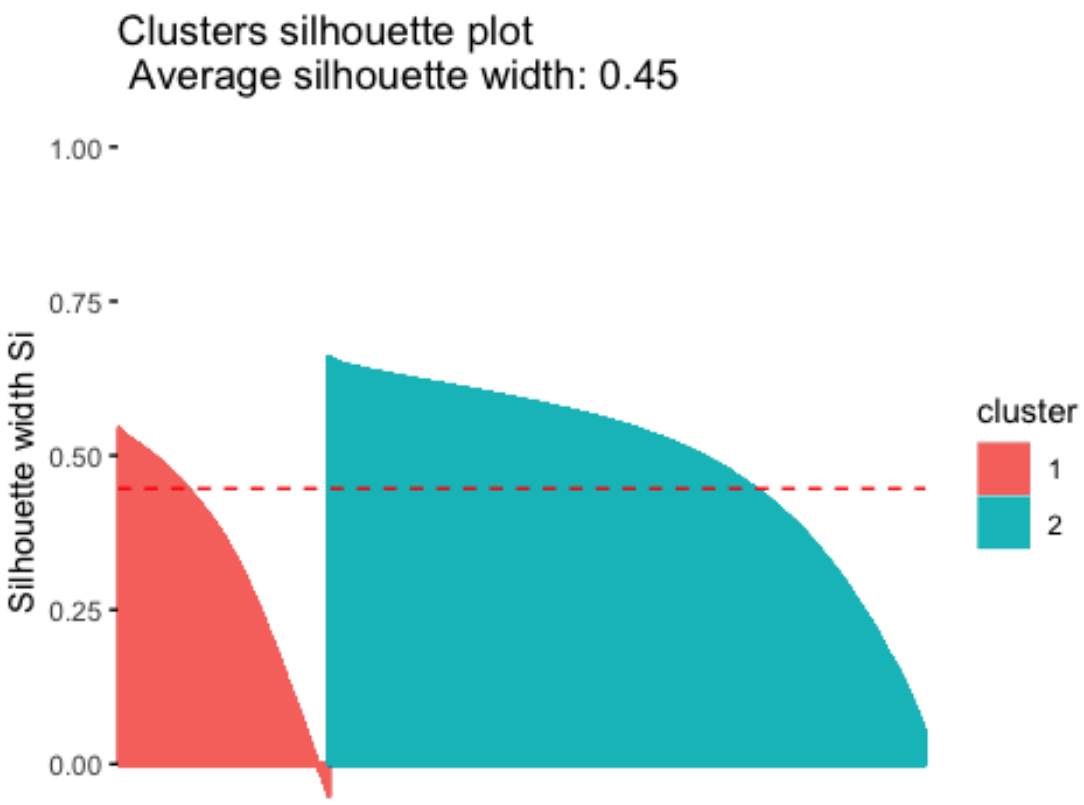


3.6.1. Évaluation de 3 clusters

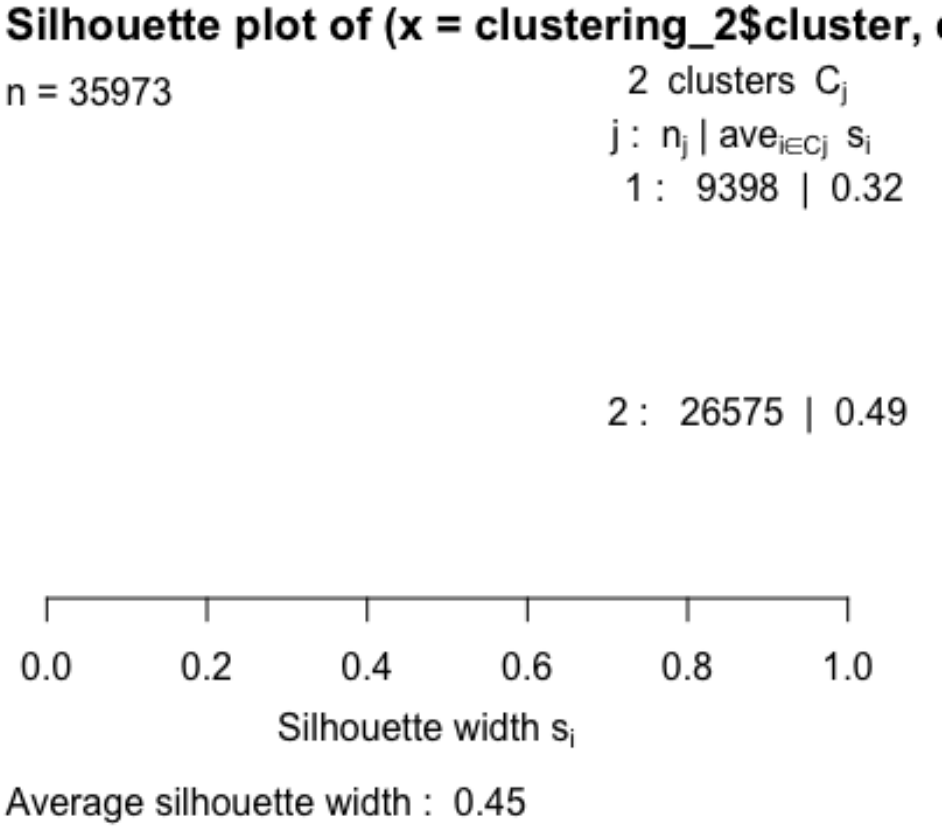


3.7. Cas de 2 clusters

```
##   cluster  size ave.sil.width
## 1       1  9398          0.32
## 2       2 26575          0.49
```


Clusters silhouette plot
Average silhouette width: 0.45
Silhouette width Si
1.00
0.75
0.50
0.25
0.00
cluster
1
2

3.7.1. Évaluation de 2 clusters




```r
#install.packages("patchwork")

# Charger les packages nécessaires
#library(cluster)
#library(factoextra)
library(patchwork) #permet de combiner plusieurs graphiques ggplot2 dans un seul

# Effectuer le clustering avec 3, 4, et 5 clusters
set.seed(123)
clustering_2 <- kmeans(res.pca$ind$coord, centers = 2)
clustering_3 <- kmeans(res.pca$ind$coord, centers = 3)
clustering_4 <- kmeans(res.pca$ind$coord, centers = 4)

# Calculer les scores de silhouette pour chaque clustering
silhouette_scores_2 <- silhouette(clustering_2$cluster,
dist(res.pca$ind$coord))
silhouette_scores_3 <- silhouette(clustering_3$cluster,
dist(res.pca$ind$coord))
silhouette_scores_4 <- silhouette(clustering_4$cluster,
dist(res.pca$ind$coord))
```

```r
# Visualiser les scores de silhouette
p_2 <- fviz_silhouette(silhouette_scores_2) + labs(title = "2 Clusters")
```

```
##   cluster  size ave.sil.width
## 1       1  9398          0.32
## 2       2 26575          0.49
```

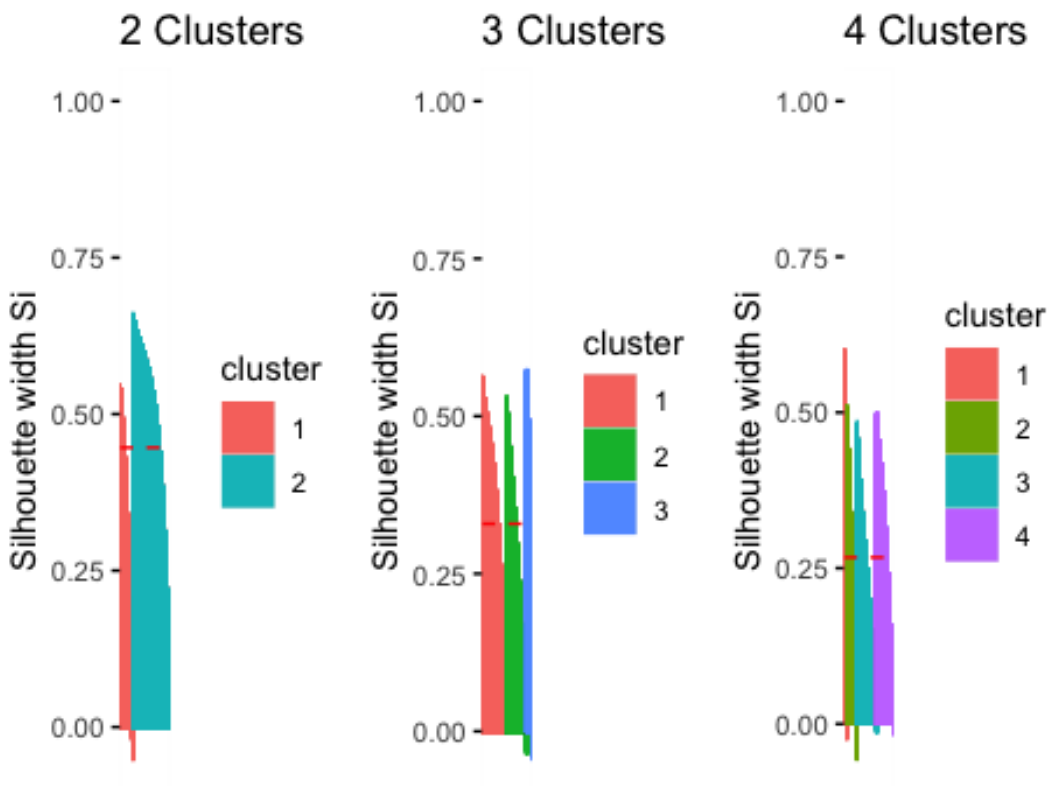
```r
p_3 <- fviz_silhouette(silhouette_scores_3) + labs(title = "3 Clusters")
```

```
##   cluster  size ave.sil.width
## 1       1 17897          0.36
## 2       2 14394          0.28
## 3       3  3682          0.35
```

```r
p_4 <- fviz_silhouette(silhouette_scores_4) + labs(title = "4 Clusters")
```

```
##   cluster  size ave.sil.width
## 1       1  2034          0.39
## 2       2  7497          0.25
## 3       3 14153          0.24
## 4       4 12289          0.29
```

```r
# Afficher les graphiques côte à côte
p_2 + p_3 + p_4
```



3.7.2. Interpretation :

Pour déterminer quel nombre de clusters est le plus approprié, nous pourrons comparer les plots de silhouette pour les différents nombres de clusters. Voici les critères que nous pouvons utiliser pour l'analyse :

1. **Largeur de silhouette moyenne** :
   - Plus la largeur moyenne est élevée, mieux c'est. La largeur de silhouette moyenne est indiquée en bas de chaque graphique. Dans notre cas :
   - Pour 2 clusters : 0.45
   - Pour 3 clusters : 0.33
   - Pour 4 clusters : 0.27

La largeur de silhouette moyenne pour 3 clusters est la plus élevée.

3.7.3. Clustering avec 2 clusters :

```
- Largeur moyenne de la silhouette : 0.45
- L'indice de silhouette moyen est le plus élevé, proche de 0.5, indiquant
une bonne séparation des clusters.
- Cette configuration présente une légère amélioration pour nos clusters, les
silhouettes sont larges et homogènes, suggérant une bonne cohésion des points
à l'intérieur des clusters.
```

3.7.4. Clustering avec 3 clusters :

```
- Largeur moyenne de la silhouette : 0.33
- L'indice de silhouette moyen diminue par rapport à 2 clusters mais reste
raisonnablement élevé, environ 0.3.
- La forme des silhouettes est moins homogène, mais il y a toujours une
certaine cohésion à l'intérieur des clusters.
```

3.7.5. Clustering avec 4 clusters :

```
- Largeur moyenne de la silhouette : 0.27
- L'indice de silhouette moyen diminue encore plus, autour de 0.25, indiquant
une qualité de clustering plus faible.
- Les silhouettes sont plus variées en taille et forme, ce qui peut suggérer
que certains points sont mal assignés ou que les clusters se chevauchent.
```

2. **Meilleur nombre de clusters** :

En comparant les indices de silhouette moyens :

- 2 clusters semblent offrir la meilleure qualité de clustering avec la meilleure cohésion et séparation.
- 3 clusters peuvent également être considérés comme une option raisonnable si nous avons besoin de plus de détails dans notre clustering, bien que la qualité soit légèrement inférieure.
- 4 clusters présentent une qualité de clustering nettement inférieure et devraient probablement être évités.

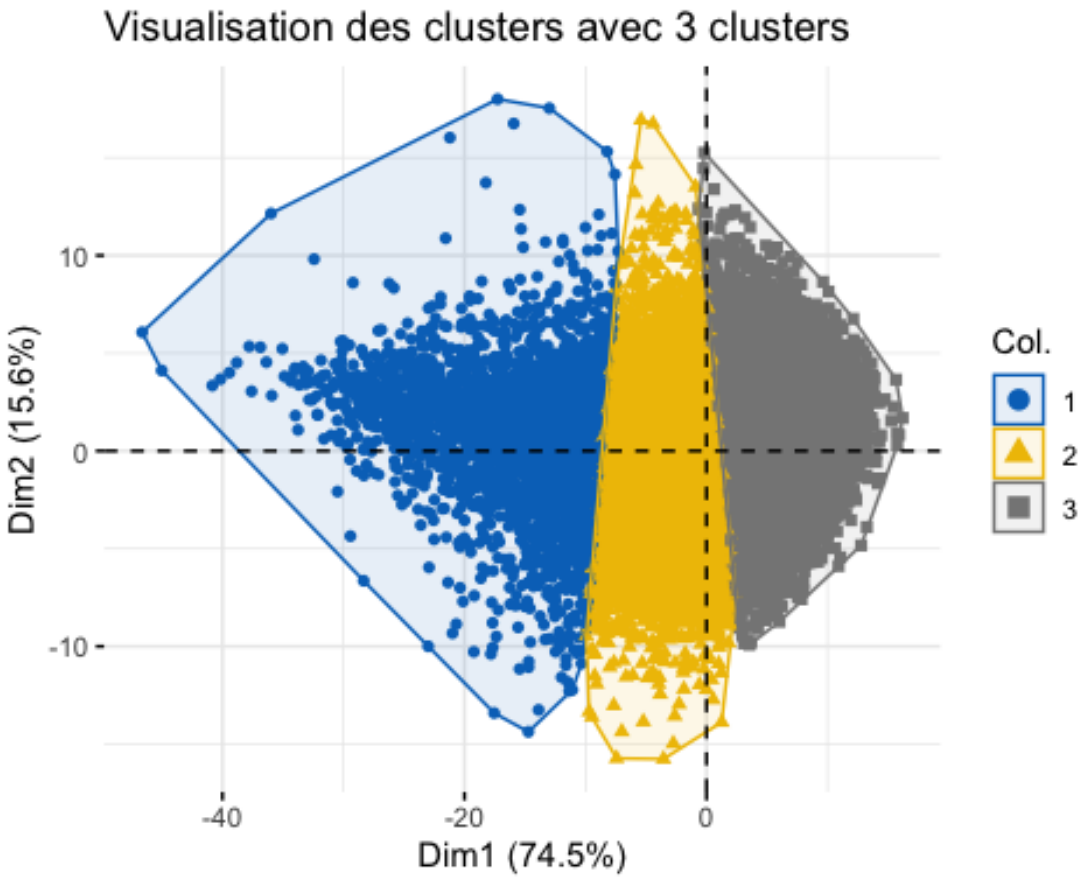
En général, une valeur moyenne de la silhouette au-dessus de 0.5 est considérée comme bonne.

3.7.6.Conclusion :

Sur la base des largeurs moyennes de silhouette, le clustering avec **2 clusters** semble être la meilleure option parmi les trois configurations testées. Il présente la plus haute valeur moyenne de silhouette, ce qui indique une meilleure séparation et compacité des clusters (0.45).

3.8. Clustering avec 3 groupes

```r
# Clustering avec 3 clusters (le meilleur selon les scores de silhouette)
clustering_3 <- kmeans(res.pca$ind$coord, centers = 3)
data_with_clusters_3 <- data.frame(res.pca$ind$coord, cluster =
as.factor(clustering_3$cluster))

# Visualiser les clusters
fviz_pca_ind(res.pca, geom.ind = "point", col.ind =
data_with_clusters_3$cluster, palette = "jco", addEllipses = TRUE,
ellipse.type = "convex", repel = TRUE) +
  labs(title = "Visualisation des clusters avec 3 clusters")
```



3.8.2. Interpretation :

Ce graphique montre la visualisation des clusters en utilisant l'analyse en composantes principales (PCA) et le clustering K-means avec 3 clusters.

1. **Axes Dim1 et Dim2** :
   - **Dim1** (Dimension 1) explique 74.5% de la variance totale dans les données.
   - **Dim2** (Dimension 2) explique 15.6% de la variance totale dans les données.
   - Ensemble, ces deux dimensions expliquent 90.1% de la variance totale, ce qui est une bonne représentation des données en deux dimensions.
2. **Points et Clusters** :
   - Les points représentent les observations (communes, dans notre cas).
   - Les couleurs et les formes des points indiquent à quel cluster appartient chaque observation.
   - Il y a trois clusters, chacun représenté par une couleur différente :
     - **Cluster 1** en bleu (points circulaires)
     - **Cluster 2** en jaune (triangles)
     - **Cluster 3** en gris (carrés)
3. **Interprétation des Clusters** :

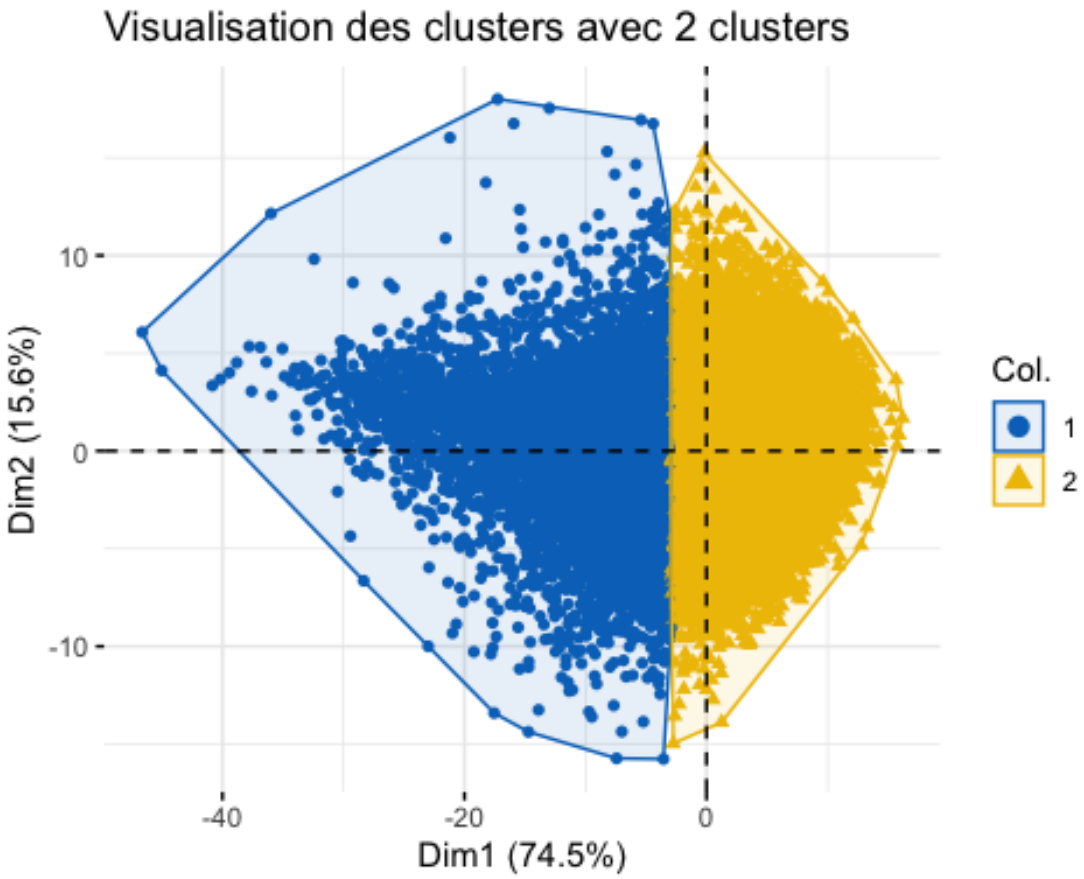
- Les clusters sont relativement bien séparés, avec des groupes distincts de points.
- Les ellipses montrent que chaque cluster a une certaine dispersion, mais les clusters sont globalement bien délimités.
- Le cluster bleu (1) semble plus dispersé que les autres, couvrant une plus grande aire dans les deux dimensions.
- Le cluster jaune (2) et le cluster gris (3) sont plus compacts.

3.8.3. Conclusion

Ce graphique montre une bonne séparation des clusters, ce qui suggère que le choix de 3 clusters est raisonnable pour nos données. Chaque cluster représente un groupe distinct de communes avec des caractéristiques similaires en termes de taux de propriété. Les deux premières dimensions de la PCA expliquent la majorité de la variance, ce qui permet une bonne visualisation en deux dimensions.

3.9. Clustering avec 2 groupes

```
# Clustering avec 3 clusters (le meilleur selon les scores de silhouette)
clustering_2 <- kmeans(res.pca$ind$coord, centers = 2)
data_with_clusters_2 <- data.frame(res.pca$ind$coord, cluster =
as.factor(clustering_2$cluster))

# Visualiser les clusters
fviz_pca_ind(res.pca, geom.ind = "point", col.ind =
data_with_clusters_2$cluster, palette = "jco", addEllipses = TRUE,
ellipse.type = "convex", repel = TRUE) +
  labs(title = "Visualisation des clusters avec 2 clusters")
```



3.9.1. Interpretation :

Bonne Qualité de Clustering : Le graphique montre que les clusters sont bien séparés et cohérents. Cela suggère que l'utilisation de 2 clusters pour ce jeu de données est une bonne option. Réduction Dimensionnelle Efficace : La PCA a réussi à réduire les données à deux dimensions tout en capturant la majorité de la variance, facilitant ainsi la visualisation et l'interprétation. On garde alors nos 2 clusters pour la suite de l'analyse.

Analyse des tendances temporelles des taux de propriété au sein de chaque cluster.

3.10. Cas du Taux de propriété comparatif 2000 et 2022 en box

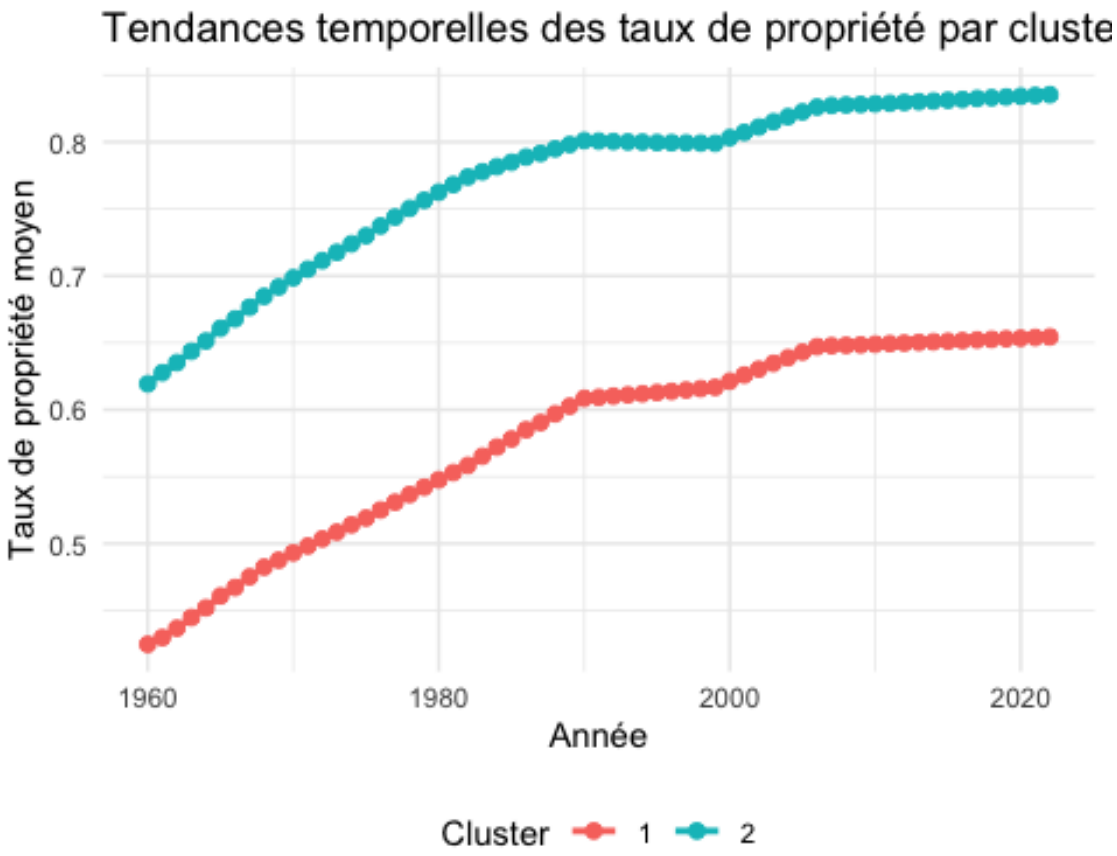
3.10.1. Interpretation :

Les box suivent légèrement le même alignement sur ces 20 années.

- **Variation des Taux de Propriété par Cluster :**
  - Cluster 1 a une distribution plus large avec une médiane plus basse en 2000 et en 2022. Cela pourrait indiquer que les zones dans ce cluster ont des taux de propriété plus variables et généralement plus bas.

- Cluster 4 montre des taux de propriété plus élevés et moins de dispersion par rapport aux autres clusters en 2000 et 2022, suggérant une homogénéité et des taux de propriété élevés dans ces zones.

- **Changements au Fil du Temps :**
  - Comparer les boxplots de 2000 et 2022 nous permet de voir comment les taux de propriété ont changé. Par exemple, si un cluster montre une augmentation de la médiane, cela indique une augmentation générale des taux de propriété dans ces zones.
  - Observer les changements dans la dispersion des données (largeur des boxplots et présence d'outliers) peut donner des indications sur la stabilité ou la variabilité des taux de propriété au fil du temps. Dans notre cas, les box gardent les mêmes proportions.



3.10.2. Interprétation du Graphique :

Ce graphique montre les tendances temporelles des taux de propriété moyen par cluster, de 1960 à 2022. Chaque ligne représente l'évolution du taux de propriété moyen pour un cluster spécifique.

1. **Clusters Définis** :
   - Les clusters sont représentés par trois couleurs différentes :

- **Cluster 1** en rouge.
- **Cluster 2** en bleu.

2. **Évolution des Taux de Propriété** :
   - **Cluster 1 (Rouge)** :
     - Ce cluster montre un taux de propriété moyen qui commence autour de 0.40 en 1960 et augmente régulièrement pour atteindre environ 0.6 en 1990 pour atteindre 0.65 en 2022
   - **Cluster 2 (Bleu)** :
     - Le taux de propriété moyen pour ce cluster commence autour de 0.60 en 1960 et augmente régulièrement pour atteindre environ 0.8 en 1990 et se stabilise 10ans puis une augmentation légère.
3. **Observations Générales** :
   - **Tendance Générale à la Hausse** :
     - Tous les clusters montrent une tendance générale à la hausse des taux de propriété au fil du temps, indiquant une augmentation générale de la propriété des logements dans toutes les régions.
     - Le cluster 2 a systématiquement un taux de propriété moyen plus élevé que le cluster 1 sur toute la période.
     - Les deux clusters montrent une tendance à la hausse, indiquant une augmentation générale du taux de propriété au fil du temps dans les deux groupes.
     - Les courbes subissent un ralentissement dans les années 1990 et semblent se stabiliser après 2000, suggérant que l'augmentation du taux de propriété ralentit.
   - **Différences entre les Clusters** :
     - Le cluster 2 a systématiquement les taux de propriété les plus élevés tout au long de la période étudiée.
     - Le cluster 1 a les taux de propriété les plus bas au début de la période et continue de l'être tout au long de la période.

3.10.3. Conclusion

Le graphique révèle des tendances distinctes dans les taux de propriété parmi les clusters identifiés. Les régions appartenant au cluster 2 ont les taux de propriété les plus élevés, tandis que celles du cluster 1 ont les taux les plus bas. Les taux de propriété augmentent pour tous les clusters, indiquant une tendance globale à la hausse de la propriété des logements au fil du temps. Cette information peut être utilisée pour approfondir l'analyse des facteurs spécifiques qui influencent les taux de propriété dans ces clusters et pour formuler des recommandations politiques adaptées.

4. Analyse des Facteurs Contributifs :

4.1. Introduction des coordonnées communale et régionale de la france

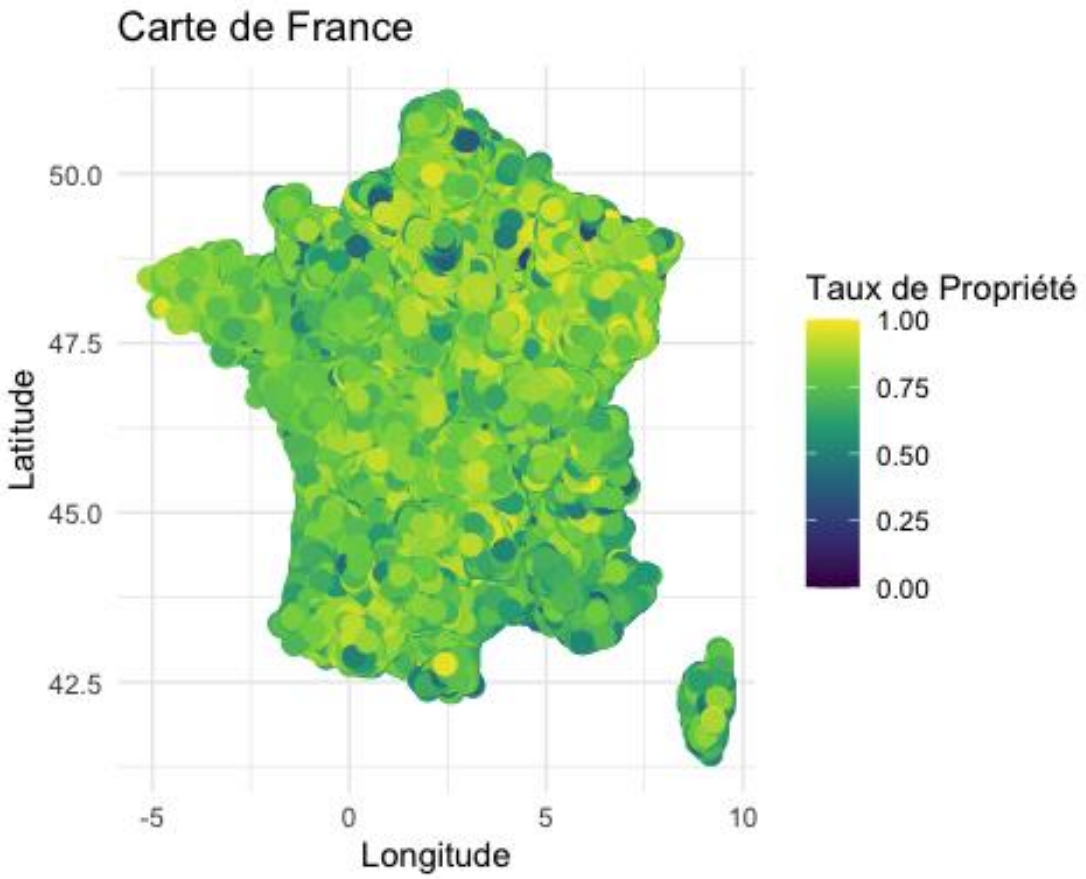
```
# Chargement des données
citiesexport <- read.csv("citiesexport.csv")
```

```r
# Renommage des colonnes de citiesexport pour correspondre aux colonnes de proprietaires_communes
names(citiesexport) <- c("codecommune", "city_code", "zip_code", "label", "latitude", "longitude",
                         "department_name", "department_number", "region_name", "region_geojson_name")

# Vérification des premières lignes des données pour s'assurer que les noms de colonnes sont corrects
head(citiesexport)
```

```
##   codecommune           city_code zip_code               label latitude
## 1       25620       ville du pont    25650       ville du pont 46.99987
## 2       25624      villers grelot    25640      villers grelot 47.36151
## 3       25615 villars les blamont    25310 villars les blamont 47.36838
## 4       25619       les villedieu    25240       les villedieu 46.71391
## 5       25622       villers buzon    25170       villers buzon 47.22856
## 6       25625    villers la combe    25510    villers la combe 47.24081
##   longitude department_name department_number             region_name
## 1  6.498147           doubs                25 bourgogne-franche-comté
## 2  6.235167           doubs                25 bourgogne-franche-comté
## 3  6.871415           doubs                25 bourgogne-franche-comté
## 4  6.265831           doubs                25 bourgogne-franche-comté
## 5  5.852187           doubs                25 bourgogne-franche-comté
## 6  6.473842           doubs                25 bourgogne-franche-comté
##       region_geojson_name
## 1 Bourgogne-Franche-Comté
## 2 Bourgogne-Franche-Comté
## 3 Bourgogne-Franche-Comté
## 4 Bourgogne-Franche-Comté
## 5 Bourgogne-Franche-Comté
## 6 Bourgogne-Franche-Comté
```

```r
#head(data_f)

# Fusionn des données par codecommune
merged_data <- merge(data_f, citiesexport, by = "codecommune")

# Vérification des premières lignes des données fusionnées
#head(merged_data)
```

(dd031722-d22b-45ce-a05b-f31fa07cf1b6)



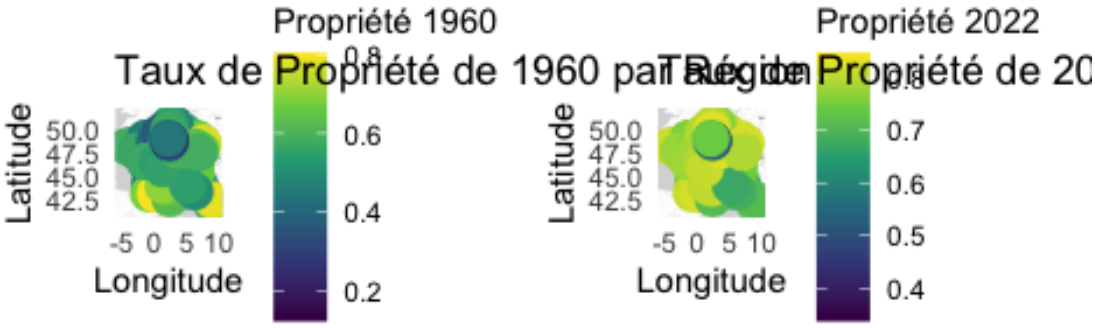
Cette carte nous permettra de mettre en évidence les taux de propriété par commune. Aucune communee ne se détache par son taux de propriété. Il faudait étudier la variation de ces taux sur plusieurs années ou regrouper par région pour mieux identifer les nuances.

4.2. Visualisation des Variations du taux de propriété sur une Carte

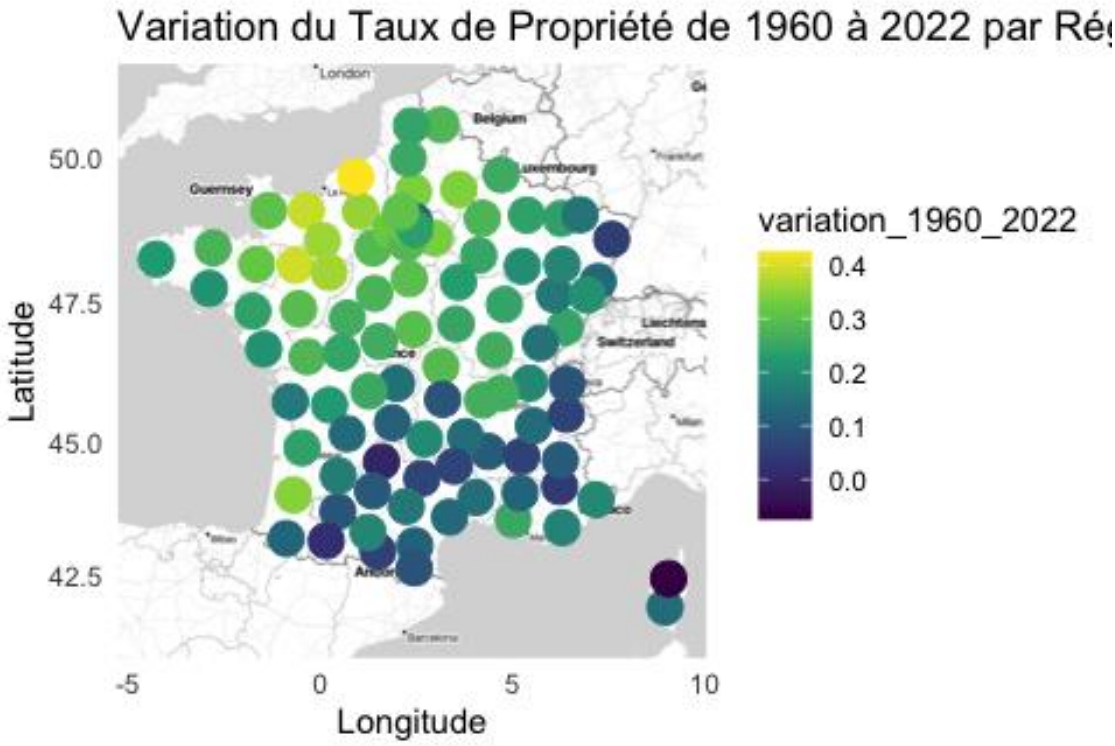
Taux de propriété en 1960 et 2022

```
library(patchwork)

#calcul de la variation du taux de propriété entre 1960 et 2022
#merged_data_region$variation_1960_2022 <- merged_data_region$`2022` -
merged_data_region$`1960`

# Visualisation des taux de propriété de 1960 et 2022 sur la carte par région
p2 <- ggmap(base_map) +
  geom_point(data = merged_data_region, aes(x = Longitude_centre, y =
Latitude_centre, color = `1960`), size = 5) +
  scale_color_viridis_c(option = "D", begin = 0, end = 1, name = "Propriété
1960") +
  theme_minimal() +
  labs(title = "Taux de Propriété de 1960 par Région", x = "Longitude", y =
"Latitude")
```

```r
p3 <- ggmap(base_map) +
  geom_point(data = merged_data_region, aes(x = Longitude_centre, y =
Latitude_centre, color = `2022`), size = 5) +
  scale_color_viridis_c(option = "D", begin = 0, end = 1, name = "Propriété
2022") +
  theme_minimal() +
  labs(title = "Taux de Propriété de 2022 par Région", x = "Longitude", y =
"Latitude")

# Combinaison des graphiques côte à côte
p2 + p3
```




```
## ℹ © Stadia Maps © Stamen Design © OpenMapTiles © OpenStreetMap
contributors.
```



4.2.1. Interprétation du Graphique

Le graphique montre la variation du taux de propriété de 1960 à 2022 par région en France. :

1. **Couleurs des Points :**
   - Les points colorés représentent la variation du taux de propriété entre 1960 et 2022 pour chaque région.
   - La légende à droite indique les valeurs de variation, où :
     - Les couleurs allant du violet au jaune représentent une augmentation du taux de propriété.
     - Les couleurs plus claires (jaune) indiquent une augmentation plus élevée.
     - Les couleurs plus foncées (violet) indiquent une augmentation plus faible.

À ce niveau, nous observons au vu de nos données que les variations (positive) du taux de propriété est plus élévé au nord de la france et encore plus au nord est. Les populations de ces régions ou d'ailleurs achètent de plus en plus dans ces régions du nord-est (le Havre et autres …) .

2. **Distribution Géographique :**

- Les points sont placés au centre géographique des régions
- La répartition des couleurs sur la carte montre comment le taux de propriété a évolué dans différentes parties de la France.

3. **Observations Clés :**
   - Certaines régions ont connu des augmentations significatives du taux de propriété (couleurs plus claires).
   - D'autres régions ont eu des augmentations moins importantes ou stables (couleurs plus foncées).

4.2.2. Analyse des Variations Régionales

1. **Régions avec Forte Augmentation :**
   - Les régions avec des points plus clairs (vers le jaune) ont connu une forte augmentation du taux de propriété entre 1960 et 2022.
   - Ces régions peuvent avoir bénéficié de politiques favorables, de développements économiques, ou de facteurs socio-économiques spécifiques.l'engouement pour ces régions est peut-être du à la qualité de vie qui s'y trouve et divers facteurs regionaux.
2. **Régions avec Faible Augmentation :**
   - Les régions avec des points plus foncés (vers le violet) ont eu une faible augmentation du taux de propriété.
   - Ces régions peuvent avoir été affectées par des défis économiques, un manque de développement immobilier, ou des politiques moins favorables.

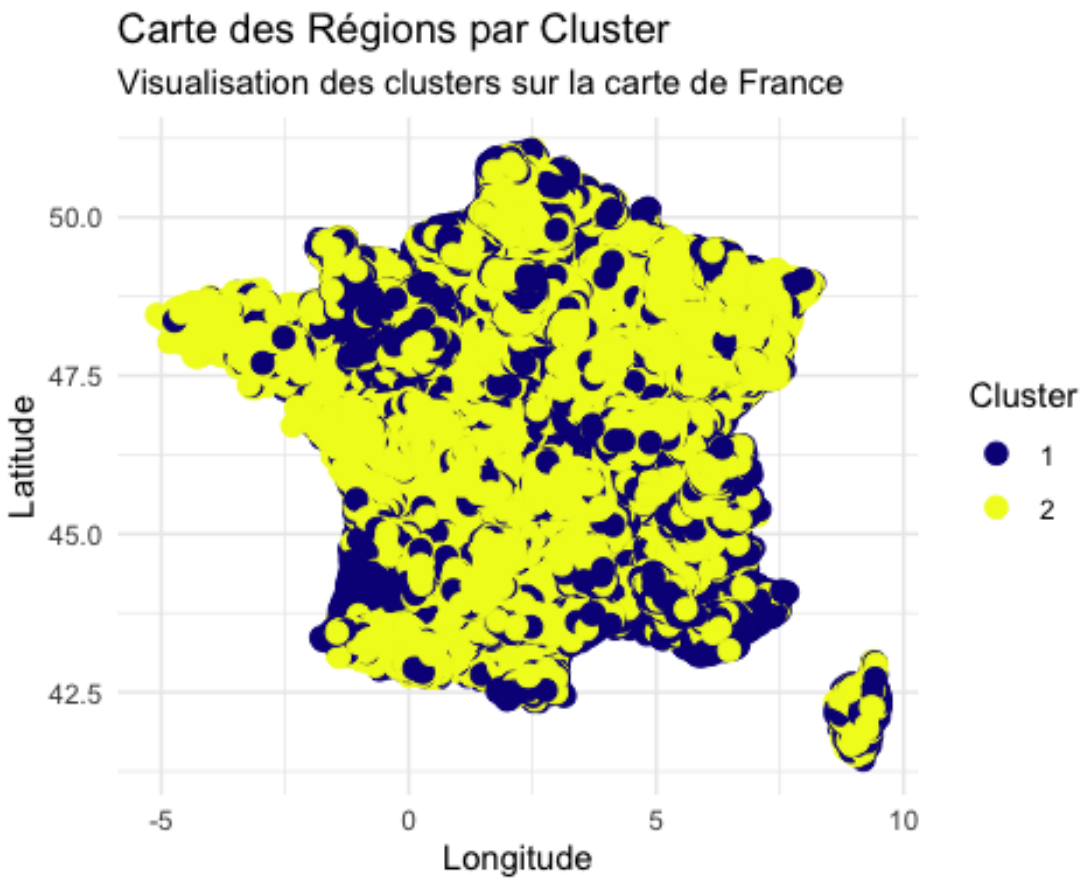
4.2.3. Conclusion

Le graphique fournit une vue d'ensemble des variations du taux de propriété en France sur une période de 62 ans. Il met en évidence les différences régionales et peut être utilisé pour identifier les zones nécessitant une attention particulière en termes de politiques de logement et de développement.

Pour une analyse plus approfondie, nous pourrions : - Comparer ces résultats avec des facteurs économiques et sociaux (par exemple, revenus, taux de chômage). - Examiner des politiques spécifiques qui ont été mises en place dans les régions avec des variations significatives. - Étudier les tendances à plus long terme pour voir si les variations observées sont récentes ou s'inscrivent dans une tendance plus large.

En résumé, ce graphique est un outil puissant pour visualiser et analyser les variations régionales du taux de propriété, aidant à comprendre les dynamiques sous-jacentes et à informer les décisions politiques et économiques.

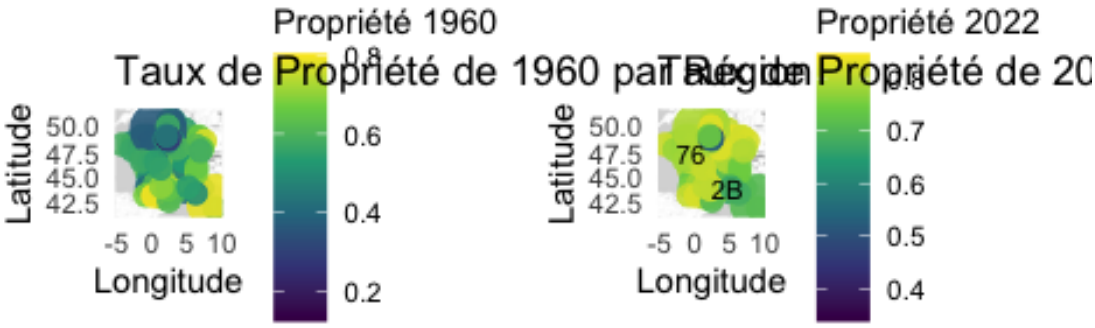
4.3. Visualisation de nos régions par Cluster :



4.3.1. Identification des Régions avec les Taux de Propriété les Plus Élevés ou les Plus Faibles

4.3.2. Calculons les moyennes des taux de propriété par région : Identifiez les régions avec les taux de propriété les plus élevés ou les plus faibles.

```
#head(regions_plus_elevees)

#regions_plus_faibles
```



4.3.3. Interpretation :

Certaines régions (représentées en jaune) avaient déjà un taux de propriété élevé en 1960. La carte de 2022 montre une augmentation générale des taux de propriété dans la plupart des régions, avec des points jaunes plus nombreux par rapport à 1960. Les régions qui étaient en violet ou en vert en 1960 mais qui sont passées au jaune en 2022 montrent une augmentation significative du taux de propriété. Les régions avec des variations minimes peuvent apparaître dans des couleurs similaires sur les deux cartes.

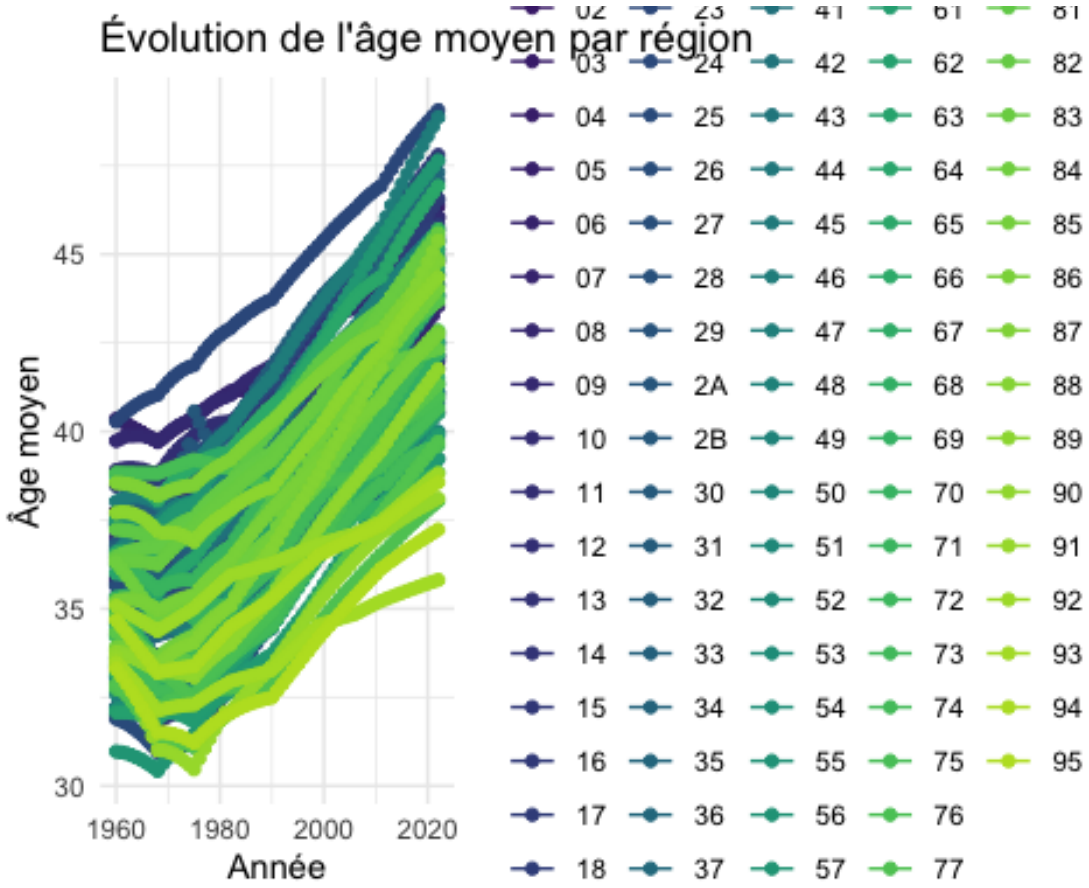
5. Introduction de l'age des populations :

Population vivant dans ces département pour expliquer la variation du taux de propriété dans certaines régions.

5.1. Prétraitement et évolution de l'âge moyen par région

```
## `summarise()` has grouped output by 'dep'. You can override using the
`.groups`
## argument.
## `summarise()` has grouped output by 'dep'. You can override using the
`.groups`
## argument.
```

```
## # A tibble: 6,048 × 3
## # Groups:   dep [96]
##    dep    year avg_age_region
##    <chr> <dbl>          <dbl>
##  1 01     1960           35.8
##  2 01     1961           35.7
##  3 01     1962           35.6
##  4 01     1963           35.5
##  5 01     1964           35.3
##  6 01     1965           35.2
##  7 01     1966           35.0
##  8 01     1967           34.8
##  9 01     1968           34.6
## 10 01     1969           34.7
## # ℹ 6,038 more rows
```



5.1.1. Interprétation du graphique de l'évolution de l'âge moyen par région

Le graphique montre l'évolution de l'âge moyen des populations de différentes régions en France de 1960 à 2022.

1. **Tendance générale à la hausse** :

- Le graphique montre une tendance générale à l'augmentation de l'âge moyen dans toutes les régions au fil des années.
- Cela peut être attribué à plusieurs facteurs tels que l'allongement de l'espérance de vie et le vieillissement de la population.

2. **Différences régionales** :
   - Certaines régions ont un âge moyen plus élevé tout au long de la période étudiée, indiquant peut-être une population plus vieillissante ou des taux de natalité plus faibles.
   - D'autres régions ont des âges moyens plus bas, ce qui pourrait indiquer une population plus jeune, peut-être en raison de taux de natalité plus élevés ou de l'immigration de populations plus jeunes.
3. **Groupements de couleurs** :
   - Les couleurs représentent différentes régions. Les régions avec des âges moyens similaires apparaissent groupées ensemble sur le graphique.
   - Les légendes à droite permettent de voir quelles couleurs correspondent à quelles régions.
4. **Variabilité au fil du temps** :
   - La pente des courbes pour chaque région montre à quelle vitesse l'âge moyen augmente dans cette région.
   - Des pentes plus raides indiquent une augmentation rapide de l'âge moyen, tandis que des pentes plus douces indiquent une augmentation plus lente.
5. **Effets des événements historiques** :
   - Des périodes spécifiques peuvent montrer des changements plus drastiques dans certaines régions, peut-être en raison d'événements historiques, économiques ou sociaux spécifiques qui ont affecté la démographie de ces régions.

5.1.2. Conclusion :

En résumé, ce graphique est un outil puissant pour visualiser les tendances démographiques au fil du temps et peut aider à identifier les régions avec des populations vieillissantes plus rapides ou plus lentes, ce qui peut être crucial pour la planification des politiques publiques et des services sociaux.

5.2. Analyse des facteurs de corrélation

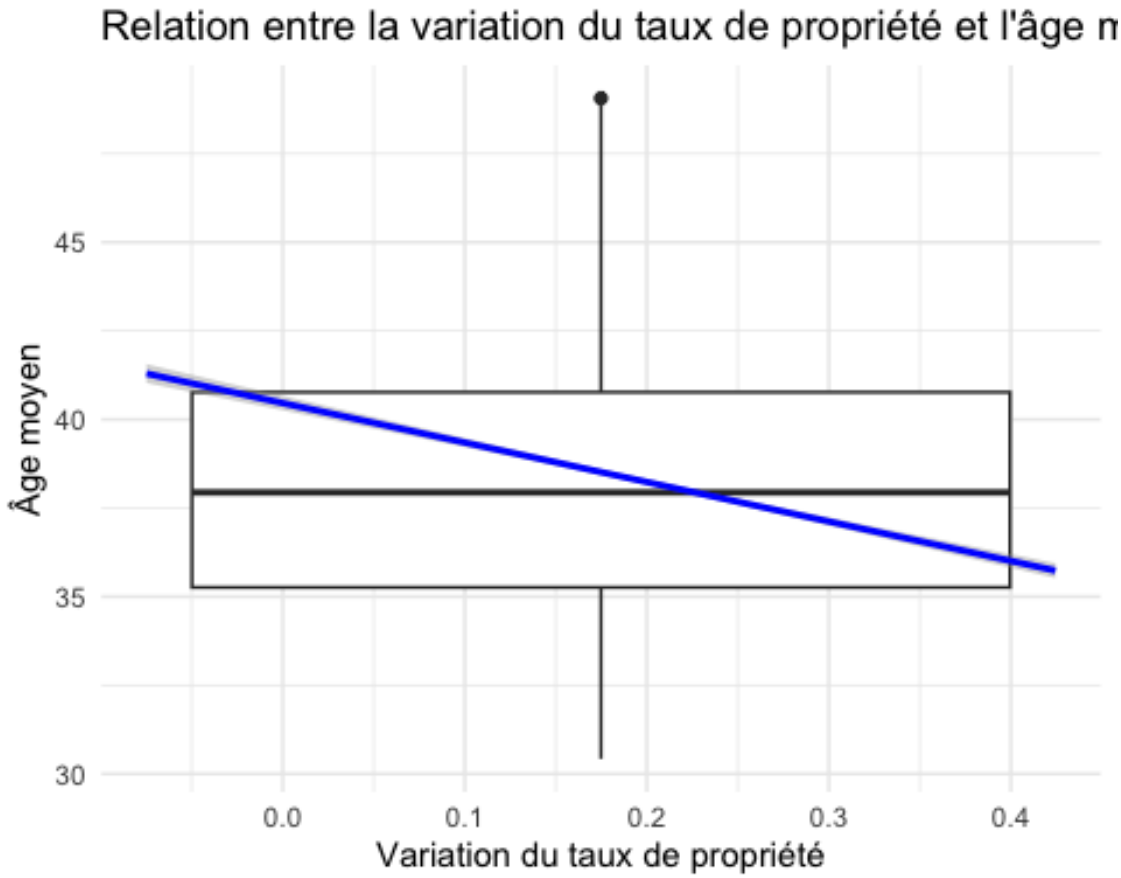
```
##   dep nomdep year property_rate
## 1  01    AIN 1960     0.3576642
## 2  01    AIN 1960     0.7560976
## 3  01    AIN 1960     0.3856955
## 4  01    AIN 1960     0.3000000
## 5  01    AIN 1960     0.7941176
## 6  01    AIN 1960     0.6375000

##                      variation_1960_2022 avg_age_region
## variation_1960_2022             1.00000       -0.29922
## avg_age_region                 -0.29922        1.00000
```

5.2.1. Interpretation de la correlation :

La corrélation négative faible à modérée entre la variation du taux de propriété entre 1960 et 2022 et l'âge moyen de la population. Ce qui signifie que lorsque l'âge moyen de la population augmente, la variation du taux de propriété tend à diminuer, et vice versa.

5.2.2. Conclusion :

L'analyse de corrélation montre une légère tendance inverse entre l'âge moyen de la population et la variation du taux de propriété entre 1960 et 2022. Cependant, la force de cette corrélation n'est pas très élevée, indiquant que l'âge moyen n'est qu'un des nombreux facteurs influençant la variation du taux de propriété. Pour une analyse plus approfondie, il serait utile d'examiner d'autres variables et de considérer des méthodes d'analyse multivariée pour mieux comprendre les facteurs influençant le taux de propriété.



5.2.3. Interpretation :

- Axe X (horizontal) : Représente la variation du taux de propriété. Les valeurs sur cet axe indiquent l'augmentation ou la diminution du taux de propriété sur la période étudiée.
- Axe Y (vertical) : Représente l'âge moyen des habitants des régions.

- Ce graphique suggère une relation inverse entre la variation du taux de propriété et l'âge moyen des habitants. Cela peut indiquer que les régions où la propriété augmente attirent une population plus jeune, peut-être en raison de meilleures opportunités économiques ou de politiques de logement favorables aux jeunes ménages.
- À l'inverse, les régions où le taux de propriété stagne ou diminue peuvent voir une population plus âgée, potentiellement due à moins d'opportunités ou à un exode des jeunes générations.

6. Intégration des Données sociaux politiques

6.1. Extractions des données Crimes_communes :

Pour explorez d'autres aspects pouvant expliquer la variation du taux de propriété de nos régions française, nous rajoutons les données "crimes_communes" refletant les conditions sociaux politiques.

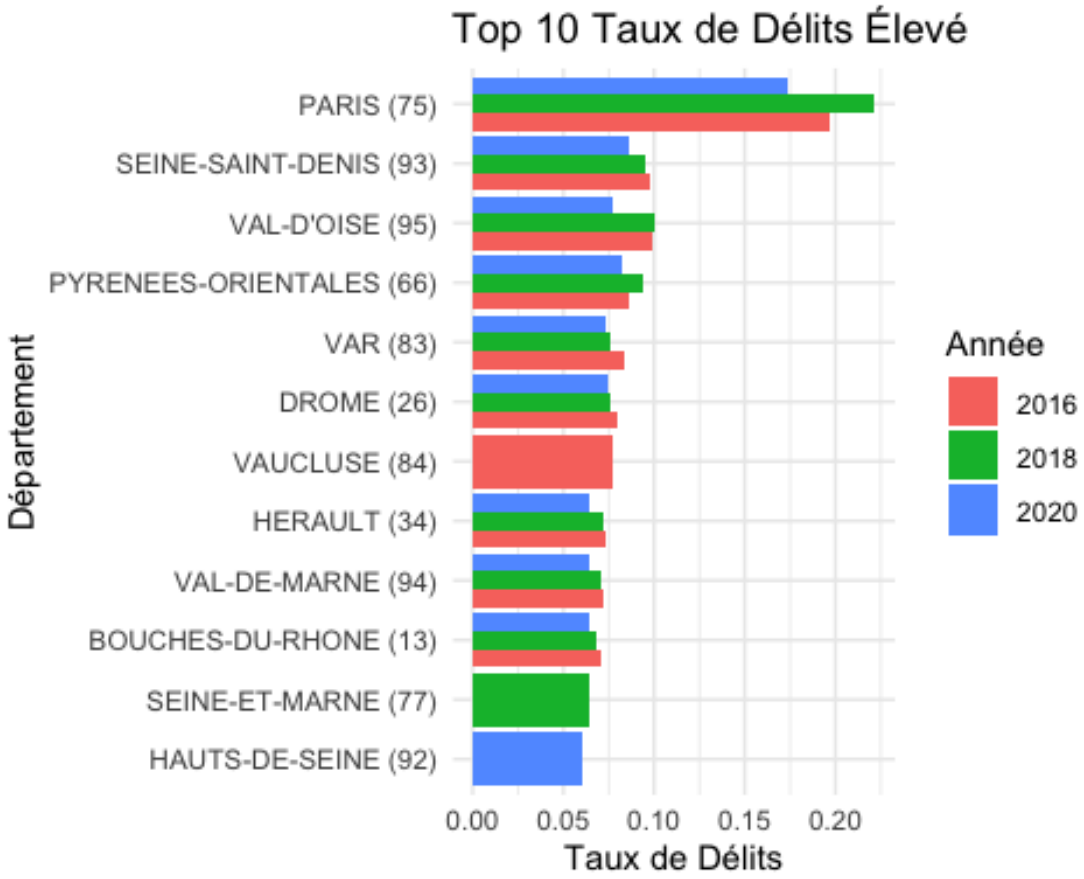
Au sein de nos données, nous retrouvons les nombres (pourcentage) de délits, de vols de voitures, de cambriolages, de violenes etc…

L'ajout de ces informations nous permettra d'identifier l'impacts du niveau de sécurité des régions et la variation des taux de propriété associés.

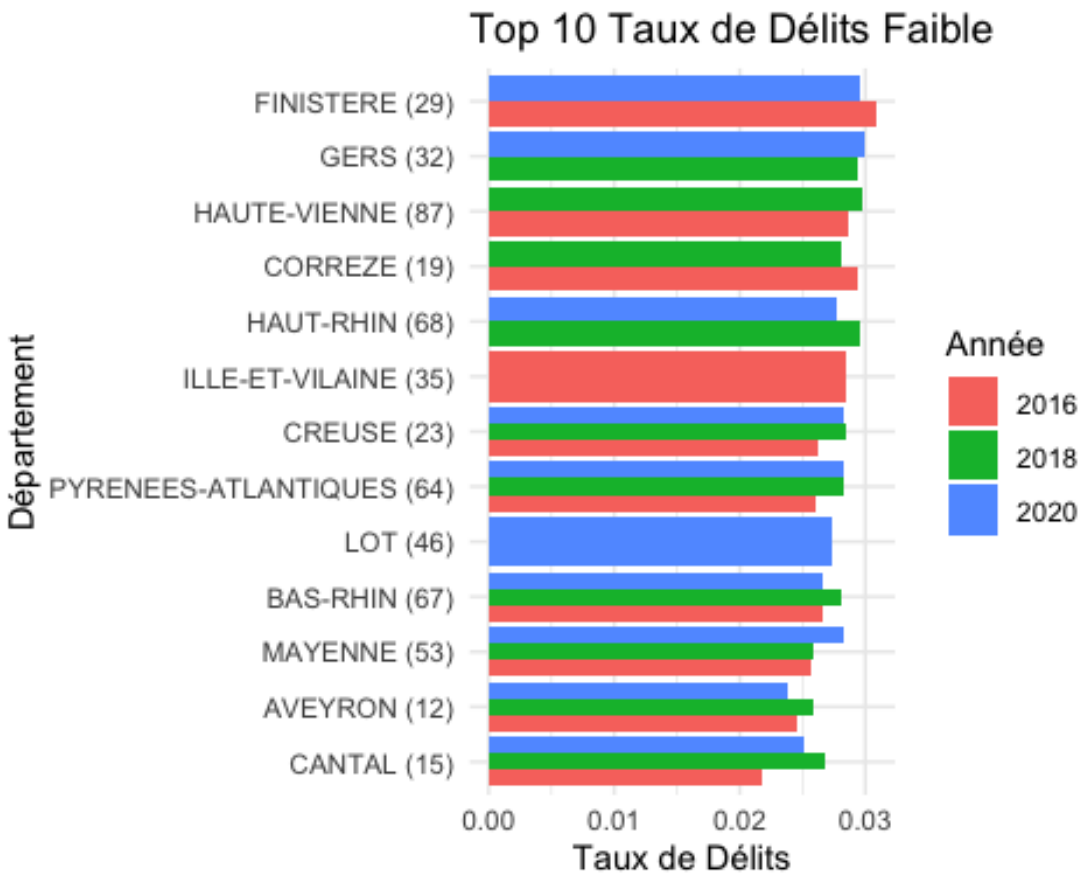
```
## # A tibble: 6 × 54
## # Groups:   dep [6]
##   dep   nomdep       paris ncrimesdelits2016 ncrimesdelits2018
ncrimesdelits2020
##   <chr> <chr>        <dbl>             <dbl>             <dbl>
<dbl>
## 1 01    AIN              0              31.9              33.9
30.1
## 2 02    AISNE            0              15.0              17.3
14.7
## 3 03    ALLIER           0              20.4              21.3
18.8
## 4 04    ALPES-DE-HA…     0              21.9              21.7
18.7
## 5 05    HAUTES-ALPES     0              21.0              19.2
16.3
## 6 06    ALPES-MARIT…     0             278.              268.
227.
## # ℹ 48 more variables: nviolences2016 <dbl>, nviolences2018 <dbl>,
## #   nviolences2020 <dbl>, ncambriolages2016 <dbl>, ncambriolages2018
<dbl>,
## #   ncambriolages2020 <dbl>, nvolsvoitures2016 <dbl>, nvolsvoitures2018
<dbl>,
## #   nvolsvoitures2020 <dbl>, nautresvols2016 <dbl>, nautresvols2018 <dbl>,
## #   nautresvols2020 <dbl>, pop2016 <dbl>, pop2018 <dbl>, pop2020 <dbl>,
## #   pcrimesdelits2016 <dbl>, pcrimesdelits2018 <dbl>, pcrimesdelits2020
```

```
<dbl>,
## #   pviolences2016 <dbl>, pviolences2018 <dbl>, pviolences2020 <dbl>, …

# bibliothèques nécessaires
library(dplyr)

#  premières lignes des données avant suppression des colonnes
#head(crimes_delits)

# Supprimer les colonnes spécifiques
data_crimes_delits <- data_crimes_delits %>%
  select(dep, nomdep, pdelits_total_2016, pdelits_total_2018,
pdelits_total_2020)

# Vérifier les premières lignes des données après suppression des colonnes
head(data_crimes_delits)

## # A tibble: 6 × 5
## # Groups:   dep [6]
##   dep   nomdep          pdelits_total_2016 pdelits_total_2018
pdelits_total_2020
##   <chr> <chr>                        <dbl>              <dbl>
<dbl>
## 1 01    AIN                         0.0457             0.0492
0.0421
## 2 02    AISNE                       0.0465             0.0524
0.0455
## 3 03    ALLIER                      0.0335             0.0352
0.0348
## 4 04    ALPES-DE-HAUTE…             0.0626             0.0628
0.0518
## 5 05    HAUTES-ALPES                0.0619             0.0537
0.0413
## 6 06    ALPES-MARITIMES             0.0608             0.0570
0.0544
```

6.2. Top 10 Départements avec le Taux de Délits le Plus Élevé sur nos 3 années

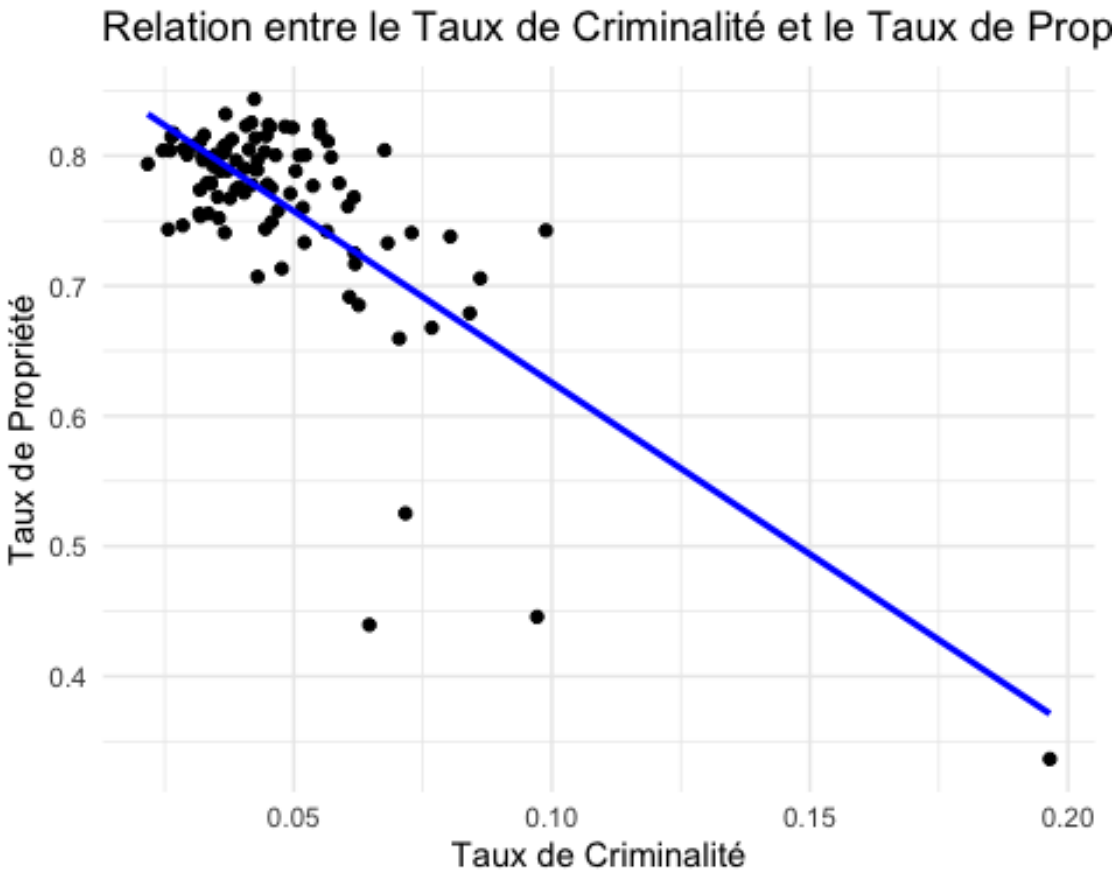


6.3. Top 10 Départements avec le Taux de Délits le Plus Faible sur nos 3 années



6.4. Relation entre le Taux de Criminalité et le Taux de Propriété

```
## [1] "Le coefficient de corrélation entre le taux de criminalité et le taux
de propriété est : -0.729377013748799"
```



####

6.4.1. Interprétation du graphique : Relation entre le Taux de Criminalité et le Taux de Propriété

1. **Axe des abscisses (x)** :
   - Représente le taux de criminalité. Les valeurs varient de 0 à 0.20, indiquant une échelle proportionnelle.
2. **Axe des ordonnées (y)** :
   - Représente le taux de propriété. Les valeurs varient de 0.4 à 0.9, indiquant également une échelle proportionnelle.
3. **Points noirs** :
   - Chaque point représente une observation (probablement une région ou un département) avec son taux de criminalité (x) et son taux de propriété (y) correspondant.
4. **Ligne de tendance (bleue)** :
   - La ligne de tendance montre la relation générale entre le taux de criminalité et le taux de propriété. La pente de cette ligne est négative.

6.4.2. Analyse de la relation :

1. **Corrélation négative** :

- Le graphique montre une corrélation négative entre le taux de criminalité et le taux de propriété. Cela signifie qu'en général, à mesure que le taux de criminalité augmente, le taux de propriété diminue.

2. **Force de la corrélation** :
   - La ligne de tendance descendante suggère que cette relation est assez prononcée, bien que le degré de dispersion des points autour de cette ligne indique qu'il existe une variation considérable qui n'est pas expliquée uniquement par cette relation linéaire.
3. **Variabilité des données** :
   - La dispersion des points autour de la ligne de tendance montre qu'il y a une variabilité significative dans les données. Certains points sont assez éloignés de la ligne de tendance, indiquant que d'autres facteurs pourraient influencer le taux de propriété en plus du taux de criminalité.

6.4.3. Conclusion :

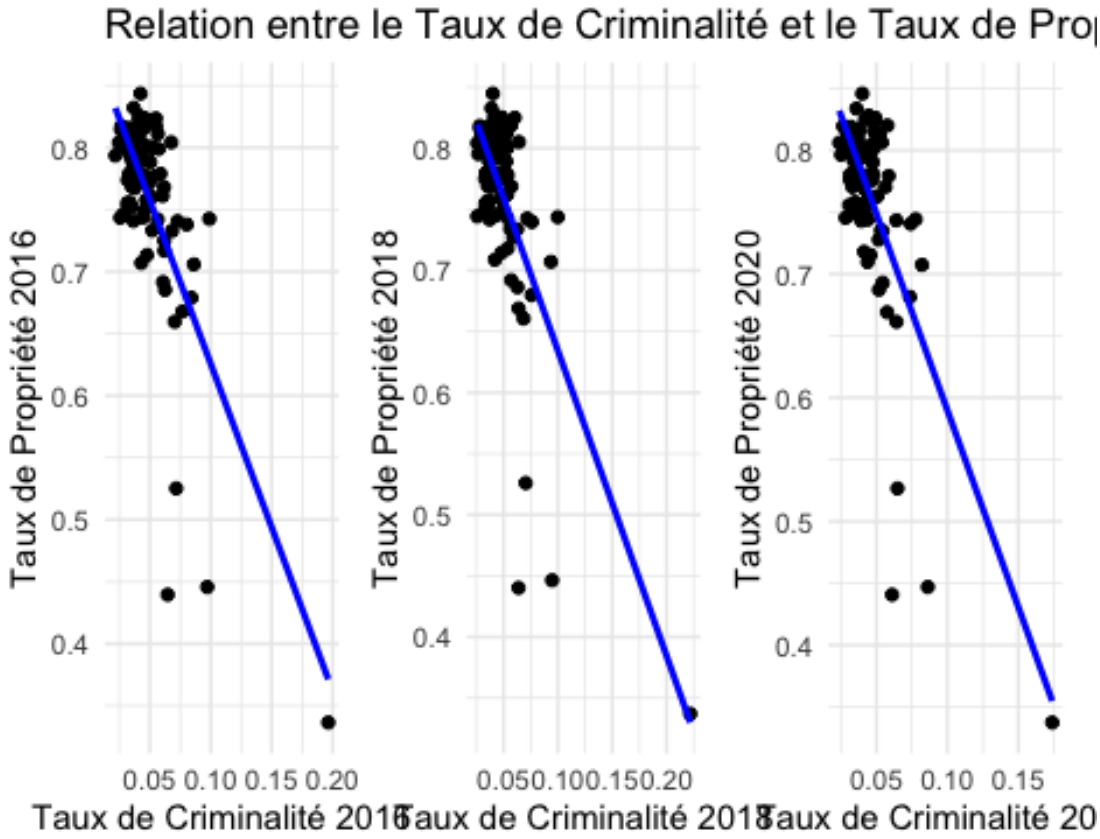
- **Impact de la criminalité sur la propriété** :
  - En général, dans les régions où le taux de criminalité est plus élevé, le taux de propriété tend à être plus bas. Cela pourrait être dû à divers facteurs, tels que la perception de sécurité, la qualité de vie, ou les décisions d'investissement immobilier.
- **Autres facteurs à considérer** :
  - Bien que la corrélation soit négative, la variabilité suggère que d'autres facteurs (tels que les politiques locales, l'économie régionale, ou les infrastructures) pourraient également jouer un rôle significatif.
- **Utilité pour la politique publique** :
  - Ces résultats peuvent aider à orienter les politiques publiques. Par exemple, les autorités locales pourraient se concentrer sur la réduction de la criminalité pour améliorer l'attractivité des régions en termes de propriété.

Ce graphique offre une vue d'ensemble utile mais nécessite une analyse plus approfondie pour comprendre pleinement les facteurs sous-jacents.

```
## [1] "Le coefficient de corrélation entre le taux de propriété et le taux
de criminalité en 2016 est : -0.729820061104887"

## [1] "Le coefficient de corrélation entre le taux de propriété et le taux
de criminalité en 2018 est : -0.726336200764371"

## [1] "Le coefficient de corrélation entre le taux de propriété et le taux
de criminalité en 2020 est : -0.728806968203325"
```

6.5. Relation entre le Taux de Criminalité et le Taux de Propriété en 2016, 2018 et 2020

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



6.5.1. Interprétation des coefficients de corrélation

- 1. Coefficient de corrélation en 2016 : -0.729820061104887
- 2. Coefficient de corrélation en 2018 : -0.726336200764371
- 3. Coefficient de corrélation en 2020 : -0.728806968203325

6.5.2. Analyse des résultats :

1. **Valeurs des coefficients** :
   - Les trois coefficients de corrélation sont négatifs et très proches les uns des autres, indiquant une relation relativement stable au fil des années.
   - Les valeurs sont toutes proches de -0.73, ce qui indique une forte corrélation négative entre le taux de propriété et le taux de criminalité.
2. **Signification de la corrélation négative** :
   - Un coefficient de corrélation négatif signifie qu'il y a une relation inverse entre les deux variables : à mesure que le taux de criminalité augmente, le taux de propriété diminue, et vice versa.
   - La valeur absolue proche de 0.73 indique que cette relation inverse est assez forte.

3. **Stabilité au fil des ans** :
   - Les coefficients de corrélation en 2016, 2018 et 2020 sont très similaires, ce qui suggère que la relation entre le taux de criminalité et le taux de propriété est stable sur cette période.
   - Cette stabilité peut indiquer que les facteurs influençant ces taux n'ont pas beaucoup changé au cours de ces années ou que les changements ont été proportionnels.

6.5.3. Implications et conclusions :

1. **Impact sur les décisions d'investissement** :
   - La forte corrélation négative suggère que les investisseurs immobiliers et les acheteurs de maisons pourraient être dissuadés d'acheter dans des régions à taux de criminalité élevé.
   - Les régions avec des taux de criminalité élevés peuvent voir une baisse de la demande de propriété, ce qui peut affecter les prix de l'immobilier et la prospérité économique locale.
2. **Politiques publiques** :
   - Les décideurs politiques peuvent utiliser ces informations pour cibler les efforts de réduction de la criminalité comme moyen d'augmenter les taux de propriété et de revitaliser les communautés.
   - Des programmes de sécurité publique renforcée pourraient être bénéfiques pour améliorer la perception de sécurité et encourager l'achat de propriétés.
3. **Considérations supplémentaires** :
   - Bien que la corrélation soit forte, elle ne démontre pas nécessairement une causalité directe. D'autres facteurs pourraient également influencer cette relation et doivent être pris en compte.
   - Des analyses supplémentaires pourraient être nécessaires pour comprendre pleinement les dynamiques entre la criminalité et la propriété, y compris des études sur d'autres variables socio-économiques.

6.5.4. Résumé :

Les coefficients de corrélation montrent une forte relation inverse stable entre le taux de propriété et le taux de criminalité sur les années étudiées. Cela suggère que les régions avec des taux de criminalité plus élevés tendent à avoir des taux de propriété plus faibles, une information cruciale pour les investisseurs et les décideurs politiques.

7. Conclusion Générale :

Notre étude a permis d'analyser le taux de propriété des logements par communes et régions sur une période de plus de 60 ans, allant de 1960 à 2022. Nous avons intégré des données sur l'âge de la population afin d'examiner un potentiel lien entre l'âge moyen des habitants et le taux de propriété. Les résultats indiquent que, pour notre échantillon, les régions avec un âge moyen plus bas présentent un taux de propriété plus élevé. Cela

pourrait s'expliquer par les commodités régionales mises à disposition des familles et des enfants, incitant ces derniers à acquérir des logements adaptés à leurs besoins.

Par la suite, nous avons incorporé des données socio-politiques, incluant les taux d'agression, de vols de voitures, de crimes, d'autres vols, de violences, etc., pour chaque département. L'objectif de l'intégration de ces nouvelles données était de visualiser l'impact du niveau de sécurité des départements sur l'acquisition de logements dans les régions. Nos analyses ont révélé que la variation du taux de propriété par région est fortement corrélée au niveau de criminalité des régions. Bien que cette corrélation négative soit assez évidente, nous reconnaissons que d'autres facteurs pourraient également expliquer les variations du niveau de propriété des logements.

Nous espérons que cette analyse éclairera nos décideurs politiques sur les actions à entreprendre pour améliorer l'attractivité de certaines régions. Cela inclut l'implantation d'infrastructures adaptées aux différentes catégories de populations, la création d'entreprises pour attirer les jeunes couples et les familles, et la garantie d'un niveau de sécurité permettant de maintenir sur le long terme les populations séduites et d'éviter les variations négatives observées dans certaines régions.