Analyse de données Bibliotheques

Martin Coulon et Laura Spatzierer

2023-11-10

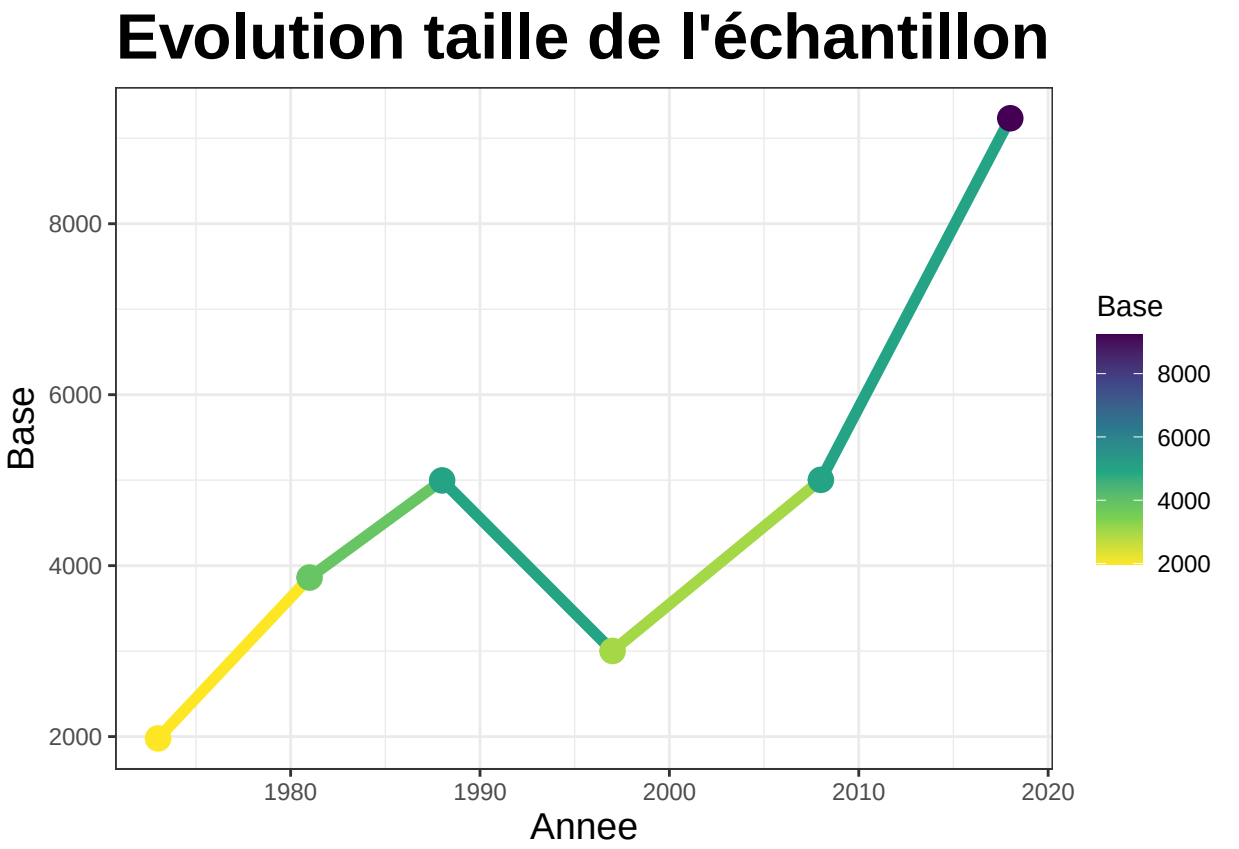
Jeux de données bibliothèque :

Introduction :

Nous avons reçu pour mission de fournir une analyse statistique des inscriptions en bibliothèques à partir des données de l'enquête sur les pratiques culturelles des Français de 1973 à 2018. Nos données ont été recueillies par questionnaire avec une récolte par interrogation au domicile. L'avantage de cette méthode est qu'elle permet de remplir le questionnaire dans son intégralité. L'échantillonnage de l'enquête se revendique comme étant représentatif de la population en France métropolitaine ayant plus de 15 ans. L'échantillon est stratifié d'après la taille de la population des régions et des catégories d'agglomération.

Les questions ont peu été modifiées au fil du temps, seule la taille de l'échantillon a varié dans le temps. Comme vous pouvez le remarquer 2018 est une année record puisque 9234 réponses ont été collectées.

En qualité d'introduction nous avons réalisé un graphique qui permet de visualiser l'évolution de la taille de l'échantillon au fil des années. Voici un aperçu :



Une différence est a notée dans l'échantillonnage.L'enquête 2018 a été tirée aléatoirement dans une base de sondage de logements alors que la méthode des quotas avait été utilisée pour les 5 premières éditions. Pour plus d'informations sur l'échantillonnage vous pouvez retrouver des informations sur le site du ministère en cliquant : ici.

Notre analyse cherche à montrer si les bibliothèques sont des lieux dans lesquels tous les individus s'inscrivent où s'il existe des inégalités de répartition au sein des populations dans la tendance à s'inscrire dans une bibliothèque. S'inscrire dans une bibliothèque est une démarche personnelle ou automatique si vous êtes étudiant dans une université, vous êtes par exemple automatiquement inscrit. Cette démarche permet de pouvoir emprunter des livres et d'accéder aux espaces dans certains cas particulier (comme la BU de Sciences Po Lille).

Ainsi, connaitre le profil des inscrits permet de connaitre le profil des emprunteurs potentiels et de savoir si toute la population a autant tendance à vouloir bénéficier de ce service.

Nous allons dans un premier temps procéder à l'importation et au nettoyage de nos données. Ensuite nous analyserons les données et effectuerons une visualisation des éléments les plus explicatifs. L'avantage de notre travail est qu'il permet par comparaison temporelle de montrer si des tendances sont en place depuis un certain temps ou émergent.

Importation des données et nettoyage par soucis de meilleure lisibilité :

Nous utlisons les outils du tidyverse pour réaliser notre nettoyage. Il est nécessaire d'effectuer un nettoyage afin d'améliorer la lisibilité et la structure de nos données. En effet, il convient de signaler que le fichier importé était extrêmement mal fait, que cela soit au niveau de sa mise en forme, avec des colonnes et lignes vides inutiles, ou des termes employés inappropriés. Ainsi, nous avons dû effectuer un long travail de nettoyage pour rendre le fichier lisible et cohérent.

Ainsi, dans les lignes suivantes, nous avons décidé d'utiliser la fonction "mutate" afin de remplacer le terme "ensemble" par "moyenne". Nous créons des sous-ensembles pour isoler les différentes variables contenu dans notre jeu de donnée. Nous précisons les lignes à sélectionner et définissons les noms des colonnes.

```
library(readxl)
library(tidyverse)
noms_colonnes <- c("variables", "1973", "1981", "1988", "1997", "2008", "2018") #Permet de corriger le

Bib_Ins <- read_excel("PC_1973-2018_Bibliotheques.xlsx", sheet= "Biblio inscrits",
                      col_names = noms_colonnes,
                      col_types = c("text", rep("numeric", 6)))
Bib_Ins <- as_tibble(Bib_Ins)
Bib_Ins <- Bib_Ins %>%
  mutate(`variables` = replace(`variables`, `variables` == "Ensemble", "Moyenne"))

#Isolation des variables dans des sous-ensembles
var_age1 <- Bib_Ins %>%
  slice(6:11) %>%
  setNames(c("age", "1973", "1981", "1988", "1997", "2008", "2018"))

var_age2 <- Bib_Ins %>%
  slice(14:18) %>%
  setNames(c("age2", "1973", "1981", "1988", "1997", "2008", "2018"))

var_sexe <- Bib_Ins %>%
  slice(22:24) %>%
  setNames(c("sexe", "1973", "1981", "1988", "1997", "2008", "2018"))

var_dipl <- Bib_Ins %>%
  slice(28:32) %>%
  setNames(c("diplome", "1973", "1981", "1988", "1997", "2008", "2018"))

var_CS_regroup <- Bib_Ins %>%
  slice(36:40) %>%
  setNames(c("pcs_regroup", "1973", "1981", "1988", "1997", "2008", "2018"))

var_ville <- Bib_Ins %>%
  slice(44:50) %>%
  setNames(c("taille_ville", "1973", "1981", "1988", "1997", "2008", "2018"))

var_age3 <- Bib_Ins %>%
  slice(54:62) %>%
  setNames(c("age_detail", "1973", "1981", "1988", "1997", "2008", "2018"))
```

Transformation des données en format long :

Par souci de cohérence avec le package ggplot nous transformons les données en format long. Dans le cadre de nos données, leur organisation en tableau croisée simplifie néanmoins la lisibilité pour un humain.

Nous utilisons une fonction pour appliquer la transformation à chaque sous-ensemble. Notre fonction s'appelle format_to_long, elle prend en premier argument le nom du sous-ensemble à modifier et en second argument le nom à donner à la première colonne du sous ensemble. La fonction va transformer le sous-ensemble en un format long et renomme le nom de la première colonne.

On utilise ensuite notre fonction pour transformer chaque sous-ensemble. Nous créons de nouvelles variables afin de pouvoir accéder facilement à nos sous-ensemble lisible.

```r
format_to_long <- function(df, name_column) {
  df %>%
    pivot_longer(
      cols = -all_of(name_column),
      names_to = "Annee",
      values_to = "Pourcentage") }

# Application de la fonction de transformation à chaque sous-ensemble

var_age1_long <- format_to_long(var_age1, "age")

var_age2_long <- format_to_long(var_age2, "age2")

var_sexe_long <- format_to_long(var_sexe, "sexe")

var_dipl_long <- format_to_long(var_dipl, "diplome")

var_CS_regroup_long <- format_to_long(var_CS_regroup, "pcs_regroup")

var_ville_long <- format_to_long(var_ville, "taille_ville")

var_age3_long <- format_to_long(var_age3, "age_detail")
```

Analyse des données

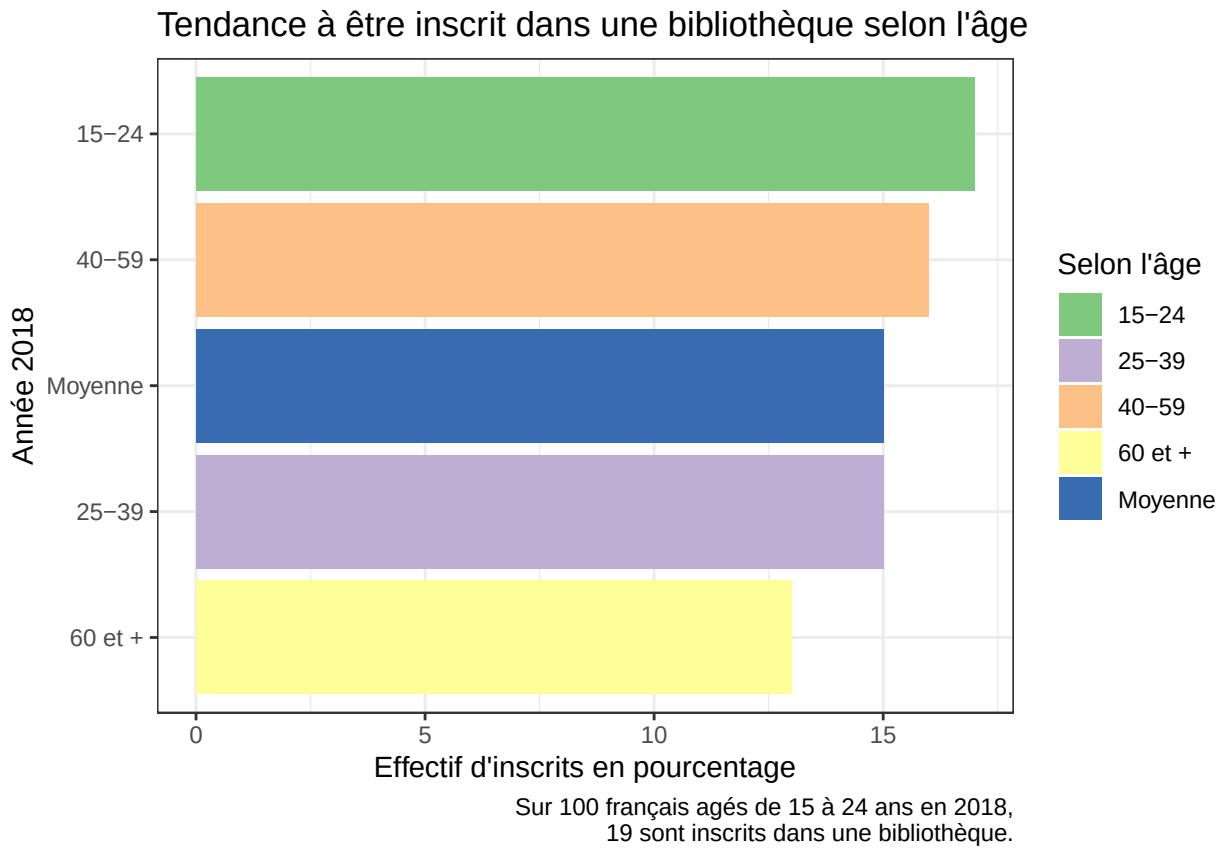
Analyse de l'âge des inscrits en bibliothèques.

Notre tableau proposait 3 sous-ensembles pour analyser la tendance selon l'âge à être inscrits en bibliothèque. La différence entre le premier et le second sous-ensemble est faible. Le premier sépare les 15-19 ans et les 20-24 ans dans deux catégories différentes. Le troisième sous-ensemble lui est plus complet. Il effectue une analyse plus fine des âges des jeunes et des personnes âgées.

Dans un premier temps, nous allons nous intéresser à l'année 2018 en utilisant le second sous-ensemble âge de notre tableau. Celui-ci à l'avantage de contenir moins de variables en regroupant l'âge des jeunes ayant entre 15 et 24 et en regroupant également les individus âgé de plus de 60 ans dans une seule catégorie. En regroupant les variables, on peut plus simplement visualiser une tendance et affiner si c'est nécessaire notre analyse en prenant nos autres sous-ensembles.

```r
var_age2_long %>%
 filter(Annee %in% "2018") %>%
 ggplot() +
  aes(
    x = reorder(age2, Pourcentage),
    fill = age2,
    group = age2,
    weight = Pourcentage
  ) +
  geom_bar(position = "dodge") +
  scale_fill_brewer(palette = "Accent", direction = 1) +
  labs(
    x = "Année 2018",
    y = "Effectif d'inscrits en pourcentage",
```

```
    title = "Tendance à être inscrit dans une bibliothèque selon l'âge",
    caption = "Sur 100 français agés de 15 à 24 ans en 2018,
    19 sont inscrits dans une bibliothèque.",
    fill = "Selon l'âge"
  ) +
  coord_flip() +
  theme_bw()
```



Notre visualisation démontre plusieurs éléments. **Les 15-24 que nous appellerons les jeunes, semblent être les français ayant le plus tendance à être inscrit dans une bibliothèque en 2018. Les populations ayant 60 ans et plus, elles semblent être celles qui ont le moins tendance à être inscrit**. Une écart de 6 point séparant les deux ensembles. Néanmoins, nos données ne reflètent pas un écart très important à la moyenne. La différence d'âge est observable néanmoins, elle ne semble pas être un élément de différenciation extrêmement important en 2018.

Pour confirmer nos observations nous allons calculer l'écart-type, la moyenne et la proportion de l'écart-type par rapport à la moyenne. La proportion écart-type par rapport à la moyenne également appelé coefficient de variation est un indicateur permettant de savoir si nos données sont dispersés ou non obtient les résultats suivant :

```
pourcentages_2018 <- var_age2_long %>%
  filter(Annee == "2018") %>%
  .$Pourcentage

ecart_type_2018 <- sd(pourcentages_2018)
moyenne_2018 <- mean(pourcentages_2018)
```

```
Proportion_ecart_type_2018 <- ecart_type_2018 / moyenne_2018

ecart_type_2018
```

```
## [1] 1.48324
```

```
moyenne_2018
```

```
## [1] 15.2
```

```
Proportion_ecart_type_2018
```

```
## [1] 0.09758156
```

Le coefficient de variation est approximativement de 0.098. Ce résultat est suffisamment faible puisqu'il est inférieur à 10%, nous pouvons donc affirmer que les écarts d'âges sont un facteur peu significatif en 2018.

Mais cette explication était-elle la même auparavant ?

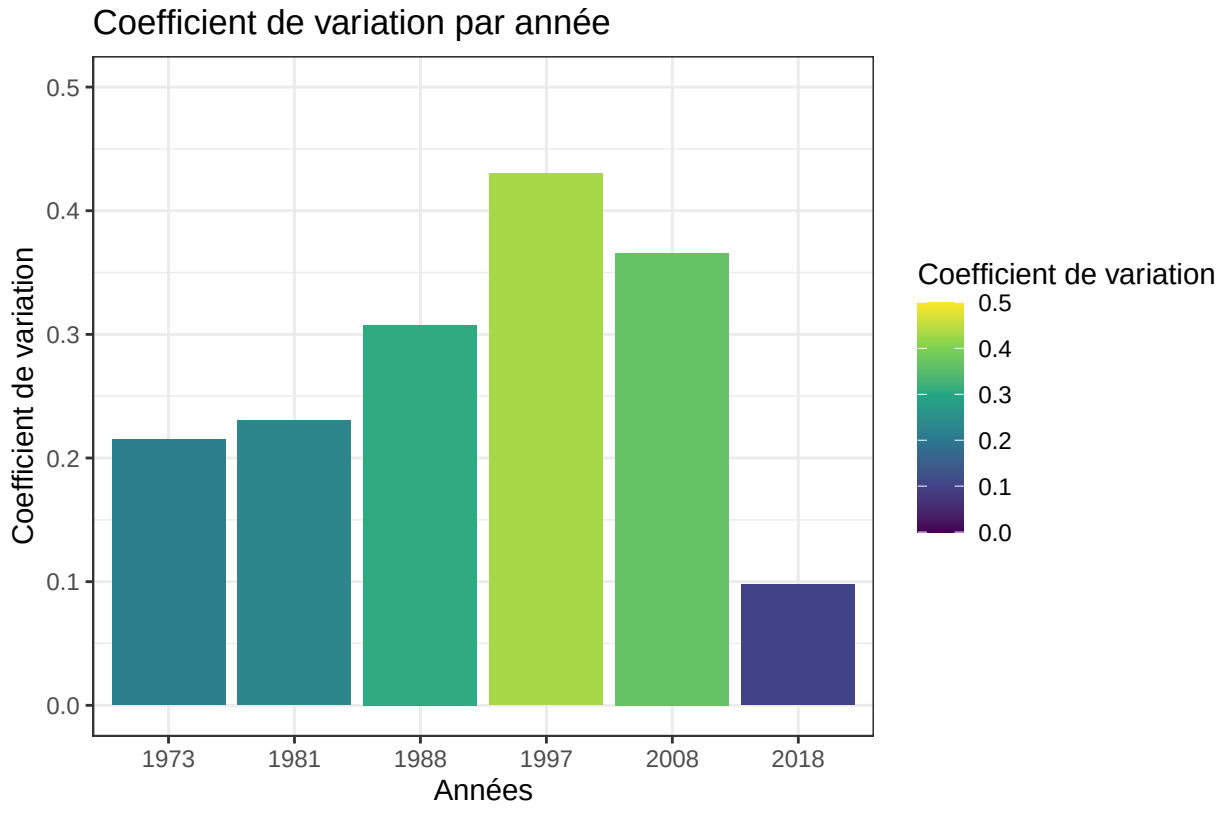
Pour répondre à cette question calculons le coefficient de variation pour chaque année et effectuons une représentation graphique.

```
# Calcul de l'écart-type et de la moyenne pour chaque année
stats_par_annee <- var_age2_long %>%
  group_by(Annee) %>%
  summarise(
    Ecart_Type = sd(Pourcentage),
    Moyenne = mean(Pourcentage)
  )

# Calcul de la proportion de l'écart-type par rapport à la moyenne
stats_par_annee <- stats_par_annee %>%
  mutate(Proportion_Ecart_Type = Ecart_Type / Moyenne)

# Afficher les statistiques
print(stats_par_annee)
```

```
## # A tibble: 6 x 4
##   Annee Ecart_Type Moyenne Proportion_Ecart_Type
##   <chr>      <dbl>   <dbl>                 <dbl>
## 1 1973        2.88    13.4                 0.215
## 2 1981        3.27    14.2                 0.230
## 3 1988        5.41    17.6                 0.308
## 4 1997        9.29    21.6                 0.430
## 5 2008        7.54    20.6                 0.366
## 6 2018        1.48    15.2                 0.0976
```

```
# Visualiser de la distribution des données
graph_cf_age2 <- ggplot(stats_par_annee) +
  aes(
```

```
    x = Annee,
    y = Proportion_Ecart_Type,
    fill = Proportion_Ecart_Type) +
  geom_col() +
  labs(
    x = "Années",
    y = "Coefficient de variation",
    title = "Coefficient de variation par année",
    caption = "En 2018, le coefficient de variation du sous-ensemble 2 est inférieur à 0,10",
    fill = "Coefficient de variation"
  ) +
  scale_y_continuous(limits = c(0, 0.5)) +
  scale_fill_viridis_c(option = "viridis", direction = 1,limits = c(0,0.5)) +
  theme_bw()

graph_cf_age2
```

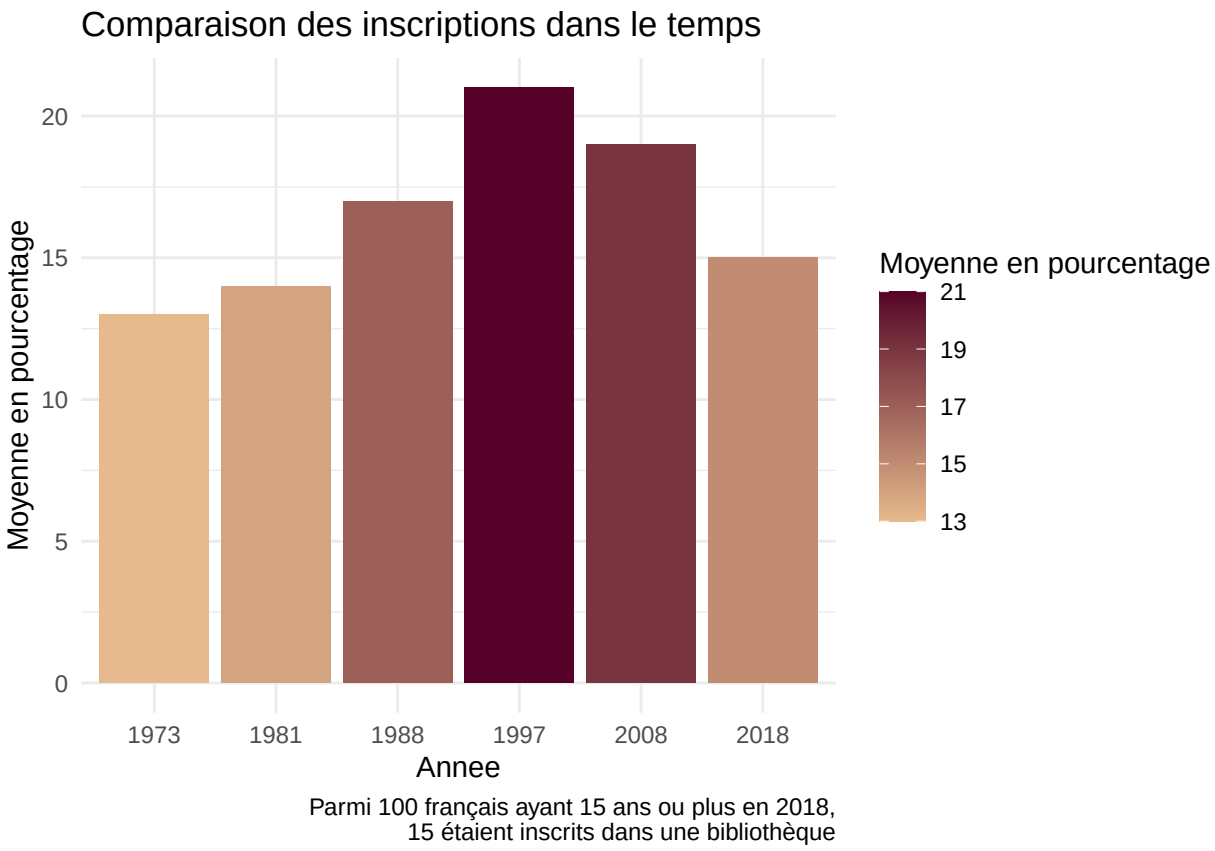


En 2018, le coefficient de variation du sous−ensemble 2 est inférieur à 0,10

Lorsque l'on s'intéresse à la distribution des enquêtes précédentes on réalise que la variable âge a pu jouer une importance particulièrement en 1997 et en 2008 où l'on observe des coefficients de variation plus importantes. Néanmoins, on observe que cette importance de la variable âge a fortement diminué en 2018 validant donc nos propos précédents.

Puisque nous commençons à nous intéresser aux variations dans le temps, nous allons maintenant observer si la tendance à être inscrit a évolué entre les années. Pour ce faire nous allons nous baser sur la visualisation de la moyenne de nos ensemble pour chaque année.

```r
var_age3_long %>%
 filter(age_detail %in% "Moyenne") %>%
 ggplot() +
  aes(x = Annee, y = Pourcentage, fill = Pourcentage) +
  geom_col() +
  scale_fill_gradient(low = "#E6BA8D", high = "#560228") +
  labs(
    y = "Moyenne en pourcentage",
    title = "Comparaison des inscriptions dans le temps",
    caption = "Parmi 100 français ayant 15 ans ou plus en 2018,
    15 étaient inscrits dans une bibliothèque",
    fill = "Moyenne en pourcentage"
  ) +
  theme_minimal()
```



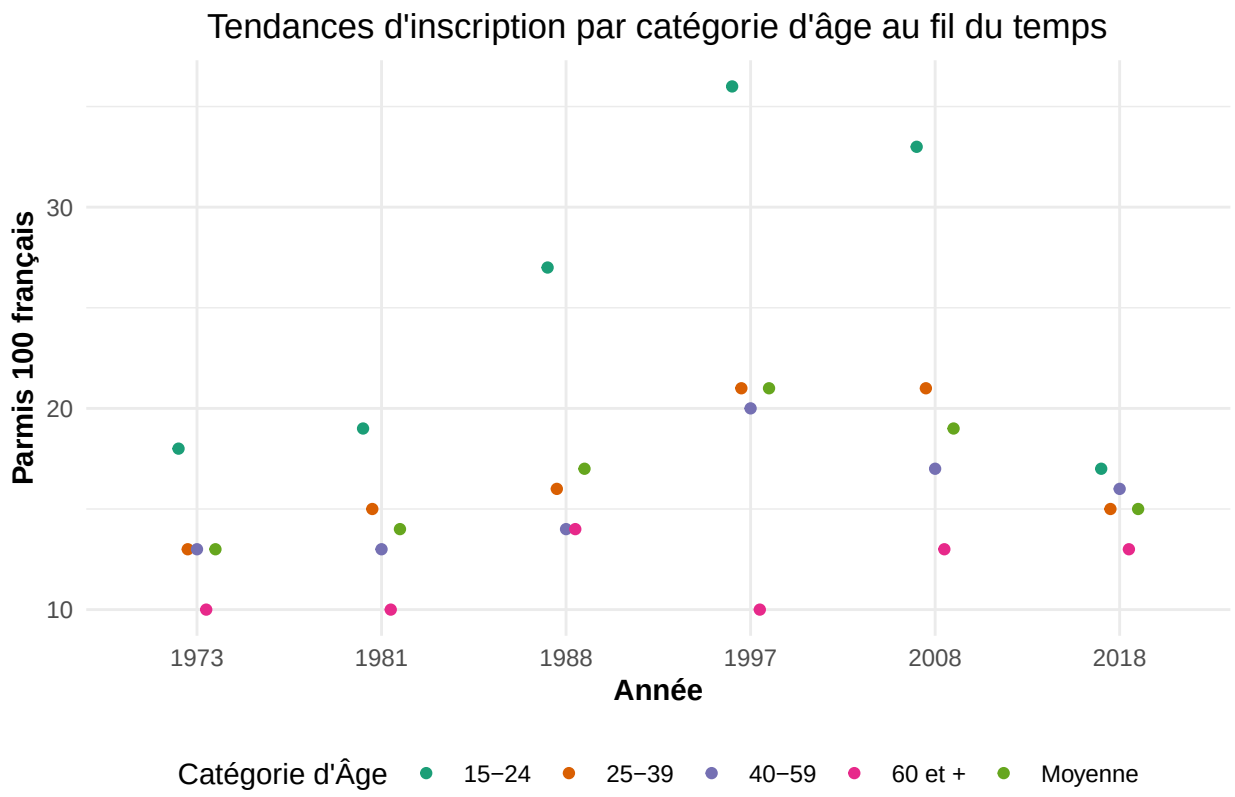
On observe des variations relativement importantes entre les années. La tendance à être inscrit dans une bibliothèque en moyenne était de 21% en 1997 quand elle est de 13% en 1973. On observe que de 1973 à 1997 la tendance générale à être inscrit augmentait. Alors qu'entre 1997 et 2018 on observe une diminution. Un ensemble de facteur explicatif très vaste peut-expliquer cette variation. Notre première hypothèse serait que le développement de l'informatique, la généralisation de l'accès à l'information par le biais d'internet et la numérisation des livres pourraient expliquer cette inversion de la tendance.

Les individus auraient pu être motivé à s'inscrire dans des bibliothèques pour accéder à des ordinateurs et parce qu'Internet n'existait pas, puis n'était pas généralisé entre 1973 et 1997. Avec la généralisation de ces outils dans les foyers, cette motivation n'existant plus les individus auraient moins tendance à s'inscrire dans une bibliothèque. Puisque l'accès à l'information peut désormais être trouvé ailleurs.

Néanmoins, pour démontrer cette hypothèse où trouver d'autres hypothèses explicatives il faudrait avoir collecter des données qualitatives expliquant les motivations des individus à se rendre dans les bibliothèques et à s'y inscrire. Notre jeu de donnée ne comportant pas ce type de donnée il nous est impossible d'affirmer ou non cette hypothèse.

Notre graphique étudiant les proportions écart-types moyenne démontrait que des variations importantes existait pour les années autres que 2018, notamment en 1997 où les valeurs sont dispersés. Il est donc pertinent d'effectuer une visualisation de chaque valeurs pour chaque année afin d'identifier si des valeurs extrêmes sont identifiables.

```
ggplot(var_age2_long, aes(x = Annee, y = Pourcentage, color = age2)) +
  geom_point(position = position_dodge(width = 0.25)) +
  scale_color_brewer(palette = "Dark2") +
  labs(
    title = "Tendances d'inscription par catégorie d'âge au fil du temps",
    x = "Année",
    y = "Parmis 100 français",
    color = "Catégorie d'Âge",
    caption = "Sur 100 français agés de 15 à 24 ans en 2018, 17 sont inscrits dans une bibliothèque. ")
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5),
    axis.title.x = element_text(face = "bold"),
    axis.title.y = element_text(face = "bold"),
    legend.position = "bottom"
  )
```



La visualisation par nuage de point nous permet de mieux comprendre les écarts de dispersions. Cette visualisation nous montre que le jeune l'âge a été une variable explicative très forte dans la tendance à être inscrit dans une bibliothèque en 1997 et en 2008. La variable avait également une importance les autres années mais moindre. On observe également que la tendance à être inscrit après 60 ans en 1997 était faible en comparaison aux autres âges.

Il convient donc maintenant d'effectuer une analyse s'intéressant au profil des jeunes et des plus de 60 ans pour comprendre si parmi ces deux ensembles des valeurs extrêmes expliquent ces différences.

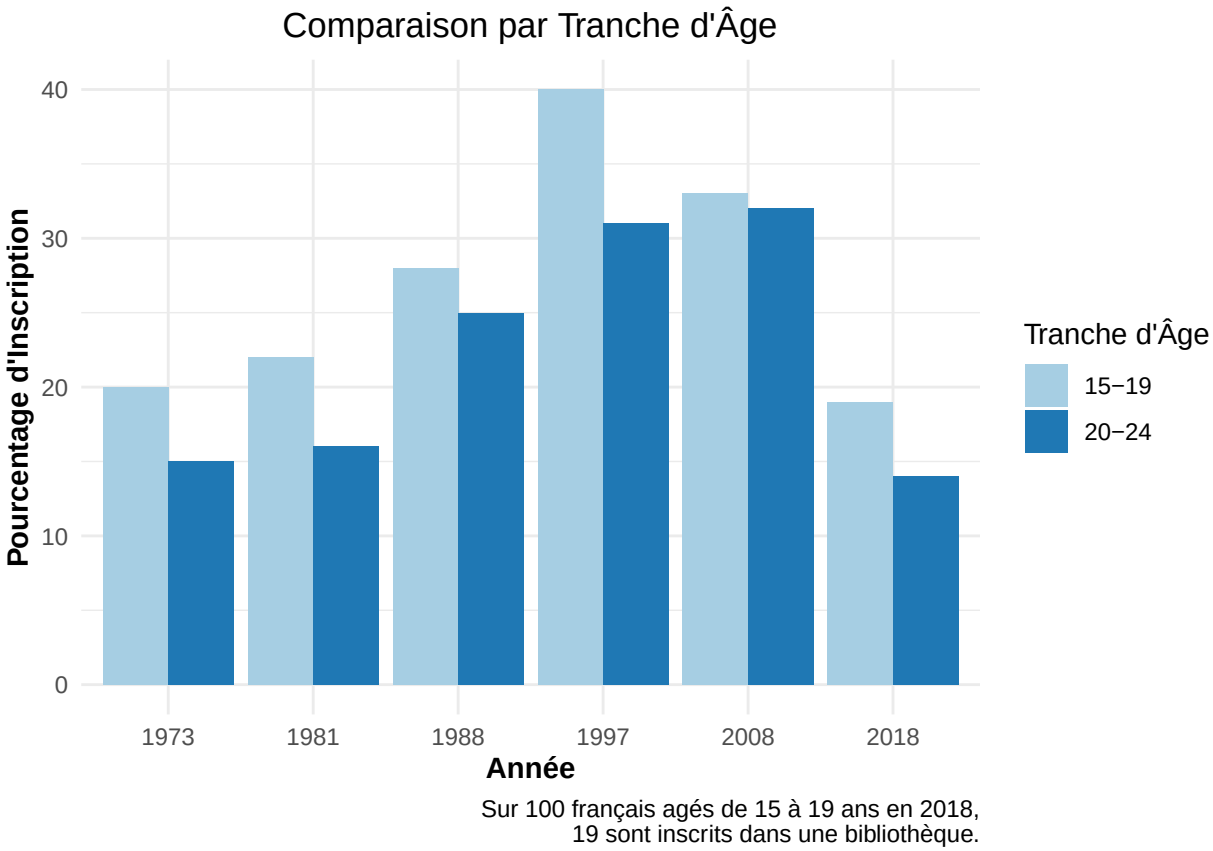
Dans le cadre des jeunes, l'ensemble des 15 à 24 est grand, nous n'avons pas les mêmes pratiques entre 15 et 19 ans et 20 à 24 ans. Les français sont lycéens entre 15 et 19 ans la plupart du temps puis étudiant, inactif ou actif après 20 ans. A première vue, il serait donc plausible que l'ensemble 15-24 soit tiré à la hausse par les étudiants ayant répondu à l'enquête, qui dans le cadre de leurs études sont souvent automatiquement inscrit dans des bibliothèques.

De même, l'ensemble des 60 et + contient un ensemble d'âge extrêmement important. Avec une durée de vie en moyenne supérieure à 80 ans en France d'après l'INSEE dans une étude sur la mortalité en 2019 disponible : ici . Cet ensemble contient 20 ans d'écarts à minima. Hors, du fait de leurs grand âge il est potentiellement plus complexe pour une personne de 70 ou 80 ans de se déplacer et donc de s'inscrire dans une bibliothèque que pour un individu de 60 ans. Ainsi, il nous faut observer plus finement cet ensemble pour vérifier si c'est vraiment le fait d'avoir plus de 60 ans qui réduit la tendance à être inscrit où si c'est à partir d'un âge plus important.

Nous allons donc désormais utiliser notre sous-ensemble 3 qui contient un plus grand nombre de sous-ensemble âge. Il va nous permettre d'observer d'éventuelles variations entre les individus de 15 à 19 et ceux de 20 à 24 ans. Mais également des variations entre les individus ayant entre 55 à 64 ans, ceux ayant entre 65 et 74 ans et enfin ceux ayant plus de 75 ans.

Intéressons nous dans un premier temps à notre ensemble des jeunes.

```
var_age3_filtered <- var_age3_long %>%
  filter(age_detail %in% c("15-19", "20-24"))

ggplot(var_age3_filtered, aes(x = Annee, y = Pourcentage, fill = age_detail)) +
  geom_bar(stat = "identity", position = position_dodge()) +
  scale_fill_brewer(palette = "Paired") +
  labs(
    title = "Comparaison par Tranche d'Âge",
    x = "Année",
    y = "Pourcentage d'Inscription",
    fill = "Tranche d'Âge",
    caption = "Sur 100 français agés de 15 à 19 ans en 2018,
    19 sont inscrits dans une bibliothèque."
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5),
    axis.title.x = element_text(face = "bold"),
    axis.title.y = element_text(face = "bold")
  )
```



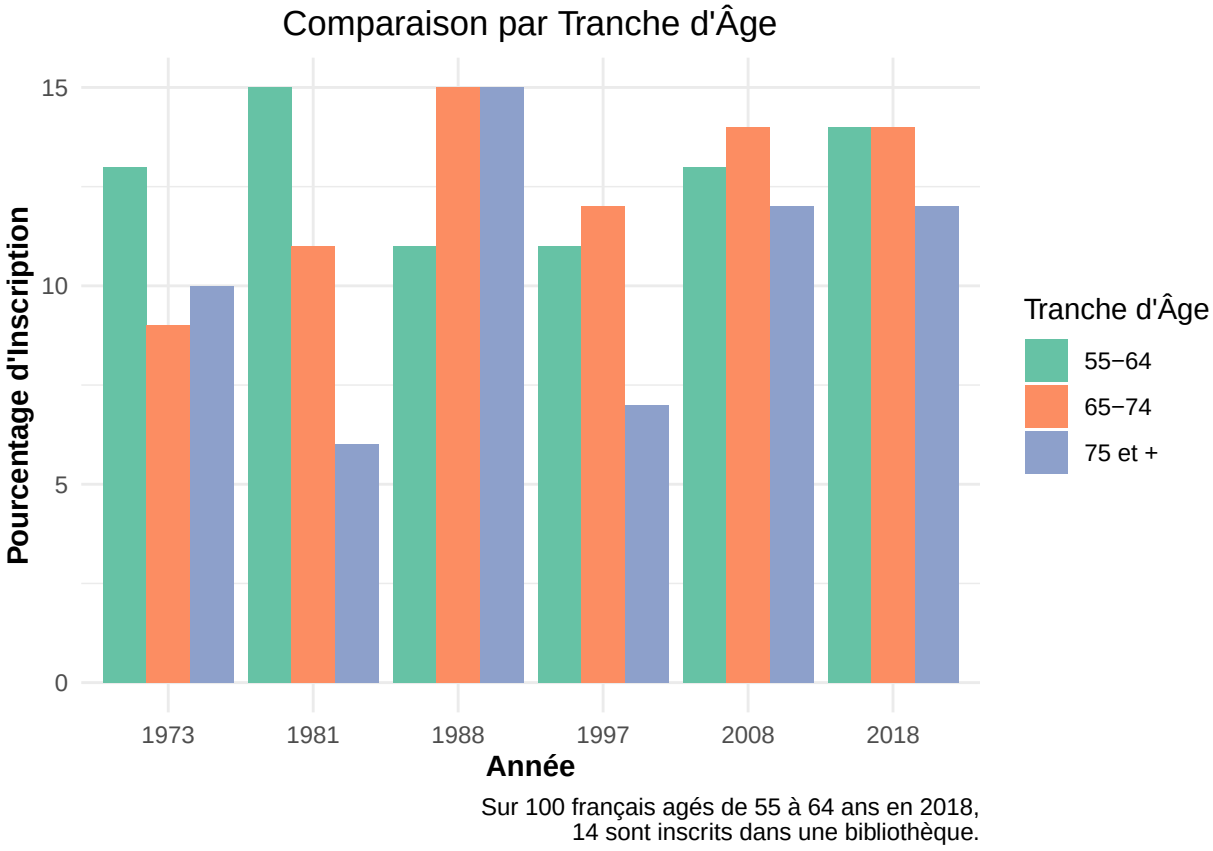
Cette histogramme démontre une tendance surprenante. Finalement ce sont les 15-19 ans qui ont la plus forte tendance à être inscrit en bibliothèque. Cette tendance est plutôt forte. en 1997 avec une différence de 9 points entre les français ayant 15 à 19 ans et ceux ayant 20 à 24 ans.

Ce sont donc à première vue les individus ayant entre 15 et 19 ans qui ont le plus tendance à être inscrit dans des bibliothèques toute année confondu. Ce qui invalide notre hypothèse précédente.

Si l'on s'intéresse désormais à aux ensembles d'âges les plus avancées:

```
var_age3_long %>%
  filter(age_detail %in% c("55-64", "65-74", "75 et +")) %>%
  ggplot(aes(x = Annee, y = Pourcentage, fill = age_detail)) +
  geom_bar(stat = "identity", position = position_dodge()) +
  scale_fill_brewer(palette = "Set2") +
  labs(
    title = "Comparaison par Tranche d'Âge",
    x = "Année",
    y = "Pourcentage d'Inscription",
    fill = "Tranche d'Âge",
    caption = "Sur 100 français agés de 55 à 64 ans en 2018,
    14 sont inscrits dans une bibliothèque."
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5),
    axis.title.x = element_text(face = "bold"),
```

```
    axis.title.y = element_text(face = "bold")
  )
```



Cette histogramme confirme notre hypothèse. **Les individus avec des âges très avancés ont moins tendance à être inscrit.** Les 75 ans et + sont pour chaque année hormis 1973 moins nombreux à être inscrit que les 65-74.

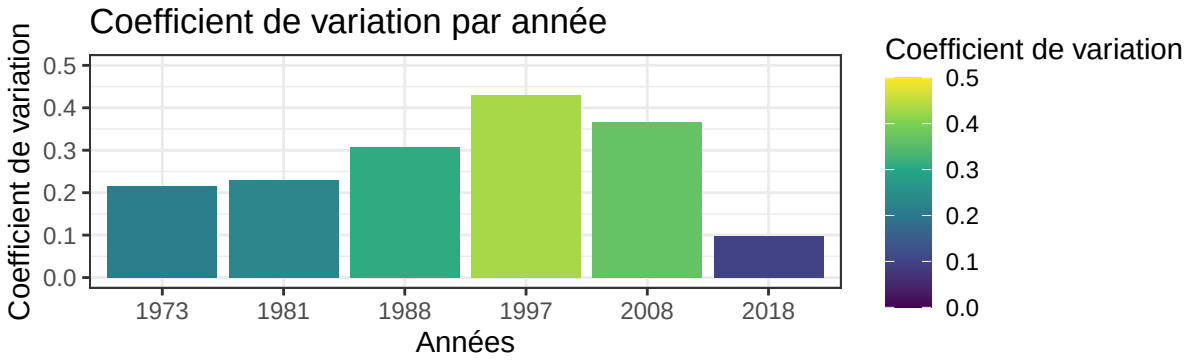
L'analyse de notre sous-ensemble 3 semble montrer des tendances surprenantes. **En poussant notre analyse à un degré plus fin, la variable âge semble devenir un facteur explicatif plus fort dans la tendance à être inscrit.** Nous devons donc réitérer nos analyses précédentes pour s'assurer que nous n'avons pas sous-estimé l'importance de la variable âge particulièrement pour 2018.

Reproduisons donc notre graphique précédent sur la mesure du coefficient de variabilité et comparons le avec le précédent.

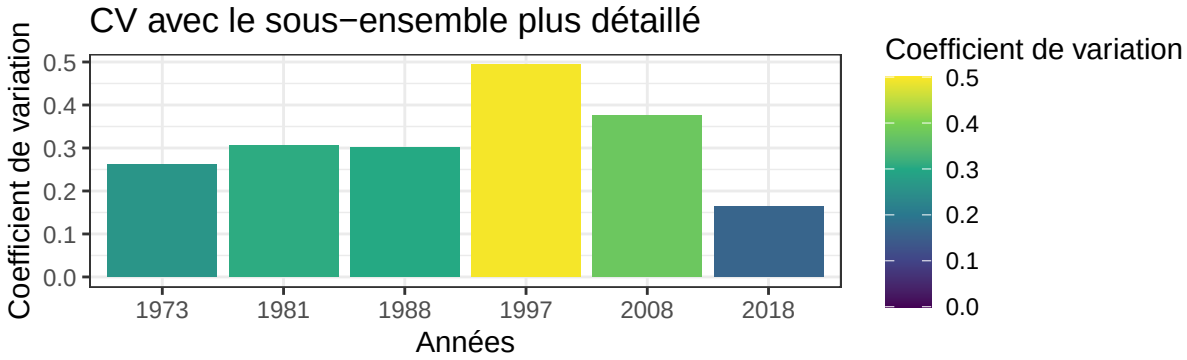
```
stats_par_annee2 <- var_age3_long %>%
  group_by(Annee) %>%
  summarise(
    Ecart_Type = sd(Pourcentage),
    Moyenne = mean(Pourcentage)
  )

stats_par_annee2 <- stats_par_annee2 %>%
  mutate(Proportion_Ecart_Type = Ecart_Type / Moyenne)

print(stats_par_annee2)
```

```
## # A tibble: 6 x 4
##   Annee Ecart_Type Moyenne Proportion_Ecart_Type
##   <chr>      <dbl>   <dbl>                 <dbl>
## 1 1973        3.39    13                   0.261
## 2 1981        4.26    13.9                 0.306
## 3 1988        5.36    17.8                 0.301
## 4 1997       10.3     20.8                 0.494
## 5 2008        7.61    20.2                 0.376
## 6 2018        2.45    15                   0.163
```

```
graph_cf_age3 <- ggplot(stats_par_annee2) +
  aes(x = Annee,
    y = Proportion_Ecart_Type,
    fill = Proportion_Ecart_Type) +
  labs(title = "CV avec le sous-ensemble plus détaillé",
       x = "Années",
        y = "Coefficient de variation",
      caption = "Sur 100 français agés de 15 à 24 ans en 2018, 19 sont inscrits dans une bibliothèque."
    fill = "Coefficient de variation") +
  geom_col() +
  scale_fill_viridis_c(option = "viridis", direction = 1,limits = c(0,0.5)) +
  theme_bw()
(graph_cf_age2 | graph_cf_age3) +
  plot_layout(ncol = 1, nrow = 2)
```



En 2018, le coefficient de variation du sous−ensemble 2 est inférieur à 0,10



L00 français agés de 15 à 24 ans en 2018, 19 sont inscrits dans une bibliothèque.

On observe des indices de variations complètement différent en améliorant la finesse de notre analyse. De

nouveau on observe que avant 2018 l'âge était une variables avec un impact important dans la tendance à être inscrit. On observe qu'en 1973, 1981, 1997 et en 2018 notre coefficientde variation augmente fortement.

Ainsi, notre hypothèse première que l'âge n'était plus un facteur très important est invalidé puisque son coefficient de variation est de 0.16 un résultat supérieur à 10%. Néanmoins il reste inférieur aux autres années.

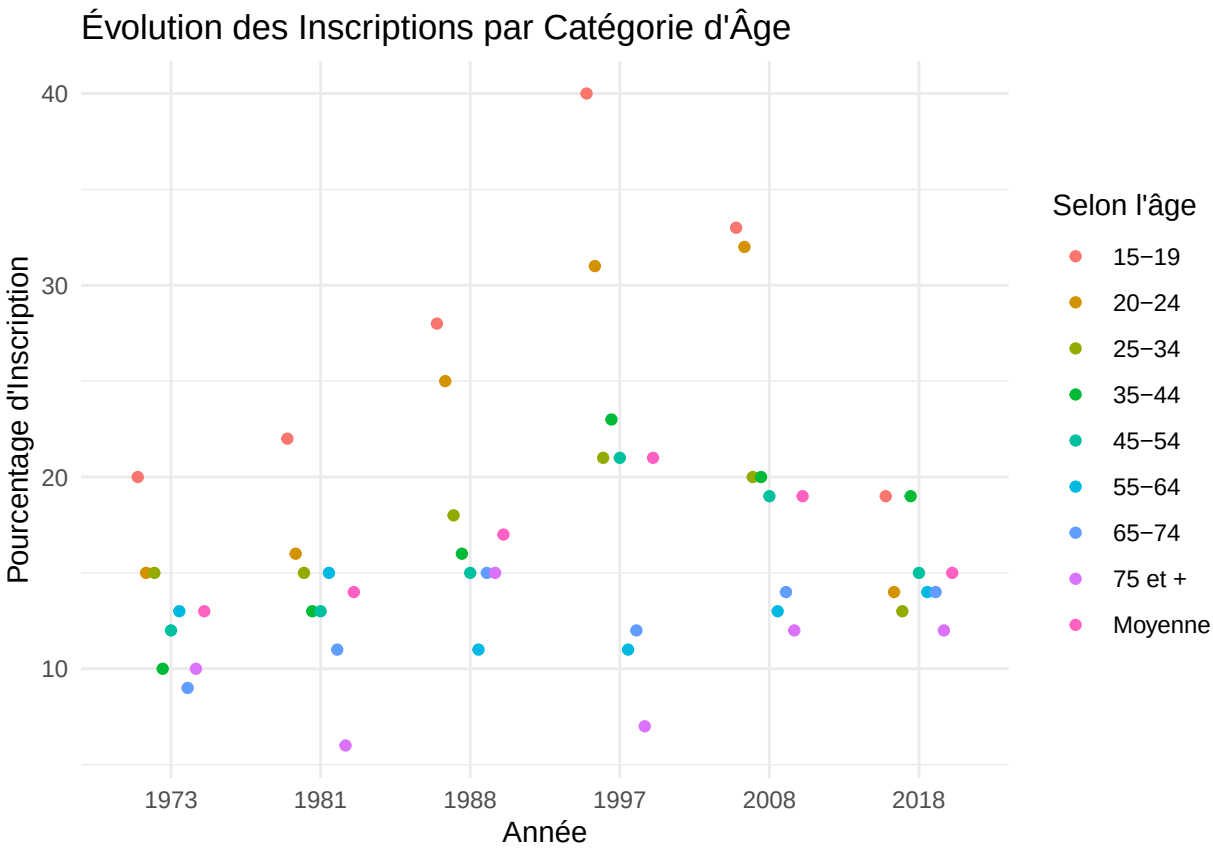
On peut donc dire qu'en 2008, l'impact de l'âge sur le taux d'inscription est moindre par rapport aux autres années. Par rapport à 1997, la variable est 3 fois moins importante.

```
print(paste("Le coefficient multiplicateur du CV de 2018 par rapport à 1997 est :", 0.4936252 / 0.163299
```

```
## [1] "Le coefficient multiplicateur du CV de 2018 par rapport à 1997 est : 3.022824960058"
```

Représentation des variables toutes années confondues:

```
ggplot(var_age3_long, aes(x = Annee, y = Pourcentage, color = age_detail)) +
  geom_point(position = position_dodge(width = 0.5)) +
  labs(title = "Évolution des Inscriptions par Catégorie d'Âge",
       x = "Année", y = "Pourcentage d'Inscription", color = "Selon l'âge") +
  theme_minimal()
```

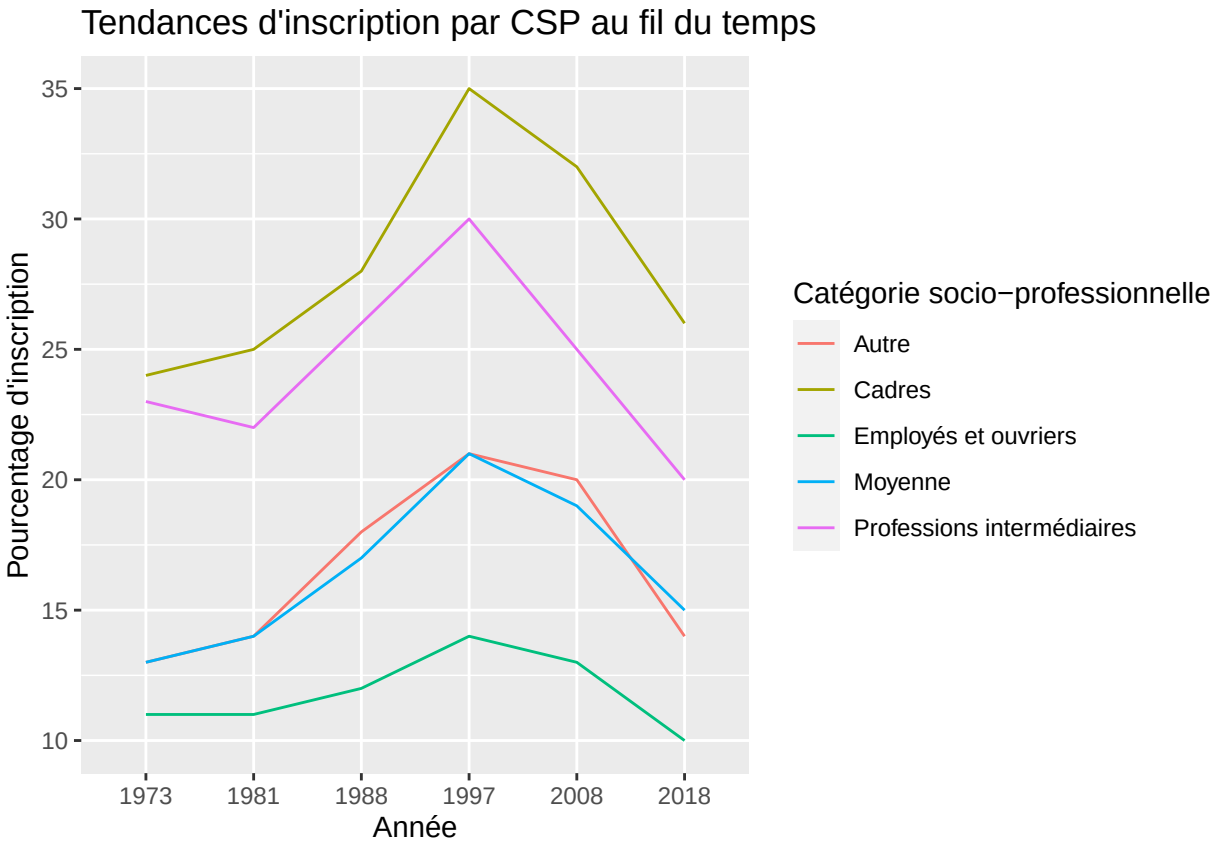


Notre nouveau graphique démontre très clairement nos affirmations, la différences d'âge est un facteur explicatif fort. Les jeunes français ayant entre 15 et 19 ans ont plus tendance à être inscrit que la moyenne. De même, les 75 ans et plus ont une tendance moindre à être inscrit.

Analyse des inscrits en bibliothèque en fonction de la CSP :

Notre analyse consistera à déterminer si la CSP est une variable déterminant l'inscription ou non dans une bibliothèque.

```
ggplot(var_CS_regroup_long, aes(x = Annee, y = Pourcentage, group = pcs_regroup, color = pcs_regroup)) +
  geom_line() +
  labs(
    title = "Tendances d'inscription par CSP au fil du temps",
    x = "Année",
    y = "Pourcentage d'inscription",
    color = "Catégorie socio-professionnelle"
  )
```

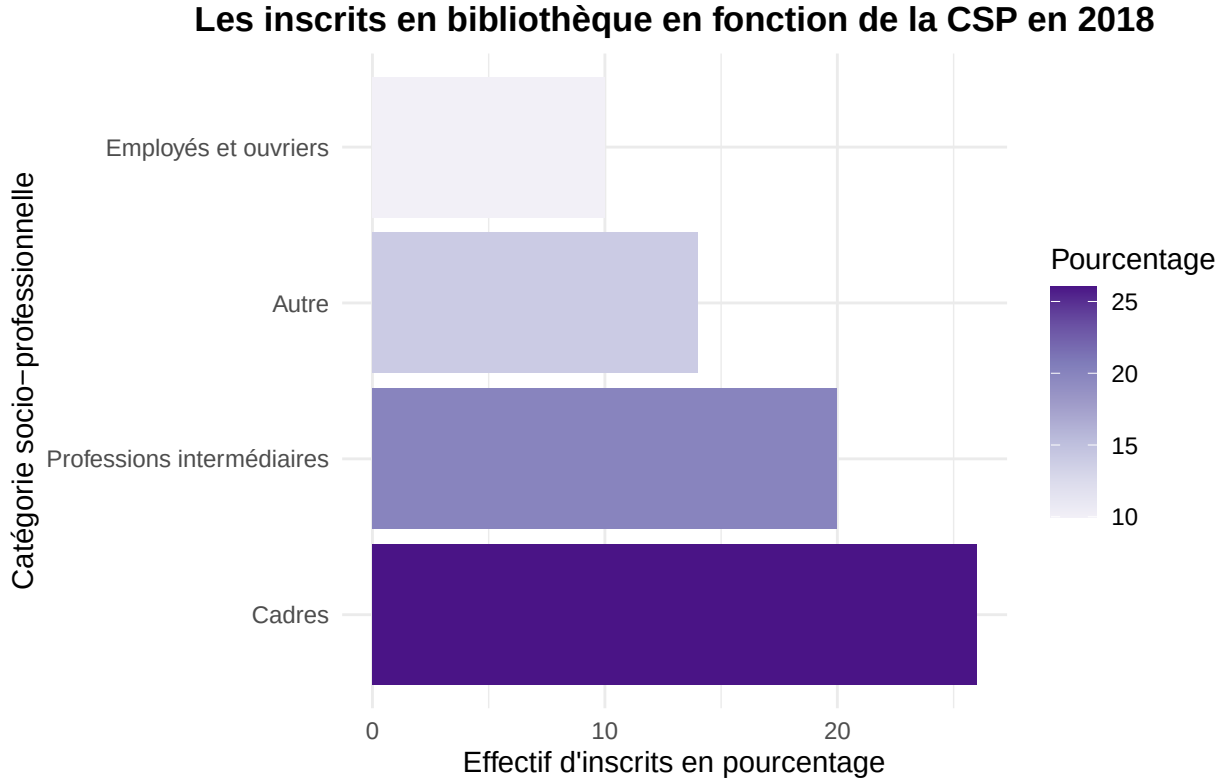


Nous faisons le même constat que pour l'âge : on observe un pic d'inscription en 1997.

Commençons par analyser l'année 2018. Pour réaliser cette analyse nous avons utilisé le package "esquisse".

```
var_CS_regroup_long %>%
  filter(!(pcs_regroup %in% "Moyenne")) %>%
  filter(Annee %in% 2018) %>%
  mutate(
    Pourcentage = as.numeric(as.character(Pourcentage)),
    pcs_regroup = ifelse(pcs_regroup %in% c("Étudiants", "Femmes/Hommes au foyer"), "Autre", pcs_regroup
  ) %>%
  ggplot() +
```

```r
  aes(
    x = reorder(pcs_regroup, -Pourcentage),
    y = Pourcentage,
    fill = Pourcentage,
    group = Annee
  ) +
  geom_col() +
  scale_fill_distiller(palette = "Purples", direction = 1) +
  labs(
    x = "Catégorie socio-professionnelle",
    y = "Effectif d'inscrits en pourcentage",
    title = "Les inscrits en bibliothèque en fonction de la CSP en 2018",
    caption = "Sur 100 français de professions intermédiaires, 20 sont inscrits en bibliothèque.\n(Autre
  ) +
  coord_flip() +
  theme_minimal() +
  theme(
    plot.title = element_text(face = "bold", hjust = 0.5),
    plot.caption = element_text(face = "bold")
  )
```



Dans un premier temps, la surepresentation des cadres dans le nombre d'inscrits en bibliothèque est assez clair. En effet, sur 100 cadres ils sont à peu près 26 à être inscrits, quand ils sont seulement 10 chez les ouvriers/employés, soit un écart de 16 points de pourcentage.

Lorsqu'on calcule l'écart-type, la moyenne et la proportion de l'écart-type par rapport à la moyenne obtient

le résultat suivant :

```
pourcentages_2018 <- var_CS_regroup_long %>%
  filter(Annee == "2018") %>%
  .$Pourcentage

ecart_type_2018 <- sd(pourcentages_2018)
moyenne_2018 <- mean(pourcentages_2018)
Proportion_ecart_type_2018 <- ecart_type_2018 / moyenne_2018

ecart_type_2018
```

```
## [1] 6.164414
```

```
moyenne_2018
```

```
## [1] 17
```

```
Proportion_ecart_type_2018
```

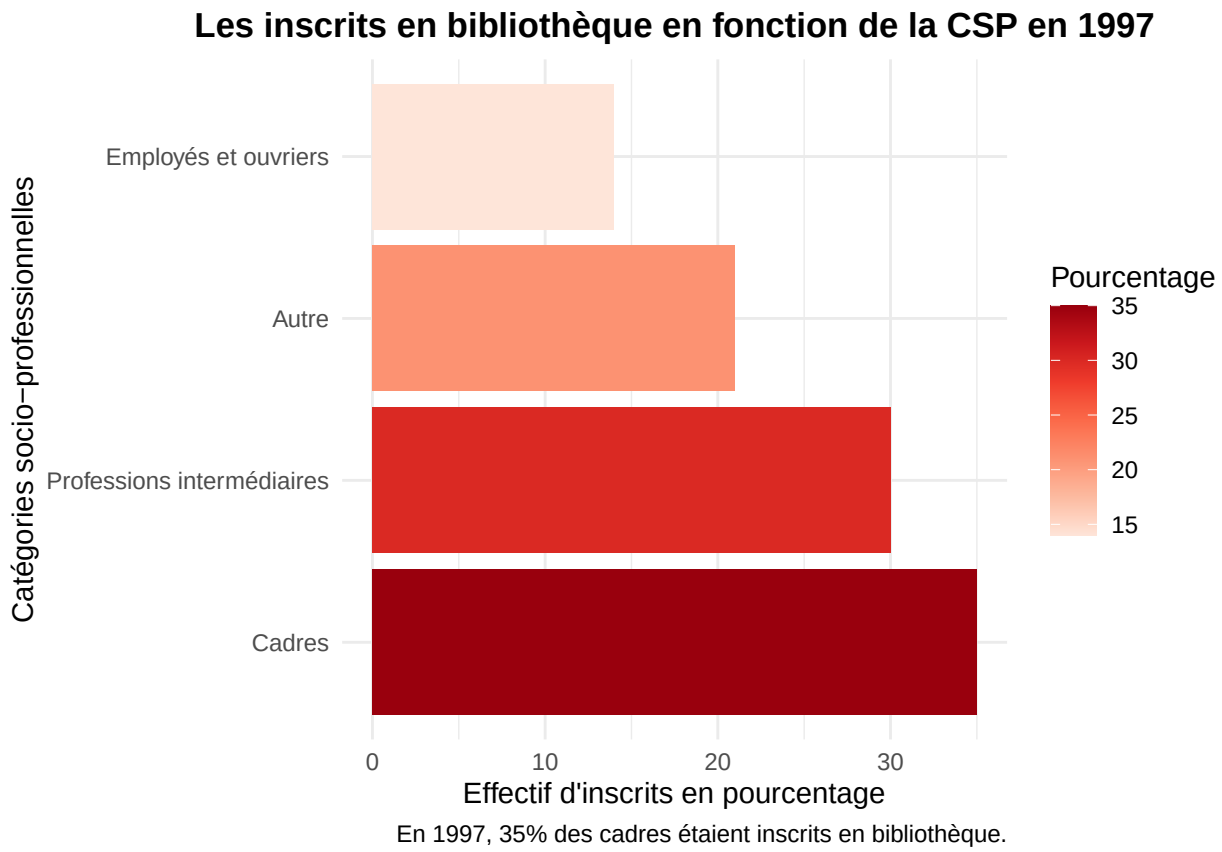
```
## [1] 0.3626126
```

En résumé, pour l'année 2018, nos données montrent une moyenne de 17% avec une certaine variabilité (écart type de 6.164414), et la variabilité relative par rapport à la moyenne est d'environ 36%. Cela suggère une dispersion relativement importante par rapport à la moyenne, avec une sureprésentation des cadres dans les inscrits en bibliothèque en 2018. La CSP semble être un facteur significatif dans l'inscription ou non en bibliothèque.

La tendance a-t-elle toujours été la même ? Comparons avec l'année 1997

```
library(ggplot2)
library(dplyr)

var_CS_regroup_long %>%
  filter(!(pcs_regroup %in% "Moyenne")) %>%
  filter(Annee %in% "1997") %>%
  mutate(Pourcentage = as.numeric(as.character(Pourcentage))) %>%
  ggplot() +
  aes(
    x = reorder(pcs_regroup, -Pourcentage),
    y = Pourcentage,
    fill = Pourcentage,
    group = Annee
  ) +
  geom_col() +
  scale_fill_distiller(palette = "Reds", direction = 1) +
  labs(
    x = "Catégories socio-professionnelles",
    y = "Effectif d'inscrits en pourcentage",
    title = "Les inscrits en bibliothèque en fonction de la CSP en 1997",
    caption = "En 1997, 35% des cadres étaient inscrits en bibliothèque."
  ) +
```

```
  coord_flip() +
  theme_minimal() +
  theme(plot.title = element_text(face = "bold", hjust = 0.5))
```



En 1997, 35% des cadres étaient inscrits en bibliothèque.

```
pourcentages_1997 <- var_CS_regroup_long %>%
  filter(Annee == "1997") %>%
  .$Pourcentage

ecart_type_1997 <- sd(pourcentages_1997)
moyenne_1997<- mean(pourcentages_1997)
Proportion_ecart_type_1997 <- ecart_type_1997 / moyenne_1997

ecart_type_1997
```

```
## [1] 8.288546
```

```
moyenne_1997
```

```
## [1] 24.2
```

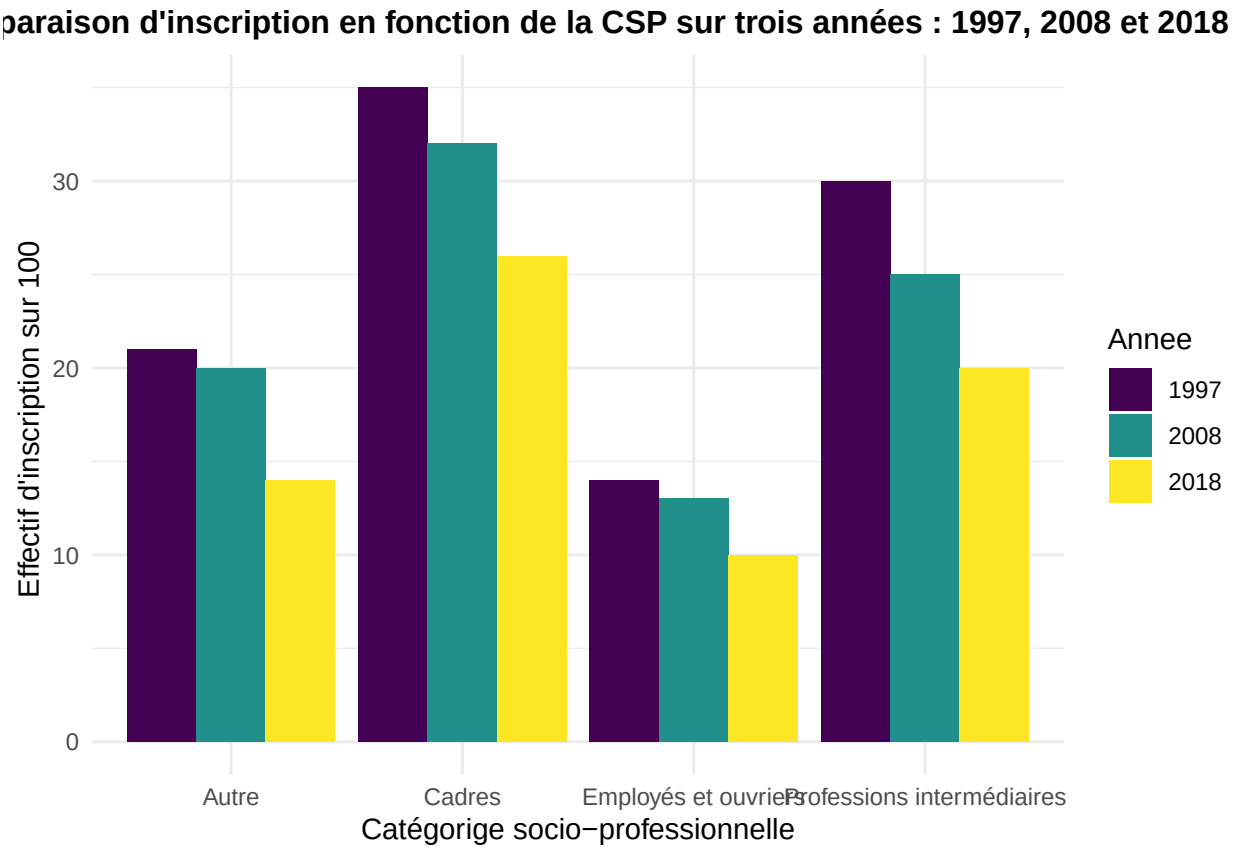
```
Proportion_ecart_type_1997
```

```
## [1] 0.3425019
```

Par rapport à l'année 2018, l'année 1997 a un écart type plus élevé ($8.288546 > 6.164414$), une moyenne plus élevée ($24.2 > 17$), mais une proportion de l'écart type par rapport à la moyenne légèrement plus basse ($0.3425019 < 0.3626126$). Cela suggère que, bien que la variabilité (écart type) soit plus importante en 1997, la variabilité relative par rapport à la moyenne est légèremen plus faible. En 1997, l'écart entre les cadres inscrits et les ouvriers/employés est d'autant plus important : il est de plus de 20 points.

Pour avoir un représentation plus globale, nous allons effectuer une comparaison du taux d'inscription par CSP sur trois années: 1997, 2008 et 2018

```
library(dplyr)
library(ggplot2)
var_CS_regroup_long %>%
 filter(!(pcs_regroup %in% "Moyenne")) %>%
 filter(Annee %in% c("2018", "2008",
"1997")) %>%
 ggplot() +
 aes(x = pcs_regroup, fill = Annee, group = Annee, weight = Pourcentage) +
 geom_bar(position = "dodge") +
 scale_fill_viridis_d(option = "viridis", direction = 1) +
 labs(x = "Catégorige socio-professionnelle",
 y = "Effectif d'inscription sur 100 ", title = "Comparaison d'inscription en fonction de la CSP sur tr
 theme_minimal() +
 theme(plot.title = element_text(size = 12L, face = "bold", hjust = 0.5))
```

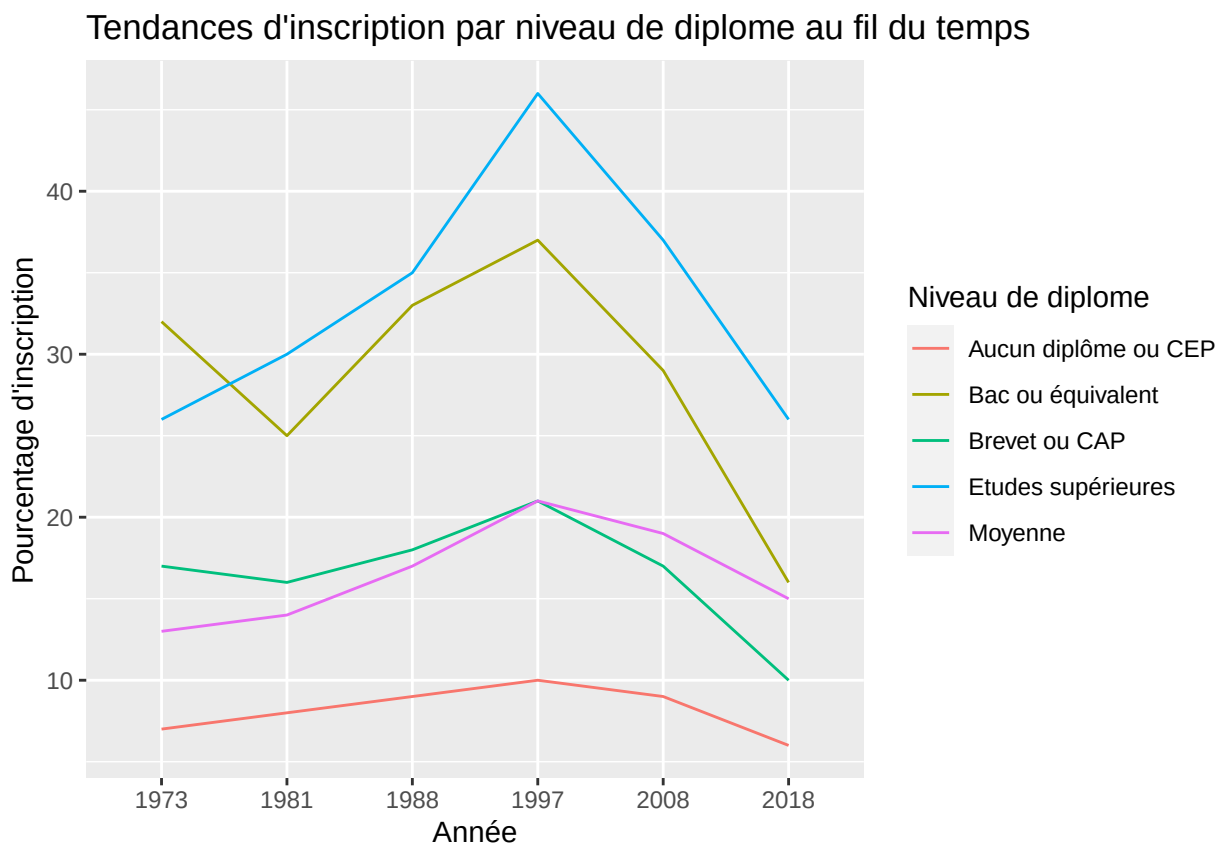


Ainsi, une première tendance s'observe: une baisse globale du taux d'inscription entre 1997 et 2018. Une diminution plus ou moins importante en fonction des CSP. Elle va être de 10 points chez les cadres et les professions intermédiaires, quand elle sera seulement de 4 chez les ouvriers/employés. Cette tendance peut

s'expliquer par plusieurs facteurs : au début des années 2000 c'est le développement de l'internet en France et le début d'une période de démocratisation des équipement de PC chez les foyers. Or on sait que ce sont les plus aisés qui pouvaient et peuvent disposer d'un PC ce qui ne nécessite pas d'inscription à la bibliothèque pour ce motif.

Analyse des inscrits en bibliothèque en fonction du diplôme

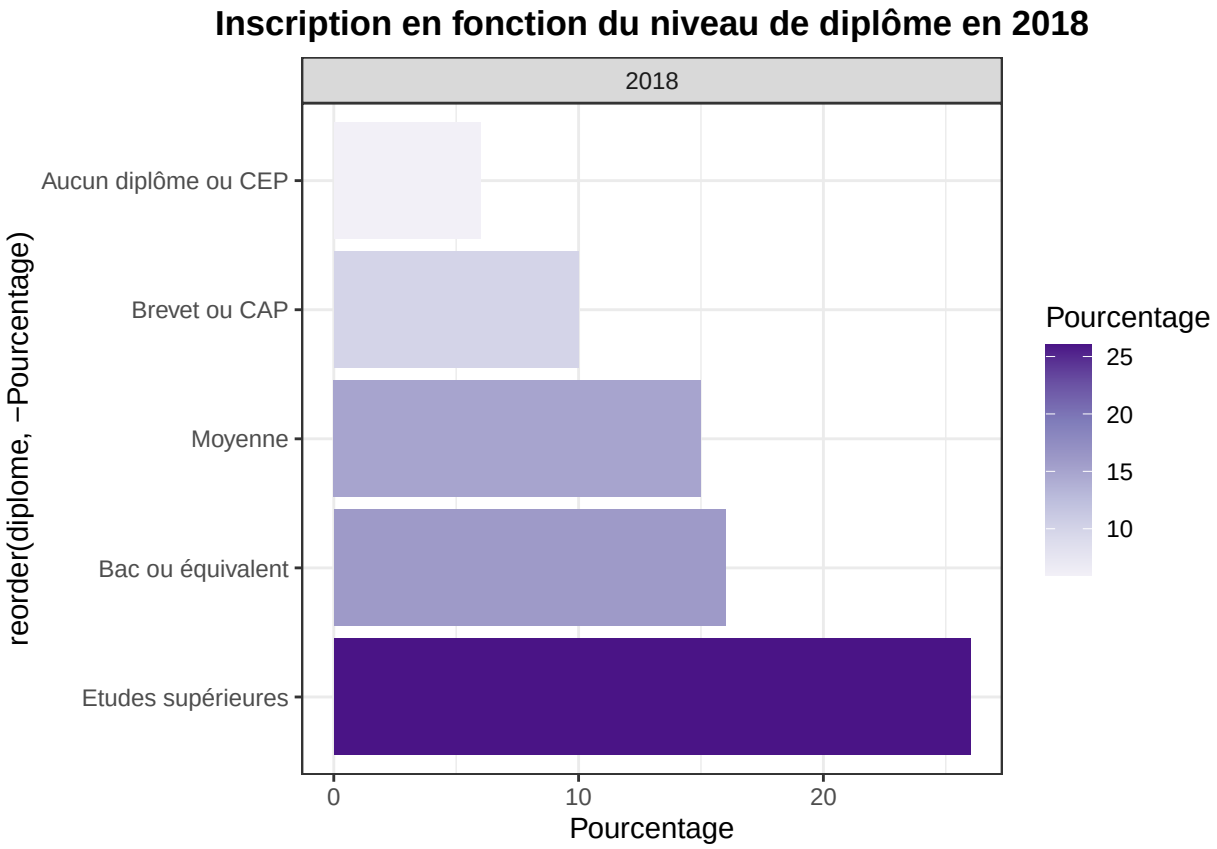
```
ggplot(var_dipl_long, aes(x = Annee, y = Pourcentage, group = diplome, color = diplome)) +
  geom_line() +
  labs(
    title = "Tendances d'inscription par niveau de diplome au fil du temps",
    x = "Année",
    y = "Pourcentage d'inscription",
    color = "Niveau de diplome"
  )
```



A première vue on observe également un pic d'inscription en 1997 : année de développement de l'ineternet et où l'équipement des foyers d'un PC n'est pas encore généralisé. L'inscription aux bibliothèques peut être un moyen d'acceder à un ordinateur.

```
library(dplyr)
library(ggplot2)

var_dipl_long %>%
```

```
  filter(Annee %in% "2018") %>%
  ggplot() +
  aes(x = reorder(diplome, -Pourcentage), y = Pourcentage, fill = Pourcentage) +
  geom_col() +
  scale_fill_distiller(palette = "Purples", direction = 1) +
  labs(title = "Inscription en fonction du niveau de diplôme en 2018") +
  coord_flip() +
  theme_bw() +
  theme(plot.title = element_text(face = "bold", hjust = 0.5)) +
  facet_wrap(vars(Annee))
```



En 2018, on voit que le niveau de diplôme influe grandement sur l'inscription en bibliothèque: les plus diplômés sont plus inscrits que la moyenne.

```
inscrits2018 <- var_dipl_long %>%
  filter(Annee == "2018") %>%
  .$Pourcentage

ETD2018 <- sd(inscrits2018)
MD2018<- mean(inscrits2018)
ProportionETD2018 <- ETD2018 / MD2018

ETD2018
```

```
## [1] 7.536577
```

```
MD2018
```

```
## [1] 14.6
```

```
ProportionETD2018
```

```
## [1] 0.5162039
```

L'écart-type (ETD2018) est d'environ 7.54, ce qui indique une certaine dispersion des taux d'inscription autour de la moyenne. La moyenne (MD2018) est de 14.6, qui représente la moyenne des taux d'inscription en 2018. La proportion de l'écart-type par rapport à la moyenne (ProportionETD2018) est d'environ 0.52, suggérant une variabilité relative importante par rapport à la moyenne.

Si on ce concentre précisément sur les inscrits qui sont diplomés du supérieur et les inscrits qui n'ont que le brevet ou un CAP, cela nous permettra de mettre en lumière l'inégale répartition en terme d'inscription en fonction du diplôme.

```
etudes_superieures_2018 <- var_dipl_long %>%
  filter(Annee == "2018", diplome == "Etudes supérieures") %>%
  .$Pourcentage

brevet_cap_2018 <- var_dipl_long %>%
  filter(Annee == "2018", diplome == "Brevet ou CAP") %>%
  .$Pourcentage

summary(etudes_superieures_2018)
```

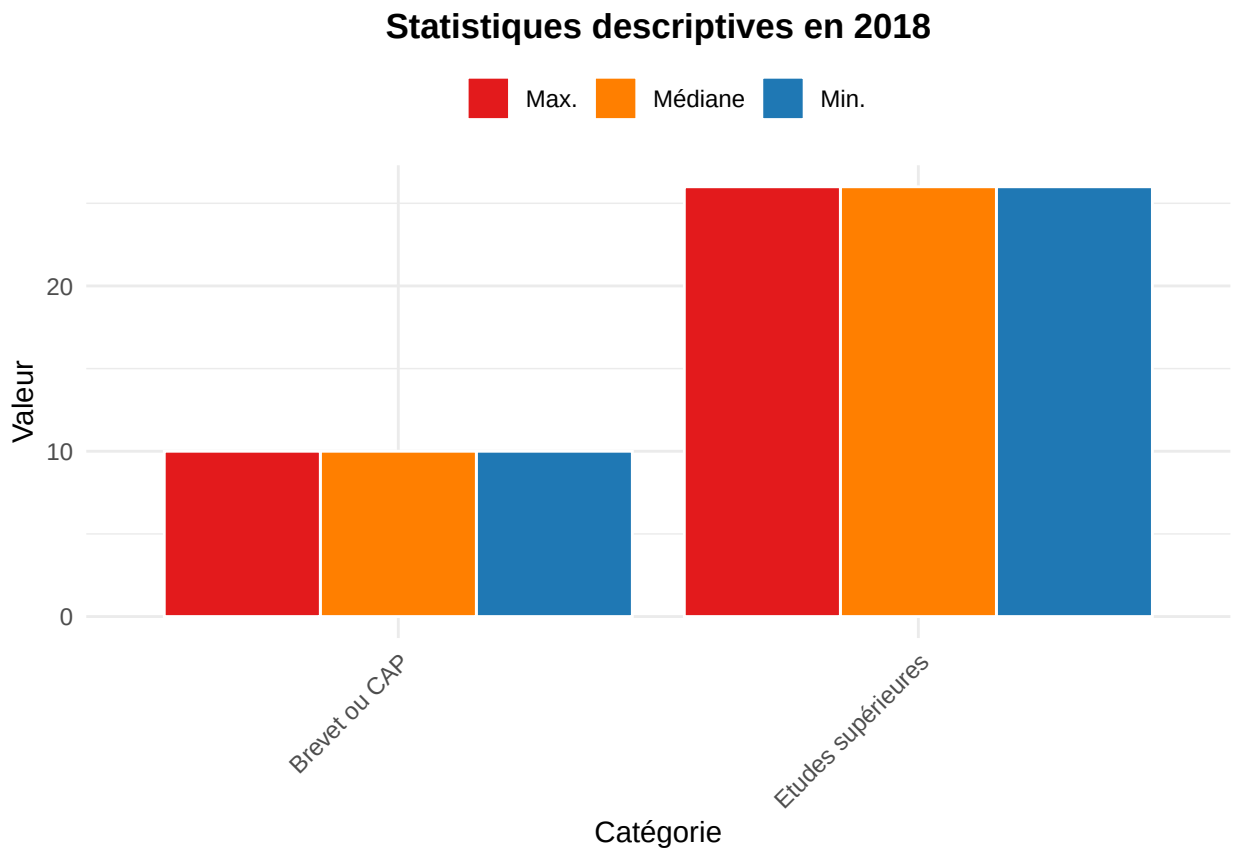
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      26      26      26      26      26      26
```

```
summary(brevet_cap_2018)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      10      10      10      10      10      10
```

```
library(ggplot2)

data_descriptive <- data.frame(
  Categorie = rep(c("Etudes supérieures", "Brevet ou CAP"), each = 3),
  Statistique = rep(c("Min.", "Médiane", "Max."), times = 2),
  Valeur = c(
    min(etudes_superieures_2018),
    median(etudes_superieures_2018),
    max(etudes_superieures_2018),
    min(brevet_cap_2018),
    median(brevet_cap_2018),
    max(brevet_cap_2018)
  )
)

couleurs <- c("Min." = "#1f78b4", "Médiane" = "#ff7f00", "Max." = "#e31a1c")
```

```r
ggplot(data_descriptive, aes(x = Categorie, y = Valeur, fill = Statistique)) +
  geom_col(position = "dodge", color = "white") +
  labs(title = "Statistiques descriptives en 2018",
       x = "Catégorie",
       y = "Valeur",
       fill = "") +
  scale_fill_manual(values = couleurs) +
  theme_minimal() +
  theme(legend.position = "top",
        axis.text.x = element_text(angle = 45, hjust = 1),
        plot.title = element_text(face = "bold", hjust = 0.5))
```

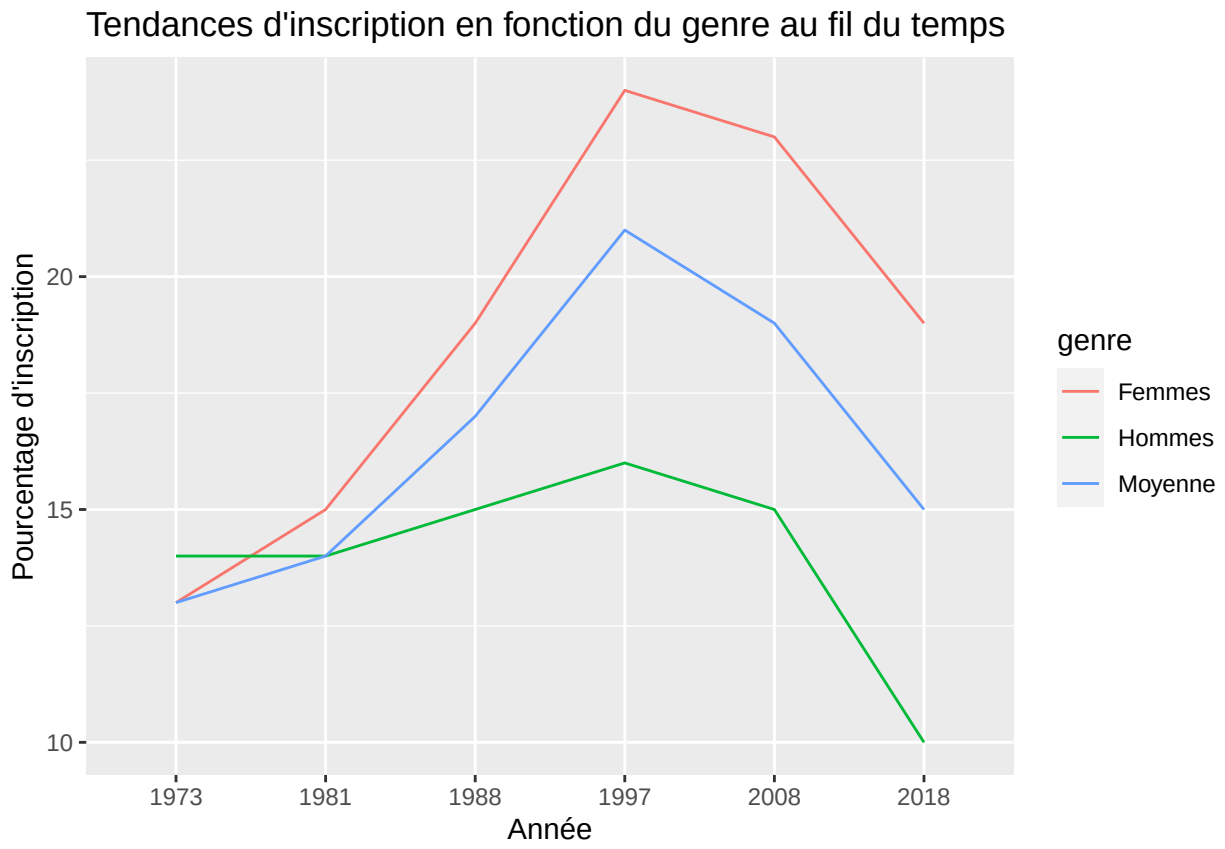


Ainsi, on voit bien qu'être diplomé du supérieur est une variable forte qui implique un fort taux d'inscription en bibliothèque par rapport à la moyenne et surtout par rapport aux inscrits ayant seulement le brevet ou un CAP. Cette observation s'applique également pour les autres années, montrant que c'est un phénomène structurel.

Cette pratique reste fortement liée au milieu social des individus, des années 1970 à aujourd'hui, malgré une légère atténuation de ces écarts : en 2018, se rendre dans une bibliothèque reste une pratique 3 fois plus courante pour les diplômés de l'enseignement supérieur par rapport aux moins diplômés. De même, les cadres sont près de 2 fois plus nombreux à s'y rendre : 37 % ont fréquenté une bibliothèque au cours de l'année, contre 19 % des ouvriers et employés.

Analyse des inscrits en bibliothèque en fonction du genre

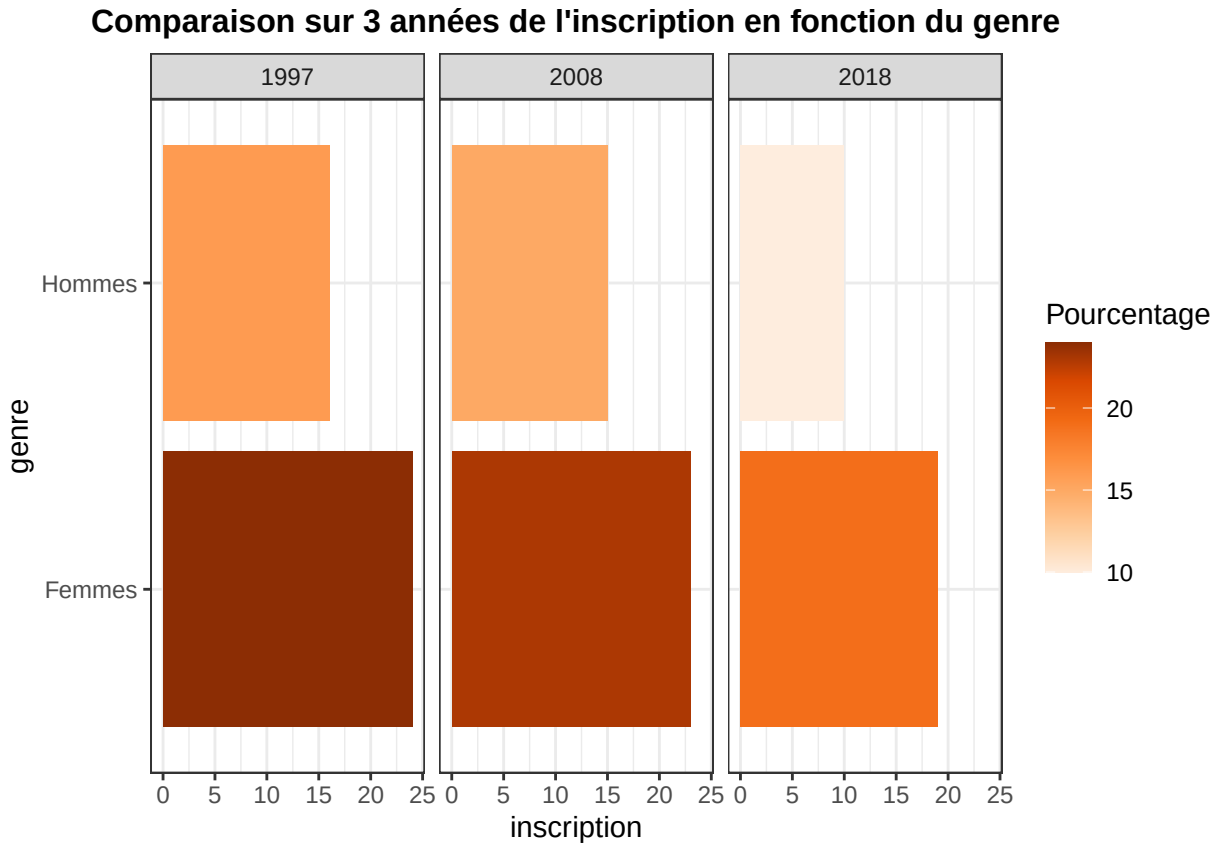
```
ggplot(var_sexe_long, aes(x = Annee, y = Pourcentage, group = sexe, color = sexe)) +
  geom_line() +
  labs(
    title = "Tendances d'inscription en fonction du genre au fil du temps",
    x = "Année",
    y = "Pourcentage d'inscription",
    color = "genre"
  )
```



Ici, on voit déjà que les femmes ont plus tendance à être inscrites en bibliothèque que les hommes au fil du temps de 1973 à 2018. On observe également ce même pic d'inscription pour les deux genres en 1997. Confirmant ainsi que 1997 marque une année particulière.

```
library(dplyr)
library(ggplot2)

var_sexe_long %>%
 filter(!(sexe %in% "Moyenne")) %>%
 filter(Annee %in% c("2018", "2008", "1997")) %>%
 ggplot() +
 aes(x = sexe, y = Pourcentage, fill = Pourcentage, group = Annee) +
 geom_col() +
 scale_fill_distiller(palette = "Oranges", direction = 1) +
```

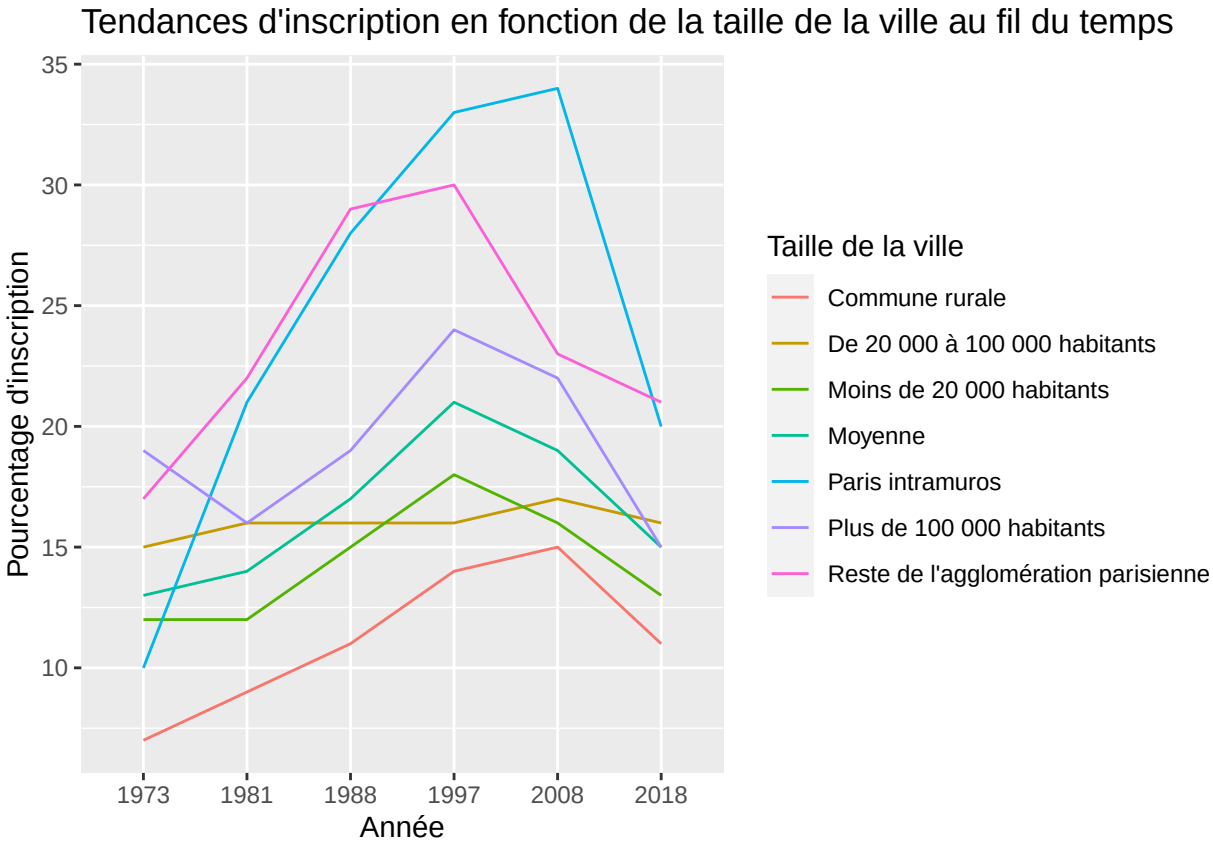
```
labs(x = "genre", y = "inscription", title = "Comparaison sur 3 années de l'inscription en fonction du
coord_flip() +
theme_bw() +
theme(plot.title = element_text(size = 12L, face = "bold", hjust = 0.5)) +
facet_wrap(vars(Annee))
```



A partir de ce graphe, un remarque que les femmes sont proportionnellement plus nombreuses que les hommes à être inscrites dans une bibliothèque et à en avoir fréquenté une, au moins une fois au cours des douze derniers mois, et ces écarts sont substantiels. En 2018, 19 % d'entre elles déclarent être inscrites dans une bibliothèque contre 10 % des hommes. Depuis 1997, pour les femmes comme pour les hommes, les inscriptions ont tendanciellement baissé.

Analyse des inscrits en bibliothèque en fonction de la taille de la ville

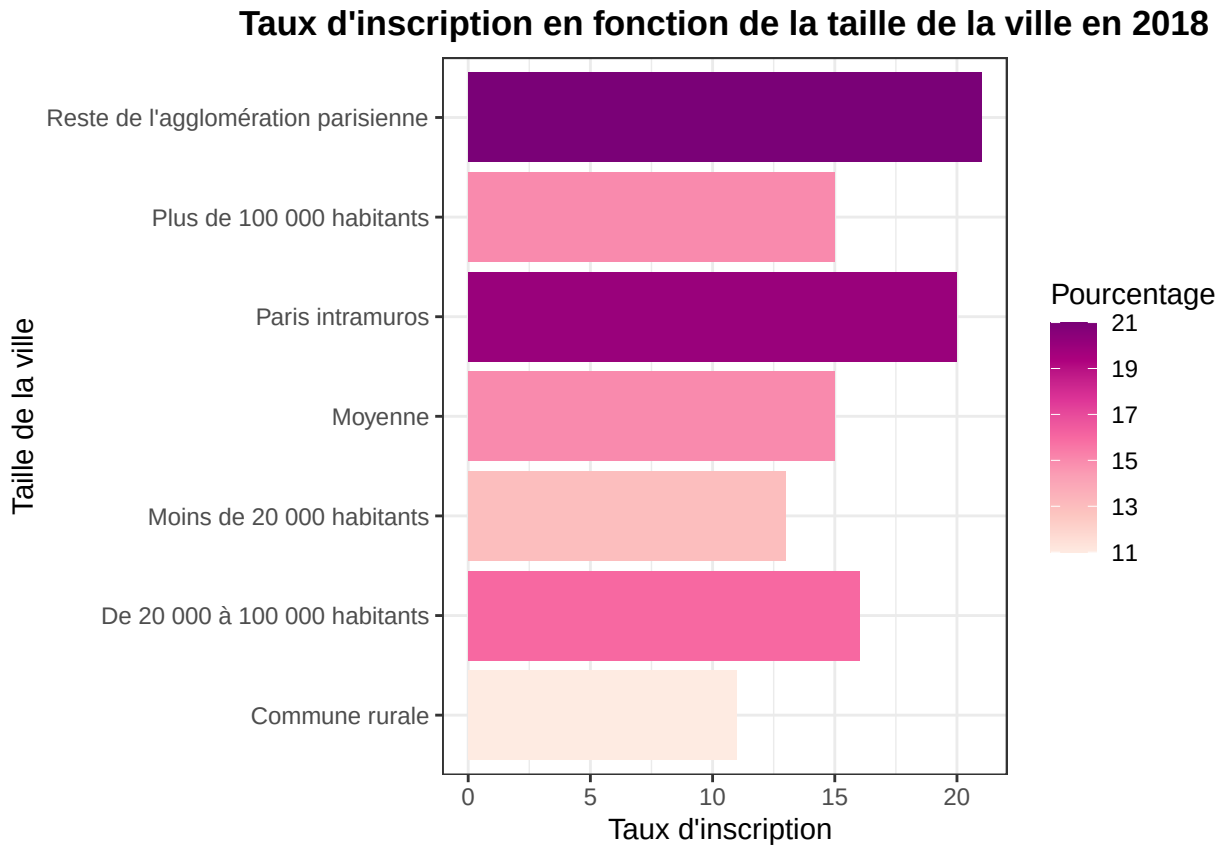
```
ggplot(var_ville_long, aes(x = Annee, y = Pourcentage, group = taille_ville, color = taille_ville)) +
  geom_line() +
  labs(
    title = "Tendances d'inscription en fonction de la taille de la ville au fil du temps",
    x = "Année",
    y = "Pourcentage d'inscription",
    color = "Taille de la ville"
  )
```



Les “Communes rurales” et les “Moins de 20 000 habitants” affichent des tendances relativement stables, tandis que les zones “Plus de 100 000 habitants” montrent une croissance plus marquée, avec des années de pointe en 1997 et 2008. Les années 1997 et 2008 semblent être des points d’inflexion importants, avec des variations notables dans plusieurs catégories. Cela pourrait indiquer des événements spécifiques, tels que des campagnes de sensibilisation réussies ou des initiatives culturelles. Paris intramuros a connu une croissance constante du taux d’inscription, tandis que le reste de l’agglomération parisienne a montré des fluctuations plus prononcées.

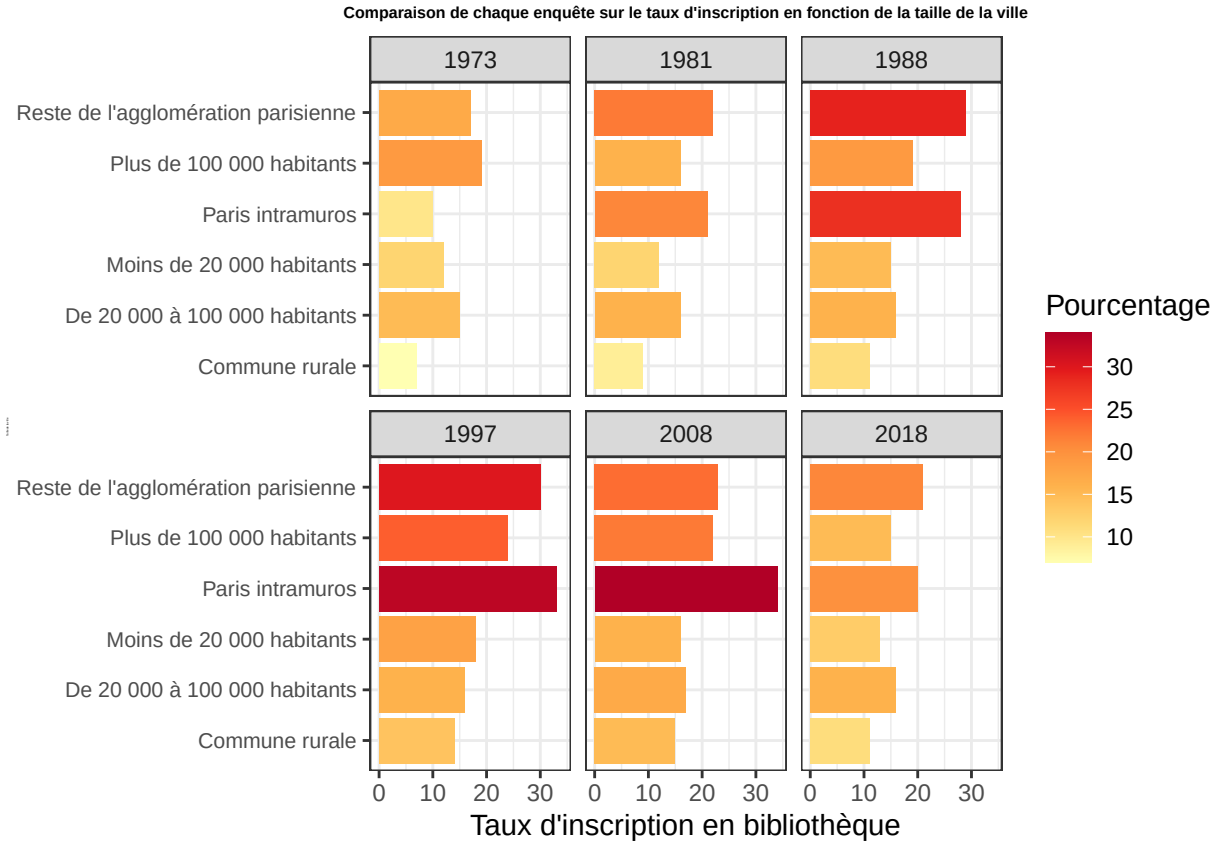
Si on se concentre sur la dernière enquête 2018 on obtient le graphe suivant:

```
library(dplyr)
library(ggplot2)

var_ville_long %>%
 filter(Annee %in% "2018") %>%
 ggplot() +
 aes(x = taille_ville, y = Pourcentage, fill = Pourcentage, group = Annee) +
 geom_col() +
 scale_fill_distiller(palette = "RdPu", direction = 1) +
 labs(x = "Taille de la ville",
 y = "Taux d'inscription ", title = "Taux d'inscription en fonction de la taille de la ville en 2018") +
 coord_flip() +
 theme_bw() +
 theme(plot.title = element_text(face = "bold", hjust = 0.5))
```



Ce graphique à barres horizontales montre le taux d'inscription en bibliothèque en 2018, en distinguant les différentes catégories de taille de ville par des couleurs de pourpre, avec des barres plus foncées indiquant des taux d'inscription plus élevés. La rotation des étiquettes de l'axe x facilite la lecture des catégories de taille de ville. On voit la surepresentation de Paris et du reste de l'agglomération par rapport à la moyenne dans le taux d'inscription en bibliothèque. L'écart entre les inscrits vivant à Paris et ceux vivant en commune rurale est de 10 points, ce qui est non négligeable.

```
var_ville_long %>%
 filter(!(taille_ville %in% "Moyenne")) %>%
 ggplot() +
 aes(x = taille_ville, y = Pourcentage, fill = Pourcentage, group = Pourcentage) +
 geom_col() +
 scale_fill_distiller(palette = "YlOrRd", direction = 1) +
 labs(x = "Taille de la ville ",
      y = "Taux d'inscription en bibliothèque",
      title = "Comparaison de chaque enquête sur le taux d'inscription en fonction de la taille de la v
 coord_flip() +
 theme_bw() +
 theme(plot.title = element_text(face = "bold", hjust = 0.5, size = 6),
       axis.text.y = element_text(size = 8),
       axis.title.y = element_text(size = 1)) +
 facet_wrap(vars(Annee))
```



Ainsi, on note une augmentation générale du taux d'inscription au fil des années sur toutes les tailles de ville. Avec pour chaque année la domination de Paris et dans une moindre mesure, des grandes villes en générale. On note cependant un recul en 2018 par rapport à 2008 du taux d'inscription à Paris avec une perte de 14 points en 10 ans.

Conclusion, observations et limites

Ces évolutions s'expliquent selon les chercheurs **Philippe Lombardo** et **Loup Wolff**par la "profonde mutation de ces équipements culturels depuis plusieurs décennies, avec une politique volontariste d'accueil des publics dans les bibliothèques, avec ou sans inscription préalable, ainsi que par le développement du réseau territorial des lieux de lecture publique au cours des années 1980 et 1990 notamment".

Ainsi, il est selon nous nécessaire de mettre en évidence les limites de ce jeu de données. Premièrement, le jeu de données est assez mal construit comme signalé en introduction. Il nous est impossible de croiser les données pour effectuer une analyse plus fine. En effet il aurait été pertinent d'avoir des informations multiples sur chaque individu : sa CSP, son genre, son âge, la taille de sa ville etc... Ce qui nous aurait permis de supperposer ces variables afin de mener une réelle étude sociologique et plus pertinente. Enfin, nous avons constaté que les questions posées (dans la méthodologie de l'enquête) induisent certains biais des les réponses, ne permettant une parfaite compréhension des données. La question posée est "est-vous personnellement inscrit", la précision "personnellement" n'est pas claire. Est-ce que les étudiants inscrits automatiquement dans une bibliothèque universitaire ont répondu "oui" ou "non" à cette question ? Car ils ne sont pas personnellement inscrits mais sont inscrits par l'université. Par manque de précision, certains répondant ont pu répondre "oui" ou "non" en fonction de leur interprétation de la question posée.

Enfin, nous pouvons conclure en avançant que la plupart des résultats s'expliquent par une logique structurelle : les ouvriers vont moins à la bibliothèque par manque de temps de loisir disponible comparé aux cadres. Les personnes disposant seulement du brevet ou d'un CAP n'ont pas la même socialisation que les diplomés

du supérieur, qui de part leur capital culturel et social sont plus enclin à être socialiser à fréquenter une bibliothèque. Sur la différence d'inscription en fonction de la taille de la ville, il est évident que les grandes villes sont dôtées de beaucoup plus de bibliothèques que les petites ou villages, dans lesquels le service public se désinvestit de plus en plus. Les habitants de petites villes ont une possibilité moindre de se rendre en bibliothèque par rapport aux habitants de grandes villes. Ainsi, ne pas préciser ces informations dans le jeu de donnée empêche, selon nous , d'engager des politiques d'action pertinentes pour palier à ces problèmes. Selon nous, ce sont des politiques structurelles qui doivent être mises en place et non des simples actions au niveau des bibliothèques. La dépolitisation du jeu de données empêche donc d'arriver à des conclusions pertinentes.