Etude des salaires des étudiants de Polytech Montpellier à la sortie des études

HANNOU Assia GOYON Nicolas PHILIPPE Tom

lien de l'application web interactive: https://projetds.shinyapps.io/ProjetDataScience/

Introduction

Notre projet de data science, centré sur les ingénieurs diplômés de Polytech Montpellier, vise à décortiquer l'évolution de leur insertion professionnelle, en mettant un accent particulier sur la dynamique salariale à 6, 18 et 30 mois après l'obtention du diplôme. En s'appuyant sur une riche collection de données compilées par l'école, cette étude cherche à élucider les facteurs déterminants qui influencent les trajectoires professionnelles et salariales post-diplôme.

L'objectif principal est d'identifier et d'analyser les paramètres qui ont un impact significatif sur les salaires des jeunes diplômés. Pour y parvenir, notre analyse s'étendra sur diverses variables, incluant les spécialités d'études, les zones géographiques d'exercice (tant en France qu'à l'international), l'année d'obtention du diplôme, le secteur d'activité (privé/public), la nature des entreprises employeuses et le genre des diplômés.

Notre approche méthodologique repose sur des études de corrélation annuelles, enrichies par une exploration approfondie des facteurs identifiés comme influents. L'ambition est de dévoiler les éléments clés qui déterminent le salaire des ingénieurs issus de Polytech Montpellier, fournissant ainsi des insights précieux pour les étudiants, l'école, et les employeurs potentiels.

Librairies

```
if (!requireNamespace("ggplot2", quietly = TRUE)) {
  install.packages("ggplot2")
}
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.2.3
```

```
if(!requireNamespace("FactoMineR", quietly = TRUE)){
  install.packages("FactoMineR")
}
```

Etude données 2018

Récupération des données

```
# Trying ISO-8859-1
raw_data <- read.csv("data/data_2018.csv", sep = ";")
```

Définir toutes les données comme caractère (variable qualitative) sauf le salaire

```
# Lists specifying which columns to convert to factors and numeric
factor_cols <- c("Date", "Identifiant", "Genre", "Annee_diplome","Type_formation","Filiere","Situation"
```

```r
numeric_cols <- c("Anciennete", "Salaire_annuel", "Responsabilite_hierarchique", "Responsabilite_budget
```

```r
# Convert columns to factors
raw_data[factor_cols] <- lapply(raw_data[factor_cols], as.factor)

# Convert columns to numeric
raw_data[numeric_cols] <- lapply(raw_data[numeric_cols], as.numeric)
```

```
## Warning in lapply(raw_data[numeric_cols], as.numeric): NAs introduced by
## coercion

## Warning in lapply(raw_data[numeric_cols], as.numeric): NAs introduced by
## coercion

## Warning in lapply(raw_data[numeric_cols], as.numeric): NAs introduced by
## coercion

## Warning in lapply(raw_data[numeric_cols], as.numeric): NAs introduced by
## coercion
```

```r
summary(raw_data)
```

Avec ces valeurs on peut déduire beaucoup de choses...

Fonction de filtre

Ceci est une fonction utilisé après pour retirer différentes lignes en fonction des valeurs dans une certain colonne. (équivalent d'un select where)

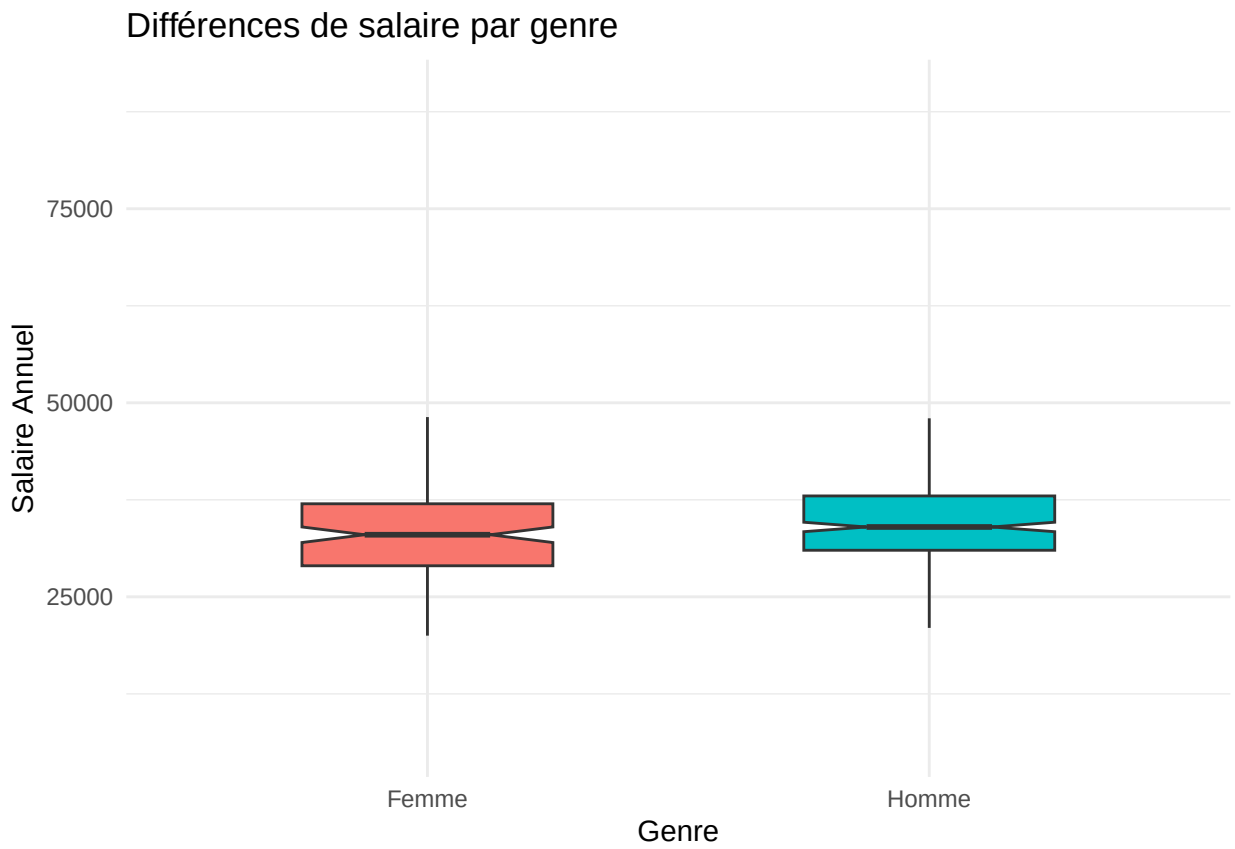
```r
remove_rows_by_value <- function(data, column_name, value_to_remove) {
  data_filtered <- subset(data, !(data[[column_name]] == value_to_remove))
  return(data_filtered)
}
```

Première étude

Disparité homme/femme - toute formation confondue

```r
ggplot(raw_data, aes(x = Genre, y = Salaire_annuel, fill = Genre)) +
  geom_boxplot(width = 0.5, notch = TRUE, outlier.shape = NA) +
  labs(title = "Différences de salaire par genre",
       x = "Genre",
       y = "Salaire Annuel") +
  theme_minimal() +
  theme(legend.position = "none")
```

```
## Warning: Removed 235 rows containing non-finite values (`stat_boxplot()`).
```



Le graphique intitulé “Différences de salaire par genre” représente un boxplot classique, qui met en évidence la distribution des salaires annuels entre les femmes et les hommes. À première vue, il suggère une légère disparité salariale entre les genres, avec les médianes indiquant que les hommes ont tendance à gagner plus que les femmes dans cet échantillon.

Les médianes de chaque groupe — représentées par la ligne centrale des boîtes — sont cruciales pour cette observation. La médiane pour le groupe des femmes semble être inférieure à celle du groupe des hommes. Il est important de noter que la médiane est souvent préférée à la moyenne pour une telle analyse, car elle est moins sensible aux valeurs extrêmes qui pourraient fausser les résultats.

En examinant la taille des boîtes, qui illustrent l’écart interquartile, nous observons une similitude dans la dispersion des salaires entre les deux groupes. Cela signifie que la moitié centrale des salaires s’étend sur une plage similaire pour les deux genres, suggérant que, mis à part la médiane, les distributions des salaires sont relativement comparables.

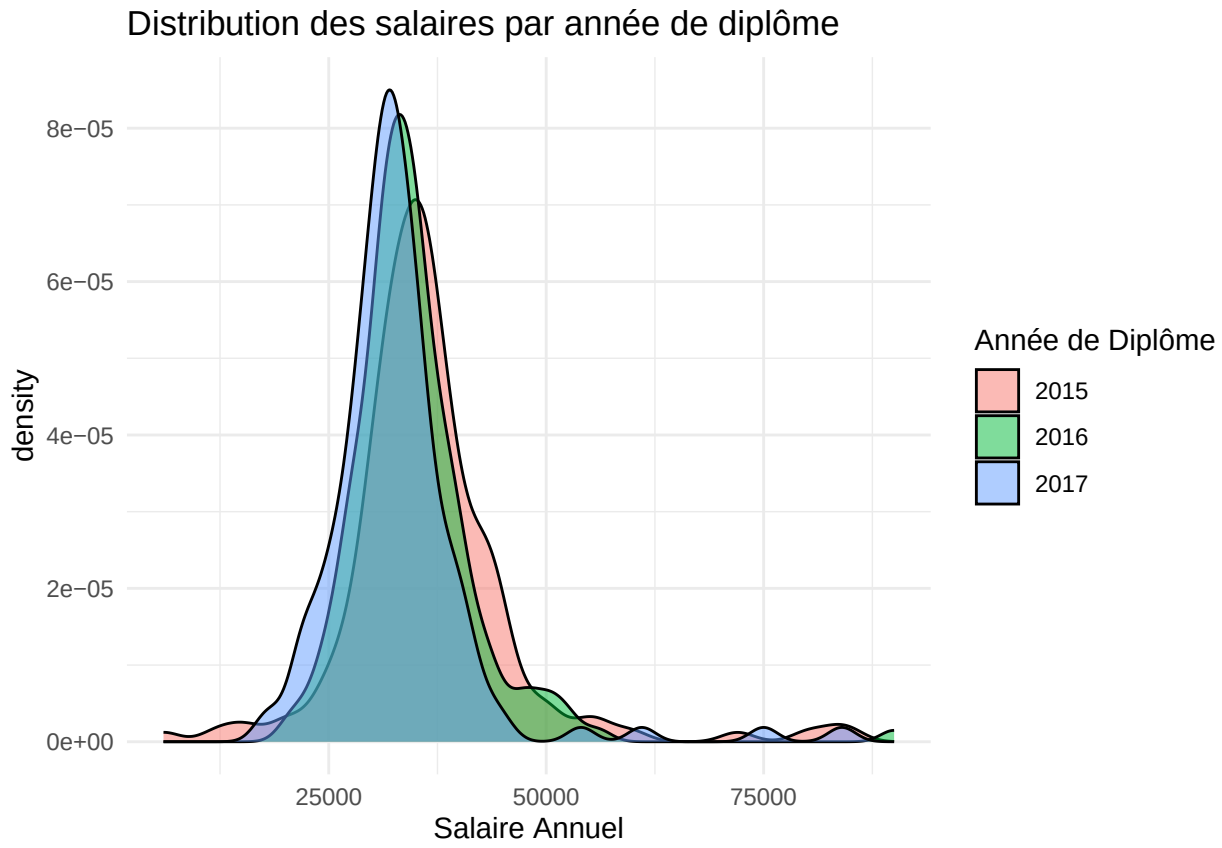
En conclusion, ce graphique révèle une tendance où les hommes semblent avoir un avantage salarial par rapport aux femmes dans l’échantillon étudié. Néanmoins, la similarité des distributions suggère que les écarts de salaires, au-delà de la médiane, ne sont pas marqués par des différences extrêmes. Pour une analyse complète, il serait judicieux de prendre en compte d’autres variables qui pourraient influencer ces résultats et de réaliser des tests statistiques pour évaluer la significativité des différences observées.

Disparité sur les dates d’optention du diplôme

```
ggplot(raw_data, aes(x = Salaire_annuel, fill = as.factor(Annee_diplome))) +
  geom_density(alpha = 0.5) +
  labs(title = "Distribution des salaires par année de diplôme",
       x = "Salaire Annuel",
```

```
      fill = "Année de Diplôme") +
  theme_minimal()
```

```
## Warning: Removed 235 rows containing non-finite values (`stat_density()`).
```



Le graphique en question, “Distribution des salaires par année de diplôme”, met en lumière les nuances subtiles des trajectoires salariales des diplômés de Polytech Montpellier, sur une période couvrant trois années distinctes. Observons d’abord que les courbes de densité pour les années 2016 et 2017 montrent une forte concentration autour d’une valeur modale, suggérant que la plupart des diplômés de ces années se regroupent autour d’un salaire annuel commun. Ce phénomène pourrait indiquer une certaine stabilité dans les conditions du marché de l’emploi ou dans les politiques de rémunération qui prévalent peu de temps après l’obtention du diplôme.

En revanche, la distribution de 2015 se caractérise par une étendue plus large et une densité plus faible. Ce profil suggère une hétérogénéité salariale plus marquée, qui pourrait être le résultat d’évolutions de carrière diversifiées parmi les diplômés. Après quelques années dans le monde professionnel, il est courant que les diplômés connaissent des chemins différents, influencés par une variété de facteurs tels que les progressions de carrière individuelles, les transitions sectorielles, les formations complémentaires ou les opportunités internationales.

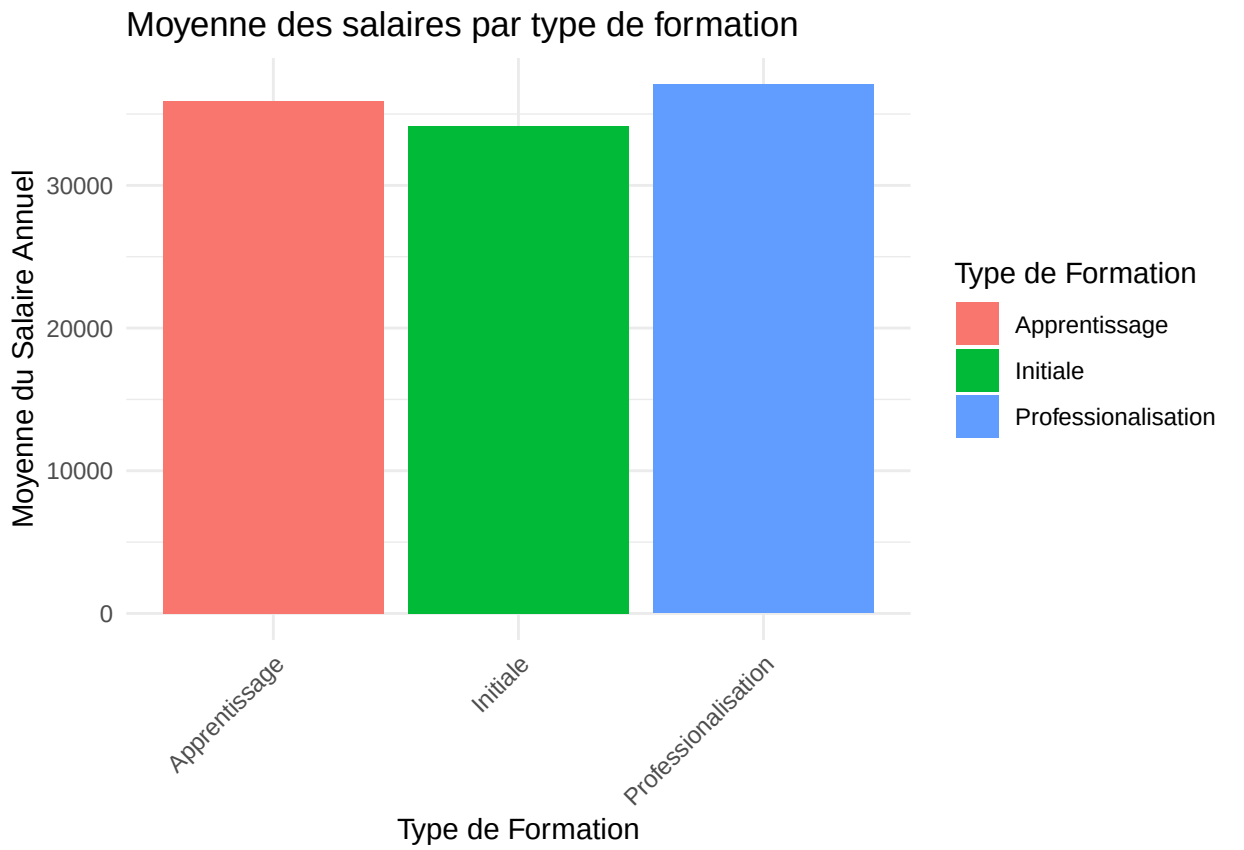
Cette variabilité accrue pour la cohorte de 2015 pourrait également refléter le fait que, avec le temps, certains diplômés bénéficient d’augmentations salariales grâce à des promotions ou des changements d’emploi qui reconnaissent leur expérience et leurs compétences accrues. Parallèlement, d’autres peuvent choisir des voies qui privilégient l’équilibre travail-vie personnelle ou l’engagement social, ce qui pourrait se traduire par des salaires moins élevés. De plus, il est possible que les secteurs d’activité qui embauchaient ces diplômés aient subi des transformations économiques, avec des secteurs en croissance offrant des rémunérations supérieures et d’autres, en décroissance ou en mutation, exerçant une pression à la baisse sur les salaires.

L'analyse de ces distributions révèle donc non seulement les conditions initiales du marché de l'emploi pour les nouveaux diplômés, mais aussi l'impact à long terme des décisions professionnelles, des opportunités de marché et des trajectoires personnelles sur les résultats financiers. Cela souligne l'importance pour les institutions éducatives et les étudiants de considérer les tendances à long terme et les facteurs d'influence sur les salaires, au-delà des premières années suivant la graduation.

Disparité formation initiale/apprentissage/contrat de professionalisation

```
library(ggplot2)

# Supprimer les lignes où 'Type_formation' est NA ou vide
clean_data <- raw_data[!is.na(raw_data$Type_formation) & raw_data$Type_formation != "", ]

# Créer le graphique avec les données nettoyées
ggplot(clean_data, aes(x = as.factor(Type_formation), y = Salaire_annuel, fill = as.factor(Type_formati
  stat_summary(fun = "mean", geom = "col") +
  labs(title = "Moyenne des salaires par type de formation",
       x = "Type de Formation",
       y = "Moyenne du Salaire Annuel",
       fill = "Type de Formation") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
## Warning: Removed 233 rows containing non-finite values (`stat_summary()`).
```



Le graphique "Moyenne des salaires par type de formation" compare les salaires moyens annuels obtenus à travers trois voies d'éducation différentes : l'apprentissage, la formation initiale et la professionnalisation.

La barre représentant l'apprentissage, en rouge, est la plus élevée, ce qui suggère que les individus qui ont suivi ce parcours bénéficient en moyenne de salaires annuels supérieurs. Cette observation pourrait indiquer que les programmes d'apprentissage, qui allient formation pratique en entreprise et enseignement théorique, sont bien valorisés par le marché du travail et peuvent mener à des emplois bien rémunérés.

La barre verte, correspondant à la formation initiale, traditionnellement considérée comme l'achèvement d'études avant l'entrée sur le marché du travail, montre des salaires moyens moins élevés que ceux de l'apprentissage. Cela pourrait refléter une tendance du marché de l'emploi à privilégier l'expérience pratique ou des spécificités liées aux secteurs d'emploi des apprentis.

La professionnalisation, indiquée par la barre bleue, présente des salaires moyens qui se situent entre ceux de l'apprentissage et de la formation initiale. Les parcours de professionnalisation, qui sont souvent ciblés vers un métier spécifique et incluent une composante de formation en milieu professionnel, semblent offrir des opportunités salariales intermédiaires.

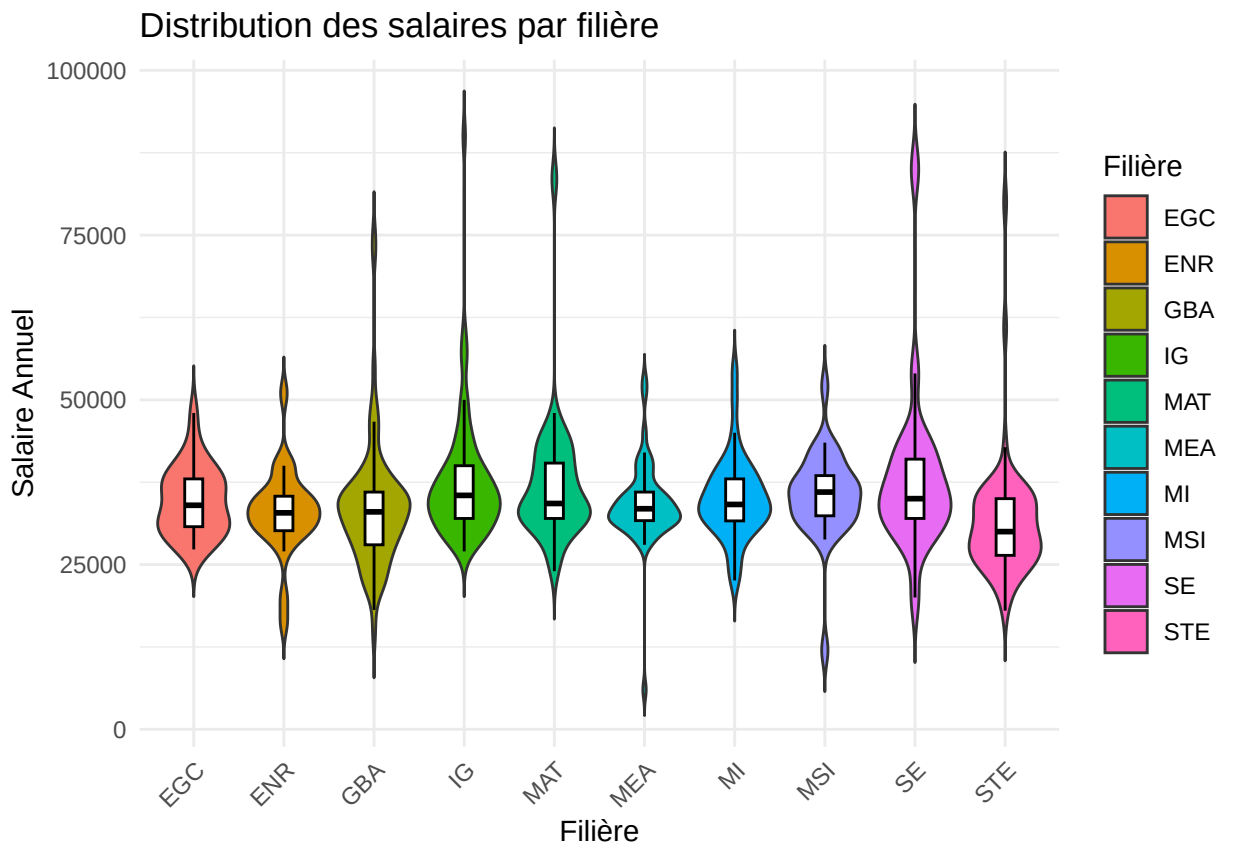
Cette disposition des moyennes salariales peut suggérer une évolution ou un changement dans la perception et la récompense des différents types de formation par le marché du travail. Les employeurs peuvent reconnaître la valeur ajoutée par l'apprentissage, qui combine expérience professionnelle et éducation, et être prêts à offrir des salaires plus attractifs pour ces compétences.

Disparité des filières

```
ggplot(raw_data, aes(x = Filiere, y = Salaire_annuel, fill = Filiere)) +
  geom_violin(trim = FALSE, scale = "width", width = 0.8) +
  geom_boxplot(width = 0.2, fill = "white", color = "black", outlier.shape = NA) +
  labs(title = "Distribution des salaires par filière",
       x = "Filière",
       y = "Salaire Annuel",
       fill = "Filière") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
## Warning: Removed 235 rows containing non-finite values (`stat_ydensity()`).
```

```
## Warning: Removed 235 rows containing non-finite values (`stat_boxplot()`).
```



Le diagramme en violons offre une représentation visuelle de la distribution des salaires annuels pour différentes filières. Chaque violon combine les caractéristiques d'une boîte à moustaches et d'un diagramme de densité de probabilité. La largeur d'un violon à un niveau de salaire donné est proportionnelle au nombre d'observations (salariés) pour ce salaire, ce qui donne une indication de la densité des données à ce niveau.

Analyse des médianes par filière : - Les filières IG et MSI présentent des médianes élevées, ce qui suggère que les salaires moyens dans ces domaines sont supérieurs à ceux des autres filières. - Les filières MI et SE suivent de près avec des médianes légèrement inférieures, indiquant que les salaires moyens sont également relativement élevés dans ces secteurs. - Les filières STE et EGC ont des médianes inférieures, ce qui indique que les salaires moyens sont plus bas dans ces domaines par rapport aux autres filières présentées.

Analyse de la distribution par filière : - Les violons pour les filières IG, MI, GBA, STE sont relativement larges en haut, indiquant une plus grande variabilité des salaires élevés et la présence de salaires très hauts comparativement aux autres filières. - Pour les filières MSI et SE, les violons sont également larges, mais avec moins d'observations aux salaires les plus élevés, ce qui peut suggérer une répartition plus homogène des salaires autour de la médiane. - Les filières STE et EGC présentent des violons plus étroits, particulièrement en haut de la distribution, ce qui pourrait signifier une concentration des salaires autour de la médiane avec moins de salariés recevant des salaires très élevés.

Les points en dehors des violons représentent les valeurs extrêmes, qui sont significativement plus élevées ou plus basses que la majorité des salaires dans la filière.

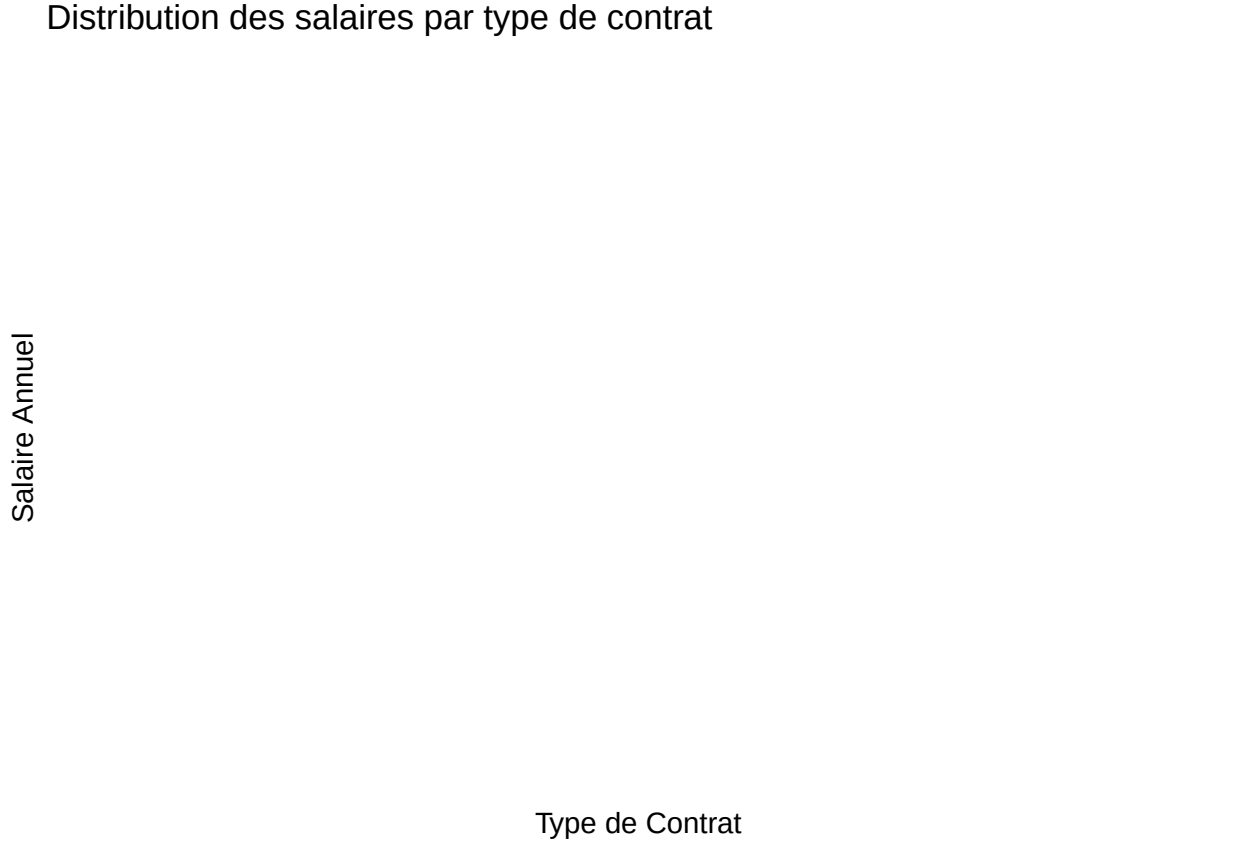
En somme, ce diagramme montre que certaines filières comme IG et MSI sont associées à des salaires plus élevés et une plus grande hétérogénéité des rémunérations, tandis que d'autres comme STE et EGC sont caractérisées par des salaires plus homogènes et généralement plus modestes. Ces informations peuvent être utiles pour les individus évaluant les perspectives de carrière dans chaque domaine, ainsi que pour les employeurs et les responsables de l'éducation lors de l'élaboration de stratégies de formation et de recrutement.

Filtrage des personnes en activité

```
filtered_data <- remove_rows_by_value(raw_data, "Situation", "En recherche emploi")
filtered_data <- remove_rows_by_value(filtered_data, "Situation", "En poursuite d'études (hors thèse) /
filtered_data <- remove_rows_by_value(filtered_data, "Situation", "These")
filtered_data <- remove_rows_by_value(filtered_data, "Situation", "Sans activité")
filtered_data <- remove_rows_by_value(filtered_data, "Situation", "Volontariat")
filtered_data <- remove_rows_by_value(filtered_data, "Situation", "En création d'entreprise /reprise d'
filtered_data <- remove_rows_by_value(filtered_data, "Situation", "Etudes")

activite_data <- filtered_data
```

Disparité par nature du contrat

```
library(ggplot2)

# Filtrage des données en excluant certaines catégories et en supprimant les lignes vides
filtered_data <- subset(raw_data, !Nature_contrat %in% c("Apprentissage", "Autre", "ile-de-France", ""))
filtered_data <- na.omit(filtered_data)

# Supposons que les catégories de contrats sont 'CDD', 'CDI', 'Intérim', etc.
# Vous pouvez définir manuellement les couleurs pour chaque catégorie de contrat
couleurs_contrats <- c("CDD" = "#66c2a5", "CDI" = "#fc8d62", "Intérim" = "#8da0cb",
                       "Stage" = "#e78ac3", "Alternance" = "#a6d854", "CTT" = "#e78ac3", "Contrat local

# Diagramme en boîte avec couleurs
ggplot(filtered_data, aes(x = Nature_contrat, y = Salaire_annuel, fill = Nature_contrat)) +
  geom_boxplot() +
  scale_fill_manual(values = couleurs_contrats) +  # Utiliser les couleurs définies précédemment
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 8)) +
  labs(title = "Distribution des salaires par type de contrat", x = "Type de Contrat", y = "Salaire Annu
```



Le graphique “Distribution des salaires par type de contrat” présente les salaires annuels en fonction du type de contrat de travail, à travers des boxplots qui indiquent la médiane, les quartiles et la variabilité des salaires.

Le boxplot pour les contrats à durée déterminée (CDD), en vert, montre une médiane de salaire relativement basse comparée aux autres types de contrats, avec une distribution étendue vers le bas, indiquant que de nombreux salariés en CDD ont des salaires inférieurs à la médiane.

Les contrats à durée indéterminée (CDI), colorés en orange, ont une médiane supérieure à celle des CDD, ce qui est cohérent avec la nature plus stable et pérenne des CDI. Bien que la répartition soit similaire à celle des CDD, avec des valeurs basses, il y a moins de dispersion vers les salaires inférieurs, suggérant une concentration plus forte autour de la médiane.

Les contrats d’expatriation, en bleu, affichent une médiane significativement plus élevée que les CDD et les CDI, reflétant la tendance des entreprises à offrir des packages de rémunération plus compétitifs pour attirer des employés à travailler à l’étranger. Cependant, la répartition est large, indiquant que pour certains expatriés, les salaires peuvent être extrêmement élevés.

Les contrats locaux, en rose, présentent une médiane plus basse que les contrats d’expatriation mais toujours plus élevée que les CDD et semblable aux CDI, suggérant que les salaires peuvent varier grandement en fonction du pays et du marché local du travail.

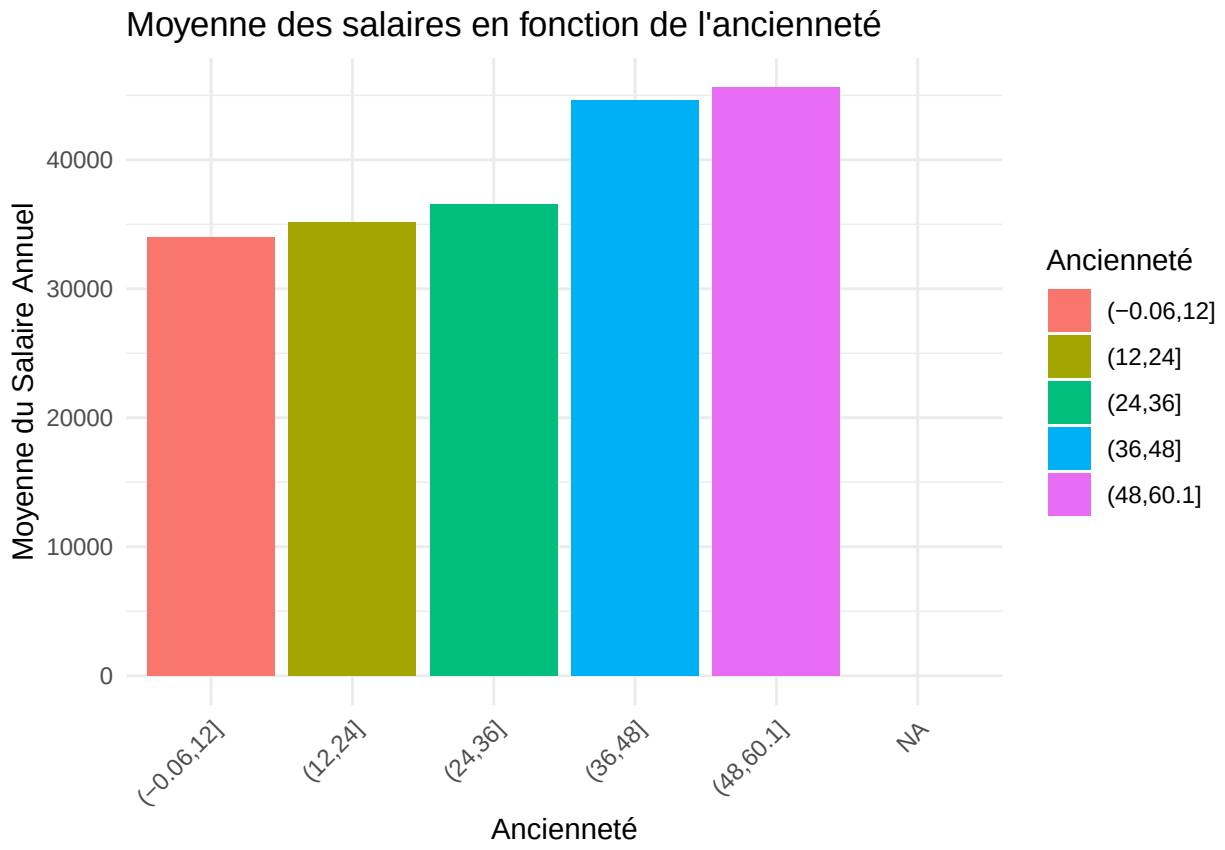
Enfin, les contrats de travail temporaire (CTT), en violet, ont la médiane la plus basse et la plus petite dispersion des salaires, ce qui indique que ces postes offrent généralement des salaires plus modestes et sont moins susceptibles de conduire à des salaires très élevés.

Globalement, ce graphique met en lumière les différences substantielles dans les salaires moyens associés à chaque type de contrat, avec des contrats d’expatriation offrant les salaires les plus élevés en moyenne et les CTT les plus bas, tandis que les CDI et les contrats locaux se situent dans une gamme intermédiaire.

Distribution du salaire en fonction de l'ancienneté

```
library(ggplot2)

ordered_data <- raw_data[order(as.numeric(raw_data$Anciennete)), ]

# Création de catégories d'ancienneté (facultatif)
ordered_data$Anciennete_category <- cut(as.numeric(ordered_data$Anciennete), breaks = 5)

# Diagramme en barres pour la moyenne des salaires en fonction de l'ancienneté
ggplot(ordered_data, aes(x = Anciennete_category, y = Salaire_annuel, fill = Anciennete_category)) +
  geom_bar(stat = "summary", fun = "mean", position = "dodge") +
  labs(title = "Moyenne des salaires en fonction de l'ancienneté",
       x = "Ancienneté",
       y = "Moyenne du Salaire Annuel",
       fill = "Ancienneté") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
## Warning: Removed 235 rows containing non-finite values (`stat_summary()`).
```



Le graphique "Moyenne des salaires en fonction de l'ancienneté" illustre comment l'expérience professionnelle accumulée influence la rémunération moyenne des employés. L'ancienneté, divisée en segments d'un an, sert de mesure pour établir une corrélation potentielle entre le temps passé dans une organisation ou un secteur et le salaire moyen perçu.

Les employés avec moins d'un an d'expérience, représentés par la barre rouge, perçoivent les salaires les plus

bas de l'échelle, ce qui est attendu pour des positions souvent associées à des rôles débutants ou en phase d'apprentissage. À mesure que l'ancienneté augmente, les barres s'élèvent progressivement : la tranche d'un à deux ans (verte) montre une augmentation modeste du salaire, suggérant un premier palier d'évolution salariale.

Ce phénomène de croissance se poursuit de manière plus marquée pour les tranches de deux à trois ans (bleue) et de trois à quatre ans (violet), où l'on observe un bond significatif dans les salaires moyens. Cela peut refléter les promotions, les augmentations méritées ou l'acquisition de compétences spécialisées qui sont souvent récompensées par des ajustements salariaux positifs.

Curieusement, la tranche des employés les plus anciens, ceux ayant entre quatre et cinq ans d'expérience (rose), bien qu'affichant un salaire moyen élevé, ne représente pas le sommet de l'échelle salariale. Ceci peut indiquer une stabilisation des salaires ou peut-être une saturation dans certaines industries où l'expérience supplémentaire au-delà d'un certain point n'entraîne pas nécessairement une augmentation significative de la rémunération.

En synthèse, ce graphique démontre un lien apparent entre l'ancienneté et le salaire moyen, soulignant l'importance de l'expérience dans la progression de carrière. Il suggère que l'accumulation d'années de service peut être un facteur déterminant dans l'évolution des salaires, bien que l'impact de l'ancienneté puisse plafonner après un certain temps.

###Distribution du salaire selon le secteur

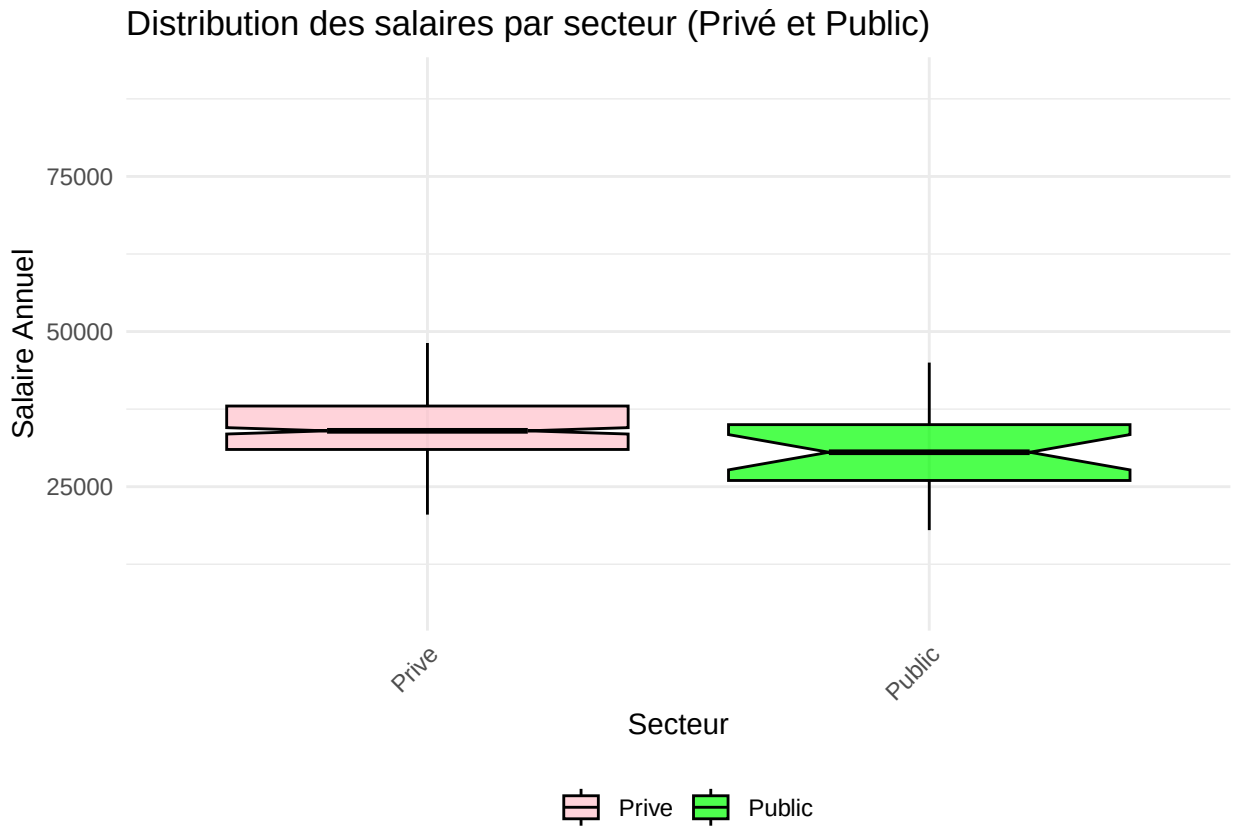
```r
library(ggplot2)

# Supposons que 'Secteur' est la colonne qui indique si une personne travaille dans le secteur privé ou
# Nous allons exclure les non-salariés qui, dans cet exemple, sont marqués comme 'Non salarie(e)' dans

# Vérifiez toutes les valeurs uniques pour identifier les valeurs non désirées
unique(raw_data$Secteur)
```

```
## [1] Prive                          Public         Non salarie(e)
## Levels:  Non salarie(e) Prive Public
```

```r
# Filtrez le DataFrame pour ne garder que les lignes avec "Privé" et "Public"
filtered_data <- raw_data[raw_data$Secteur %in% c("Prive", "Public"), ]

# Continuez avec la création du graphique ggplot
ggplot(filtered_data, aes(x = Secteur, y = Salaire_annuel, fill = Secteur)) +
  geom_boxplot(
    width = 0.8,
    notch = TRUE,
    outlier.shape = NA,
    color = "black",
    alpha = 0.7
  ) +
  scale_fill_manual(values = c("Prive" = "pink", "Public" = "green")) +
  labs(
    title = "Distribution des salaires par secteur (Privé et Public)",
    x = "Secteur",
    y = "Salaire Annuel"
  ) +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1),
    legend.position = "bottom",
    legend.title = element_blank()
```

```
)
```

```
## Warning: Removed 77 rows containing non-finite values (`stat_boxplot()`).
```



Le graphique “Distribution des salaires par secteur (Privé et Public)” présente deux boxplots qui comparent les distributions des salaires annuels dans les secteurs privé et public. Les boxplots offrent une vue d’ensemble des médianes, des quartiles et des étendues des salaires au sein de chaque secteur.

Le boxplot du secteur privé, coloré en rose, a une médiane qui se situe au-dessus de la moitié inférieure de la boîte, indiquant que la médiane des salaires est relativement élevée. Cependant, la dispersion des salaires, illustrée par la longueur des moustaches, est modeste, ce qui suggère une variabilité moins importante des salaires dans le secteur privé.

D’autre part, le boxplot du secteur public, en vert, montre une médiane plus basse que celle du secteur privé, indiquant que les salaires médians sont inférieurs dans le secteur public. Néanmoins, l’étendue de la distribution des salaires est plus large dans le secteur public, comme le montre l’amplitude des moustaches, ce qui indique une plus grande variété dans les salaires des employés du secteur public.

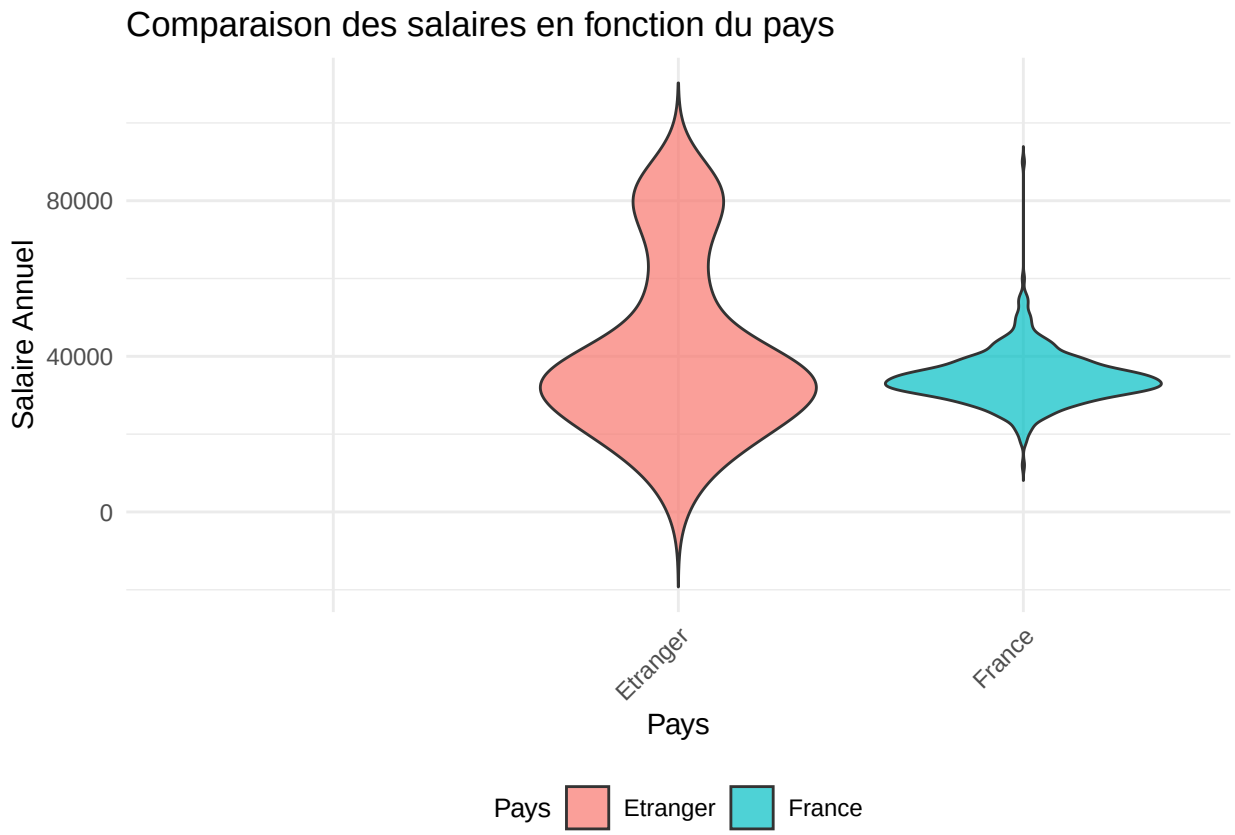
La présence de valeurs extrêmes ou aberrantes n’est pas notable dans ce graphique, ce qui indique que la majorité des salaires se situent dans une gamme relativement prévisible pour les deux secteurs.

En conclusion, ce graphique suggère que bien que le secteur privé puisse offrir en moyenne des salaires plus élevés, le secteur public présente une plus grande hétérogénéité dans la rémunération de ses employés. Cela peut refléter une diversité des rôles et des niveaux de responsabilité plus marquée dans le secteur public ou des politiques salariales différentes.

Distribution du salaire en fonction de France/étranger

```
ggplot(raw_data, aes(x = factor(France), y = Salaire_annuel, fill = factor(France))) +
  geom_violin(trim = FALSE, scale = "width", width = 0.8, alpha = 0.7) +
  labs(
    title = "Comparaison des salaires en fonction du pays",
    x = "Pays",
    y = "Salaire Annuel",
    fill = "Pays"
  ) +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1),
    legend.position = "bottom",
    legend.title = element_text(size = 10)
  )
```

```
## Warning: Removed 235 rows containing non-finite values (`stat_ydensity()`).
```



Le graphique “Comparaison des salaires en fonction du pays” présente une analyse visuelle de la distribution des salaires annuels entre la France et l’étranger à l’aide de graphiques en violon. Ces derniers illustrent non seulement la médiane et la gamme des salaires, mais aussi leur densité à différents niveaux de rémunération.

Le violon rouge, représentant les salaires à l’étranger, est large et présente un sommet élevé, indiquant une concentration de salaires autour d’un niveau plus élevé que celui observé pour la France. La base plus large de ce violon suggère également que les salaires à l’étranger peuvent atteindre des niveaux significativement plus élevés que ceux en France.

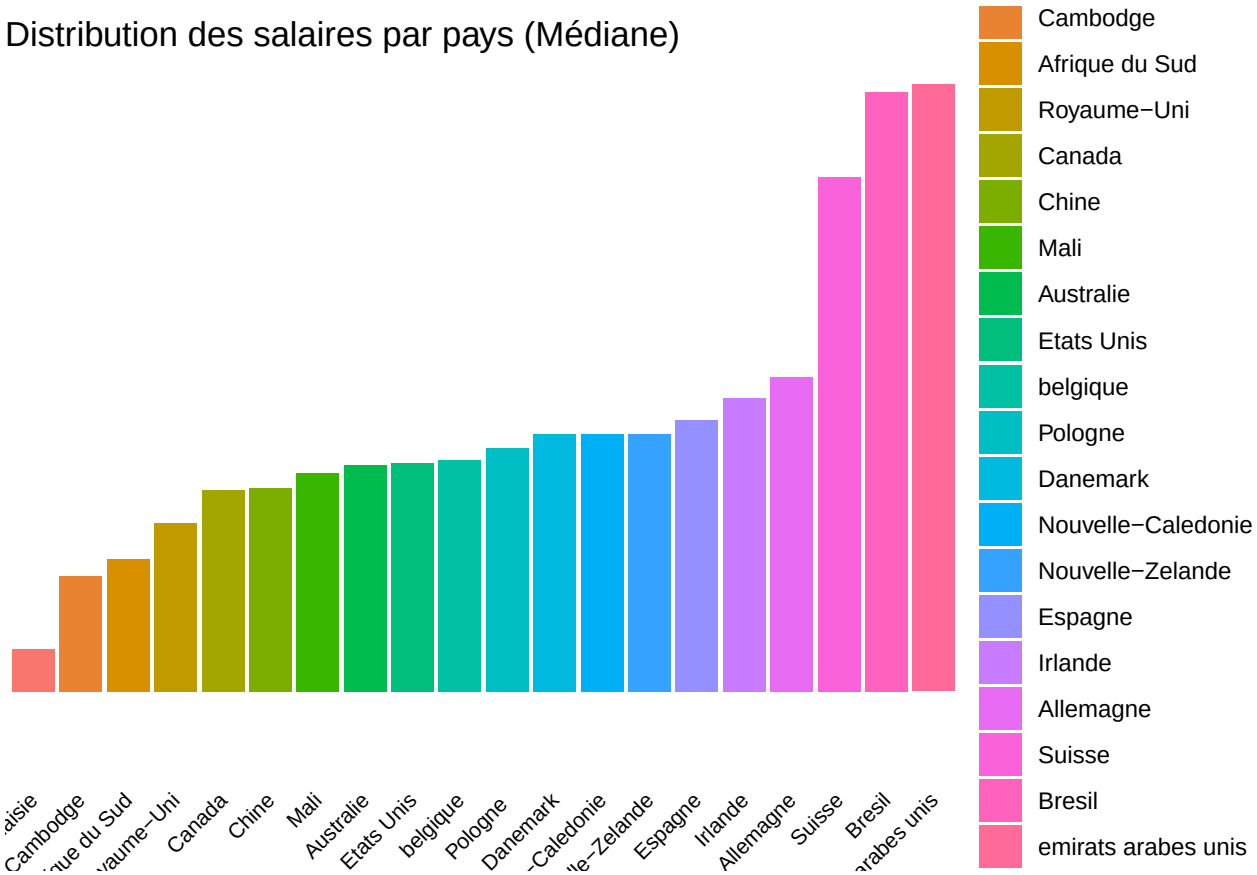
Le violon bleu-vert, correspondant à la France, montre une distribution plus étroite avec une médiane inférieure à celle de l'étranger. Cela suggère que les salaires en France sont généralement plus bas et moins variés que ceux observés à l'étranger. En outre, la concentration des données autour de la médiane est moins prononcée pour la France, comme l'indique la forme plus aiguisée du violon.

Cette visualisation suggère que travailler à l'étranger pourrait potentiellement offrir de meilleurs salaires que travailler en France. Cela peut être dû à une variété de facteurs, y compris des différences dans les normes du marché du travail, les niveaux de vie, ou les stratégies de rémunération des entreprises à l'international.

En résumé, la comparaison des distributions des salaires montre que l'expatriation pourrait être financièrement avantageuse, avec des salaires médians et des potentiels de rémunération supérieurs. Toutefois, il est important de considérer d'autres facteurs tels que le coût de la vie et les avantages sociaux, qui peuvent varier considérablement d'un pays à l'autre et influencer l'attractivité financière globale des opportunités d'emploi à l'étranger.

Distribution du salaire en fonction du pays à l'étranger

```r
library(ggplot2)

# Filtrage des données pour exclure les valeurs manquantes ou vides
raw_data_filtered <- raw_data[complete.cases(raw_data$Pays) & raw_data$Pays != "" & !grepl("^\\s*$", ra

# Calcul des salaires médians par pays
median_salaries <- tapply(raw_data_filtered$Salaire_annuel, raw_data_filtered$Pays, median)
median_order <- names(sort(median_salaries))

# Création d'un facteur ordonné pour les pays en fonction des salaires médians
emploiFactor <- factor(raw_data_filtered$Pays, levels = median_order)

# Création du graphique en barres coloré sans le fond gris
ggplot(raw_data_filtered, aes(x = emploiFactor, y = Salaire_annuel, fill = emploiFactor)) +
  geom_bar(stat = "summary", fun = "median") +
  theme_void() +  # Supprime le fond gris
  theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 8)) +
  labs(title = "Distribution des salaires par pays (Médiane)",
       y = "Salaire Médian",
       fill = "Pays")
```



Le graphique “Médiane des Salaires par Pays” illustre la médiane des salaires annuels exprimés en euros pour une sélection de pays, allant de l’Afrique du Sud à l’Émirats arabes unis. Chaque barre, colorée différemment pour représenter un pays spécifique, indique la médiane salariale dans ce pays.

À partir de ce graphique, nous pouvons observer que :

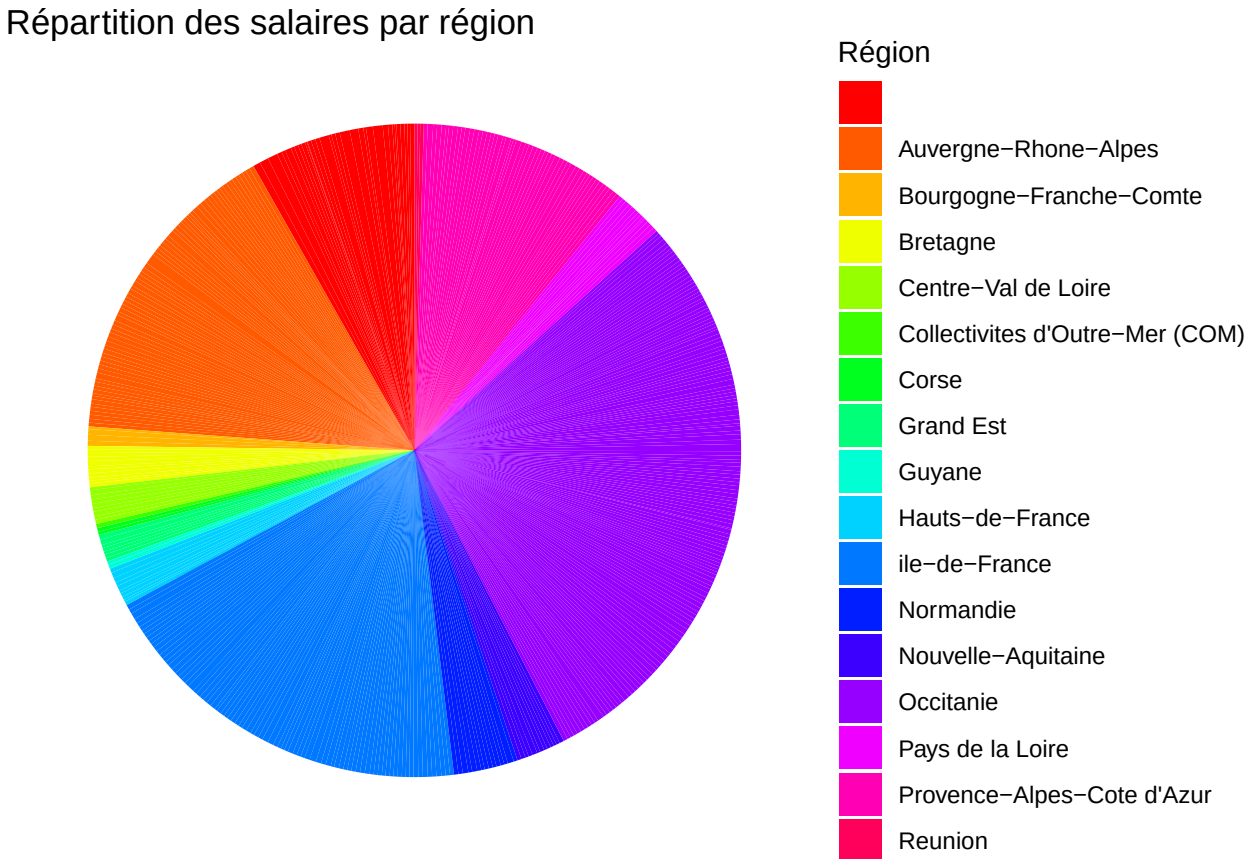
- Les Émirats arabes unis et le Brésil se distinguent avec la médiane salariale la plus élevée parmi les pays listés, ce qui pourrait refléter une économie riche en opportunités de haute rémunération.
- La Suisse suit, avec des médianes de salaire également élevées.
- Les pays de l’Union Européenne, tels que l’Allemagne, l’Irlande et la Belgique, présentent des médianes salariales moyennes.
- En comparaison, des pays comme la Cambodge et le Malaisie affichent des médianes plus basses, ce qui pourrait être lié à des différences dans le coût de la vie, les structures économiques ou les niveaux de développement industriel.

Il est important de noter que la médiane salariale ne reflète pas nécessairement la répartition des revenus au sein du pays ou la parité de pouvoir d’achat. Un salaire élevé dans un pays avec un coût de la vie élevé peut ne pas avoir la même valeur que le même salaire dans un pays où le coût de la vie est plus bas.

Ce graphique offre ainsi une perspective comparative utile pour comprendre les dynamiques économiques mondiales et peut éclairer les décisions des professionnels envisageant de travailler à l’étranger. Toutefois, une évaluation approfondie devrait également prendre en compte d’autres facteurs tels que la qualité de vie, les avantages sociaux, les impôts et le coût de la vie pour une compréhension complète du pouvoir d’achat.

Distribution des salaires en fonction de la région

```r
library(ggplot2)

# Remove rows with non-finite values in Salaire_annuel
raw_data <- raw_data[!is.na(raw_data$Salaire_annuel) & is.finite(raw_data$Salaire_annuel), ]

# Remove rows with non-finite values in other numeric columns if needed
#raw_data <- raw_data[complete.cases(raw_data[, numeric_cols]), ]

# Calculate average salary by region
avg_salary_by_region <- tapply(raw_data$Salaire_annuel, raw_data$Region, mean)
max_avg_salary_region <- names(which.max(avg_salary_by_region))

# Define a color palette for each region
color_palette <- rainbow(length(unique(raw_data$Region)))

# Plotting the distribution of salaries based on region as a pie chart
ggplot(raw_data, aes(x = "", y = Salaire_annuel, fill = Region)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar("y") +
  theme_void() +  # Removes unnecessary elements
  labs(title = "Répartition des salaires par région",
       fill = "Région",
       y = "Salaire Annuel") +
  scale_fill_manual(values = ifelse(raw_data$Region == max_avg_salary_region, "red", color_palette))  #
```



Le diagramme “Répartition des salaires par région” illustre les moyennes salariales annuelles à travers différentes régions de France, mettant en exergue des différences notables. L'Occitanie et l'Île-de-France se démarquent avec des moyennes salariales parmi les plus élevées, tandis que la Bourgogne-Franche-Comté et les Pays de la Loire affichent des moyennes inférieures.

Ces variations peuvent refléter des différences dans les marchés du travail régionaux et les secteurs d'activité prédominants. L'économie plus large et diversifiée de l'Île-de-France et de l'Occitanie pourrait contribuer à leurs moyennes salariales plus élevées. En revanche, les moyennes plus basses en Bourgogne-Franche-Comté et dans les Pays de la Loire pourraient être le reflet d'une concentration différente d'industries ou d'une demande variée en compétences spécifiques. Ces constats suggèrent des disparités économiques régionales à travers le pays.

Etude données 2019

Récupération des données

```
# Trying ISO-8859-1
raw_data <- read.csv("data/data_2019.csv", sep = ";")
```

Définir toutes les données comme caractère (variable qualitative) sauf le salaire

```
# Lists specifying which columns to convert to factors and numeric
factor_cols <- c("Date", "identifiant", "Genre", "Annee_diplome","Type_formation","Filiere","Situation"
numeric_cols <- c("Anciennete", "Salaire_annuel", "Responsabilite_hierarchique", "Responsabilite_budget

# Convert columns to factors
raw_data[factor_cols] <- lapply(raw_data[factor_cols], as.factor)

# Convert columns to numeric
raw_data[numeric_cols] <- lapply(raw_data[numeric_cols], as.numeric)
```

```
## Warning in lapply(raw_data[numeric_cols], as.numeric): NAs introduced by
## coercion

## Warning in lapply(raw_data[numeric_cols], as.numeric): NAs introduced by
## coercion

## Warning in lapply(raw_data[numeric_cols], as.numeric): NAs introduced by
## coercion

## Warning in lapply(raw_data[numeric_cols], as.numeric): NAs introduced by
## coercion
```

```
summary(raw_data)
```

Nous allons analyser le salaire en fonction des différents paramètres.

Fonction de filtre

Ceci est une fonction utilisé après pour retirer différentes lignes en fonction des valeurs dans une certain colonne. (équivalent d'un select where)

```
remove_rows_by_value <- function(data, column_name, value_to_remove) {
  data_filtered <- subset(data, !(data[[column_name]] == value_to_remove))
  return(data_filtered)
}
```

Première étude

Disparité homme/femme - toute formation confondue

```
ggplot(raw_data, aes(x = Genre, y = Salaire_annuel, fill = Genre)) +
  geom_boxplot(width = 0.5, notch = TRUE, outlier.shape = NA) +
  labs(title = "Différences de salaire par genre",
       x = "Genre",
       y = "Salaire Annuel") +
  theme_minimal() +
  theme(legend.position = "none")
```

```
## Warning: Removed 222 rows containing non-finite values (`stat_boxplot()`).
```

Le graphique intitulé “Différences de salaire par genre” présente deux boxplots qui comparent les salaires annuels entre les femmes et les hommes. Chaque boxplot montre la médiane des salaires, les quartiles et les variations au sein de chaque genre.

Du côté des femmes, le boxplot rouge indique une médiane qui se situe plutôt au centre de la boîte, avec des “moustaches” qui s’étendent pour montrer la gamme complète des salaires. La répartition semble relativement étroite, suggérant que la majorité des salaires féminins se concentre autour de la médiane sans variations extrêmes.

Pour les hommes, le boxplot cyan montre également une médiane bien définie, et l’étendue des salaires semble légèrement plus large que celle des femmes, comme indiqué par les moustaches plus longues. Cela pourrait impliquer que les salaires masculins varient plus et que certains hommes peuvent atteindre des salaires plus élevés par rapport à leurs homologues féminins.

La comparaison des deux boxplots révèle une médiane de salaire légèrement plus élevée pour les hommes que

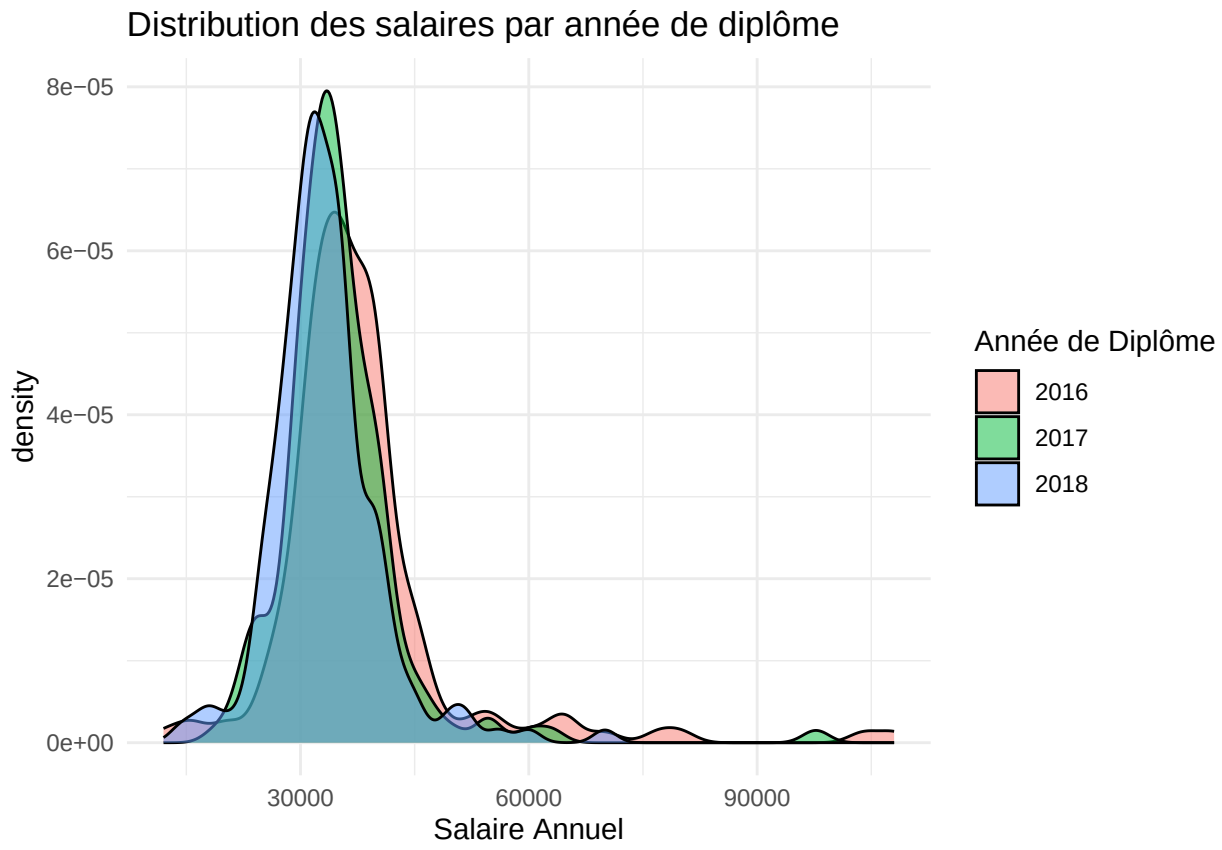
pour les femmes, ce qui soulève des questions sur l'égalité des salaires et la représentation des genres dans différents rôles ou niveaux de responsabilité au sein de l'économie.

Ce graphique illustre l'existence d'un écart salarial entre les genres dans le contexte étudié. Alors que les boxplots ne fournissent pas les raisons sous-jacentes à cet écart, ils signalent une différence qui pourrait être due à une variété de facteurs, y compris mais non limités à la discrimination de genre, les différences dans les choix de carrière, l'éducation, l'expérience professionnelle ou le temps de travail.

Disparité sur les dates d'optention du diplôme

```
ggplot(raw_data, aes(x = Salaire_annuel, fill = as.factor(Annee_diplome))) +
  geom_density(alpha = 0.5) +
  labs(title = "Distribution des salaires par année de diplôme",
       x = "Salaire Annuel",
       fill = "Année de Diplôme") +
  theme_minimal()
```

```
## Warning: Removed 222 rows containing non-finite values (`stat_density()`).
```



Le graphique "Distribution des salaires par année de diplôme" dépeint les distributions des salaires annuels pour trois cohortes de diplômés: 2016, 2017 et 2018. Les courbes de densité illustrent où les salaires se concentrent pour chaque groupe.

Pour les diplômés de 2018, la courbe de densité est centrée vers des salaires plus bas, indiquant que ceux qui sont le plus récemment entrés sur le marché du travail ont tendance à commencer avec des salaires inférieurs. Cela peut s'expliquer par le fait que les individus les plus récemment diplômés sont souvent au début de leur carrière et peuvent ne pas avoir encore bénéficié d'augmentations ou de promotions.

En remontant dans le temps, les courbes des années 2017 et 2016 démontrent que les pics se déplacent progressivement vers des salaires plus élevés. Cette tendance suggère que, avec le temps, les diplômés gagnent en expérience et progressent dans leur carrière, ce qui se traduit par des salaires plus élevés. Ce phénomène reflète une évolution rapide dans les premières années suivant l'obtention du diplôme, où les augmentations de salaire peuvent être plus fréquentes et substantielles.

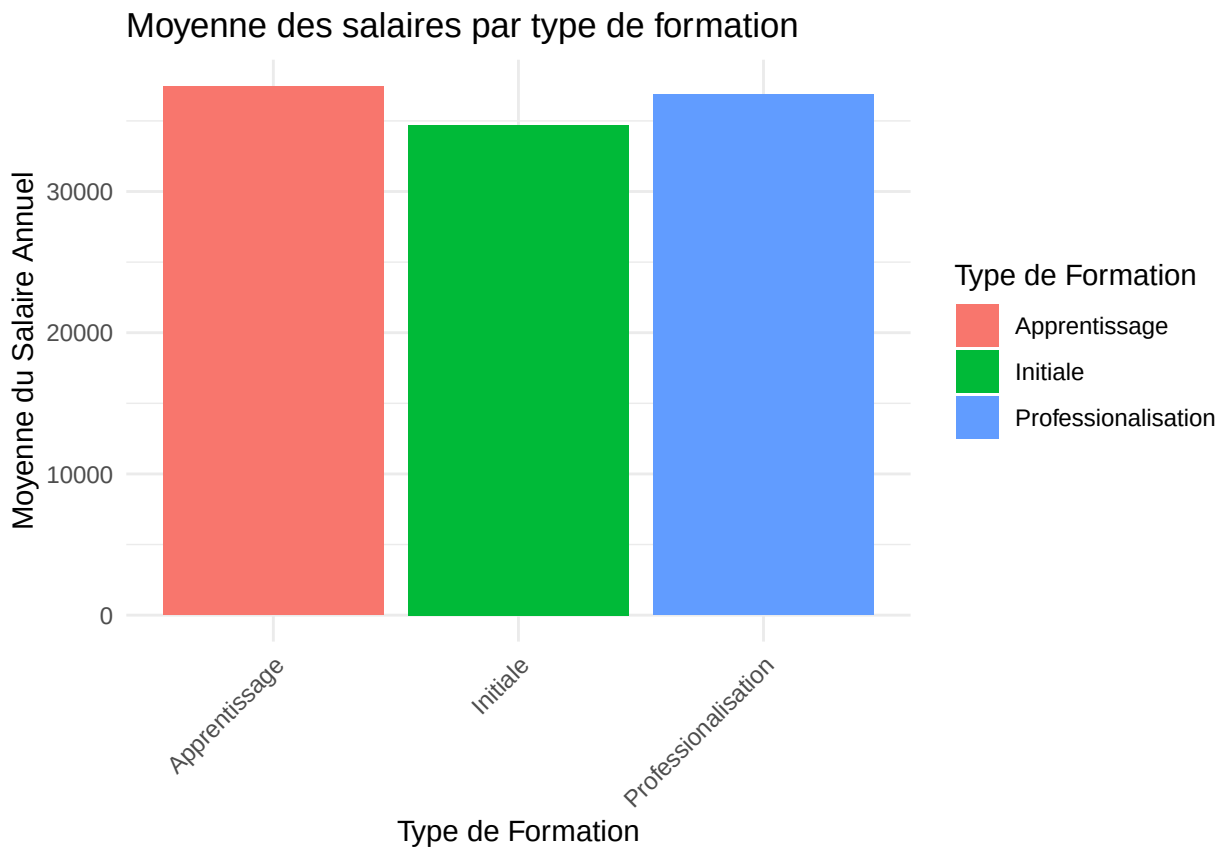
De plus, d'année en année, les courbes montrent des pics moins aigus et sont plus étalées, indiquant une distribution des salaires plus semblable parmi les diplômés. Pour les diplômés de 2018, la courbe est relativement pointue, ce qui suggère une grande variabilité des salaires au sein de cette cohorte, avec moins de concentration autour d'un salaire typique ou médian. Cela pourrait signifier que le marché du travail est devenu plus dynamique ou incertain, avec des opportunités et des trajectoires professionnelles variées qui se traduisent par une variabilité accrue des salaires.

En résumé, cette analyse suggère qu'avec le temps, les diplômés tendent à voir une amélioration de leur situation salariale et que le marché du travail offre une diversité croissante de rémunérations à mesure que les diplômés s'établissent et avancent dans leurs carrières respectives.

Disparité formation initiale/apprentissage/contrat de professionalisation

```
library(ggplot2)

# Supprimer les lignes où 'Type_formation' est NA ou vide
clean_data <- raw_data[!is.na(raw_data$Type_formation) & raw_data$Type_formation != "", ]

# Créer le graphique avec les données nettoyées
ggplot(clean_data, aes(x = as.factor(Type_formation), y = Salaire_annuel, fill = as.factor(Type_formati
  stat_summary(fun = "mean", geom = "col") +
  labs(title = "Moyenne des salaires par type de formation",
       x = "Type de Formation",
       y = "Moyenne du Salaire Annuel",
       fill = "Type de Formation") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
## Warning: Removed 222 rows containing non-finite values (`stat_summary()`).
```



Le graphique “Moyenne des salaires par type de formation” présente une comparaison visuelle des salaires moyens pour trois types de formation : apprentissage, formation initiale et professionnalisation. Chaque barre, colorée différemment, représente la moyenne salariale annuelle associée à chaque type de formation.

La barre rouge, correspondant à l’apprentissage, affiche une moyenne salariale qui semble être la plus haute des trois. Cela peut s’expliquer par le fait que l’apprentissage combine souvent travail et études. Ainsi, l’acquisition d’expérience durant les études favorise un salaire plus élevé.

La formation initiale, représentée en vert, montre une moyenne salariale légèrement supérieure à celle de l’apprentissage. Cette catégorie inclut typiquement les étudiants sortant directement de l’éducation secondaire ou supérieure et entrant sur le marché du travail, ce qui peut conduire à des salaires de début de carrière standard.

La professionnalisation, en bleu, présente la moyenne salariale la plus élevée des trois catégories. Cela peut refléter la valeur du marché pour les formations professionnelles spécialisées, qui sont souvent conçues pour répondre directement aux besoins des industries et peuvent conduire à une insertion professionnelle avec des salaires plus compétitifs.

Ce graphique suggère que le type de formation a un impact significatif sur les salaires moyens. Les formations professionnelles, axées sur l’acquisition de compétences spécifiques et immédiatement applicables, semblent offrir les meilleures perspectives salariales moyennes, tandis que les parcours plus traditionnels ou les formations en apprentissage offrent des salaires moyens inférieurs.

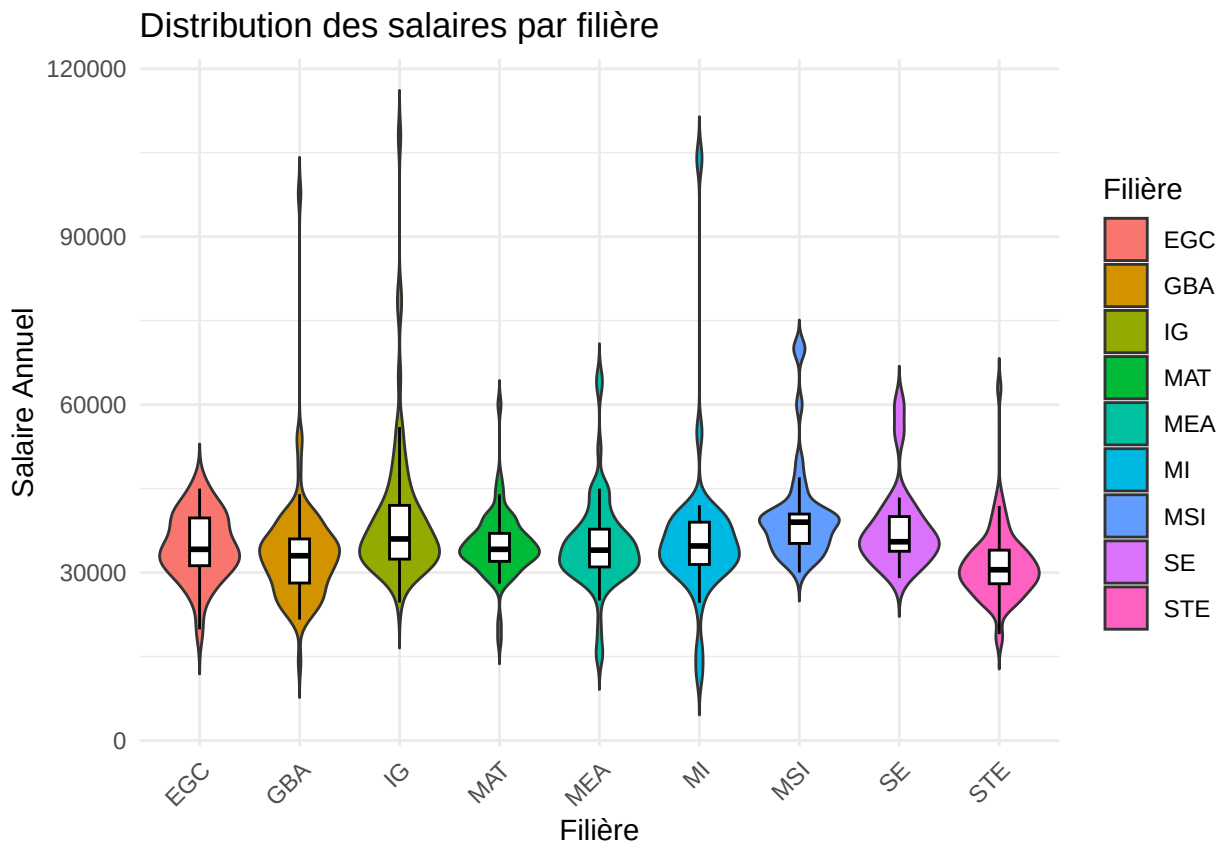
Disparité des filières

```
ggplot(raw_data, aes(x = Filiere, y = Salaire_annuel, fill = Filiere)) +
  geom_violin(trim = FALSE, scale = "width", width = 0.8) +
  geom_boxplot(width = 0.2, fill = "white", color = "black", outlier.shape = NA) +
```

```
  labs(title = "Distribution des salaires par filière",
       x = "Filière",
       y = "Salaire Annuel",
       fill = "Filière") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
## Warning: Removed 222 rows containing non-finite values (`stat_ydensity()`).
```

```
## Warning: Removed 222 rows containing non-finite values (`stat_boxplot()`).
```



Le diagramme en violons offre une représentation visuelle de la distribution des salaires annuels pour différentes filières. Chaque violon combine les caractéristiques d'une boîte à moustaches et d'un diagramme de densité de probabilité. La largeur d'un violon à un niveau de salaire donné est proportionnelle au nombre d'observations (salariés) pour ce salaire, ce qui donne une indication de la densité des données à ce niveau.

Analyse des médianes par filière : - Les filières IG et MSI présentent des médianes élevées, ce qui suggère que les salaires moyens dans ces domaines sont supérieurs à ceux des autres filières. - Les filières MI et SE suivent de près avec des médianes légèrement inférieures, indiquant que les salaires moyens sont également relativement élevés dans ces secteurs. - Les filières STE et EGC ont des médianes inférieures, ce qui indique que les salaires moyens sont plus bas dans ces domaines par rapport aux autres filières présentées.

Analyse de la distribution par filière : - Les violons pour les filières IG, MI et GBA sont relativement larges en haut, indiquant une plus grande variabilité des salaires élevés et la présence de salaires très hauts comparativement aux autres filières. - Pour les filières MSI et SE, les violons sont également larges, mais avec moins d'observations aux salaires les plus élevés, ce qui peut suggérer une répartition plus homogène des salaires autour de la médiane. - Les filières STE et EGC présentent des violons plus étroits, particulièrement en haut de la distribution, ce qui pourrait signifier une concentration des salaires autour de la médiane avec

moins de salariés recevant des salaires très élevés.

Les points en dehors des violons représentent les valeurs extrêmes, qui sont significativement plus élevées ou plus basses que la majorité des salaires dans la filière.

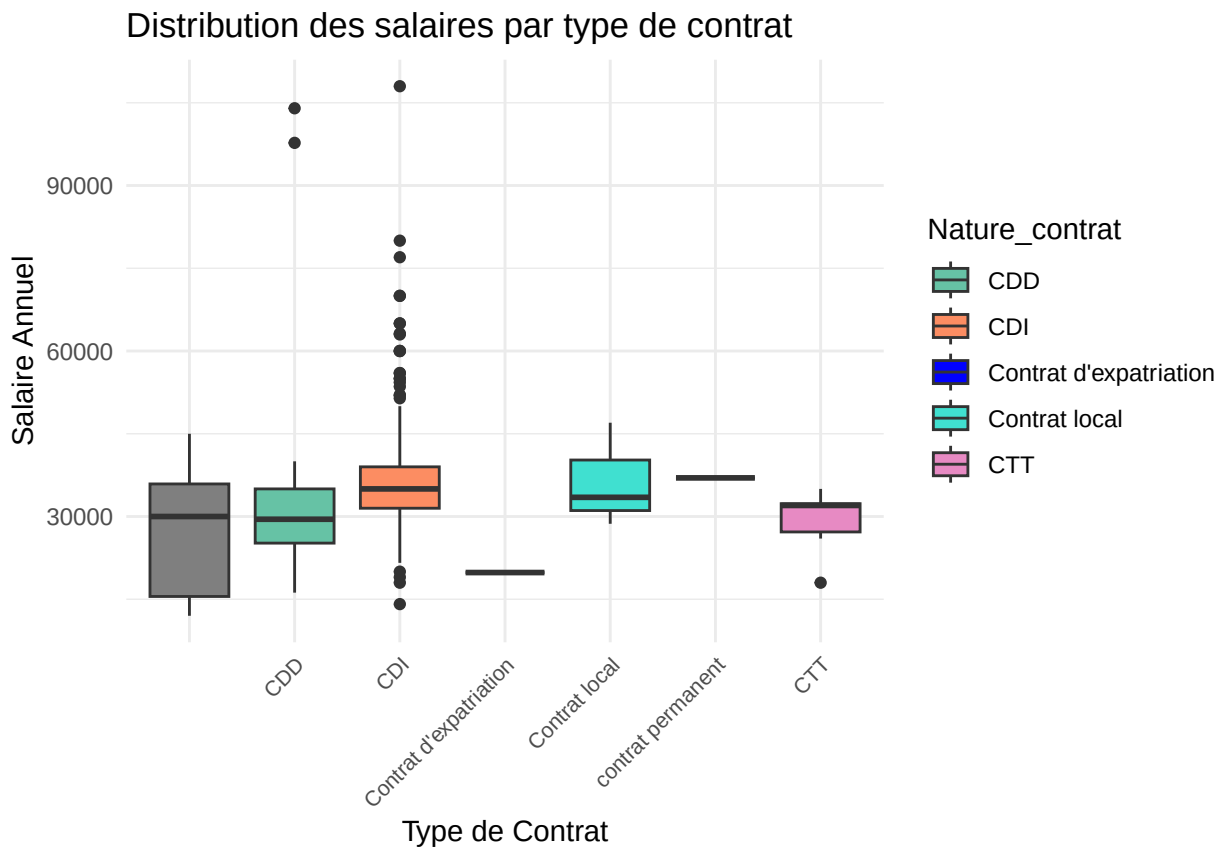
En somme, ce diagramme montre que certaines filières comme IG et MSI sont associées à des salaires plus élevés et une plus grande hétérogénéité des rémunérations, tandis que d'autres comme STE et EGC sont caractérisées par des salaires plus homogènes et généralement plus modestes. Ces informations peuvent être utiles pour les individus évaluant les perspectives de carrière dans chaque domaine, ainsi que pour les employeurs et les responsables de l'éducation lors de l'élaboration de stratégies de formation et de recrutement.

Filtrage des personnes en activité

```
filtered_data <- remove_rows_by_value(raw_data, "Situation", "En recherche emploi")
filtered_data <- remove_rows_by_value(filtered_data, "Situation", "En poursuite d'études (hors thèse) /
filtered_data <- remove_rows_by_value(filtered_data, "Situation", "These")
filtered_data <- remove_rows_by_value(filtered_data, "Situation", "Sans activité")
filtered_data <- remove_rows_by_value(filtered_data, "Situation", "Volontariat")
filtered_data <- remove_rows_by_value(filtered_data, "Situation", "En création d'entreprise /reprise d'
filtered_data <- remove_rows_by_value(filtered_data, "Situation", "Etudes")

activite_data <- filtered_data
```

Disparité par nature du contrat

```
library(ggplot2)

# Filtrage des données en excluant certaines catégories
filtered_data <- subset(raw_data, !Nature_contrat %in% c("Apprentissage", "Autre", "ile-de-France"))

# Supposons que les catégories de contrats sont 'CDD', 'CDI', 'Intérim', etc.
# Vous pouvez définir manuellement les couleurs pour chaque catégorie de contrat
couleurs_contrats <- c("CDD" = "#66c2a5", "CDI" = "#fc8d62", "Intérim" = "#8da0cb",
                       "Stage" = "#e78ac3", "Alternance" = "#a6d854", "CTT" = "#e78ac3", "Contrat local

# Diagramme en boîte avec couleurs
ggplot(filtered_data, aes(x = Nature_contrat, y = Salaire_annuel, fill = Nature_contrat)) +
  geom_boxplot() +
  scale_fill_manual(values = couleurs_contrats) +  # Utiliser les couleurs définies précédemment
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 8)) +
  labs(title = "Distribution des salaires par type de contrat", x = "Type de Contrat", y = "Salaire Annu
```



Le graphique intitulé “Distribution des salaires par type de contrat” fournit une comparaison des salaires annuels en fonction de la nature du contrat de travail. À travers des boxplots, il distingue clairement trois types de contrats : les CDD (contrats à durée déterminée), les CDI (contrats à durée indéterminée) et les CTT (contrats de travail temporaire), chacun représenté par une couleur distincte.

Le CDI, illustré en orange, présente une médiane de salaire supérieure à celle des deux autres types de contrats, indiquant que les employés sous CDI ont en moyenne des salaires plus élevés. La dispersion des salaires dans cette catégorie est également plus large, ce qui suggère une gamme de salaires allant des niveaux débutants aux positions très rémunérées. Les points isolés au-dessus du boxplot principal indiquent l’existence de salaires bien au-delà de la norme, soulignant la présence d’opportunités hautement rémunératrices au sein des CDI.

Les CDD, représentés en turquoise, montrent une concentration plus étroite des salaires, avec une médiane inférieure à celle des CDI. Cette concentration révèle une moindre variabilité salariale, ce qui peut être attribué à la nature temporaire et souvent plus précaire de ces contrats.

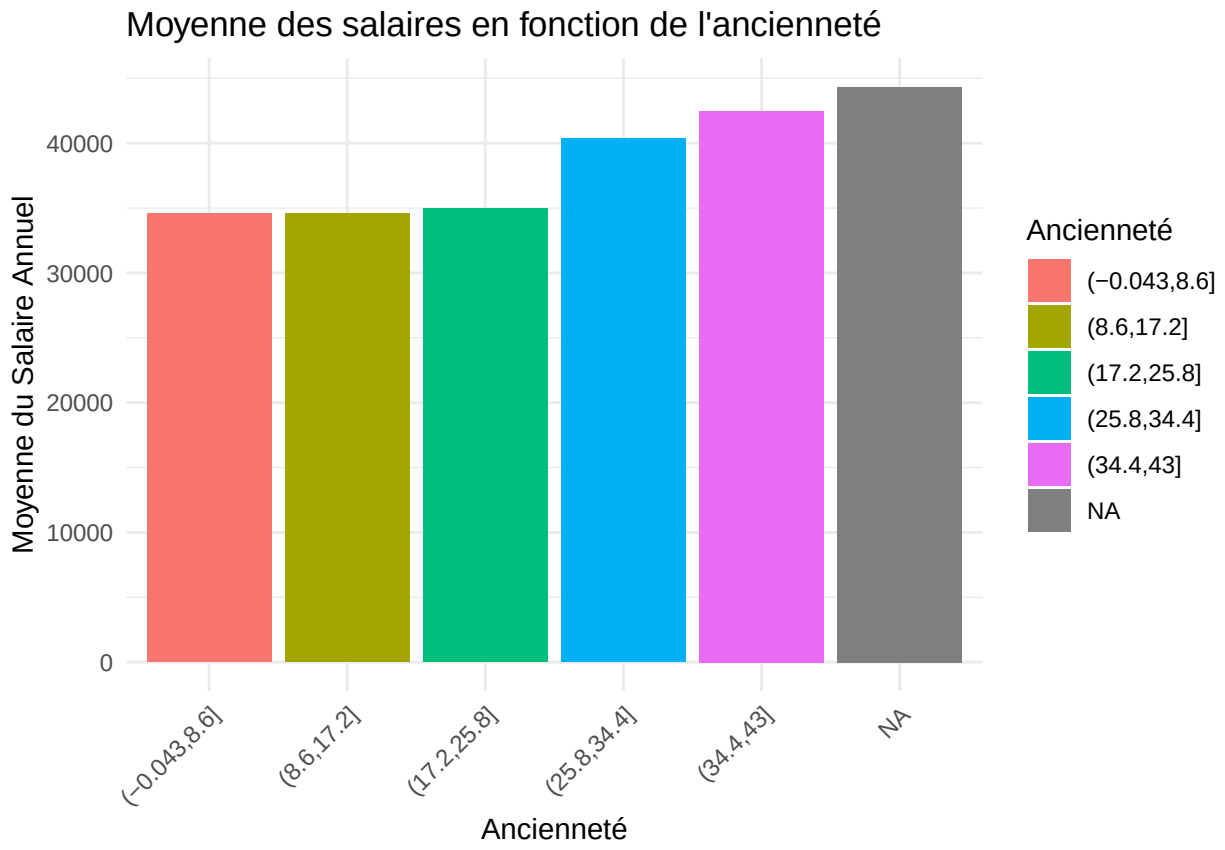
Quant aux CTT, affichés en rose, ils arborent la médiane la plus basse et la distribution des salaires la plus resserrée, ce qui indique non seulement des salaires généralement inférieurs mais aussi une moindre variation dans les rémunérations offertes. Cela pourrait refléter les limitations inhérentes aux emplois temporaires, qui offrent moins d’opportunités pour des salaires élevés.

En résumé, le type de contrat apparaît comme un facteur déterminant des perspectives salariales. Les CDI semblent offrir des salaires plus élevés et une plus grande variabilité, reflétant la sécurité et les possibilités de carrière à long terme. Les CDD et les CTT présentent des profils de rémunération plus modestes, avec des CTT particulièrement restreints dans leur potentiel salarial. Pour les individus qui évaluent leurs options d’emploi, cette visualisation des données salariales pourrait être un outil précieux pour orienter leurs décisions professionnelles.

Distribution du salaire en fonction de l'ancienneté

```
library(ggplot2)

ordered_data <- raw_data[order(as.numeric(raw_data$Anciennete)), ]

# Création de catégories d'ancienneté (facultatif)
ordered_data$Anciennete_category <- cut(as.numeric(ordered_data$Anciennete), breaks = 5)

# Diagramme en barres pour la moyenne des salaires en fonction de l'ancienneté
ggplot(ordered_data, aes(x = Anciennete_category, y = Salaire_annuel, fill = Anciennete_category)) +
  geom_bar(stat = "summary", fun = "mean", position = "dodge") +
  labs(title = "Moyenne des salaires en fonction de l'ancienneté",
       x = "Ancienneté",
       y = "Moyenne du Salaire Annuel",
       fill = "Ancienneté") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
## Warning: Removed 222 rows containing non-finite values (`stat_summary()`).
```



Le graphique "Moyenne des salaires en fonction de l'ancienneté" illustre comment l'expérience professionnelle accumulée influence la rémunération moyenne des employés. L'ancienneté, divisée en segments d'un an, sert de mesure pour établir une corrélation potentielle entre le temps passé dans une organisation ou un secteur et le salaire moyen perçu.

Les employés avec moins d'un an d'expérience, représentés par la barre rouge, perçoivent les salaires les plus

bas de l'échelle, ce qui est attendu pour des positions souvent associées à des rôles débutants ou en phase d'apprentissage. À mesure que l'ancienneté augmente, les barres s'élèvent progressivement : la tranche d'un à deux ans (verte) montre une augmentation modeste du salaire, suggérant un premier palier d'évolution salariale.

Ce phénomène de croissance se poursuit de manière plus marquée pour les tranches de deux à trois ans (bleue) et de trois à quatre ans (violet), où l'on observe un bond significatif dans les salaires moyens. Cela peut refléter les promotions, les augmentations méritées ou l'acquisition de compétences spécialisées qui sont souvent récompensées par des ajustements salariaux positifs.

Curieusement, la tranche des employés les plus anciens, ceux ayant entre quatre et cinq ans d'expérience (rose), bien qu'affichant un salaire moyen élevé, ne représente pas le sommet de l'échelle salariale. Ceci peut indiquer une stabilisation des salaires ou peut-être une saturation dans certaines industries où l'expérience supplémentaire au-delà d'un certain point n'entraîne pas nécessairement une augmentation significative de la rémunération.

En synthèse, ce graphique démontre un lien apparent entre l'ancienneté et le salaire moyen, soulignant l'importance de l'expérience dans la progression de carrière. Il suggère que l'accumulation d'années de service peut être un facteur déterminant dans l'évolution des salaires, bien que l'impact de l'ancienneté puisse plafonner après un certain temps.

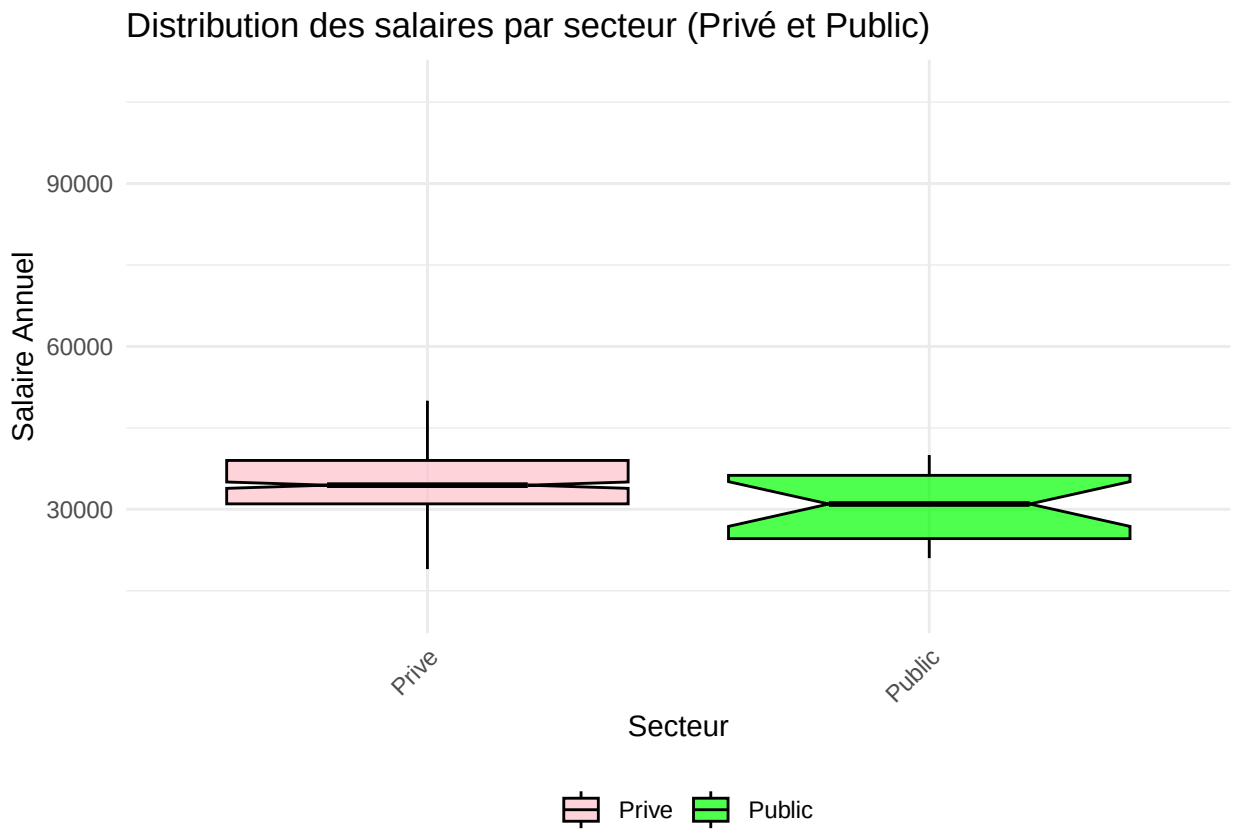
###Distribution du salaire selon le secteur

```
library(ggplot2)

# Vérifiez toutes les valeurs uniques pour identifier les valeurs non désirées
unique(raw_data$Secteur)
```

```
## [1]                Public         Prive          Non salarie(e)
## Levels:  Non salarie(e) Prive Public
```

```
# Filtrez le DataFrame pour ne garder que les lignes avec "Privé" et "Public"
filtered_data <- raw_data[raw_data$Secteur %in% c("Prive", "Public"), ]

# Continuez avec la création du graphique ggplot
ggplot(filtered_data, aes(x = Secteur, y = Salaire_annuel, fill = Secteur)) +
  geom_boxplot(
    width = 0.8,
    notch = TRUE,
    outlier.shape = NA,
    color = "black",
    alpha = 0.7
  ) +
  scale_fill_manual(values = c("Prive" = "pink", "Public" = "green")) +
  labs(
    title = "Distribution des salaires par secteur (Privé et Public)",
    x = "Secteur",
    y = "Salaire Annuel"
  ) +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1),
    legend.position = "bottom",
    legend.title = element_blank()
  )
```

```
## Warning: Removed 62 rows containing non-finite values (`stat_boxplot()`).
```



Le graphique “Distribution des salaires par secteur (Privé et Public)” présente deux boxplots qui comparent les distributions des salaires annuels dans les secteurs privé et public. Les boxplots offrent une vue d’ensemble des médianes, des quartiles et des étendues des salaires au sein de chaque secteur.

Le boxplot du secteur privé, coloré en rose, a une médiane qui se situe au-dessus de la moitié inférieure de la boîte, indiquant que la médiane des salaires est relativement élevée. Cependant, la dispersion des salaires, illustrée par la longueur des moustaches, est modeste, ce qui suggère une variabilité moins importante des salaires dans le secteur privé.

D’autre part, le boxplot du secteur public, en vert, montre une médiane plus basse que celle du secteur privé, indiquant que les salaires médians sont inférieurs dans le secteur public. Néanmoins, l’étendue de la distribution des salaires est plus large dans le secteur public, comme le montre l’amplitude des moustaches, ce qui indique une plus grande variété dans les salaires des employés du secteur public.

La présence de valeurs extrêmes ou aberrantes n’est pas notable dans ce graphique, ce qui indique que la majorité des salaires se situent dans une gamme relativement prévisible pour les deux secteurs.

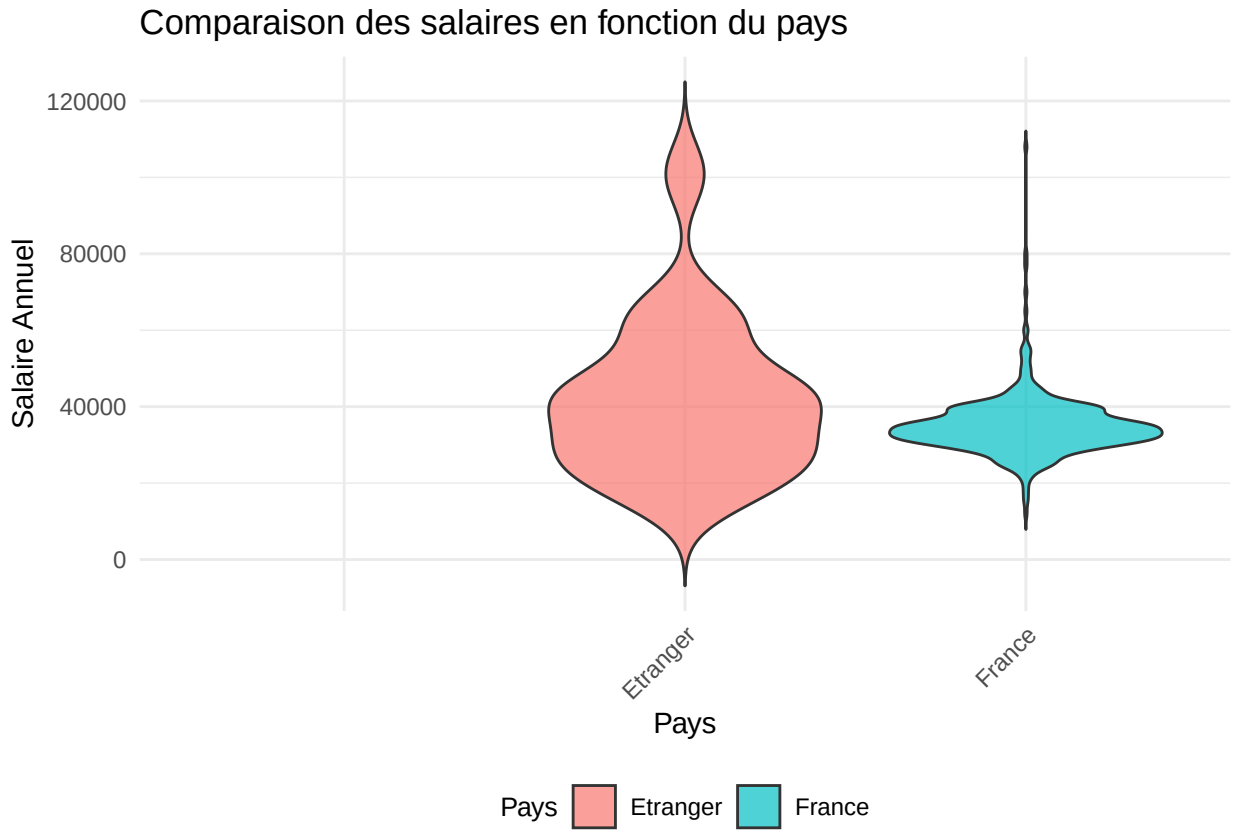
En conclusion, ce graphique suggère que bien que le secteur privé puisse offrir en moyenne des salaires plus élevés, le secteur public présente une plus grande hétérogénéité dans la rémunération de ses employés. Cela peut refléter une diversité des rôles et des niveaux de responsabilité plus marquée dans le secteur public ou des politiques salariales différentes.

Distribution du salaire en fonction de France/étranger

```
ggplot(raw_data, aes(x = factor(France), y = Salaire_annuel, fill = factor(France))) +
  geom_violin(trim = FALSE, scale = "width", width = 0.8, alpha = 0.7) +
  labs(
    title = "Comparaison des salaires en fonction du pays",
```

```r
    x = "Pays",
    y = "Salaire Annuel",
    fill = "Pays"
  ) +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1),
    legend.position = "bottom",
    legend.title = element_text(size = 10)
  )
```

```
## Warning: Removed 222 rows containing non-finite values (`stat_ydensity()`).
```



Le graphique “Comparaison des salaires en fonction du pays” montre deux graphiques en violon pour les salaires annuels en France et à l’étranger. Ces graphiques en violon combinent des éléments de boîtes à moustaches et de densité de probabilité pour montrer la distribution des salaires.

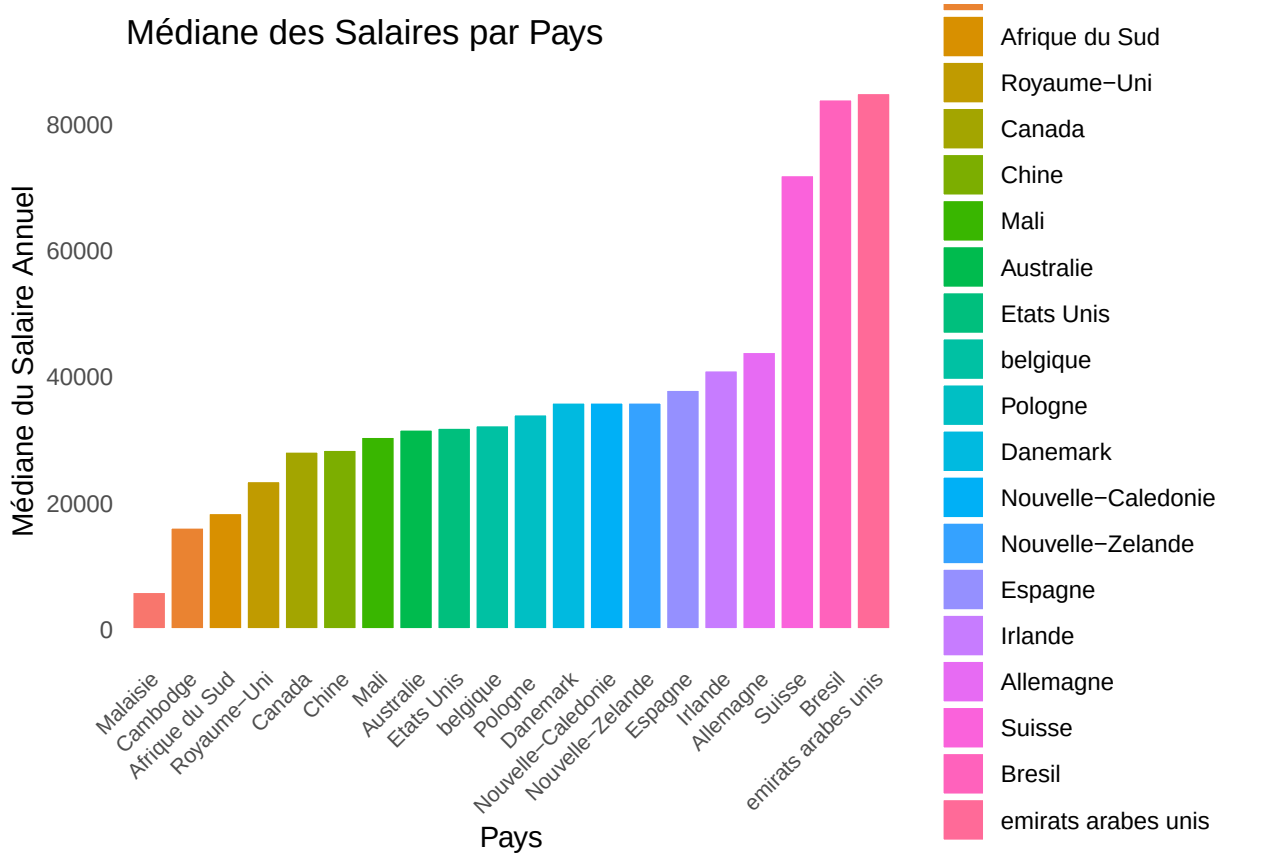
Le violon pour “Étranger” est plus large et présente une forme centrée vers le haut, indiquant une distribution avec une moyenne plus élevée et une variabilité plus grande des salaires. Cela suggère que les salaires à l’étranger ne sont pas seulement variés, mais comprennent également des valeurs élevées, ce qui peut refléter des différences dans les normes économiques, les structures de rémunération, ou des niveaux de vie plus élevés qui nécessitent des salaires supérieurs.

Pour la France, le violon est plus étroit avec une médiane plus basse et une distribution des salaires moins large, indiquant que les salaires sont généralement plus bas et moins dispersés. Cela pourrait être le reflet de politiques salariales plus homogènes ou de différences dans les secteurs économiques entre la France et les pays étrangers.

La distribution plus large et la médiane plus élevée à l'étranger pourraient également indiquer l'existence de secteurs d'emploi hautement rémunérateurs ou le fait que des expatriés français puissent bénéficier de primes et d'avantages liés à leur statut à l'étranger. En revanche, la concentration des salaires plus basse et la médiane inférieure en France pourraient suggérer une cohésion plus forte des politiques salariales ou une prévalence de secteurs moins rémunérateurs.

Distribution du salaire en fonction du pays à l'étranger

```r
library(ggplot2)

# Calcul de la médiane des salaires par pays
median_salaries <- tapply(raw_data_filtered$Salaire_annuel, raw_data_filtered$Pays, median)
median_order <- names(sort(median_salaries))

# Création d'un facteur pour l'ordre des pays par médiane
emploiFactor <- factor(raw_data_filtered$Pays, levels = median_order)

# Création du diagramme en bâtonnets
ggplot(raw_data_filtered, aes(x = emploiFactor, y = Salaire_annuel, fill = emploiFactor)) +
  geom_bar(stat = "summary", fun = "median", color = "white", size = 0.5) +
  theme_minimal() +  # Fond blanc
  theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 8),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank()) +  # Suppression des lignes de grille
  labs(title = "Médiane des Salaires par Pays", x = "Pays", y = "Médiane du Salaire Annuel",
       fill = "Pays")  # Légende pour les couleurs des pays
```



Le graphique “Comparaison des salaires en fonction du pays” utilise des graphiques en violon pour illustrer la distribution des salaires annuels en France et à l’étranger.

Le graphique en violon pour “Étranger”, teinté de rouge, montre une large base et un sommet élevé, indiquant une grande variété dans la distribution des salaires avec une densité significative autour de la médiane, qui se situe dans la partie supérieure de la distribution. Cela suggère que les salaires à l’étranger sont non seulement variés mais ont également tendance à être plus élevés en moyenne, avec un nombre notable d’individus gagnant des salaires considérablement plus élevés, comme en témoigne la partie supérieure étendue du violon.

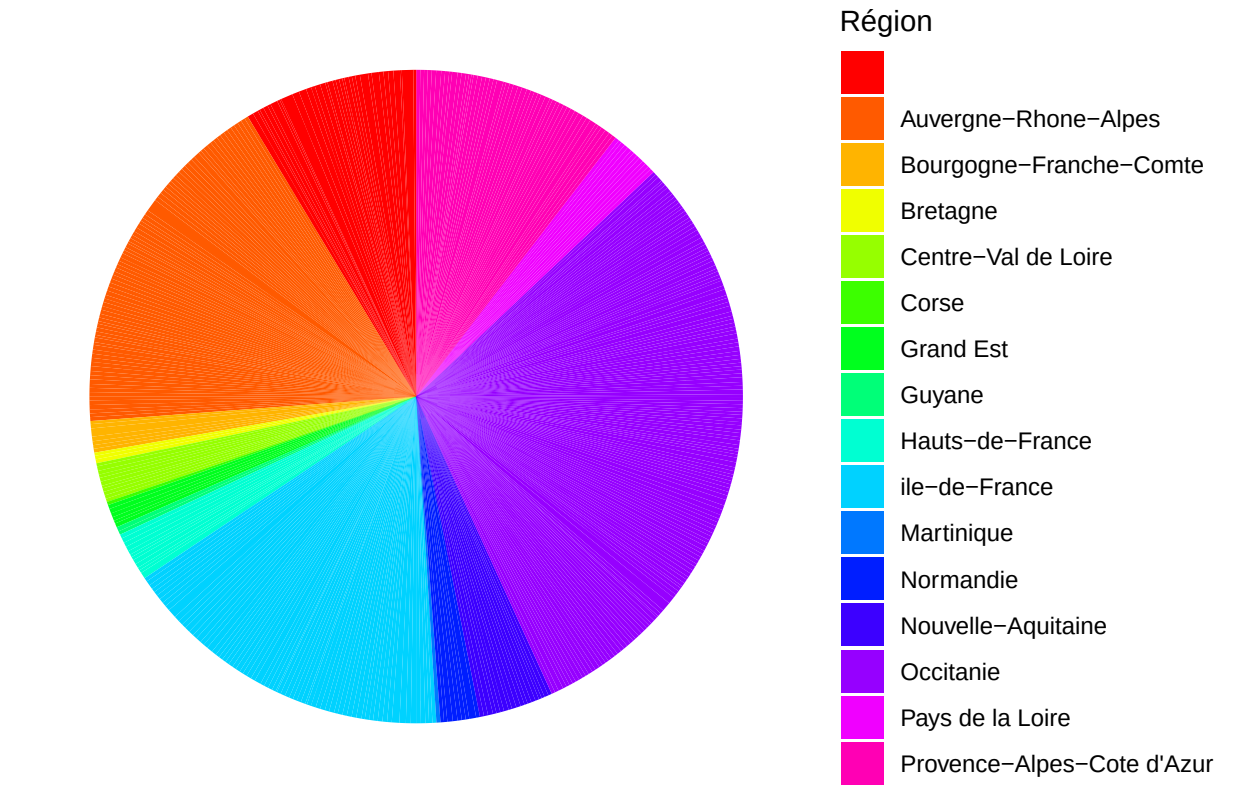
En comparaison, le graphique en violon pour la France, coloré en cyan, présente une médiane plus basse, une distribution plus resserrée et une queue supérieure moins étendue, ce qui implique que bien que les salaires soient globalement plus bas en France, ils sont également moins dispersés. La concentration des salaires autour de la médiane est plus marquée, indiquant moins de variation dans les niveaux de rémunération.

L’analyse de ces deux distributions souligne les différences potentielles en matière de rémunération entre travailler en France et à l’étranger. Les opportunités à l’étranger peuvent offrir des rémunérations plus élevées, ce qui pourrait être dû à des facteurs économiques variés, y compris le coût de la vie, la demande de compétences spécialisées, ou les politiques salariales des entreprises internationales. Cependant, il est important de noter que le coût de la vie et les avantages sociaux, qui ne sont pas représentés dans ce graphique, jouent également un rôle crucial dans l’évaluation globale des salaires et de la qualité de vie.

Distribution des salaires en fonction de la région

```r
library(ggplot2)

# Assurez-vous d'abord d'exclure les lignes où Region est NA ou a une valeur non désirée
# Remplacez 'valeur non désirée' par la valeur réelle qui représente le secteur non affilié
raw_data <- raw_data[!is.na(raw_data$Region) & raw_data$Region != 'valeur non désirée', ]

# Continuez avec les étapes précédentes
# Calculate average salary by region
avg_salary_by_region <- tapply(raw_data$Salaire_annuel, raw_data$Region, mean)

# Define a color palette for each region
color_palette <- rainbow(length(unique(raw_data$Region)))

# Plotting the distribution of salaries based on region as a pie chart
ggplot(raw_data, aes(x = "", y = Salaire_annuel, fill = Region)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar("y") +
  theme_void() +  # Removes unnecessary elements
  labs(title = "Répartition des salaires par région",
       fill = "Région",
       y = "Salaire Annuel") +
  scale_fill_manual(values = color_palette)  # Use the color palette without conditionally assigning re
```

```
## Warning: Removed 222 rows containing missing values (`position_stack()`).
```



Le diagramme en camembert “Répartition des salaires par région” met en évidence les différences dans les salaires moyens à travers diverses régions de France, avec une attention particulière portée à l'Occitanie, l'Île-de-France, l'Auvergne-Rhône-Alpes, la Bretagne et les Pays de la Loire.

L'Occitanie se distingue avec le segment le plus large, indiquant que cette région a la moyenne des salaires la plus élevée parmi celles représentées. Cela pourrait suggérer une présence de secteurs d'activité bien rémunérés ou une concentration de professions spécialisées qui valorisent les compétences de la main-d'œuvre locale.

L'Île-de-France, généralement considérée comme un centre économique majeur en raison de la présence de Paris, est représentée par le deuxième plus grand segment. Malgré son statut de capitale économique, elle ne dépasse pas l'Occitanie dans ce graphique, ce qui pourrait être dû à une gamme plus étendue de salaires dans la région.

L'Auvergne-Rhône-Alpes suit de près, avec un autre segment important. Cette région, connue pour sa diversité économique, incluant l'industrie et le tourisme, montre des salaires moyens élevés, ce qui est cohérent avec son dynamisme économique et industriel.

En contraste, la Bretagne et les Pays de la Loire sont représentés par des segments plus petits, ce qui indique des moyennes de salaires inférieures dans ces régions. Cela pourrait refléter une économie moins orientée vers les industries à haute valeur ajoutée ou un tissu économique axé sur des secteurs offrant des salaires moins élevés.

Dans l'ensemble, cette visualisation offre une perspective utile sur la distribution géographique des salaires en France, montrant que certaines régions, comme l'Occitanie, surpassent d'autres régions plus traditionnellement associées à la richesse économique, telles que l'Île-de-France. Elle met également en lumière les régions qui peuvent être considérées comme moins prospères du point de vue salarial, comme la Bretagne et les Pays de la Loire, et qui pourraient bénéficier de stratégies de développement économique pour améliorer la rémunération des travailleurs.

Etude données 2020

Récupération des données

```
# Trying ISO-8859-1
raw_data <- read.csv("data/data_2020.csv", sep = ";")
```

Définir toutes les données comme caractère (variable qualitative) sauf le salaire

```
# Lists specifying which columns to convert to factors and numeric
factor_cols <- c("Date", "identifiant", "Genre", "Annee_diplome","Type_formation","Filiere","Situation"
numeric_cols <- c("Anciennete", "Salaire_annuel", "Responsabilite_hierarchique", "Responsabilite_budget

# Convert columns to factors
raw_data[factor_cols] <- lapply(raw_data[factor_cols], as.factor)

# Convert columns to numeric
raw_data[numeric_cols] <- lapply(raw_data[numeric_cols], as.numeric)
```

```
## Warning in lapply(raw_data[numeric_cols], as.numeric): NAs introduced by
## coercion

## Warning in lapply(raw_data[numeric_cols], as.numeric): NAs introduced by
## coercion

## Warning in lapply(raw_data[numeric_cols], as.numeric): NAs introduced by
## coercion

## Warning in lapply(raw_data[numeric_cols], as.numeric): NAs introduced by
## coercion
```

```
summary(raw_data)
```

Avec ces valeurs on peut déduire beaucoup de choses...

Fonction de filtre

Ceci est une fonction utilisé après pour retirer différentes lignes en fonction des valeurs dans une certain colonne. (équivalent d'un select where)

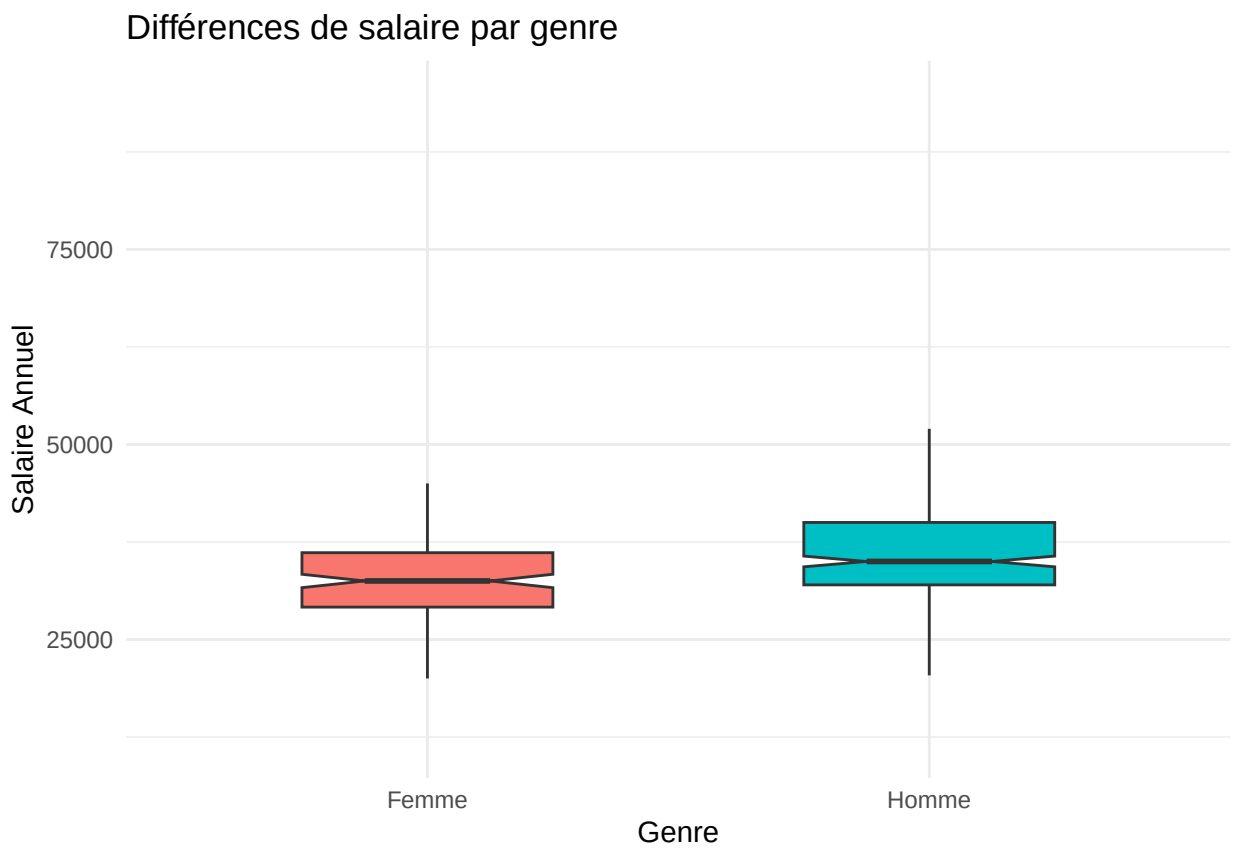
```
remove_rows_by_value <- function(data, column_name, value_to_remove) {
  data_filtered <- subset(data, !(data[[column_name]] == value_to_remove))
  return(data_filtered)
}
```

Première étude

Disparité homme/femme - toute formation confondue

```
ggplot(raw_data, aes(x = Genre, y = Salaire_annuel, fill = Genre)) +
  geom_boxplot(width = 0.5, notch = TRUE, outlier.shape = NA) +
  labs(title = "Différences de salaire par genre",
       x = "Genre",
       y = "Salaire Annuel") +
  theme_minimal() +
  theme(legend.position = "none")
```

```
## Warning: Removed 208 rows containing non-finite values (`stat_boxplot()`).
```



Le graphique intitulé “Différences de salaire par genre” représente un boxplot classique, qui met en évidence la distribution des salaires annuels entre les femmes et les hommes. À première vue, il suggère une légère disparité salariale entre les genres, avec les médianes indiquant que les hommes ont tendance à gagner plus que les femmes dans cet échantillon.

Les médianes de chaque groupe — représentées par la ligne centrale des boîtes — sont cruciales pour cette observation. La médiane pour le groupe des femmes semble être inférieure à celle du groupe des hommes. Il est important de noter que la médiane est souvent préférée à la moyenne pour une telle analyse, car elle est moins sensible aux valeurs extrêmes qui pourraient fausser les résultats.

En examinant la taille des boîtes, qui illustrent l’écart interquartile, nous observons une similitude dans la dispersion des salaires entre les deux groupes. Cela signifie que la moitié centrale des salaires s’étend sur une plage similaire pour les deux genres, suggérant que, mis à part la médiane, les distributions des salaires sont relativement comparables.

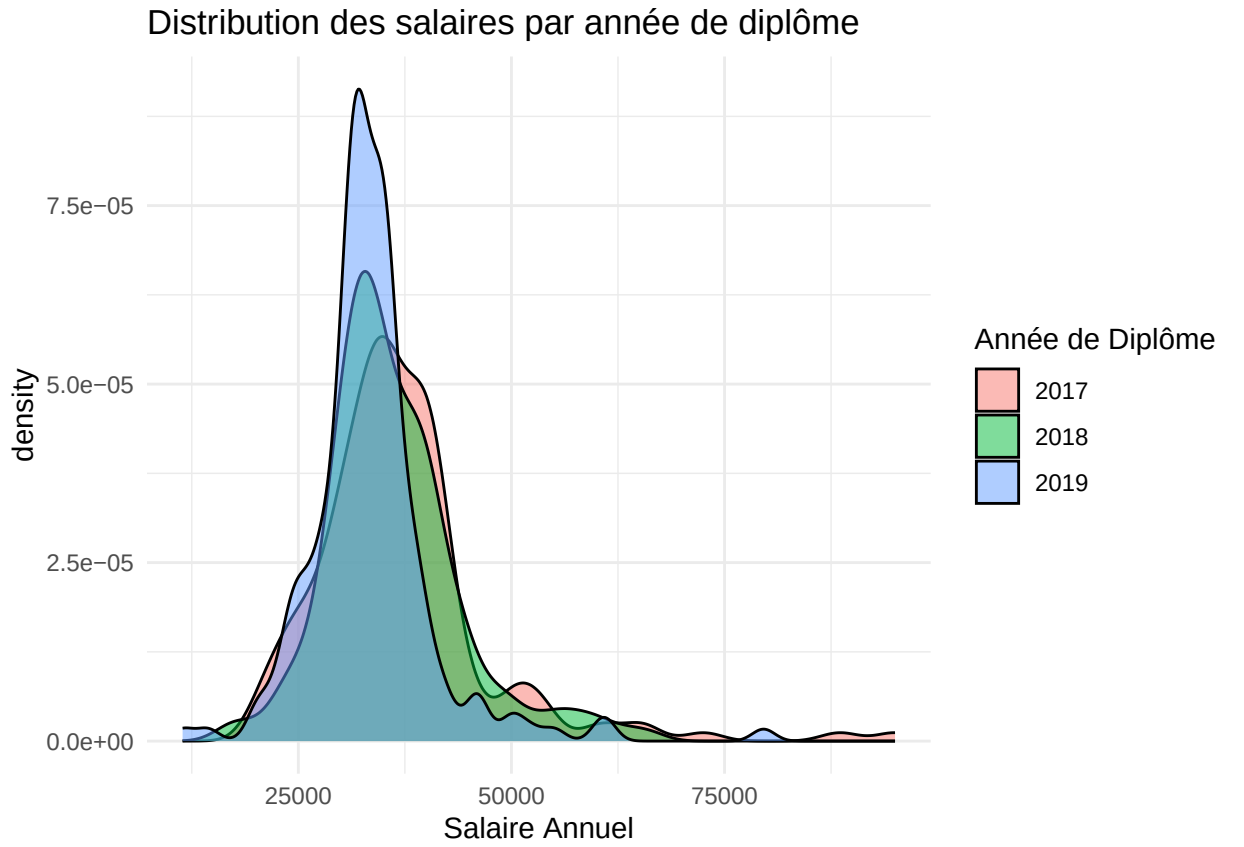
En conclusion, ce graphique révèle une tendance où les hommes semblent avoir un avantage salarial par rapport aux femmes dans l’échantillon étudié. Néanmoins, la similarité des distributions suggère que les écarts de salaires, au-delà de la médiane, ne sont pas marqués par des différences extrêmes. Pour une analyse complète, il serait judicieux de prendre en compte d’autres variables qui pourraient influencer ces résultats et de réaliser des tests statistiques pour évaluer la significativité des différences observées.

Disparité sur les dates d’optention du diplôme

```
ggplot(raw_data, aes(x = Salaire_annuel, fill = as.factor(Annee_diplome))) +
  geom_density(alpha = 0.5) +
  labs(title = "Distribution des salaires par année de diplôme",
```

```
       x = "Salaire Annuel",
       fill = "Année de Diplôme") +
  theme_minimal()
```

```
## Warning: Removed 208 rows containing non-finite values (`stat_density()`).
```



Le graphique “Distribution des salaires par année de diplôme” illustre les distributions des salaires annuels pour les diplômés des années 2017, 2018 et 2019 à l’aide de courbes de densité.

La courbe de 2019, en bleu, présente la densité la plus élevée autour des salaires inférieurs, suggérant que les diplômés les plus récents ont commencé avec des salaires plus bas. Cela pourrait refléter une entrée sur le marché du travail à des positions de niveau débutant, qui ont tendance à offrir des rémunérations plus modestes.

En revanche, les courbes de 2017 et 2018 montrent que les salaires ont tendance à être centrés sur des montants plus élevés, avec la courbe de 2017 ayant le pic le plus haut. Cela indique que les diplômés de ces années pourraient avoir bénéficié d’une augmentation de salaire due à l’accumulation d’expérience ou à des progressions de carrière survenues depuis leur entrée sur le marché du travail.

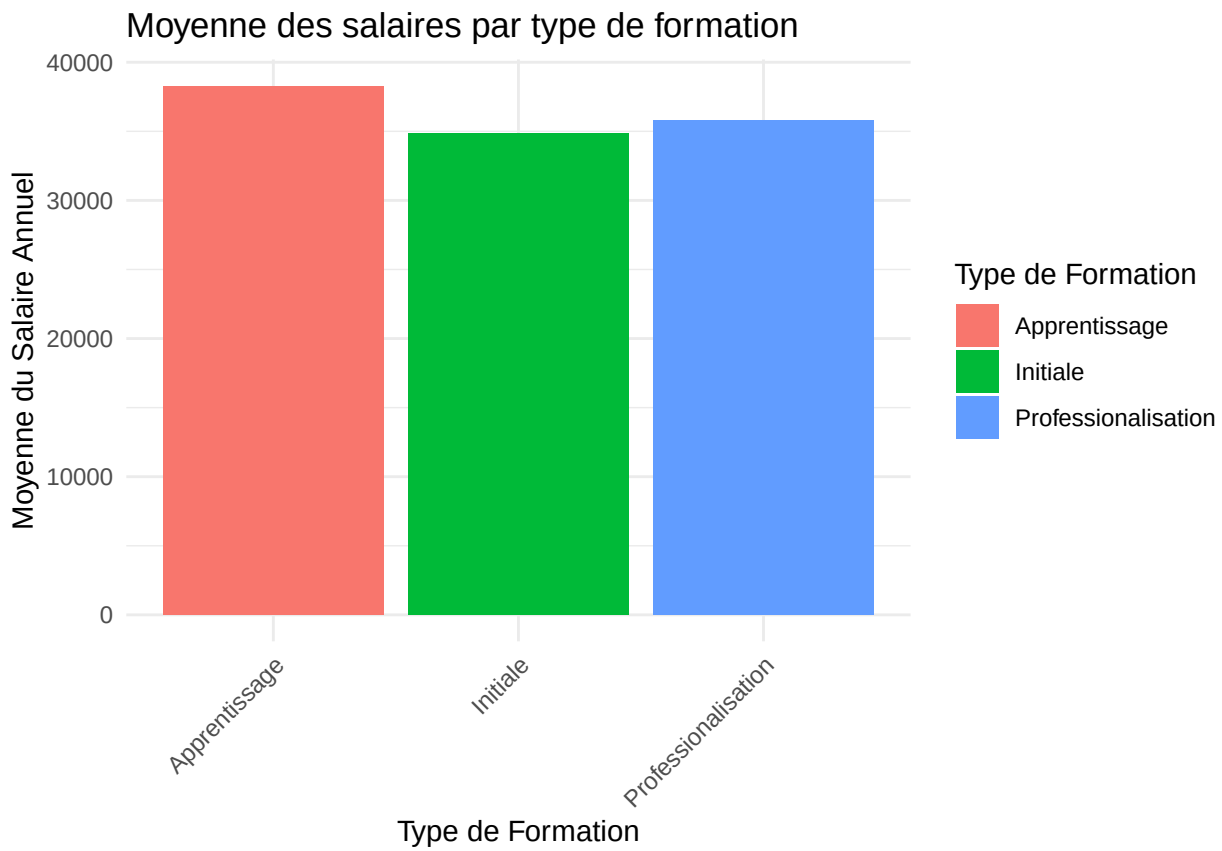
De plus, la courbe de 2018 montre une distribution légèrement plus large que celle de 2017, ce qui peut indiquer une plus grande variabilité des salaires parmi les diplômés de cette année-là. Cela pourrait être dû à une diversité plus grande dans les carrières ou à des changements dans les secteurs d’emploi entre ces années.

L’élargissement progressif des courbes de densité d’une année sur l’autre pourrait également suggérer que les diplômés avec plus d’années sur le marché du travail ont eu l’opportunité de progresser vers des postes mieux rémunérés, ce qui augmente la variabilité des salaires au sein de leur groupe. Cela peut également refléter l’évolution des conditions économiques qui influencent les structures salariales au fil du temps.

Disparité formation initiale/apprentissage/contrat de professionalisation

```
library(ggplot2)

# Supprimer les lignes où 'Type_formation' est NA ou vide
clean_data <- raw_data[!is.na(raw_data$Type_formation) & raw_data$Type_formation != "", ]

# Créer le graphique avec les données nettoyées
ggplot(clean_data, aes(x = as.factor(Type_formation), y = Salaire_annuel, fill = as.factor(Type_formati
  stat_summary(fun = "mean", geom = "col") +
  labs(title = "Moyenne des salaires par type de formation",
       x = "Type de Formation",
       y = "Moyenne du Salaire Annuel",
       fill = "Type de Formation") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
## Warning: Removed 208 rows containing non-finite values (`stat_summary()`).
```



Le graphique intitulé “Moyenne des salaires par type de formation” présente une comparaison des salaires moyens annuels associés à trois types de formation : l’apprentissage, la formation initiale et la professionnalisation.

La barre rouge, représentant l’apprentissage, affiche la moyenne salariale la plus élevée parmi les trois catégories. Cela suggère que les individus ayant suivi une formation en apprentissage peuvent bénéficier de salaires annuels moyens supérieurs, ce qui pourrait être attribué à la combinaison de formation pratique en entreprise et d’enseignement théorique. Il est souvent perçu que l’apprentissage conduit directement à des

emplois dans le domaine d'étude, ce qui peut justifier ces salaires plus élevés.

La barre verte, correspondant à la formation initiale, montre une moyenne salariale inférieure à celle de l'apprentissage. La formation initiale est généralement suivie par les étudiants qui entrent sur le marché du travail après avoir complété leurs études sans expérience professionnelle significative, ce qui pourrait expliquer pourquoi les salaires moyens sont plus bas par rapport à l'apprentissage.

La formation professionnelle, indiquée par la barre bleue, a une moyenne salariale légèrement inférieure à celle de l'apprentissage, mais supérieure à celle de la formation initiale. Ce type de formation vise souvent à doter les étudiants de compétences pratiques et spécialisées pour des emplois spécifiques, ce qui peut conduire à des salaires moyens relativement élevés.

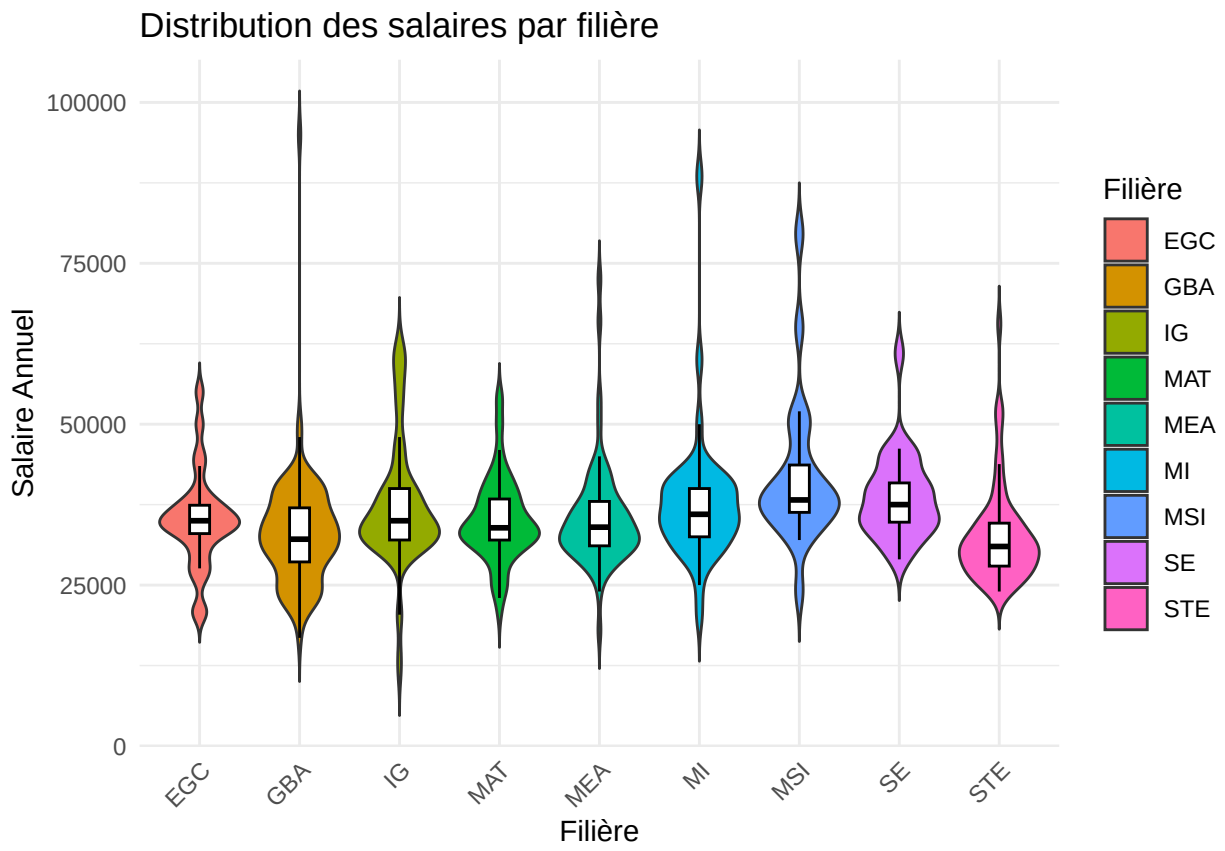
Ces différences peuvent refléter l'accent mis par les employeurs sur les expériences pratiques et les compétences techniques directement applicables dans le milieu professionnel. Les résultats indiquent que les parcours qui intègrent des expériences en milieu de travail, comme l'apprentissage et la professionnalisation, peuvent offrir un avantage salarial sur la formation purement académique ou théorique.

Disparité des filières

```
ggplot(raw_data, aes(x = Filiere, y = Salaire_annuel, fill = Filiere)) +
  geom_violin(trim = FALSE, scale = "width", width = 0.8) +
  geom_boxplot(width = 0.2, fill = "white", color = "black", outlier.shape = NA) +
  labs(title = "Distribution des salaires par filière",
       x = "Filière",
       y = "Salaire Annuel",
       fill = "Filière") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
## Warning: Removed 208 rows containing non-finite values (`stat_ydensity()`).
```

```
## Warning: Removed 208 rows containing non-finite values (`stat_boxplot()`).
```



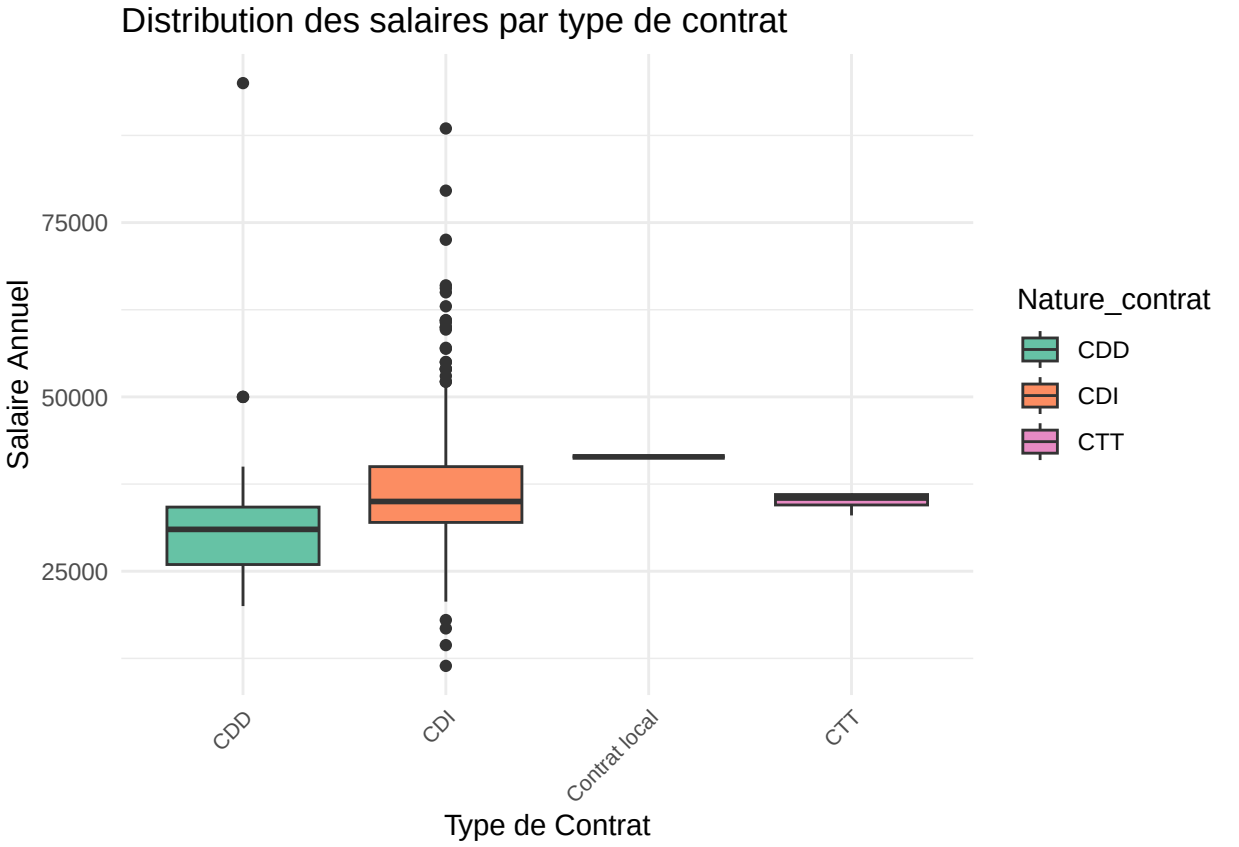
Filtrage des personnes en activité

```
filtered_data <- remove_rows_by_value(raw_data, "Situation", "En recherche emploi")
filtered_data <- remove_rows_by_value(filtered_data, "Situation", "En poursuite d'études (hors thèse) /
filtered_data <- remove_rows_by_value(filtered_data, "Situation", "These")
filtered_data <- remove_rows_by_value(filtered_data, "Situation", "Sans activité")
filtered_data <- remove_rows_by_value(filtered_data, "Situation", "Volontariat")
filtered_data <- remove_rows_by_value(filtered_data, "Situation", "En création d'entreprise /reprise d'
filtered_data <- remove_rows_by_value(filtered_data, "Situation", "Etudes")

activite_data <- filtered_data
```

Disparité par nature du contrat

```
library(ggplot2)

# Remplacer 'activite_data' par votre dataframe réel
# Filtrage de certains contrats peu représentatifs
filtered_data <- raw_data[!raw_data$Nature_contrat %in% c("Service à la personne (cours particulier de n

# Supposons que les catégories de contrats sont 'CDD', 'CDI', 'Intérim', etc.
# Vous pouvez définir manuellement les couleurs pour chaque catégorie de contrat
couleurs_contrats <- c("CDD" = "#66c2a5", "CDI" = "#fc8d62", "Intérim" = "#8da0cb",
                       "Stage" = "#e78ac3", "Alternance" = "#a6d854", "CTT" = "#e78ac3", "contart local
```

```r
# Diagramme en boîte avec couleurs
ggplot(filtered_data, aes(x = Nature_contrat, y = Salaire_annuel, fill = Nature_contrat)) +
  geom_boxplot() +
  scale_fill_manual(values = couleurs_contrats) +  # Utiliser les couleurs définies précédemment
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 8)) +
  labs(title = "Distribution des salaires par type de contrat", x = "Type de Contrat", y = "Salaire Annu
```



Le graphique “Distribution des salaires par type de contrat” présente des boîtes à moustaches pour trois types de contrats de travail : CDD (contrat à durée déterminée), CDI (contrat à durée indéterminée), et CTT (contrat de travail temporaire), chacun représenté par une couleur différente.

La boîte à moustaches pour les CDD, en turquoise, montre une médiane relativement basse et une dispersion modérée des salaires, avec quelques valeurs extrêmes qui s’étendent vers le haut. Cela indique que bien que la majorité des salariés en CDD aient des salaires plus bas, il y a des cas où les salaires peuvent être significativement plus élevés.

Pour les CDI, colorés en orange, la médiane est plus élevée que celle des CDD, suggérant que les postes permanents offrent en général des salaires supérieurs. La distribution est également plus large, avec une variabilité plus grande des salaires, ce qui reflète peut-être la diversité des rôles et des niveaux d’expérience au sein des postes en CDI.

Les CTT, représentés en violet, ont la médiane la plus basse et la distribution la plus serrée des salaires, sans valeurs extrêmes élevées. Cela peut suggérer que les emplois temporaires sont généralement caractérisés par des salaires inférieurs et moins de variation.

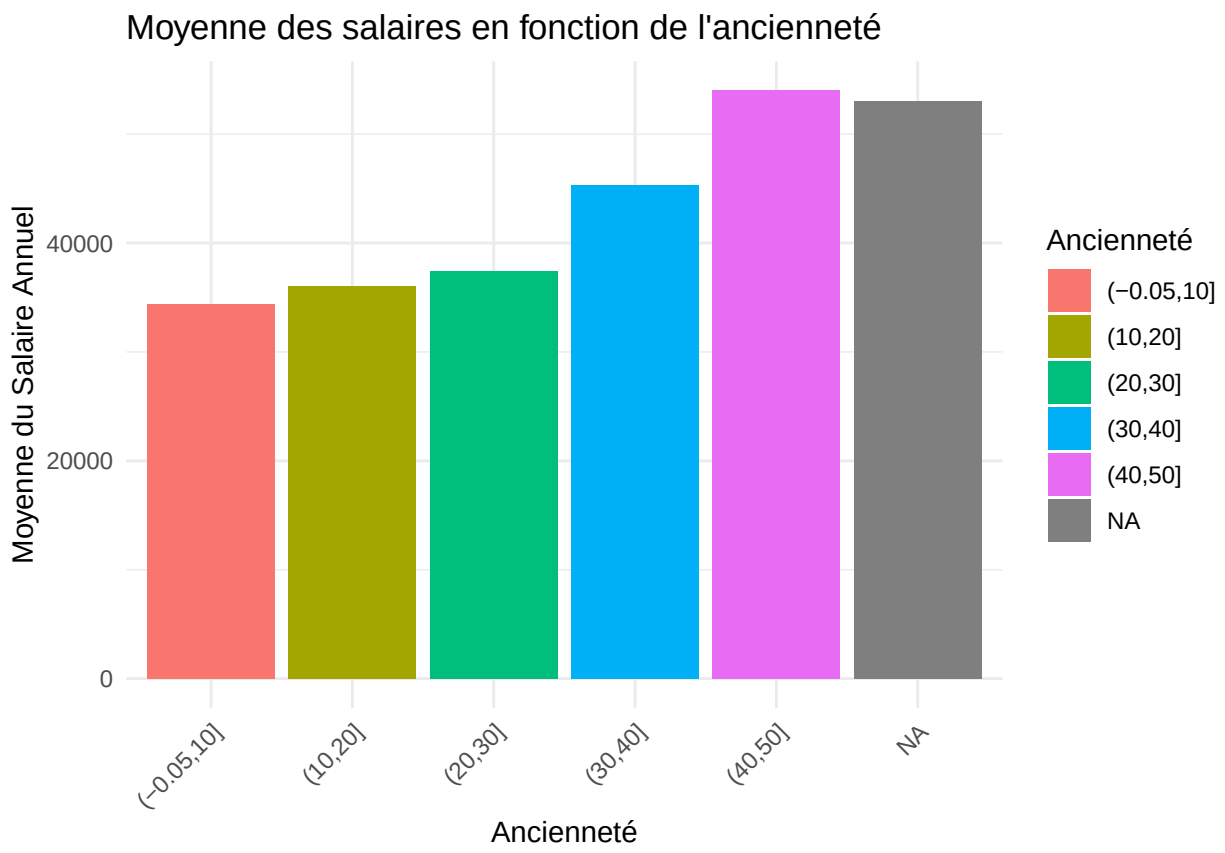
En somme, ce graphique illustre les différences de rémunération associées à la nature du contrat de travail. Les CDI semblent offrir les meilleures conditions salariales, tandis que les CTT sont associés aux salaires

les plus bas. Ces informations sont utiles pour comprendre comment les types de contrats influencent la rémunération sur le marché du travail.

Distribution du salaire en fonction de l'ancienneté

```
library(ggplot2)

ordered_data <- raw_data[order(as.numeric(raw_data$Anciennete)), ]

# Création de catégories d'ancienneté (facultatif)
ordered_data$Anciennete_category <- cut(as.numeric(ordered_data$Anciennete), breaks = 5)

# Diagramme en barres pour la moyenne des salaires en fonction de l'ancienneté
ggplot(ordered_data, aes(x = Anciennete_category, y = Salaire_annuel, fill = Anciennete_category)) +
  geom_bar(stat = "summary", fun = "mean", position = "dodge") +
  labs(title = "Moyenne des salaires en fonction de l'ancienneté",
       x = "Ancienneté",
       y = "Moyenne du Salaire Annuel",
       fill = "Ancienneté") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
## Warning: Removed 208 rows containing non-finite values (`stat_summary()`).
```



Le graphique "Moyenne des salaires en fonction de l'ancienneté" illustre comment l'expérience professionnelle accumulée influence la rémunération moyenne des employés. L'ancienneté, divisée en segments d'un an, sert de mesure pour établir une corrélation potentielle entre le temps passé dans une organisation ou un secteur et

le salaire moyen perçu.

Les employés avec moins d'un an d'expérience, représentés par la barre rouge, perçoivent les salaires les plus bas de l'échelle, ce qui est attendu pour des positions souvent associées à des rôles débutants ou en phase d'apprentissage. À mesure que l'ancienneté augmente, les barres s'élèvent progressivement : la tranche d'un à deux ans (verte) montre une augmentation modeste du salaire, suggérant un premier palier d'évolution salariale.

Ce phénomène de croissance se poursuit de manière plus marquée pour les tranches de deux à trois ans (bleue) et de trois à quatre ans (violet), où l'on observe un bond significatif dans les salaires moyens. Cela peut refléter les promotions, les augmentations méritées ou l'acquisition de compétences spécialisées qui sont souvent récompensées par des ajustements salariaux positifs.

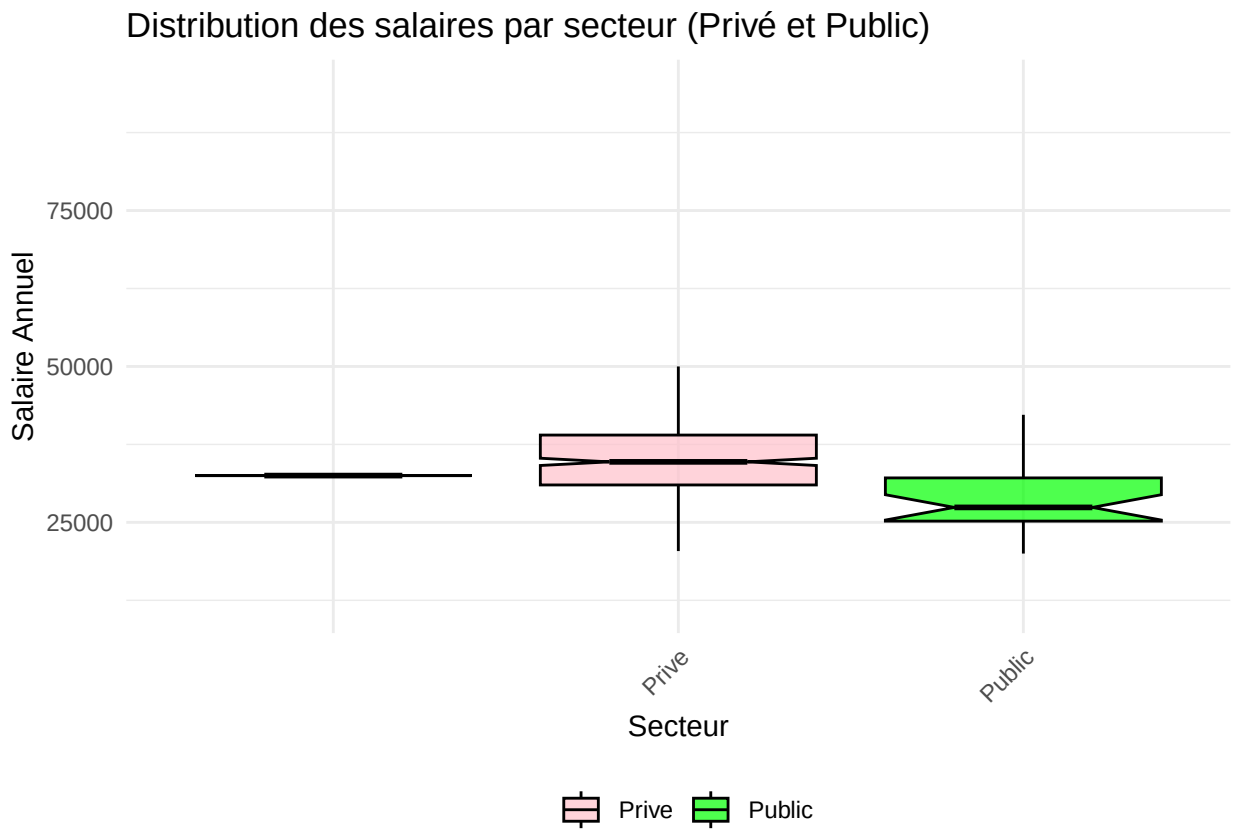
Curieusement, la tranche des employés les plus anciens, ceux ayant entre quatre et cinq ans d'expérience (rose), bien qu'affichant un salaire moyen élevé, ne représente pas le sommet de l'échelle salariale. Ceci peut indiquer une stabilisation des salaires ou peut-être une saturation dans certaines industries où l'expérience supplémentaire au-delà d'un certain point n'entraîne pas nécessairement une augmentation significative de la rémunération.

En synthèse, ce graphique démontre un lien apparent entre l'ancienneté et le salaire moyen, soulignant l'importance de l'expérience dans la progression de carrière. Il suggère que l'accumulation d'années de service peut être un facteur déterminant dans l'évolution des salaires, bien que l'impact de l'ancienneté puisse plafonner après un certain temps.

###Distribution du salaire selon le secteur

```
library(ggplot2)

# Supposons que 'Secteur' est la colonne qui indique si une personne travaille dans le secteur privé ou
# Nous allons exclure les non-salariés qui, dans cet exemple, sont marqués comme 'Non salarie(e)' dans

# Correction du code de filtrage pour exclure les non-salariés
filtered_data <- raw_data[raw_data$Secteur != 'Non_salarie',  ]

# Graphique Boîte à moustaches avec deux couleurs pour le secteur privé et le secteur public
ggplot(filtered_data, aes(x = Secteur, y = Salaire_annuel, fill = Secteur)) +
  geom_boxplot(
    width = 0.8,
    notch = TRUE,
    outlier.shape = NA,
    color = "black", # La couleur des lignes du boxplot
    alpha = 0.7
  ) +
  scale_fill_manual(values = c("Prive" = "pink", "Public" = "green")) + # Deux couleurs pour privé et p
  labs(
    title = "Distribution des salaires par secteur (Privé et Public)",
    x = "Secteur",
    y = "Salaire Annuel"
  ) +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1),
    legend.position = "bottom",
    legend.title = element_blank() # Enlève le titre de la légende
  )
```

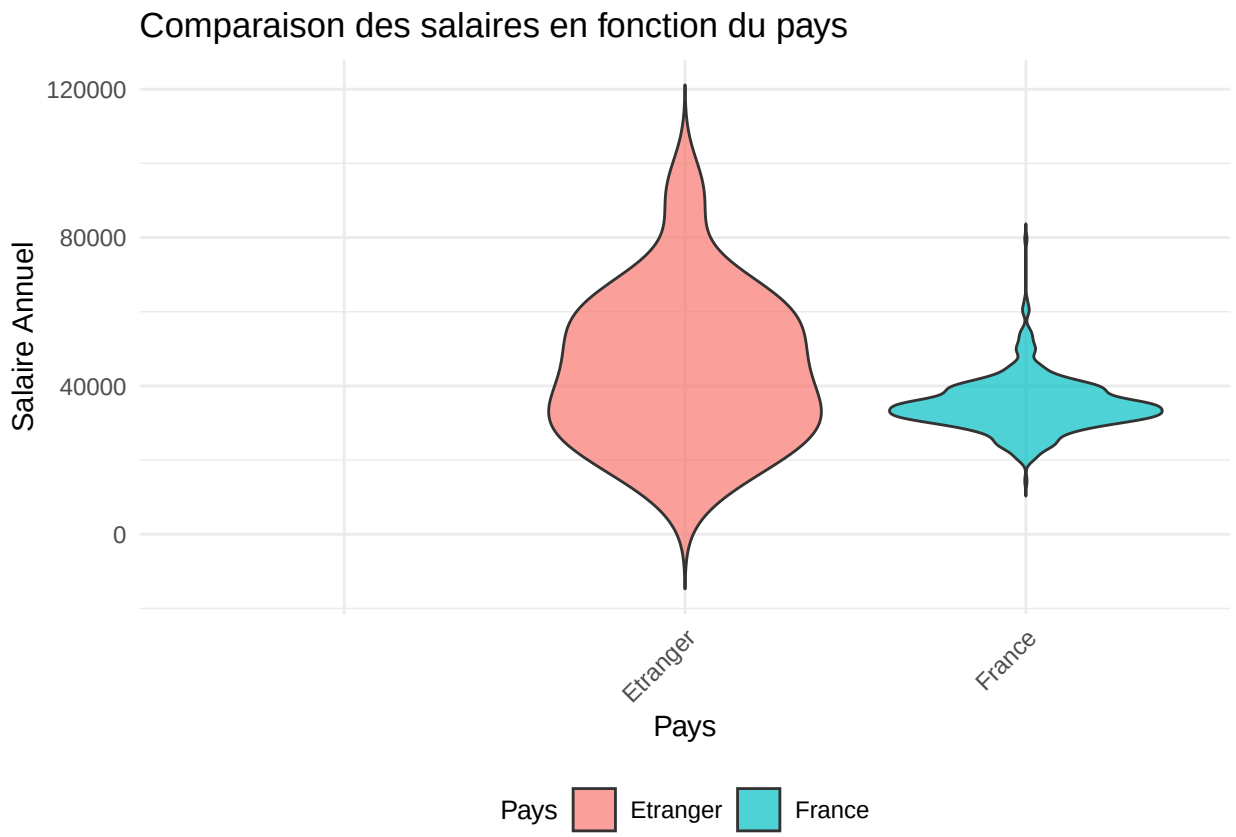
```
## Warning: Removed 206 rows containing non-finite values (`stat_boxplot()`).
```



Distribution du salaire en fonction de France/étranger

```
ggplot(raw_data, aes(x = factor(France), y = Salaire_annuel, fill = factor(France))) +
  geom_violin(trim = FALSE, scale = "width", width = 0.8, alpha = 0.7) +
  labs(
    title = "Comparaison des salaires en fonction du pays",
    x = "Pays",
    y = "Salaire Annuel",
    fill = "Pays"
  ) +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1),
    legend.position = "bottom",
    legend.title = element_text(size = 10)
  )
```

```
## Warning: Removed 208 rows containing non-finite values (`stat_ydensity()`).
```



Le graphique “Comparaison des salaires en fonction du pays” présente les distributions des salaires en France et à l’étranger sous forme de diagrammes en violon. Le violon plus large pour “Étranger” indique une plus grande variété dans les salaires, avec un éventail plus large tant vers les bas que vers les hauts salaires. Cela suggère que travailler à l’étranger peut offrir des opportunités de salaires plus élevés, possiblement en raison de marchés du travail différents ou de primes d’expatriation.

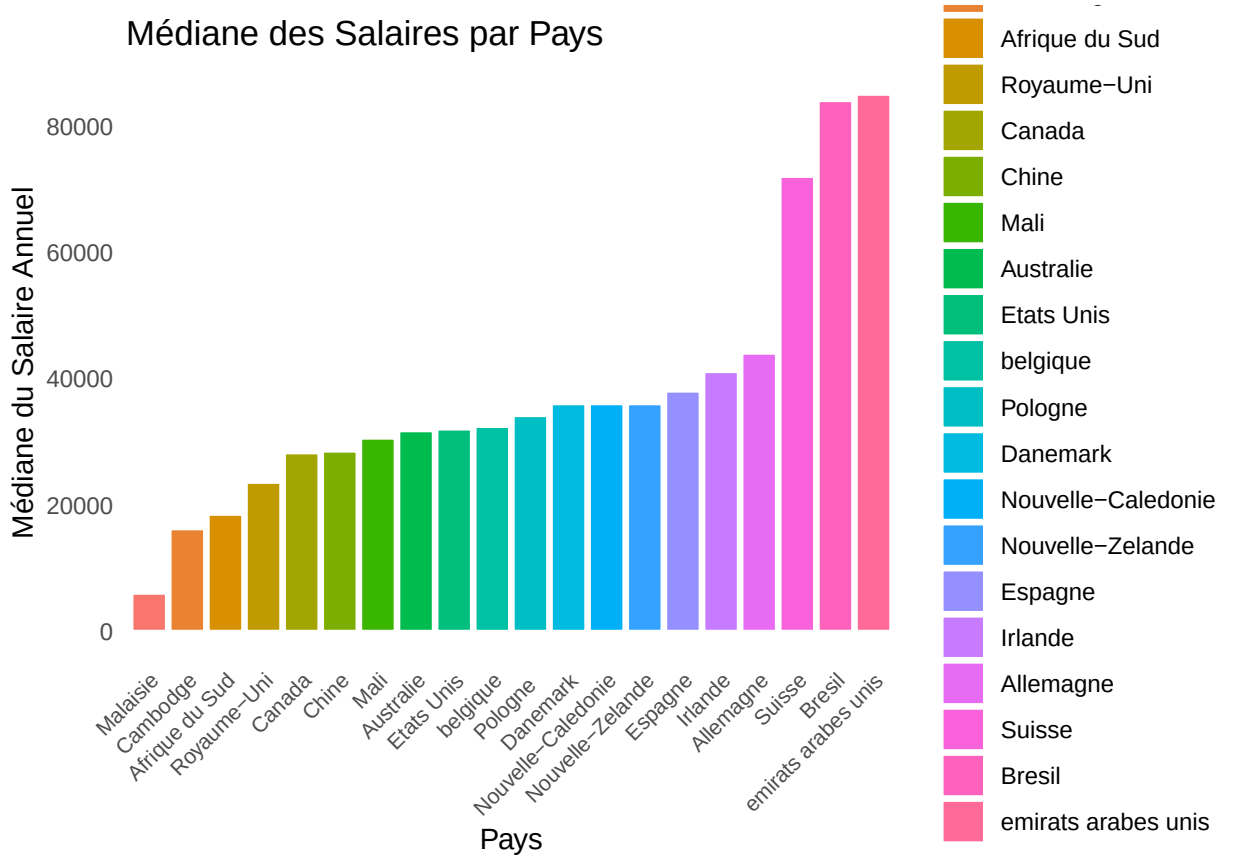
À l’inverse, le violon pour la France est plus étroit, impliquant que les salaires sont plus concentrés autour d’une certaine médiane, avec moins de salaires extrêmement élevés ou bas. Cela pourrait refléter une structure de rémunération plus uniforme ou une cohérence dans les niveaux de salaire au sein du marché du travail français.

Ces observations peuvent aider à comprendre les dynamiques de rémunération et à évaluer l’impact de la géographie sur les salaires.

Distribution du salaire en fonction du pays à l’étranger

```
library(ggplot2)

# Calcul de la médiane des salaires par pays
median_salaries <- tapply(raw_data_filtered$Salaire_annuel, raw_data_filtered$Pays, median)
median_order <- names(sort(median_salaries))

# Création d'un facteur pour l'ordre des pays par médiane
emploiFactor <- factor(raw_data_filtered$Pays, levels = median_order)

# Création du diagramme en bâtonnets
ggplot(raw_data_filtered, aes(x = emploiFactor, y = Salaire_annuel, fill = emploiFactor)) +
```

```r
  geom_bar(stat = "summary", fun = "median", color = "white", size = 0.5) +
  theme_minimal() +  # Fond blanc
  theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 8),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank()) +  # Suppression des lignes de grille
  labs(title = "Médiane des Salaires par Pays", x = "Pays", y = "Médiane du Salaire Annuel",
       fill = "Pays")  # Légende pour les couleurs des pays
```

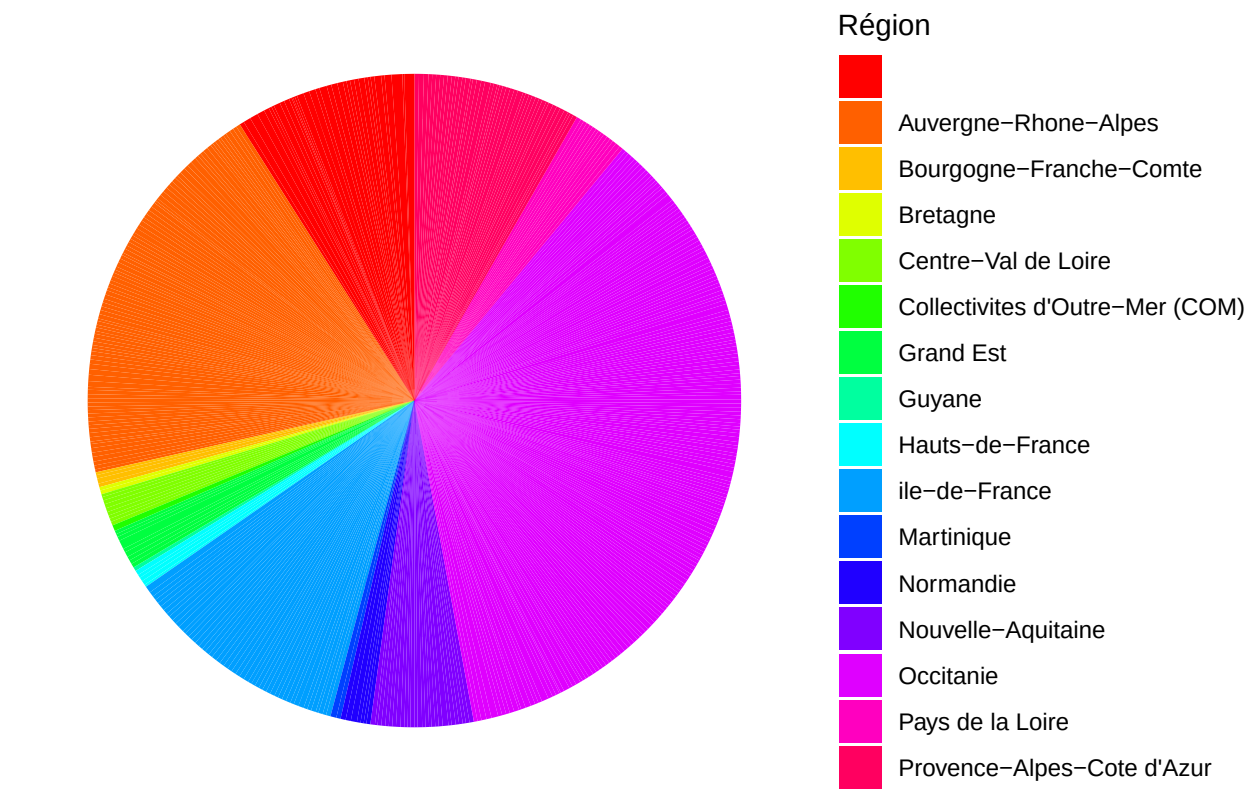


Idem que pour les années précédentes.

Distribution des salaires en fonction de la région

```r
library(ggplot2)

# Remove rows with non-finite values in Salaire_annuel
raw_data <- raw_data[!is.na(raw_data$Salaire_annuel) & is.finite(raw_data$Salaire_annuel), ]

# Remove rows with non-finite values in other numeric columns if needed
# raw_data <- raw_data[complete.cases(raw_data[, numeric_cols]), ]

# Calculate average salary by region
avg_salary_by_region <- tapply(raw_data$Salaire_annuel, raw_data$Region, mean)
max_avg_salary_region <- names(which.max(avg_salary_by_region))

# Define a color palette for each region
color_palette <- rainbow(length(unique(raw_data$Region)))
```

```r
# Plotting the distribution of salaries based on region as a pie chart
ggplot(raw_data, aes(x = "", y = Salaire_annuel, fill = Region)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar("y") +
  theme_void() +  # Removes unnecessary elements
  labs(title = "Répartition des salaires par région",
       fill = "Région",
       y = "Salaire Annuel") +
  scale_fill_manual(values = ifelse(raw_data$Region == max_avg_salary_region, "red", color_palette))  #
```



Idem que pour les années précédentes.

Etude données 2021

Récupération des données

```r
# Trying ISO-8859-1
raw_data <- read.csv("data/data_2021.csv", sep = ",", fileEncoding = "UTF-8")
```

Définir toutes les données comme caractère (variable qualitative) sauf le salaire

```r
# Lists specifying which columns to convert to factors and numeric
factor_cols <- c("Date", "Identifiant", "Genre", "Annee_diplome","Type_formation","Filiere","Situation"
numeric_cols <- c("Anciennete", "Salaire_annuel", "Responsabilite_budget", "Responsabilite_equipe", "Res

# Convert columns to factors
```

```r
raw_data[factor_cols] <- lapply(raw_data[factor_cols], as.factor)

# Convert columns to numeric
raw_data[numeric_cols] <- lapply(raw_data[numeric_cols], as.numeric)
```

```
## Warning in lapply(raw_data[numeric_cols], as.numeric): NAs introduced by
## coercion

## Warning in lapply(raw_data[numeric_cols], as.numeric): NAs introduced by
## coercion

## Warning in lapply(raw_data[numeric_cols], as.numeric): NAs introduced by
## coercion
```

```r
summary(raw_data)
```

Avec ces valeurs on peut déduire beaucoup de choses...

Fonction de filtre

Ceci est une fonction utilisé après pour retirer différentes lignes en fonction des valeurs dans une certain colonne. (équivalent d'un select where)

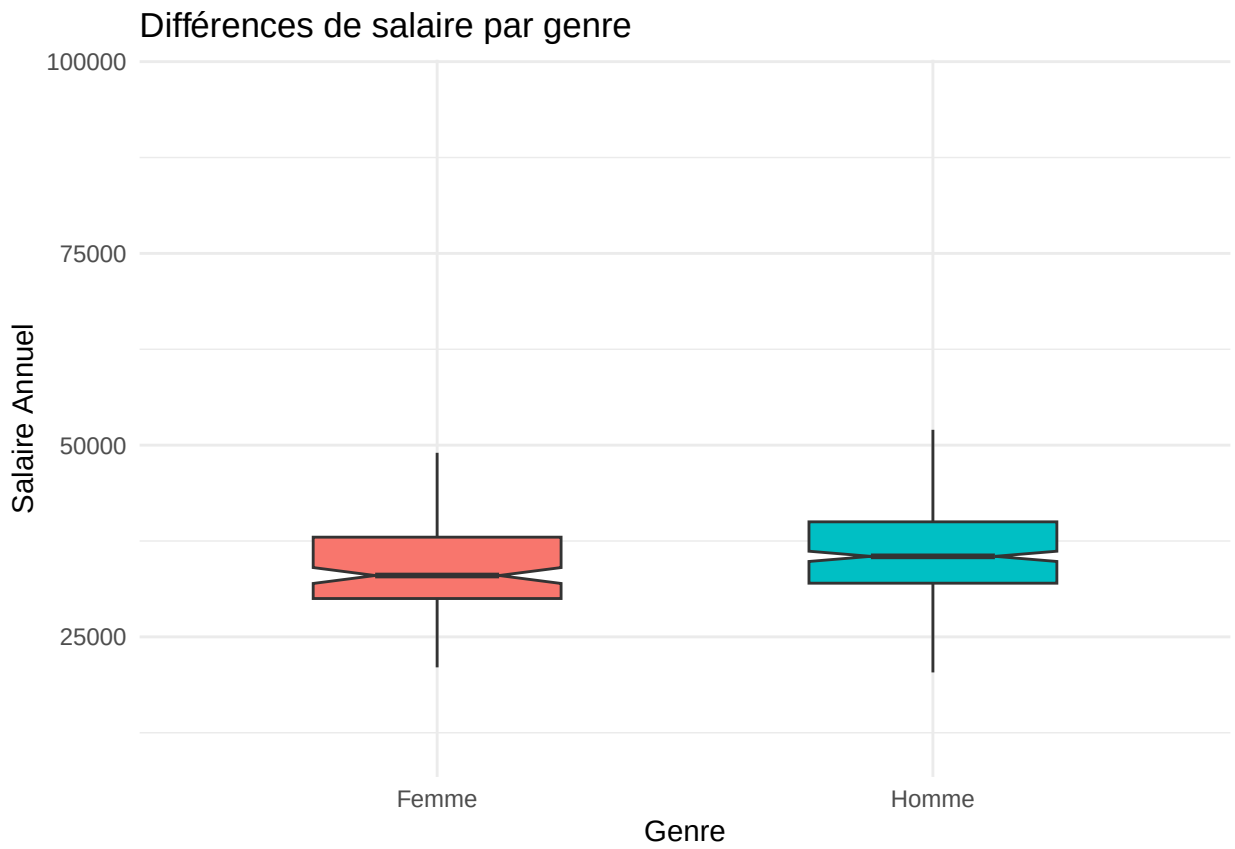
```r
remove_rows_by_value <- function(data, column_name, value_to_remove) {
  data_filtered <- subset(data, !(data[[column_name]] == value_to_remove))
  return(data_filtered)
}
```

Première étude

Disparité homme/femme - toute formation confondue

```r
ggplot(raw_data, aes(x = Genre, y = Salaire_annuel, fill = Genre)) +
  geom_boxplot(width = 0.5, notch = TRUE, outlier.shape = NA) +
  labs(title = "Différences de salaire par genre",
       x = "Genre",
       y = "Salaire Annuel") +
  theme_minimal() +
  theme(legend.position = "none")
```

```
## Warning: Removed 253 rows containing non-finite values (`stat_boxplot()`).
```



En observant les boîtes à moustaches fournies pour les différentes années, une caractéristique notable est l'étendue plus grande de la partie supérieure des boîtes, particulièrement pour les hommes. Cette étendue supérieure, aussi connue sous le nom de troisième quartile, montre le salaire au-dessus duquel se situent les 25 % des salariés les mieux payés.

Pour les hommes, il est visible que la distribution des salaires s'étale sur une plage plus large au-dessus de la médiane, ce qui indique que non seulement il y a des hommes qui gagnent plus que la médiane, mais aussi que parmi eux, il existe une variété considérable de salaires élevés. En comparaison, pour les femmes, bien que la distribution au-dessus de la médiane soit présente, elle est moins étendue, suggérant que les salaires élevés sont moins fréquents ou moins dispersés.

Cette observation peut signaler que, dans la population étudiée, il y a une plus grande hétérogénéité dans les niveaux de rémunération des hommes au-dessus de la médiane salariale. Cela pourrait indiquer une présence accrue d'hommes dans des rôles de haute rémunération ou des positions de leadership, ou encore refléter des différences dans les opportunités de carrière ou les négociations salariales qui favorisent les hommes.

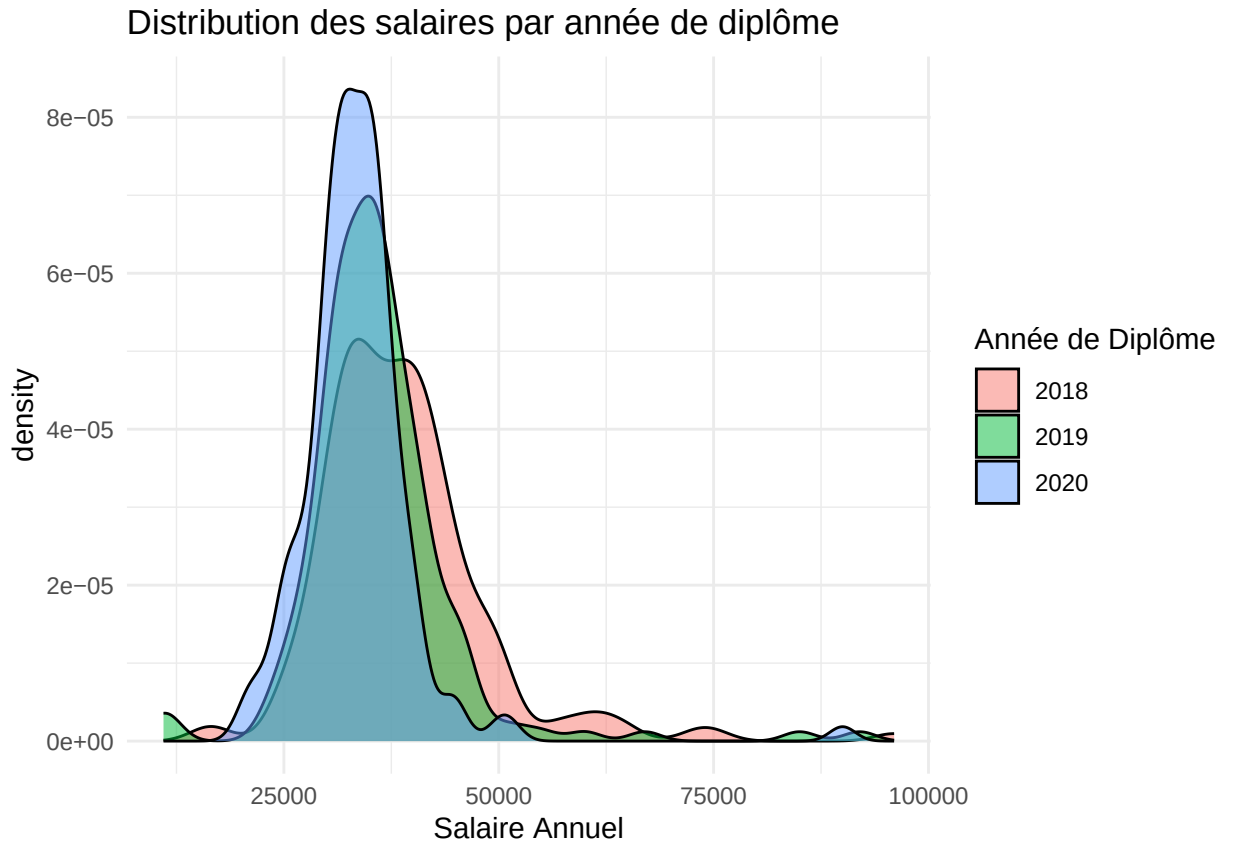
Il est également possible que des facteurs tels que l'ancienneté, le secteur d'activité ou la spécialisation professionnelle influencent cette distribution. Les femmes peuvent être sous-représentées dans des postes ou des industries à haute rémunération, ou des barrières systémiques pourraient limiter leur progression salariale au-delà d'un certain seuil. Ces graphiques suggèrent l'importance d'une analyse plus approfondie pour comprendre et aborder les dynamiques qui conduisent à ces disparités de salaire entre les genres.

Disparité sur les dates d'optention du diplôme

```
ggplot(raw_data, aes(x = Salaire_annuel, fill = as.factor(Annee_diplome))) +
  geom_density(alpha = 0.5) +
  labs(title = "Distribution des salaires par année de diplôme",
```

```
      x = "Salaire Annuel",
      fill = "Année de Diplôme") +
  theme_minimal()
```

```
## Warning: Removed 253 rows containing non-finite values (`stat_density()`).
```



Le premier diagramme, qui présente les données des années de diplôme 2018 à 2020, montre une tendance à la diversification des salaires annuels, particulièrement pour l'année 2020 où la distribution est visiblement plus étendue. Ceci contraste avec les distributions plus concentrées des années 2017 à 2019 et 2016 à 2018, observées dans les autres diagrammes, où les courbes de densité sont plus pointues et moins dispersées.

La courbe élargie pour les diplômés de 2020 suggère une variabilité accrue dans les salaires de départ, ce qui pourrait indiquer une diversification des rôles disponibles pour les nouveaux diplômés ou refléter l'impact de facteurs économiques tels que les fluctuations du marché du travail.

La cohorte de 2018, qui apparaît dans tous les graphiques, offre une référence de comparaison intéressante. La stabilité relative de sa distribution suggère que les conditions du marché pour cette cohorte n'ont pas changé de manière significative au fil du temps, du moins en termes de salaires de départ.

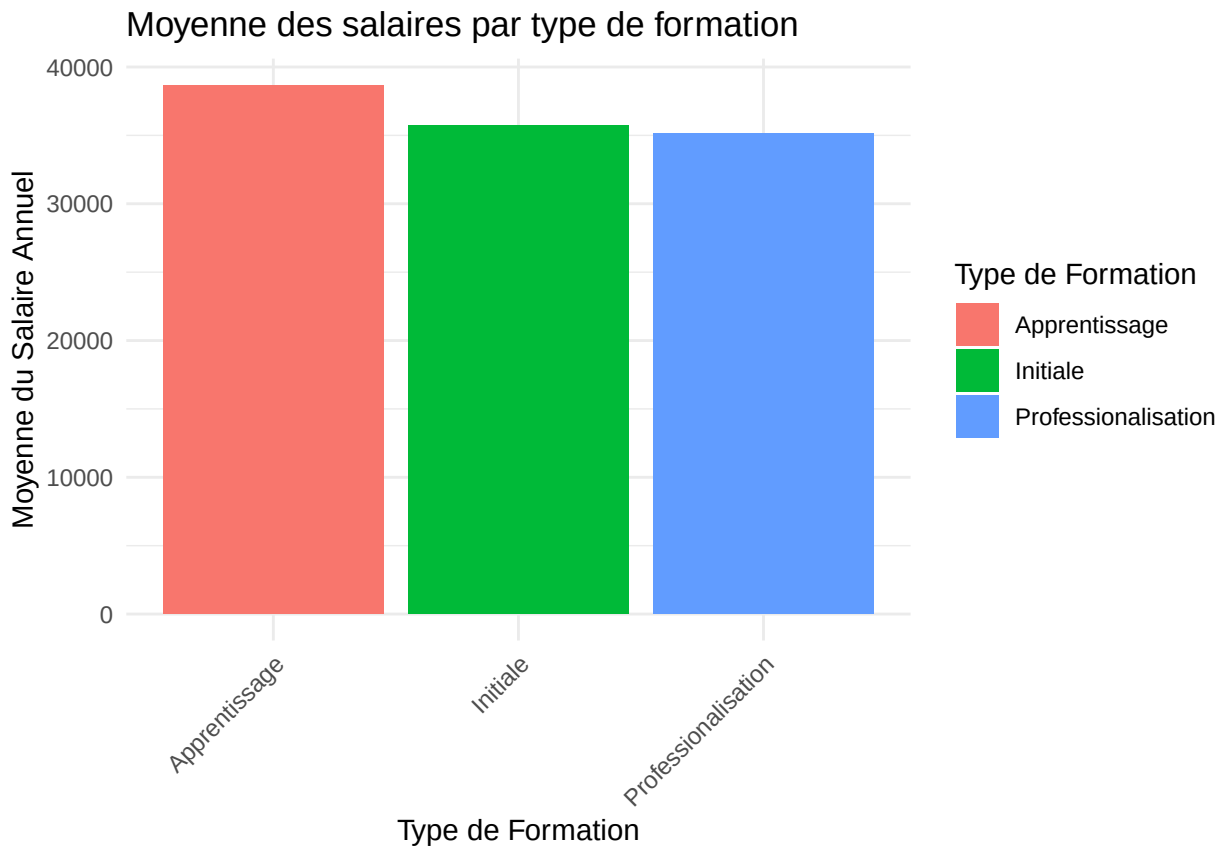
Cette analyse globale indique que bien que les conditions du marché pour les diplômés de 2018 soient restées constantes, les diplômés de 2020 sont entrés sur un marché du travail potentiellement différent, avec une gamme de salaires initiaux plus large, pointant vers une évolution des conditions économiques ou des politiques d'embauche qui pourraient avoir influencé la structure salariale des nouveaux entrants.

Disparité formation initiale/apprentissage/contrat de professionalisation

```
library(ggplot2)
```

```r
# Supprimer les lignes où 'Type_formation' est NA ou vide
clean_data <- raw_data[!is.na(raw_data$Type_formation) & raw_data$Type_formation != "", ]

# Créer le graphique avec les données nettoyées
ggplot(clean_data, aes(x = as.factor(Type_formation), y = Salaire_annuel, fill = as.factor(Type_formati
  stat_summary(fun = "mean", geom = "col") +
  labs(title = "Moyenne des salaires par type de formation",
       x = "Type de Formation",
       y = "Moyenne du Salaire Annuel",
       fill = "Type de Formation") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
## Warning: Removed 253 rows containing non-finite values (`stat_summary()`).
```



En examinant le premier diagramme et en le comparant aux deux autres, nous constatons une inversion dans la hiérarchie des salaires moyens par type de formation. Dans le premier diagramme, le type de formation par apprentissage montre le salaire moyen le plus élevé, suivi de la professionnalisation et enfin de la formation initiale, qui présente le salaire moyen le plus bas.

Cette tendance est à l'opposé de ce qui est observé dans les deux autres diagrammes, où la professionnalisation mène généralement avec le salaire moyen le plus élevé, et l'apprentissage est le plus bas. Cette inversion pourrait être le résultat de plusieurs facteurs, comme des changements dans la perception de la valeur de l'apprentissage sur le marché du travail, des modifications dans les structures de rémunération, ou des évolutions dans les opportunités de carrière post-apprentissage.

Il est également possible que cette différence soit due à des variations annuelles spécifiques, des changements

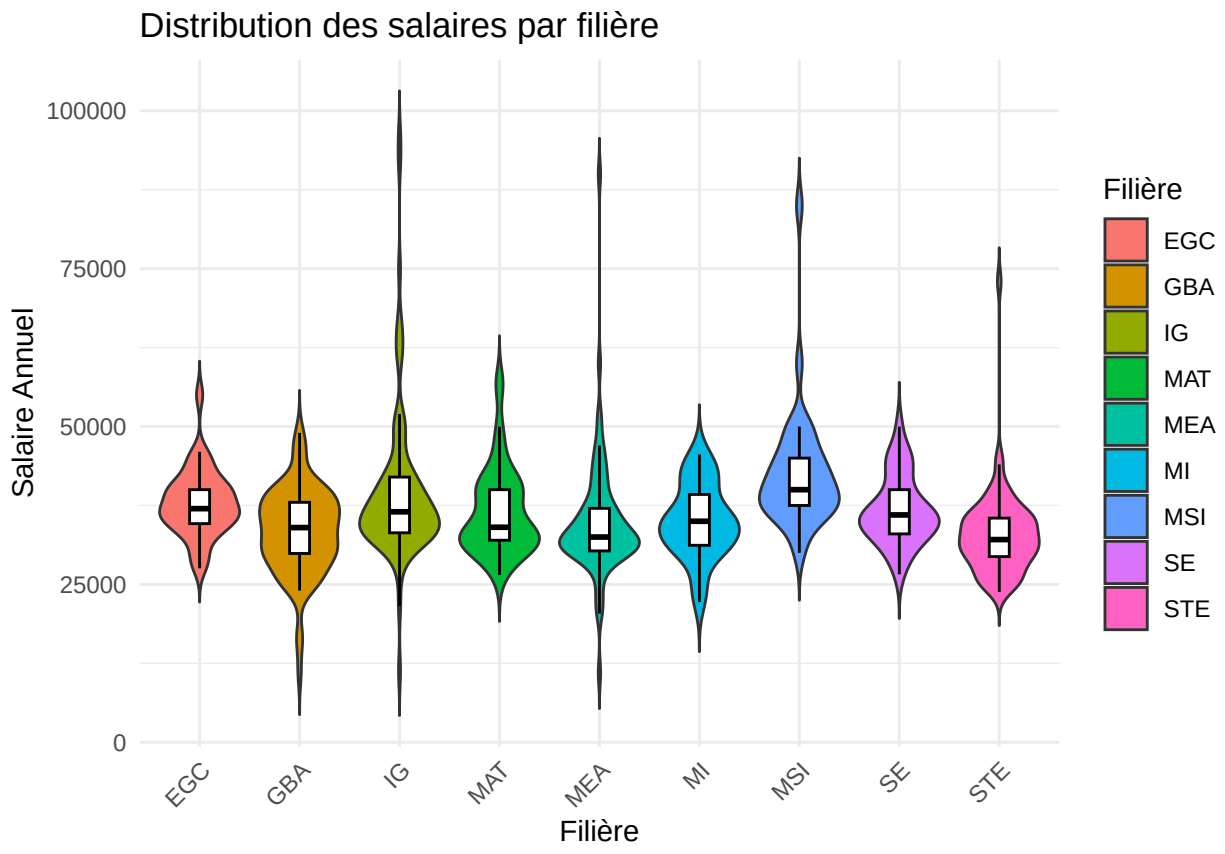
dans les secteurs économiques qui embauchent des apprentis, ou des politiques incitatives qui valorisent l'apprentissage. Pour comprendre pleinement ces tendances, une analyse plus détaillée qui inclut d'autres variables et contextes de marché serait nécessaire.

Disparité des filières

```
ggplot(raw_data, aes(x = Filiere, y = Salaire_annuel, fill = Filiere)) +
  geom_violin(trim = FALSE, scale = "width", width = 0.8) +
  geom_boxplot(width = 0.2, fill = "white", color = "black", outlier.shape = NA) +
  labs(title = "Distribution des salaires par filière",
       x = "Filière",
       y = "Salaire Annuel",
       fill = "Filière") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
## Warning: Removed 253 rows containing non-finite values (`stat_ydensity()`).
```

```
## Warning: Removed 253 rows containing non-finite values (`stat_boxplot()`).
```



La distribution des salaires par filière, représentée sous forme de diagrammes en violons, fournit une vue d'ensemble des tendances salariales au sein d'un échantillon de travailleurs. Chaque violon révèle non seulement la médiane des salaires pour une filière donnée mais aussi la répartition complète des salaires, de la base à la cime.

Le violon de la filière MSI (en violet) se distingue par une médiane élevée, indiquant que les salaires médians dans cette filière sont parmi les plus hauts de l'échantillon. La largeur du violon en son sommet suggère également une dispersion significative des salaires les plus élevés, ce qui pourrait correspondre à des postes

hautement qualifiés ou à des responsabilités managériales importantes.

La filière IG (en vert) présente des caractéristiques similaires à la filière MSI en termes de médiane et de distribution des salaires élevés, ce qui indique que les compétences acquises dans cette filière sont également très valorisées sur le marché du travail.

Les filières GBA (en jaune) et STE (en rose) se caractérisent par des médianes inférieures, reflétant des salaires globalement plus modestes. La forme resserrée du violon pour GBA indique une concentration des salaires autour de la médiane, tandis que le violon de la filière STE montre une certaine variabilité, avec des salaires s'étendant vers le haut malgré une médiane plus basse.

Quant à la filière MEA (en bleu), elle affiche une médiane centrale avec une queue supérieure étendue, ce qui dénote la présence de salaires élevés, peut-être dus à des spécialisations ou à des postes de niveau supérieur au sein de la filière.

Les filières MAT, MI, et SE montrent des médianes intermédiaires et une variabilité des salaires moins marquée que pour les filières MSI et IG, mais avec une présence de salaires supérieurs qui reste notable.

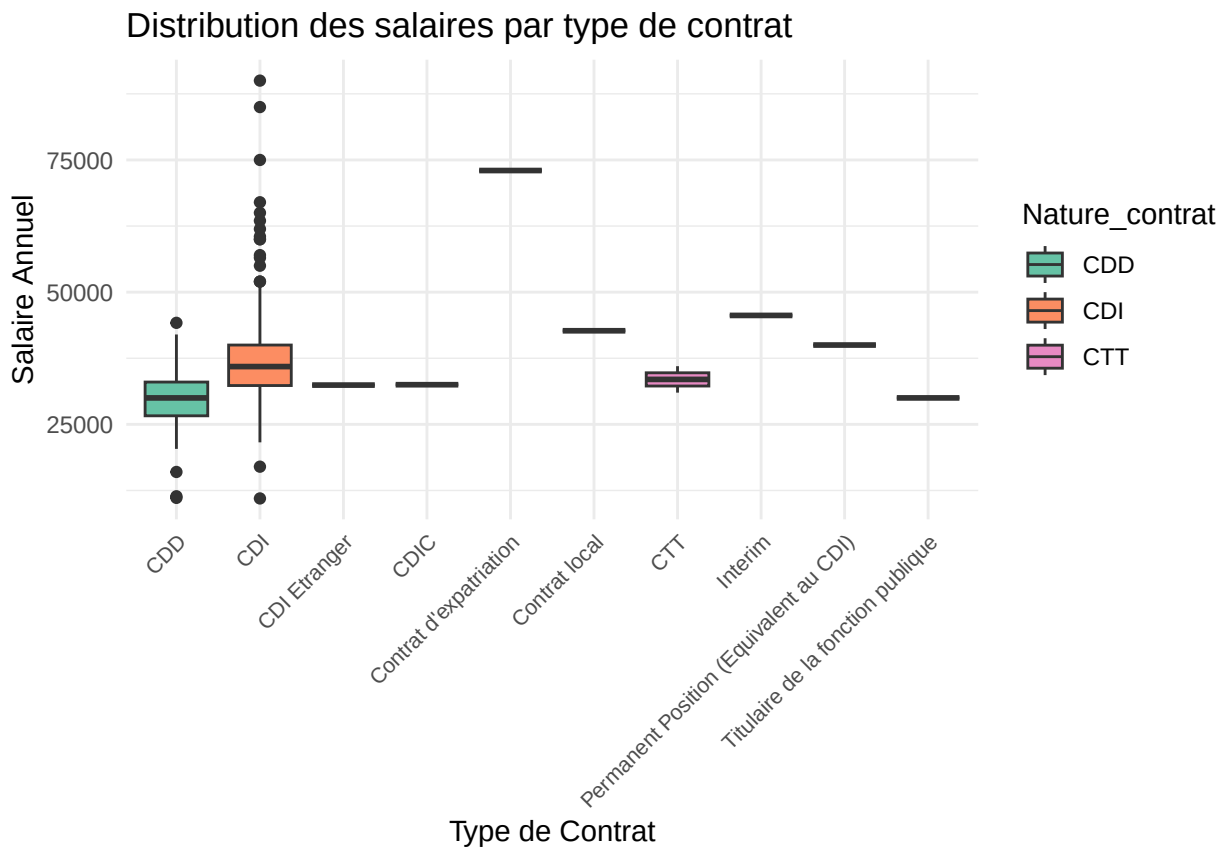
En somme, ces diagrammes en violons illustrent non seulement les niveaux de salaires moyens par filière mais aussi l'étendue et la dispersion des salaires au sein de chacune. Ils mettent en lumière les filières où les salaires sont non seulement élevés en médiane mais aussi largement distribués vers le haut, signalant la présence de parcours professionnels avantageux et diversifiés. Cela souligne l'importance des compétences spécialisées et de l'expérience dans la détermination des salaires dans le monde professionnel.

Filtrage des personnes en activité

```
filtered_data <- remove_rows_by_value(raw_data, "Situation", "En recherche emploi")
filtered_data <- remove_rows_by_value(filtered_data, "Situation", "En poursuite d'études (hors thèse) /
filtered_data <- remove_rows_by_value(filtered_data, "Situation", "These")
filtered_data <- remove_rows_by_value(filtered_data, "Situation", "Sans activité")
filtered_data <- remove_rows_by_value(filtered_data, "Situation", "Volontariat")
filtered_data <- remove_rows_by_value(filtered_data, "Situation", "En création d'entreprise /reprise d'
filtered_data <- remove_rows_by_value(filtered_data, "Situation", "Etudes")

activite_data <- filtered_data
```

Disparité par nature du contrat

```
library(ggplot2)

# Remplacer 'activite_data' par votre dataframe réel
# Filtrage de certains contrats peu représentatifs
filtered_data <- raw_data[!raw_data$Nature_contrat %in% c("Service à la personne (cours particulier de r

# Supposons que les catégories de contrats sont 'CDD', 'CDI', 'Intérim', etc.
# Vous pouvez définir manuellement les couleurs pour chaque catégorie de contrat
couleurs_contrats <- c("CDD" = "#66c2a5", "CDI" = "#fc8d62", "Intérim" = "#8da0cb",
                       "Stage" = "#e78ac3", "Alternance" = "#a6d854", "CTT" = "#e78ac3")

# Diagramme en boîte avec couleurs
ggplot(filtered_data, aes(x = Nature_contrat, y = Salaire_annuel, fill = Nature_contrat)) +
  geom_boxplot() +
  scale_fill_manual(values = couleurs_contrats) +  # Utiliser les couleurs définies précédemment
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 8)) +
  labs(title = "Distribution des salaires par type de contrat", x = "Type de Contrat", y = "Salaire Annu
```



Sur le diagramme “Distribution des salaires par type de contrat”, nous observons les salaires annuels associés à différents types de contrats de travail. La médiane des salaires, illustrée par la ligne au centre de chaque boîte, permet de comparer les niveaux de rémunération médians entre les types de contrats.

Les contrats à durée déterminée (CDD) présentent une médiane inférieure et une distribution relativement concentrée, indiquant une uniformité dans les salaires mais également des niveaux de rémunération généralement plus bas. Ceci est attendu étant donné la nature temporaire et souvent précaire des CDD.

Les contrats à durée indéterminée (CDI) affichent une médiane plus élevée, soulignant le salaire plus stable et souvent plus avantageux associé à la sécurité de l'emploi à long terme. La variabilité des salaires est également plus large, reflétant une diversité de postes et de niveaux de responsabilité.

Les CDI à l'étranger semblent offrir des salaires encore plus élevés, ce qui pourrait être attribué aux primes d'expatriation ou aux différences de marchés du travail entre pays.

Les contrats d'expatriation montrent une distribution similaire aux CDI étrangers, avec des salaires médians élevés, ce qui est cohérent avec les avantages financiers souvent associés à l'expatriation.

Les contrats locaux, possiblement des contrats standards dans le pays d'emploi, présentent une médiane plus basse, suggérant des salaires moins compétitifs ou des rôles moins spécialisés.

Les contrats de travail temporaire (CTT) et les postes intérimaires indiquent des médianes plus faibles et des distributions étroites, alignées sur la nature éphémère et les rémunérations moins élevées de ces emplois.

Les postes permanents équivalents à un CDI révèlent une médiane et une distribution comparables aux CDI, ce qui suggère des conditions d'emploi et de rémunération similaires.

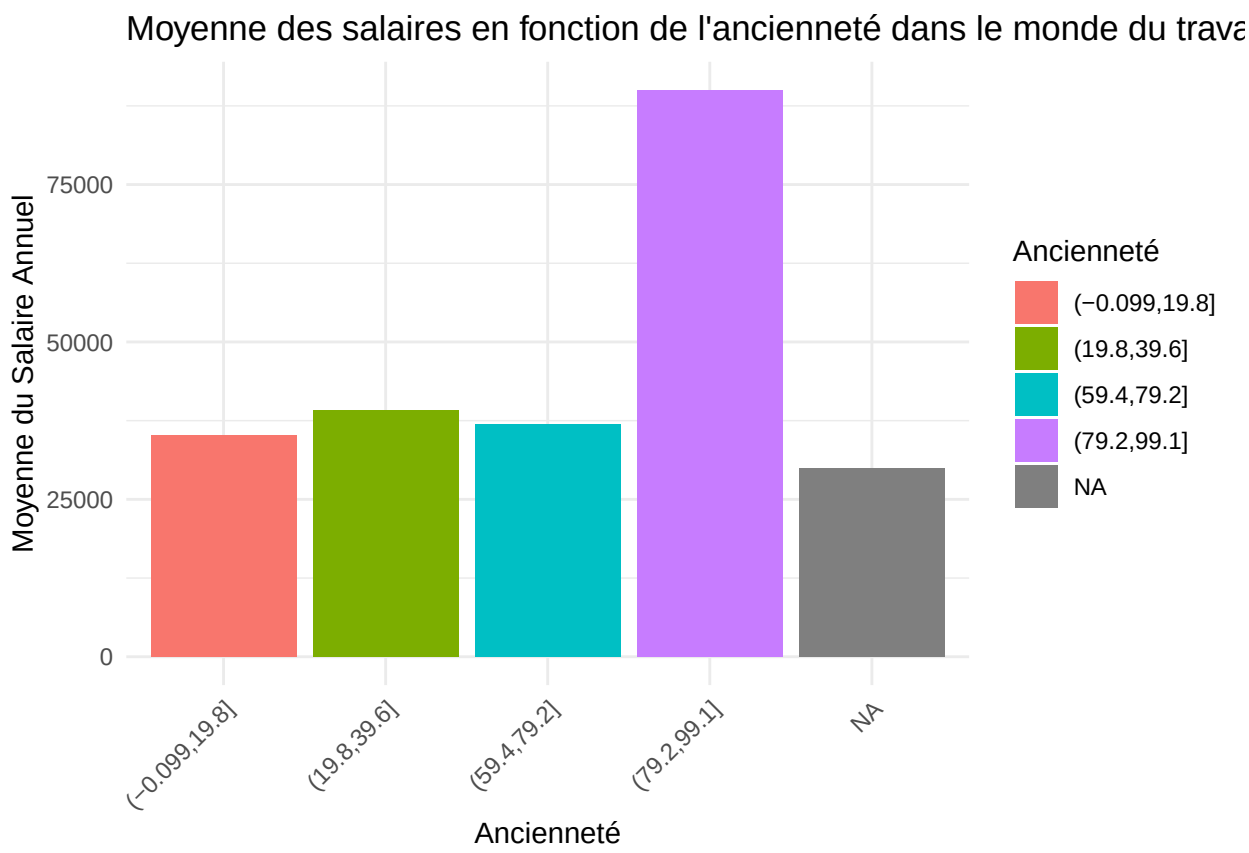
Les titulaires de la fonction publique présentent une distribution des salaires plus resserrée, ce qui peut refléter une grille salariale plus standardisée dans le secteur public.

En résumé, le graphique met en exergue le fait que les CDI, tant nationaux qu'internationaux, ainsi que les postes permanents et les fonctionnaires, tendent à offrir des salaires médians supérieurs par rapport aux contrats temporaires et locaux, illustrant l'influence significative de la stabilité de l'emploi et de la localisation géographique sur les niveaux de salaire.

Distribution du salaire en fonction de l'ancienneté

```
library(ggplot2)


ordered_data <- raw_data[order(as.numeric(raw_data$Anciennete)), ]

# Création de catégories d'ancienneté (facultatif)
ordered_data$Anciennete_category <- cut(as.numeric(ordered_data$Anciennete), breaks = 5)

# Diagramme en barres pour la moyenne des salaires en fonction de l'ancienneté
ggplot(ordered_data, aes(x = Anciennete_category, y = Salaire_annuel, fill = Anciennete_category)) +
  geom_bar(stat = "summary", fun = "mean", position = "dodge") +
  labs(title = "Moyenne des salaires en fonction de l'ancienneté dans le monde du travail",
       x = "Ancienneté",
       y = "Moyenne du Salaire Annuel",
       fill = "Ancienneté") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
## Warning: Removed 253 rows containing non-finite values (`stat_summary()`).
```



Le graphique “Moyenne des salaires en fonction de l’ancienneté” illustre comment l’expérience professionnelle accumulée influence la rémunération moyenne des employés. L’ancienneté, divisée en segments d’un an, sert de mesure pour établir une corrélation potentielle entre le temps passé dans une organisation ou un secteur et le salaire moyen perçu.

Les employés avec moins d’un an d’expérience, représentés par la barre rouge, perçoivent les salaires les plus bas de l’échelle, ce qui est attendu pour des positions souvent associées à des rôles débutants ou en phase d’apprentissage. À mesure que l’ancienneté augmente, les barres s’élèvent progressivement : la tranche d’un à deux ans (verte) montre une augmentation modeste du salaire, suggérant un premier palier d’évolution salariale.

Ce phénomène de croissance se poursuit de manière plus marquée pour les tranches de deux à trois ans (bleue) et de trois à quatre ans (violet), où l’on observe un bond significatif dans les salaires moyens. Cela peut refléter les promotions, les augmentations méritées ou l’acquisition de compétences spécialisées qui sont souvent récompensées par des ajustements salariaux positifs.

Curieusement, la tranche des employés les plus anciens, ceux ayant entre quatre et cinq ans d’expérience (rose), bien qu’affichant un salaire moyen élevé, ne représente pas le sommet de l’échelle salariale. Ceci peut indiquer une stabilisation des salaires ou peut-être une saturation dans certaines industries où l’expérience supplémentaire au-delà d’un certain point n’entraîne pas nécessairement une augmentation significative de la rémunération.

En synthèse, ce graphique démontre un lien apparent entre l’ancienneté et le salaire moyen, soulignant l’importance de l’expérience dans la progression de carrière. Il suggère que l’accumulation d’années de service peut être un facteur déterminant dans l’évolution des salaires, bien que l’impact de l’ancienneté puisse plafonner après un certain temps.

###Distribution du salaire selon le secteur

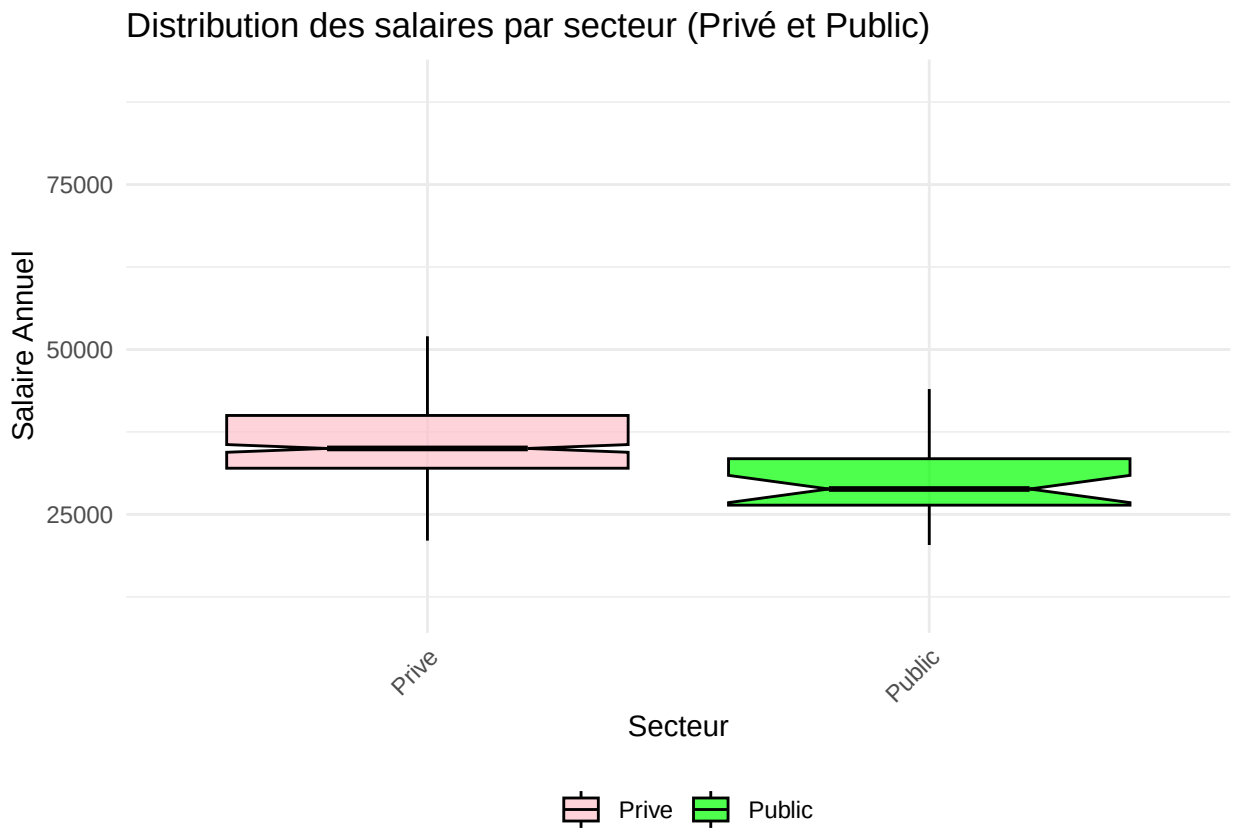
```
library(ggplot2)

# Vérifiez toutes les valeurs uniques pour identifier les valeurs non désirées
unique(raw_data$Secteur)
```

```
## [1]                Prive           Public          Non salarie(e)
## Levels:  Non salarie(e) Prive Public
```

```
# Filtrez le DataFrame pour ne garder que les lignes avec "Privé" et "Public"
filtered_data <- raw_data[raw_data$Secteur %in% c("Prive", "Public"), ]

# Continuez avec la création du graphique ggplot
ggplot(filtered_data, aes(x = Secteur, y = Salaire_annuel, fill = Secteur)) +
  geom_boxplot(
    width = 0.8,
    notch = TRUE,
    outlier.shape = NA,
    color = "black",
    alpha = 0.7
  ) +
  scale_fill_manual(values = c("Prive" = "pink", "Public" = "green")) +
  labs(
    title = "Distribution des salaires par secteur (Privé et Public)",
    x = "Secteur",
    y = "Salaire Annuel"
  ) +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1),
```

```
    legend.position = "bottom",
    legend.title = element_blank()
  )
```

```
## Warning: Removed 48 rows containing non-finite values (`stat_boxplot()`).
```

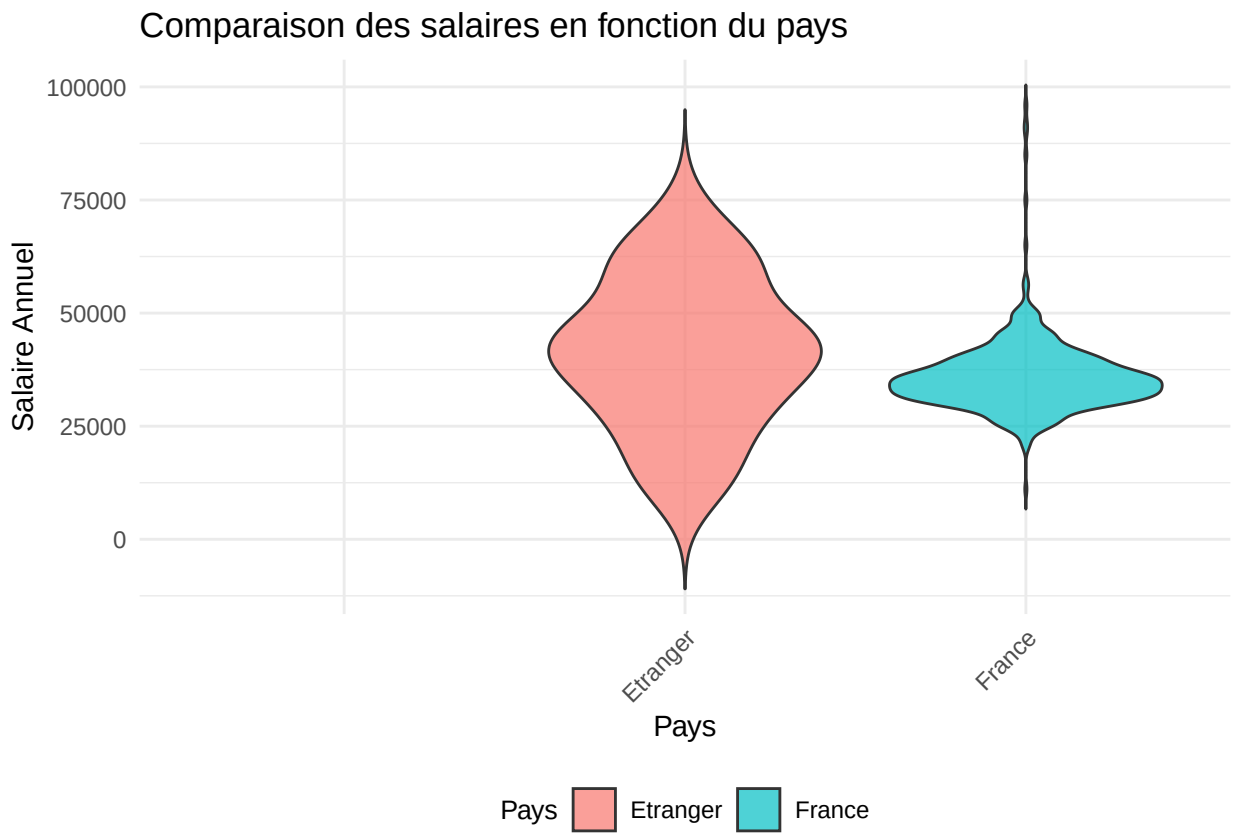


Idem que les années précédentes.

Distribution du salaire en fonction de France/étranger

```
ggplot(raw_data, aes(x = factor(France), y = Salaire_annuel, fill = factor(France))) +
  geom_violin(trim = FALSE, scale = "width", width = 0.8, alpha = 0.7) +
  labs(
    title = "Comparaison des salaires en fonction du pays",
    x = "Pays",
    y = "Salaire Annuel",
    fill = "Pays"
  ) +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1),
    legend.position = "bottom",
    legend.title = element_text(size = 10)
  )
```

```
## Warning: Removed 253 rows containing non-finite values (`stat_ydensity()`).
```



Le diagramme “Comparaison des salaires en fonction du pays” met en lumière les différences entre les salaires annuels en France et à l’étranger. Les violons représentent la densité des données, montrant où les salaires se concentrent pour chaque groupe.

Pour “Étranger”, le violon s’élargit considérablement au milieu, indiquant que la majorité des salaires se situe autour de ce point central, qui est nettement supérieur à celui de la France. Cela suggère que le salaire médian à l’étranger est plus élevé. La largeur du violon à l’étranger suggère également une plus grande variabilité des salaires, avec des valeurs extrêmes plus hautes illustrées par la longue queue supérieure. Cela peut indiquer que les opportunités de salaires élevés sont plus nombreuses à l’étranger ou que certains secteurs ou rôles internationaux offrent des primes significatives.

En revanche, le violon représentant la France montre une concentration plus étroite des salaires, avec une densité plus élevée autour de la médiane et moins de valeurs extrêmes. La queue supérieure est plus courte, ce qui indique que moins de travailleurs atteignent les salaires les plus élevés comparativement à leurs homologues à l’étranger.

Cette visualisation pourrait refléter les dynamiques du marché du travail global, où le travail à l’étranger peut être associé à des rémunérations plus compétitives ou à des avantages liés à l’expatriation. Elle pourrait aussi pointer vers des différences dans la structure des industries, des niveaux de vie ou des politiques salariales entre la France et d’autres pays.

Etude données 2022

Récupération des données

```r
# Trying ISO-8859-1
raw_data <- read.csv("data/data_2022.csv", sep = ",", fileEncoding = "UTF-8")
```

Définir toutes les données comme caractère (variable qualitative) sauf le salaire

```
# Lists specifying which columns to convert to factors and numeric
factor_cols <- c("Date", "Identifiant", "Genre", "Annee_diplome","Type_formation","Filiere","Situation_a
numeric_cols <- c("Anciennete","Salaire_annuel_brut", "Salaire_annuel_brut_primes", "Responsabilite_hie

# Convert columns to factors
raw_data[factor_cols] <- lapply(raw_data[factor_cols], as.factor)

# Convert columns to numeric
raw_data[numeric_cols] <- lapply(raw_data[numeric_cols], as.numeric)
```

```
## Warning in lapply(raw_data[numeric_cols], as.numeric): NAs introduced by
## coercion

## Warning in lapply(raw_data[numeric_cols], as.numeric): NAs introduced by
## coercion

## Warning in lapply(raw_data[numeric_cols], as.numeric): NAs introduced by
## coercion

## Warning in lapply(raw_data[numeric_cols], as.numeric): NAs introduced by
## coercion

## Warning in lapply(raw_data[numeric_cols], as.numeric): NAs introduced by
## coercion

## Warning in lapply(raw_data[numeric_cols], as.numeric): NAs introduced by
## coercion
```

```
summary(raw_data)
```

Avec ces valeurs on peut déduire beaucoup de choses...

Fonction de filtre

Ceci est une fonction utilisé après pour retirer différentes lignes en fonction des valeurs dans une certain colonne. (équivalent d'un select where)

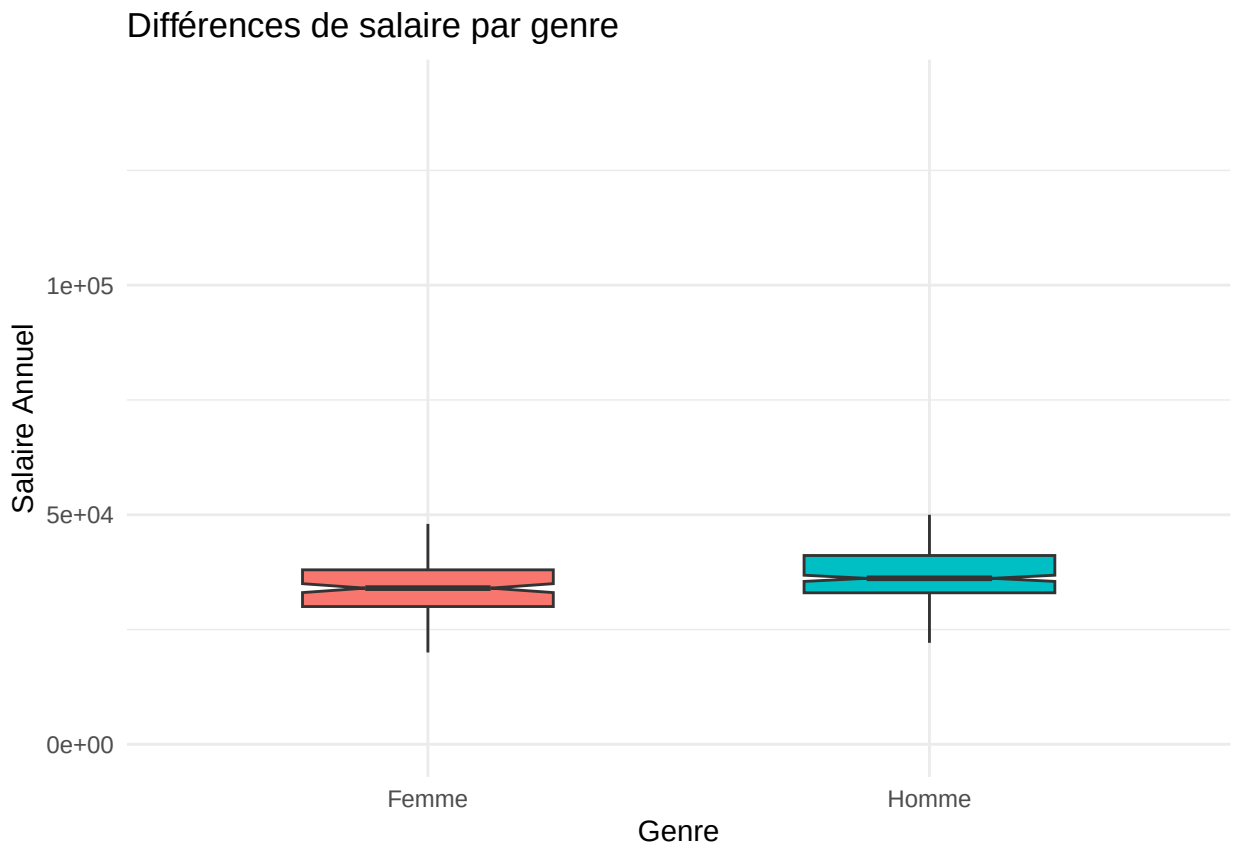
```
remove_rows_by_value <- function(data, column_name, value_to_remove) {
  data_filtered <- subset(data, !(data[[column_name]] == value_to_remove))
  return(data_filtered)
}
```

Première étude

Disparité homme/femme - toute formation confondue

```
ggplot(raw_data, aes(x = Genre, y = Salaire_annuel_brut_primes, fill = Genre)) +
  geom_boxplot(width = 0.5, notch = TRUE, outlier.shape = NA) +
  labs(title = "Différences de salaire par genre",
       x = "Genre",
       y = "Salaire Annuel") +
  theme_minimal() +
  theme(legend.position = "none")
```

```
## Warning: Removed 250 rows containing non-finite values (`stat_boxplot()`).
```

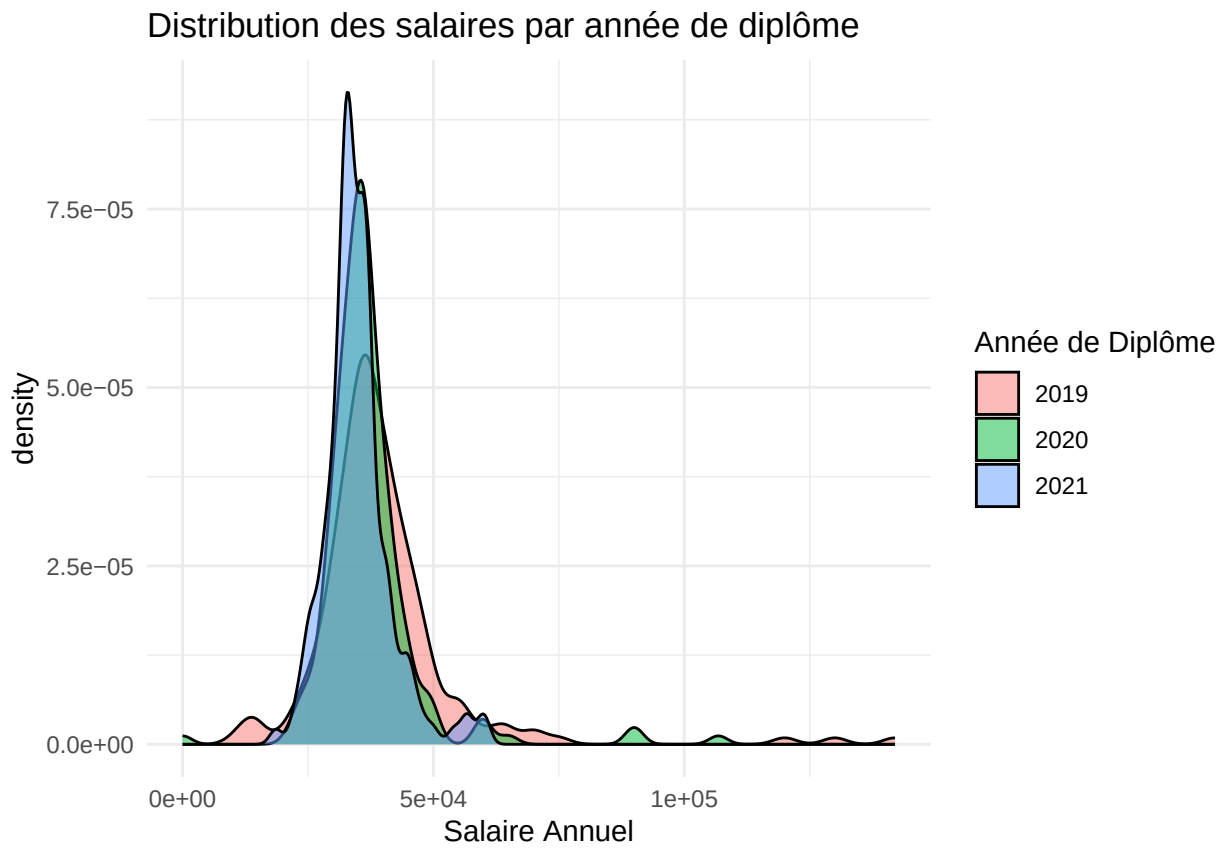


Idem que les autres années.

Disparité sur les dates d'optention du diplôme

```
ggplot(raw_data, aes(x = Salaire_annuel_brut_primes, fill = as.factor(Annee_diplome))) +
  geom_density(alpha = 0.5) +
  labs(title = "Distribution des salaires par année de diplôme",
       x = "Salaire Annuel",
       fill = "Année de Diplôme") +
  theme_minimal()
```

```
## Warning: Removed 250 rows containing non-finite values (`stat_density()`).
```



La première distribution des salaires par année de diplôme pour les années 2019, 2020 et 2021 montre une répartition des données avec un pic très prononcé pour l'année 2019, indiquant une concentration élevée des salaires autour d'une gamme spécifique. En comparaison, les années 2020 et 2021 présentent des distributions moins pointues et plus étalées, suggérant une plus grande variabilité dans les salaires pour ces cohortes.

Lorsque l'on compare cette distribution avec les deux autres distributions des années 2017, 2018, et 2019, puis 2018, 2019, et 2020, on observe une tendance intéressante. Les diplômés des années précédentes (2017 et 2018) ont des distributions avec des pics moins marqués et des densités qui s'étalent sur une plus large gamme de salaires, ce qui pourrait indiquer que les salaires ont eu le temps de s'ajuster avec l'expérience accumulée, menant à une plus grande dispersion.

En particulier, le pic pour l'année la plus récente dans chaque graphique est toujours plus prononcé, ce qui est cohérent avec des entrées récentes sur le marché du travail où les salaires sont moins dispersés. À mesure que les cohortes vieillissent, les pics deviennent moins prononcés et les distributions plus larges, reflétant probablement une diversification des parcours professionnels et des augmentations de salaire avec l'avancement en carrière.

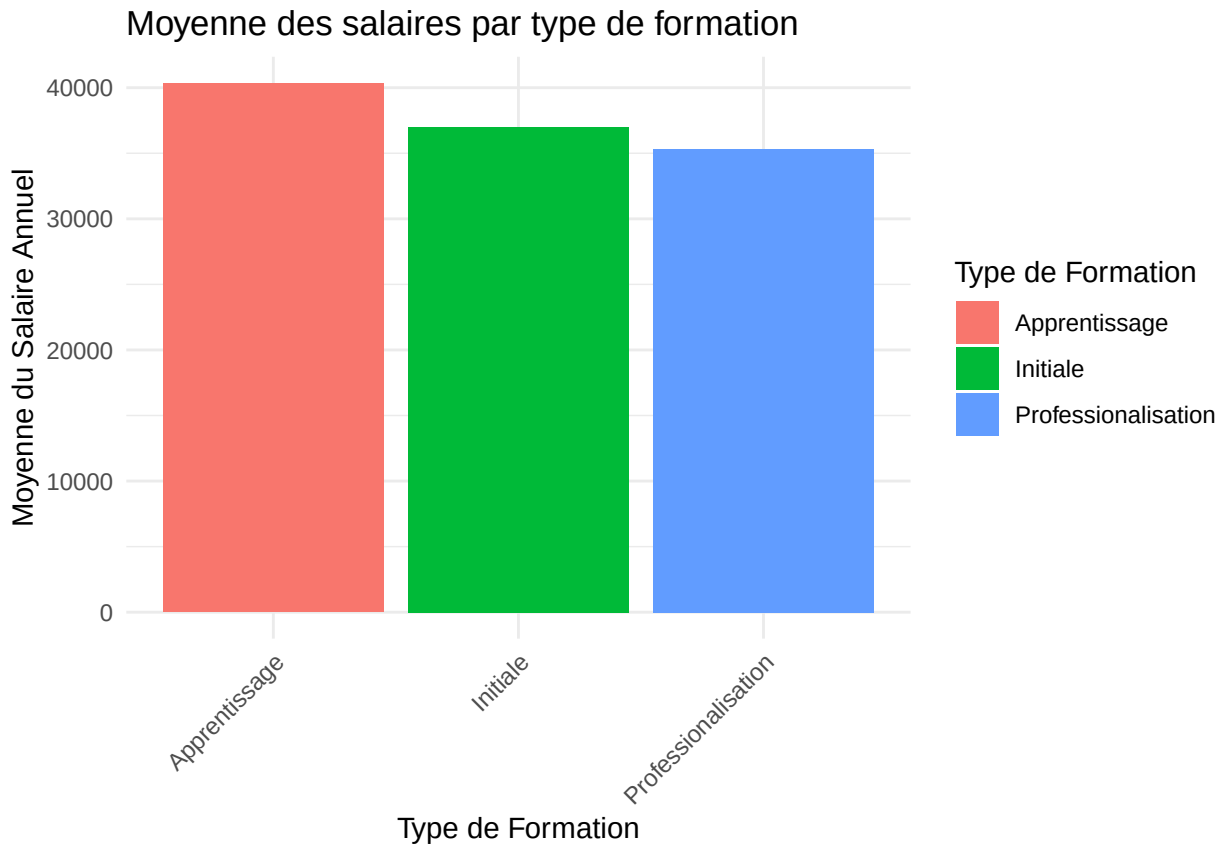
Ces observations suggèrent que les salaires évoluent avec le temps post-diplôme, avec une tendance à une plus grande hétérogénéité à mesure que les diplômés gagnent en expérience. Pour les nouveaux diplômés, la concentration des salaires pourrait également refléter des structures de rémunération d'entrée de gamme plus standardisées.

Disparité formation initiale/apprentissage/contrat de professionalisation

```
library(ggplot2)

# Supprimer les lignes où 'Type_formation' est NA ou vide
```

```r
clean_data <- raw_data[!is.na(raw_data$Type_formation) & raw_data$Type_formation != "", ]

# Créer le graphique avec les données nettoyées
ggplot(clean_data, aes(x = as.factor(Type_formation), y = Salaire_annuel_brut_primes, fill = as.factor(
  stat_summary(fun = "mean", geom = "col") +
  labs(title = "Moyenne des salaires par type de formation",
       x = "Type de Formation",
       y = "Moyenne du Salaire Annuel",
       fill = "Type de Formation") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
## Warning: Removed 250 rows containing non-finite values (`stat_summary()`).
```



Le diagramme “Moyenne des salaires par type de formation” illustre les salaires moyens annuels attribués aux individus en fonction de leur type de formation. L’apprentissage, représenté par la barre rouge, affiche le salaire moyen le plus élevé parmi les trois catégories. À l’opposé, la formation professionnalisante, indiquée par la barre bleue, montre le salaire moyen le plus bas. Entre les deux, la formation initiale, représentée en vert, se positionne en deuxième place avec un salaire moyen qui dépasse celui de la professionnalisation.

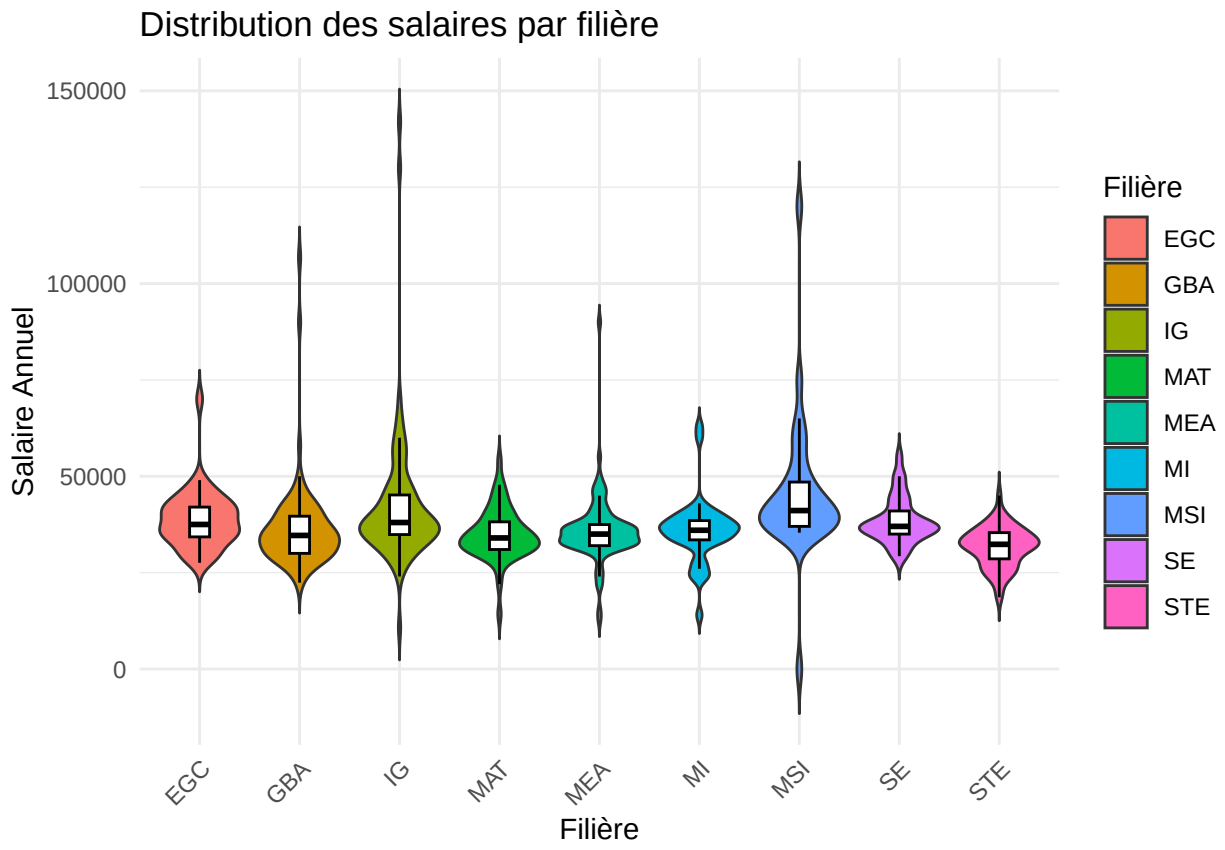
En comparaison avec les données des années antérieures, le graphique révèle une tendance à la baisse du salaire moyen pour l’apprentissage et la professionnalisation. Par contre, la formation initiale présente une augmentation de la moyenne salariale, se démarquant ainsi des autres catégories par une évolution positive de ses salaires moyens au fil des ans.

Disparité des filières

```
ggplot(raw_data, aes(x = Filiere, y = Salaire_annuel_brut_primes, fill = Filiere)) +
  geom_violin(trim = FALSE, scale = "width", width = 0.8) +
  geom_boxplot(width = 0.2, fill = "white", color = "black", outlier.shape = NA) +
  labs(title = "Distribution des salaires par filière",
       x = "Filière",
       y = "Salaire Annuel",
       fill = "Filière") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
## Warning: Removed 250 rows containing non-finite values (`stat_ydensity()`).
```

```
## Warning: Removed 250 rows containing non-finite values (`stat_boxplot()`).
```



On voit que les tendances sont les mêmes que les années précédentes.

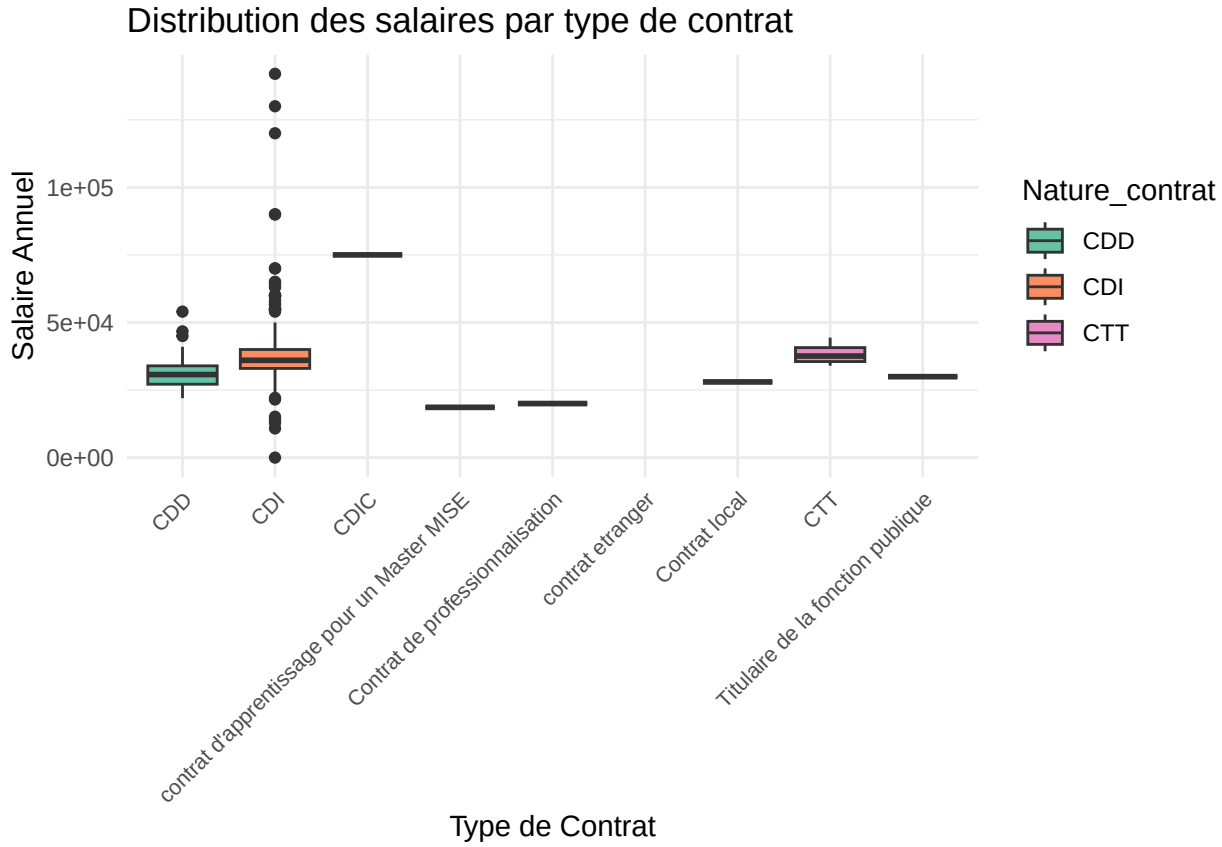
Filtrage des personnes en activité

```
filtered_data <- remove_rows_by_value(raw_data, "Situation", "En recherche emploi")
filtered_data <- remove_rows_by_value(filtered_data, "Situation", "En poursuite d'études (hors thèse) /
filtered_data <- remove_rows_by_value(filtered_data, "Situation", "These")
filtered_data <- remove_rows_by_value(filtered_data, "Situation", "Sans activité")
filtered_data <- remove_rows_by_value(filtered_data, "Situation", "Volontariat")
filtered_data <- remove_rows_by_value(filtered_data, "Situation", "En création d'entreprise /reprise d'
filtered_data <- remove_rows_by_value(filtered_data, "Situation", "Etudes")
```

```r
activite_data <- filtered_data
```

Disparité par nature du contrat

```r
library(ggplot2)

# Remplacer 'activite_data' par votre dataframe réel
# Filtrage de certains contrats peu représentatifs
filtered_data <- raw_data[!raw_data$Nature_contrat %in% c("Service à la personne (cours particulier de r

# Supposons que les catégories de contrats sont 'CDD', 'CDI', 'Intérim', etc.
# Vous pouvez définir manuellement les couleurs pour chaque catégorie de contrat
couleurs_contrats <- c("CDD" = "#66c2a5", "CDI" = "#fc8d62", "Intérim" = "#8da0cb",
                       "Stage" = "#e78ac3", "Alternance" = "#a6d854", "CTT" = "#e78ac3")

# Diagramme en boîte avec couleurs
ggplot(filtered_data, aes(x = Nature_contrat, y = Salaire_annuel_brut_primes, fill = Nature_contrat)) +
  geom_boxplot() +
  scale_fill_manual(values = couleurs_contrats) +  # Utiliser les couleurs définies précédemment
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 8)) +
  labs(title = "Distribution des salaires par type de contrat", x = "Type de Contrat", y = "Salaire Annu
```



Le graphique intitulé “Distribution des salaires par type de contrat” fournit une comparaison des salaires annuels en fonction de la nature du contrat de travail. À travers des boxplots, il distingue clairement trois

types de contrats : les CDD (contrats à durée déterminée), les CDI (contrats à durée indéterminée) et les CTT (contrats de travail temporaire), chacun représenté par une couleur distincte.

Le CDI, illustré en orange, présente une médiane de salaire supérieure à celle des deux autres types de contrats, indiquant que les employés sous CDI ont en moyenne des salaires plus élevés. La dispersion des salaires dans cette catégorie est également plus large, ce qui suggère une gamme de salaires allant des niveaux débutants aux positions très rémunérées. Les points isolés au-dessus du boxplot principal indiquent l'existence de salaires bien au-delà de la norme, soulignant la présence d'opportunités hautement rémunératrices au sein des CDI.

Les CDD, représentés en turquoise, montrent une concentration plus étroite des salaires, avec une médiane inférieure à celle des CDI. Cette concentration révèle une moindre variabilité salariale, ce qui peut être attribué à la nature temporaire et souvent plus précaire de ces contrats.

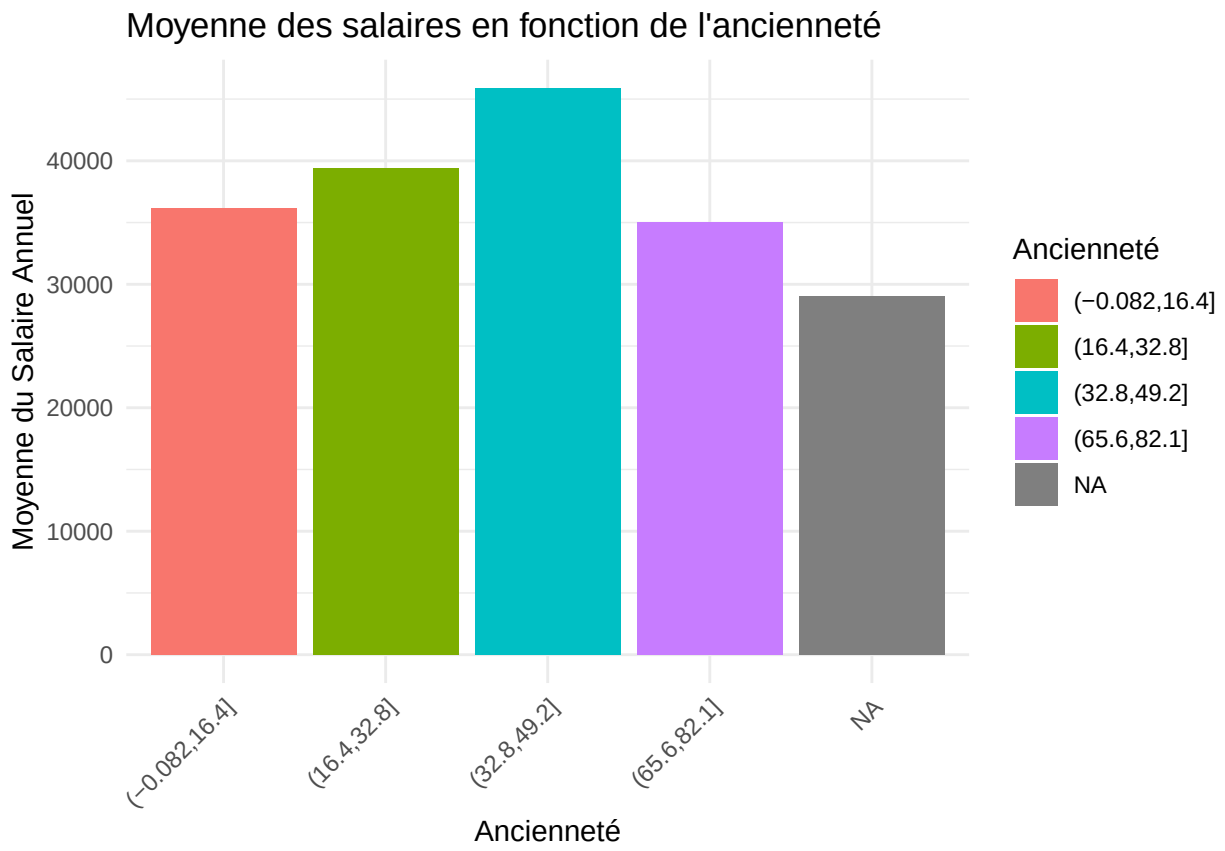
Quant aux CTT, affichés en rose, ils arborent la médiane la plus basse et la distribution des salaires la plus resserrée, ce qui indique non seulement des salaires généralement inférieurs mais aussi une moindre variation dans les rémunérations offertes. Cela pourrait refléter les limitations inhérentes aux emplois temporaires, qui offrent moins d'opportunités pour des salaires élevés.

En résumé, le type de contrat apparaît comme un facteur déterminant des perspectives salariales. Les CDI semblent offrir des salaires plus élevés et une plus grande variabilité, reflétant la sécurité et les possibilités de carrière à long terme. Les CDD et les CTT présentent des profils de rémunération plus modestes, avec des CTT particulièrement restreints dans leur potentiel salarial. Pour les individus qui évaluent leurs options d'emploi, cette visualisation des données salariales pourrait être un outil précieux pour orienter leurs décisions professionnelles.

Distribution du salaire en fonction de l'ancienneté

```r
library(ggplot2)

ordered_data <- raw_data[order(as.numeric(raw_data$Anciennete)), ]

# Création de catégories d'ancienneté
ordered_data$Anciennete_category <- cut(as.numeric(ordered_data$Anciennete), breaks = 5)

# Diagramme en barres pour la moyenne des salaires en fonction de l'ancienneté
ggplot(ordered_data, aes(x = Anciennete_category, y = Salaire_annuel_brut_primes, fill = Anciennete_cat
  geom_bar(stat = "summary", fun = "mean", position = "dodge") +
  labs(title = "Moyenne des salaires en fonction de l'ancienneté",
       x = "Ancienneté",
       y = "Moyenne du Salaire Annuel",
       fill = "Ancienneté") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
## Warning: Removed 250 rows containing non-finite values (`stat_summary()`).
```



Le diagramme présente la moyenne des salaires en fonction de l'ancienneté, avec une tendance inhabituelle par rapport aux années antérieures. Habituellement, on observe une croissance continue des salaires avec l'augmentation de l'ancienneté, mais le graphique de 2022 montre une progression suivie d'une régression.

Dans les premiers intervalles d'ancienneté, [0,82-16,4] et [16,4-32,8], il y a une augmentation progressive du salaire moyen, ce qui est cohérent avec l'expérience acquise sur le marché du travail. Cependant, dans l'intervalle suivant, [32,8-49,2], il y a un pic, où le salaire moyen atteint son apogée avant de diminuer dans la tranche d'ancienneté [65,8-82,1]. Cette diminution pourrait être due à plusieurs facteurs, notamment le départ à la retraite de travailleurs hautement rémunérés ou des changements dans les structures de rémunération qui favorisent moins l'ancienneté. De plus, il pourrait s'agir de l'impact de la pandémie de COVID-19 qui a commencé fin 2019 et ses répercussions économiques, qui ont pu influencer les structures salariales et les opportunités de carrière.

Le groupe "NA" représente les données manquantes ou non applicables, et ne peut donc pas être directement comparé aux autres groupes en termes de tendances salariales.

En conclusion, les données de 2022 suggèrent une évolution des salaires qui ne correspond pas à la progression linéaire attendue avec l'ancienneté. Cette situation pourrait refléter des ajustements économiques spécifiques à cette période, des changements dans les politiques de rémunération ou les effets de la pandémie qui ont perturbé les tendances salariales habituelles.

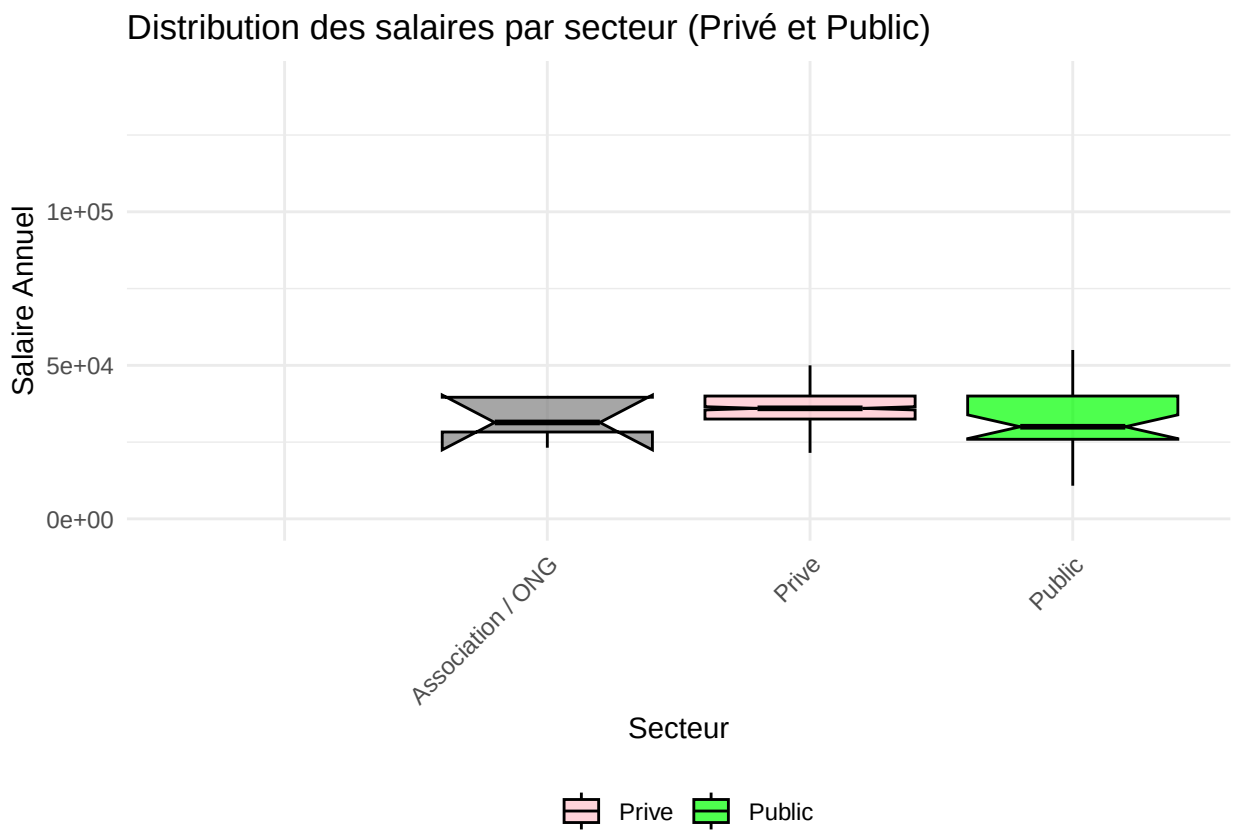
###Distribution du salaire selon le secteur

```
library(ggplot2)

# Supposons que 'Secteur' est la colonne qui indique si une personne travaille dans le secteur privé ou
# Nous allons exclure les non-salariés qui, dans cet exemple, sont marqués comme 'Non salarie(e)' dans

filtered_data <- raw_data[raw_data$Secteur != 'Non salarie(e)', ]
```

```r
# Graphique Boîte à moustaches avec deux couleurs pour le secteur privé et le secteur public
ggplot(filtered_data, aes(x = Secteur, y = Salaire_annuel_brut_primes, fill = Secteur)) +
  geom_boxplot(
    width = 0.8,
    notch = TRUE,
    outlier.shape = NA,
    color = "black", # La couleur des lignes du boxplot
    alpha = 0.7
  ) +
  scale_fill_manual(values = c("Prive" = "pink", "Public" = "green")) + # Deux couleurs pour privé et p
  labs(
    title = "Distribution des salaires par secteur (Privé et Public)",
    x = "Secteur",
    y = "Salaire Annuel"
  ) +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1),
    legend.position = "bottom",
    legend.title = element_blank() # Enlève le titre de la légende
  )
```

```
## Warning: Removed 250 rows containing non-finite values (`stat_boxplot()`).
```

```
## Notch went outside hinges
## i Do you want `notch = FALSE`?
```



Le diagramme de distribution des salaires par type de contrat fournit un aperçu visuel comparatif des échelles salariales associées à différents types de contrats de travail. Chaque boxplot représente une catégorie contractuelle distincte avec une variation des salaires allant des postes de débutants aux plus expérimentés.

Le boxplot pour le CDD (Contrat à Durée Déterminée) dénote une médiane relativement basse comparée à celle du CDI (Contrat à Durée Indéterminée), indiquant que la majorité des salariés en CDD gagnent moins que la moitié des salariés en CDI. La présence de valeurs aberrantes au-dessus de la moustache supérieure du CDD suggère que certains postes en CDD offrent des salaires exceptionnellement élevés.

Le CDI présente une médiane plus élevée, ce qui est courant étant donné la nature plus stable et à long terme de ces postes. Les valeurs aberrantes, bien que moins fréquentes que pour le CDD, indiquent toujours l'existence de postes en CDI très bien rémunérés.

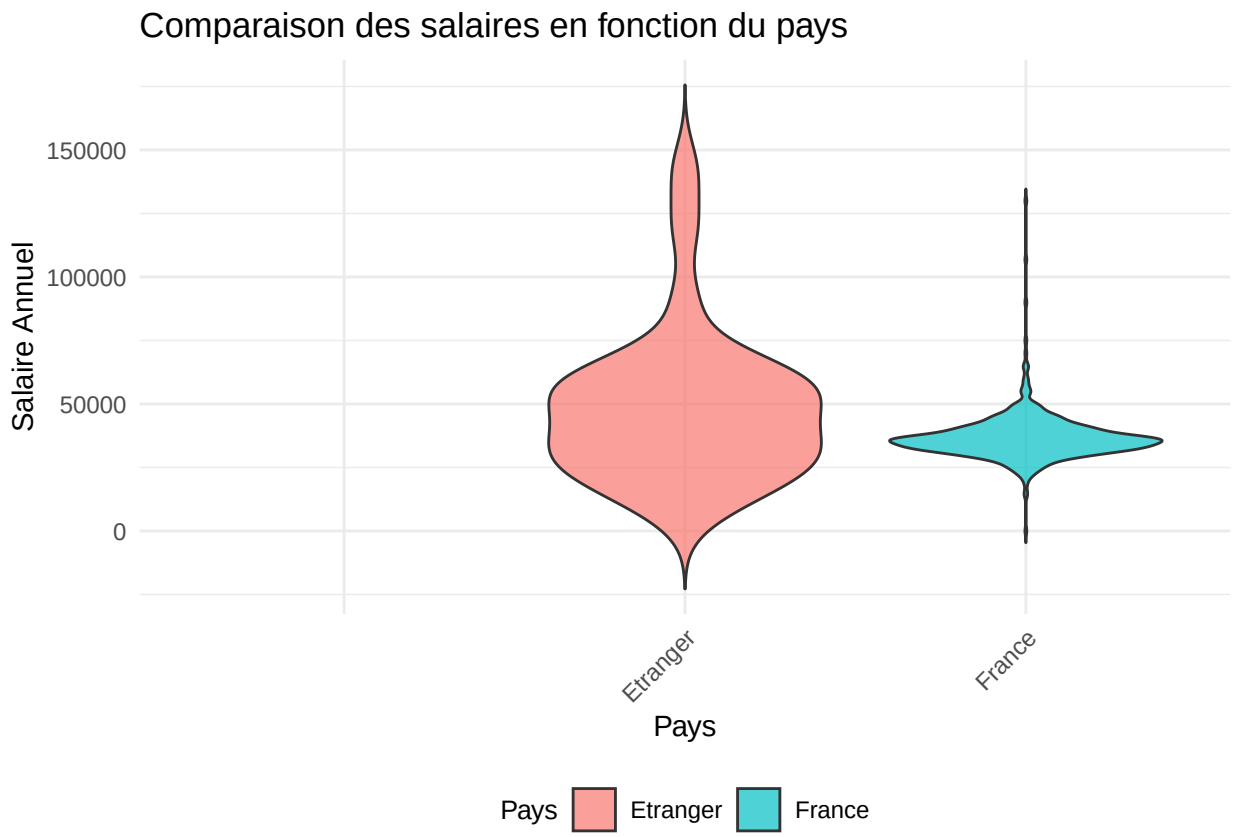
Les autres catégories de contrats, tels que le contrat d'apprentissage ou le CTT (Contrat de Travail Temporaire), montrent des médianes et des étendues différentes, reflétant la diversité des salaires au sein de ces types de contrats. Le CTT, par exemple, affiche une médiane plus basse et une variation de salaire moins importante, soulignant la nature souvent précaire et moins rémunératrice des postes temporaires.

En résumé, ce diagramme met en évidence les différences de rémunération non seulement entre les contrats à durée déterminée et indéterminée mais aussi parmi les autres formes contractuelles, offrant ainsi une compréhension de la structure salariale globale au sein du marché du travail.

Distribution du salaire en fonction de France/étranger

```
ggplot(raw_data, aes(x = factor(Pays), y = Salaire_annuel_brut_primes, fill = factor(Pays))) +
  geom_violin(trim = FALSE, scale = "width", width = 0.8, alpha = 0.7) +
  labs(
    title = "Comparaison des salaires en fonction du pays",
    x = "Pays",
    y = "Salaire Annuel",
    fill = "Pays"
  ) +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1),
    legend.position = "bottom",
    legend.title = element_text(size = 10)
  )
```

```
## Warning: Removed 250 rows containing non-finite values (`stat_ydensity()`).
```



Le diagramme en forme de violon présente une comparaison des distributions de salaires annuels en France par rapport à l'étranger. Les différences marquées entre les deux distributions sont immédiatement visibles.

Pour le "Étranger", la distribution est large avec une pointe élevée, ce qui suggère que la majorité des salaires sont concentrés autour de la médiane, mais qu'il existe également un nombre considérable de salaires très élevés, comme le montrent les longues queues de distribution. Cela pourrait indiquer que, bien que le salaire moyen à l'étranger puisse être élevé, il existe une grande disparité, avec à la fois des salaires modérés et des salaires extrêmement élevés.

En revanche, la distribution pour la France est plus étroite et le corps du violon est plus compact, impliquant une concentration des salaires autour de la médiane avec moins de variance extrême. Cela pourrait refléter une plus grande uniformité des salaires en France, avec moins de disparités extrêmes entre les salaires les plus bas et les plus élevés.

Ces observations suggèrent que travailler à l'étranger peut offrir un potentiel de gains plus élevés mais possiblement avec un risque plus grand d'inégalité salariale, tandis que la France pourrait offrir une plus grande sécurité en termes d'équité salariale, bien que potentiellement avec des salaires globalement moins élevés.

Etude données 2023

Récupération des données

```
# Trying ISO-8859-1
raw_data <- read.csv("data/data_2023.csv", sep = ",", fileEncoding = "UTF-8")
```

Définir toutes les données comme caractère (variable qualitative) sauf le salaire

```r
# Lists specifying which columns to convert to factors and numeric
factor_cols <- c("Date", "Identifiant", "Genre", "Annee_diplome","Type_formation","Filiere","Situation_
numeric_cols <- c("Anciennete","Salaire_annuel_brut", "Salaire_annuel_brut_primes", "Responsabilite_hie

# Convert columns to factors
raw_data[factor_cols] <- lapply(raw_data[factor_cols], as.factor)

# Convert columns to numeric
raw_data[numeric_cols] <- sapply(raw_data[numeric_cols], as.numeric)
```

```
## Warning in lapply(X = X, FUN = FUN, ...): NAs introduced by coercion

## Warning in lapply(X = X, FUN = FUN, ...): NAs introduced by coercion

## Warning in lapply(X = X, FUN = FUN, ...): NAs introduced by coercion

## Warning in lapply(X = X, FUN = FUN, ...): NAs introduced by coercion
```

```r
summary(raw_data)
```

Avec ces valeurs on peut déduire beaucoup de choses...

Fonction de filtre

Ceci est une fonction utilisé après pour retirer différentes lignes en fonction des valeurs dans une certain colonne. (équivalent d'un select where)

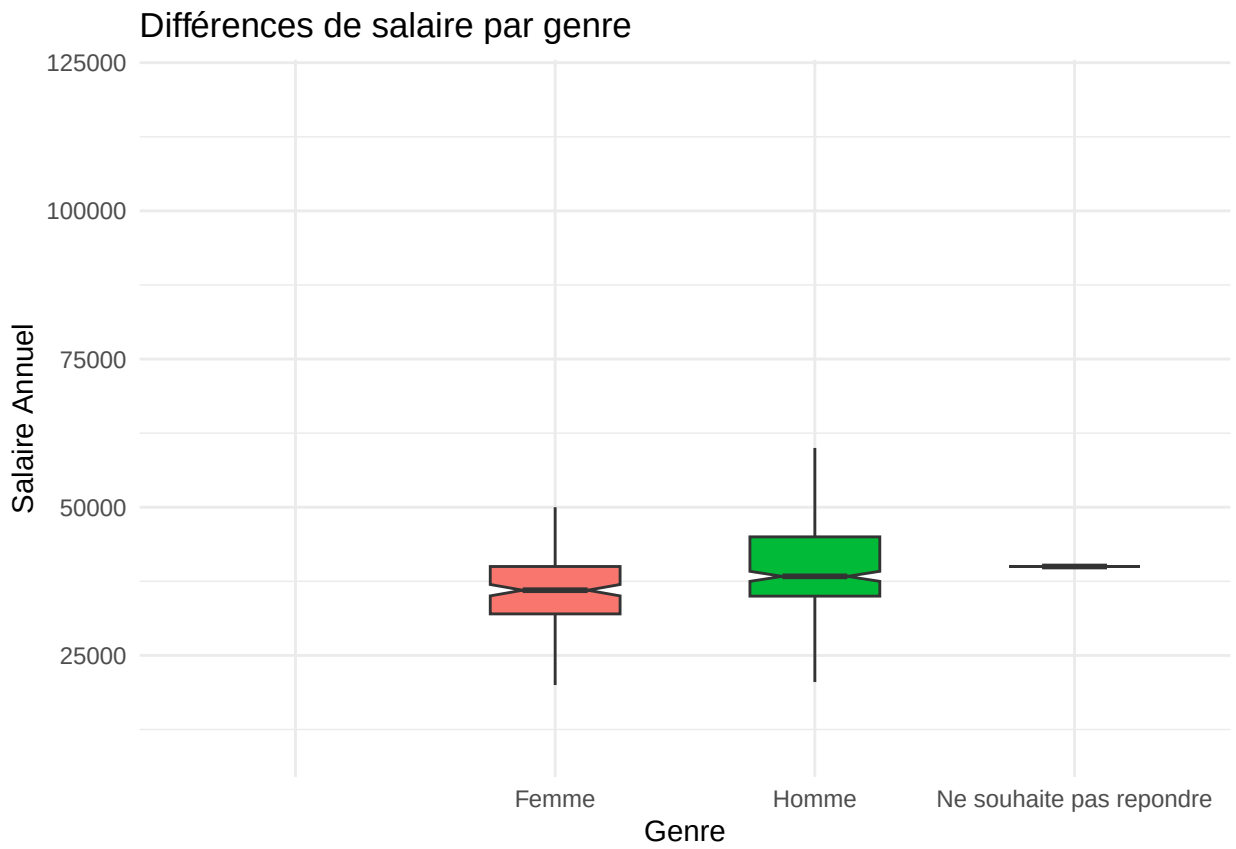
```r
remove_rows_by_value <- function(data, column_name, value_to_remove) {
  data_filtered <- subset(data, !(data[[column_name]] == value_to_remove))
  return(data_filtered)
}
```

Première étude

Disparité homme/femme - toute formation confondue

```r
ggplot(raw_data, aes(x = Genre, y = Salaire_annuel_brut_primes, fill = Genre)) +
  geom_boxplot(width = 0.5, notch = TRUE, outlier.shape = NA) +
  labs(title = "Différences de salaire par genre",
       x = "Genre",
       y = "Salaire Annuel") +
  theme_minimal() +
  theme(legend.position = "none")
```

```
## Warning: Removed 202 rows containing non-finite values (`stat_boxplot()`).
```



Le graphique intitulé “Différences de salaire par genre” représente un boxplot classique, qui met en évidence la distribution des salaires annuels entre les femmes et les hommes. À première vue, il suggère une légère disparité salariale entre les genres, avec les médianes indiquant que les hommes ont tendance à gagner plus que les femmes dans cet échantillon.

Les médianes de chaque groupe — représentées par la ligne centrale des boîtes — sont cruciales pour cette observation. La médiane pour le groupe des femmes semble être inférieure à celle du groupe des hommes. Il est important de noter que la médiane est souvent préférée à la moyenne pour une telle analyse, car elle est moins sensible aux valeurs extrêmes qui pourraient fausser les résultats.

En examinant la taille des boîtes, qui illustrent l’écart interquartile, nous observons une similitude dans la dispersion des salaires entre les deux groupes. Cela signifie que la moitié centrale des salaires s’étend sur une plage similaire pour les deux genres, suggérant que, mis à part la médiane, les distributions des salaires sont relativement comparables.

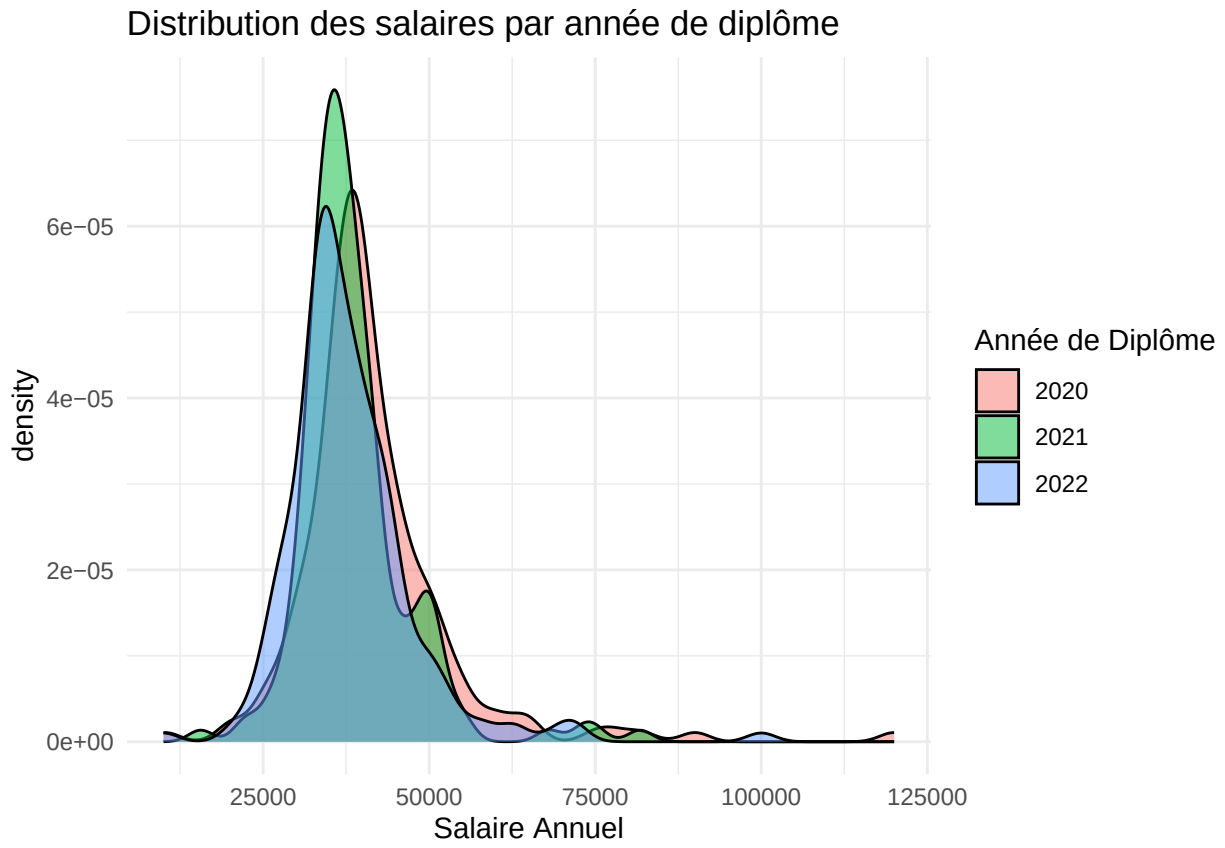
En conclusion, ce graphique révèle une tendance où les hommes semblent avoir un avantage salarial par rapport aux femmes dans l’échantillon étudié. Néanmoins, la similarité des distributions suggère que les écarts de salaires, au-delà de la médiane, ne sont pas marqués par des différences extrêmes. Pour une analyse complète, il serait judicieux de prendre en compte d’autres variables qui pourraient influencer ces résultats et de réaliser des tests statistiques pour évaluer la significativité des différences observées.

Disparité sur les dates d’optention du diplôme

```
ggplot(raw_data, aes(x = Salaire_annuel_brut_primes, fill = as.factor(Annee_diplome))) +
  geom_density(alpha = 0.5) +
  labs(title = "Distribution des salaires par année de diplôme",
       x = "Salaire Annuel",
```

```
      fill = "Année de Diplôme") +
  theme_minimal()
```

```
## Warning: Removed 202 rows containing non-finite values (`stat_density()`).
```



Le graphique “Distribution des salaires par année de diplôme” présente les tendances des salaires annuels des diplômés de Polytech Montpellier pour les années 2020, 2021 et 2022. Comme les années précédentes, les courbes de densité pour 2020 et 2021 montrent des pics bien définis, indiquant une concentration des salaires autour de valeurs modales spécifiques. Ceci suggère une certaine cohérence dans les conditions du marché de l'emploi peu après la graduation, où la plupart des diplômés gagnent des salaires similaires.

La distribution pour l'année 2022 se distingue nettement des deux précédentes. Elle présente une courbe plus aplatie avec une densité réduite, suggérant une variabilité plus grande dans les salaires des diplômés. Cela peut indiquer que les diplômés de 2022 ont connu des trajectoires professionnelles plus diversifiées ou que le marché de l'emploi a subi des changements significatifs, influençant ainsi les salaires de manière plus hétérogène. Des facteurs tels que des ajustements sectoriels, l'impact économique post-pandémie ou des tendances telles que la digitalisation et la transformation des compétences requises pourraient jouer un rôle.

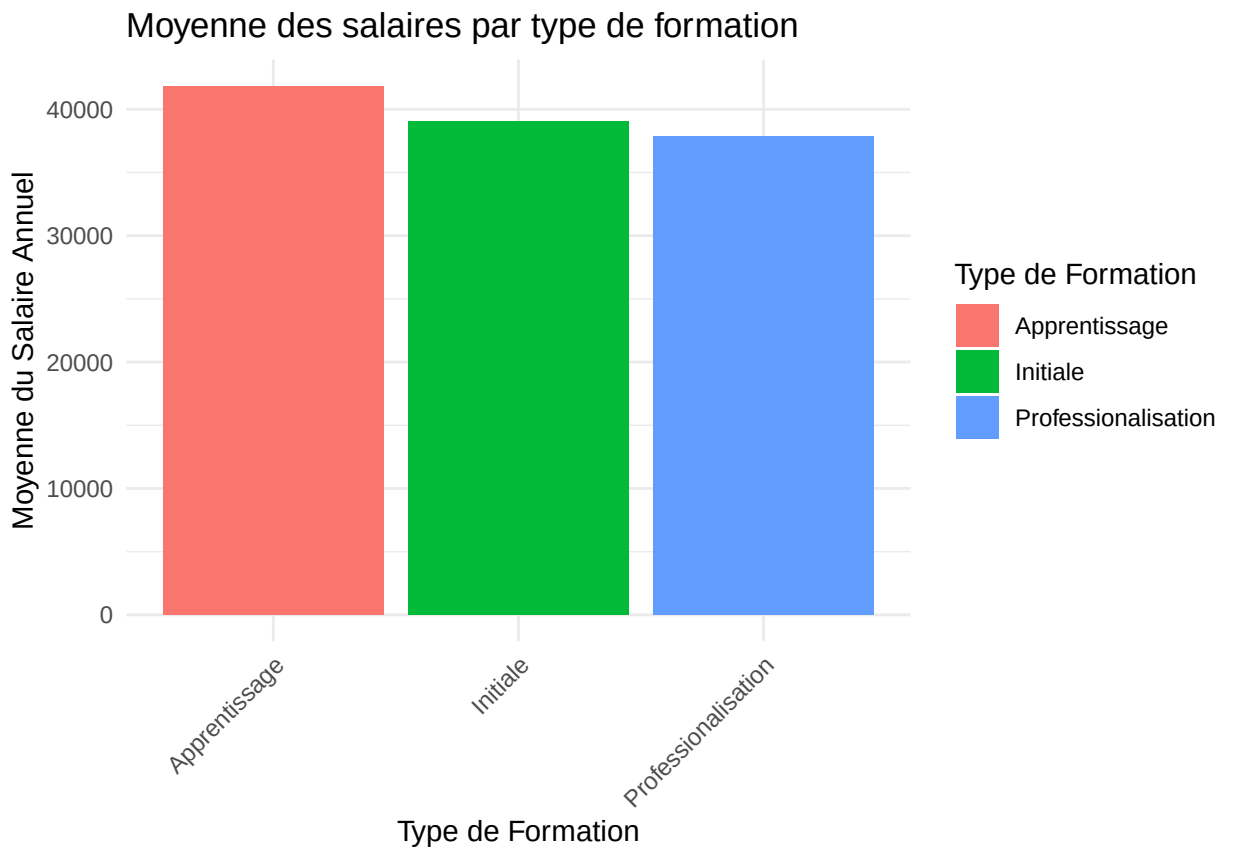
La distribution plus large pour 2022 pourrait aussi refléter des opportunités salariales variables en raison d'une plus grande mobilité professionnelle ou de choix de carrière visant un équilibre entre vie professionnelle et personnelle. Les salaires élevés pourraient être attribués à des individus ayant rapidement gravi les échelons ou ayant bénéficié d'opportunités lucratives à l'étranger, tandis que d'autres pourraient avoir opté pour des rôles moins rémunérateurs mais plus satisfaisants sur le plan personnel ou social.

En somme, l'analyse des distributions salariales des trois années révèle l'importance de considérer les conditions changeantes du marché de l'emploi et l'impact des choix professionnels individuels sur les salaires à long terme. Pour les institutions éducatives et les étudiants, il est crucial de prendre en compte ces éléments pour planifier les carrières et anticiper les potentiels financiers futurs.

Disparité formation initiale/apprentissage/contrat de professionalisation

```
library(ggplot2)

# Supprimer les lignes où 'Type_formation' est NA ou vide
clean_data <- raw_data[!is.na(raw_data$Type_formation) & raw_data$Type_formation != "", ]

# Créer le graphique avec les données nettoyées
ggplot(clean_data, aes(x = as.factor(Type_formation), y = Salaire_annuel_brut_primes, fill = as.factor(
  stat_summary(fun = "mean", geom = "col") +
  labs(title = "Moyenne des salaires par type de formation",
       x = "Type de Formation",
       y = "Moyenne du Salaire Annuel",
       fill = "Type de Formation") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
## Warning: Removed 201 rows containing non-finite values (`stat_summary()`).
```



Le diagramme représente la moyenne des salaires annuels en fonction du type de formation. Les données montrent clairement que l'apprentissage affiche le salaire moyen le plus élevé parmi les trois catégories, ce qui peut indiquer une valorisation sur le marché du travail des compétences pratiques et de l'expérience acquise à travers ce type de formation. En deuxième position, nous avons les salaires moyens pour les formations initiales, qui sont légèrement inférieurs à ceux de l'apprentissage, mais qui restent néanmoins significatifs. Cela pourrait refléter la reconnaissance de la formation académique traditionnelle. Enfin, la professionalisation affiche le salaire moyen le plus bas des trois, ce qui pourrait suggérer que ces programmes sont soit moins valorisés, soit qu'ils mènent à des secteurs avec des grilles salariales globalement plus basses.

Cette hiérarchie des salaires peut être influencée par divers facteurs, tels que la demande du marché pour certaines compétences, la nature des industries employant les diplômés de ces formations, ou encore le niveau de spécialisation requis.

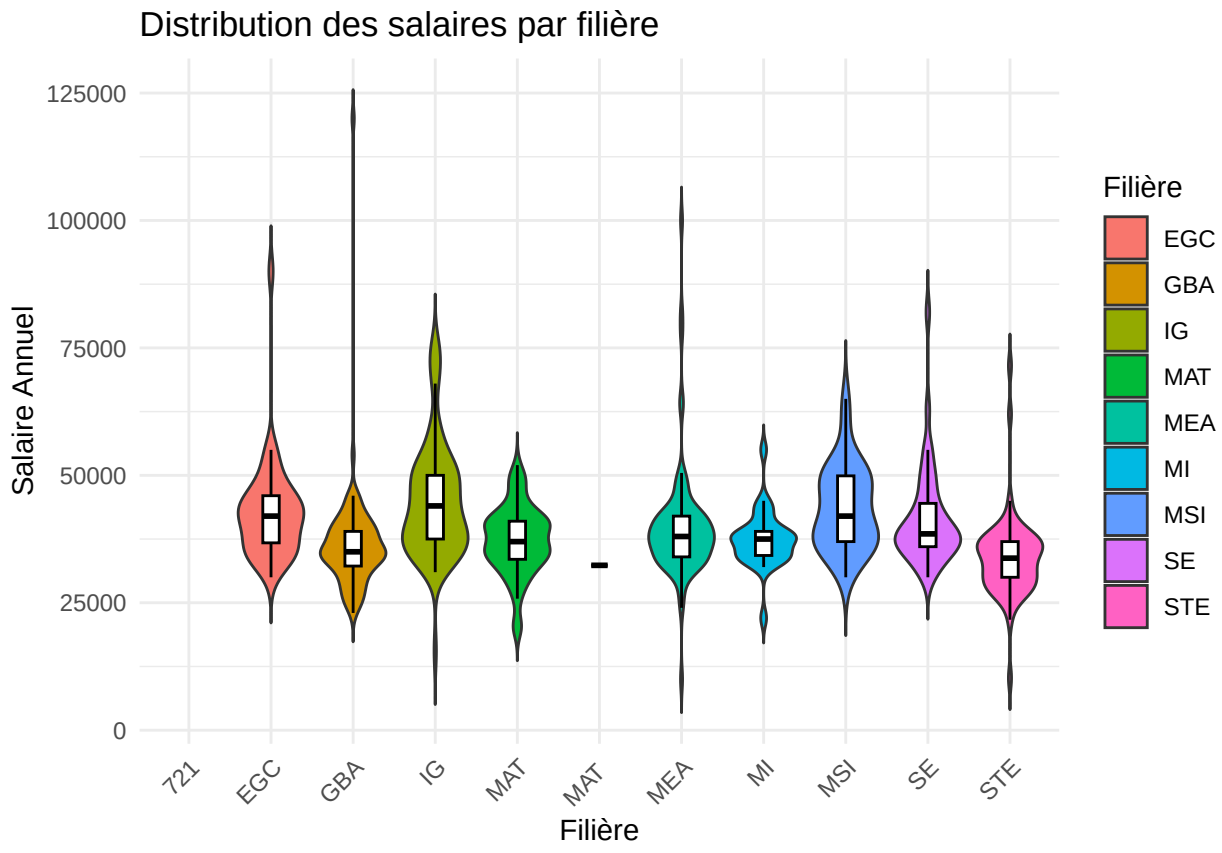
Disparité des filières

```
ggplot(raw_data, aes(x = Filiere, y = Salaire_annuel_brut_primes, fill = Filiere)) +
  geom_violin(trim = FALSE, scale = "width", width = 0.8) +
  geom_boxplot(width = 0.2, fill = "white", color = "black", outlier.shape = NA) +
  labs(title = "Distribution des salaires par filière",
       x = "Filière",
       y = "Salaire Annuel",
       fill = "Filière") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
## Warning: Removed 202 rows containing non-finite values (`stat_ydensity()`).
```

```
## Warning: Groups with fewer than two data points have been dropped.
```

```
## Warning: Removed 202 rows containing non-finite values (`stat_boxplot()`).
```



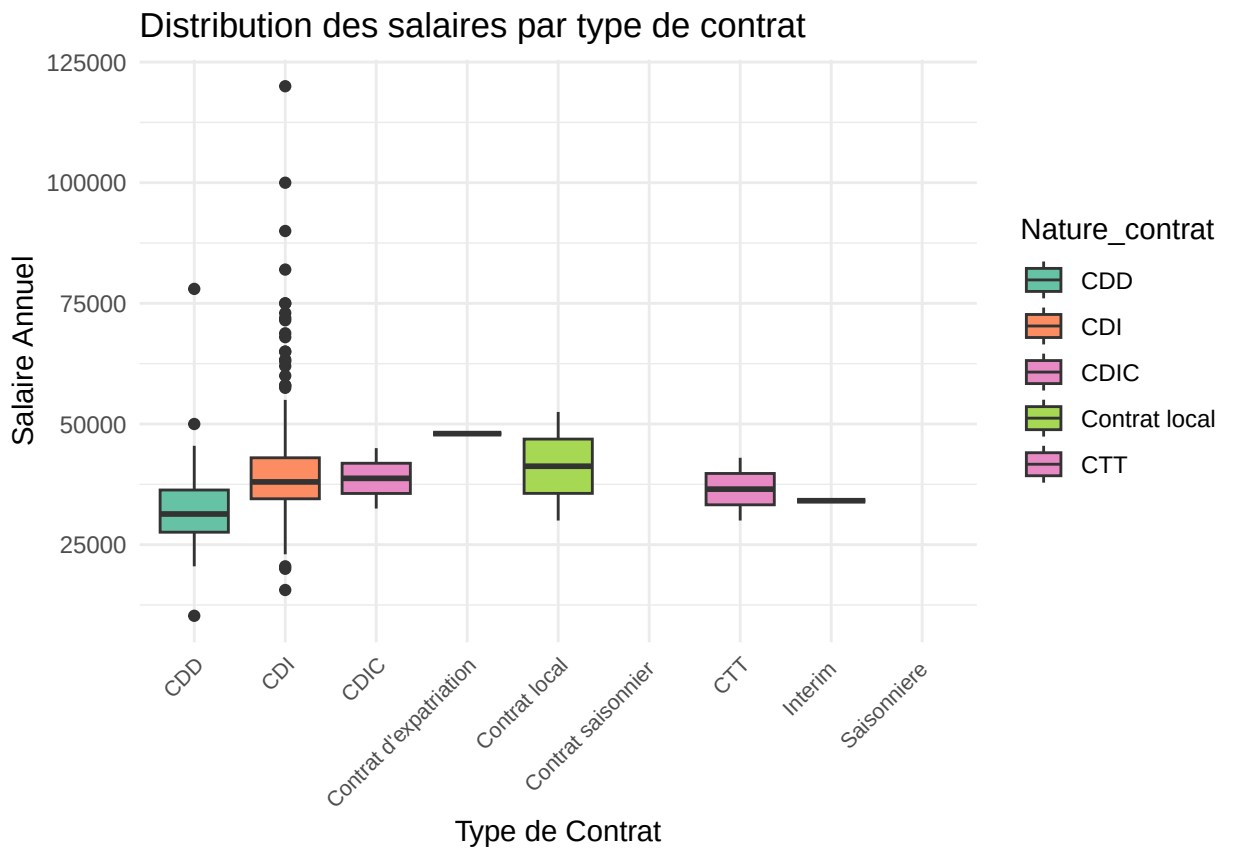
Filtrage des personnes en activité

```
filtered_data <- remove_rows_by_value(raw_data, "Situation", "En recherche emploi")
filtered_data <- remove_rows_by_value(filtered_data, "Situation", "En poursuite d'études (hors thèse) /
filtered_data <- remove_rows_by_value(filtered_data, "Situation", "These")
```

```r
filtered_data <- remove_rows_by_value(filtered_data, "Situation", "Sans activité")
filtered_data <- remove_rows_by_value(filtered_data, "Situation", "Volontariat")
filtered_data <- remove_rows_by_value(filtered_data, "Situation", "En création d'entreprise /reprise d'
filtered_data <- remove_rows_by_value(filtered_data, "Situation", "Etudes")

activite_data <- filtered_data
```

Disparité par nature du contrat

```r
library(ggplot2)

# Remplacer 'activite_data' par votre dataframe réel
# Filtrage de certains contrats peu représentatifs
filtered_data <- raw_data[!raw_data$Nature_contrat %in% c("Service à la personne (cours particulier de r

# Supposons que les catégories de contrats sont 'CDD', 'CDI', 'Intérim', etc.
# Vous pouvez définir manuellement les couleurs pour chaque catégorie de contrat
couleurs_contrats <- c("CDD" = "#66c2a5", "CDI" = "#fc8d62", "Intérim" = "#8da0cb",
                       "CDIC" = "#e78ac3", "Contrat local" = "#a6d854", "CTT" = "#e78ac3")

# Diagramme en boîte avec couleurs
ggplot(filtered_data, aes(x = Nature_contrat, y = Salaire_annuel_brut_primes, fill = Nature_contrat)) +
  geom_boxplot() +
  scale_fill_manual(values = couleurs_contrats) +  # Utiliser les couleurs définies précédemment
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 8)) +
  labs(title = "Distribution des salaires par type de contrat", x = "Type de Contrat", y = "Salaire Annu
```



Le diagramme “Distribution des salaires par type de contrat” présente une comparaison des salaires annuels en fonction des différents contrats de travail. Les contrats à durée indéterminée (CDI) et les contrats à durée déterminée (CDD) sont les plus représentés, avec une médiane indiquant le salaire annuel moyen pour chaque type. Les CDI, généralement associés à une plus grande sécurité de l'emploi, ont une médiane plus élevée et une distribution plus large des salaires, ce qui suggère une variété de postes et de niveaux d'ancienneté. Les CDD, en revanche, montrent une distribution plus serrée, ce qui peut indiquer des salaires plus uniformes et potentiellement plus bas.

On observe également d'autres types de contrats comme les contrats d'apprentissage et les contrats de professionnalisation, qui ont tendance à présenter des salaires inférieurs, conformément à leur nature de formation en milieu professionnel.

Les contrats locaux et les contrats de travail temporaire (CTT) ont des médianes plus basses, ce qui est cohérent avec la nature temporaire et potentiellement moins rémunératrice de ces emplois. Les stages, avec une médiane encore plus basse, indiquent le statut d'emploi non salarié ou peu rémunéré, souvent lié à l'apprentissage et à la formation.

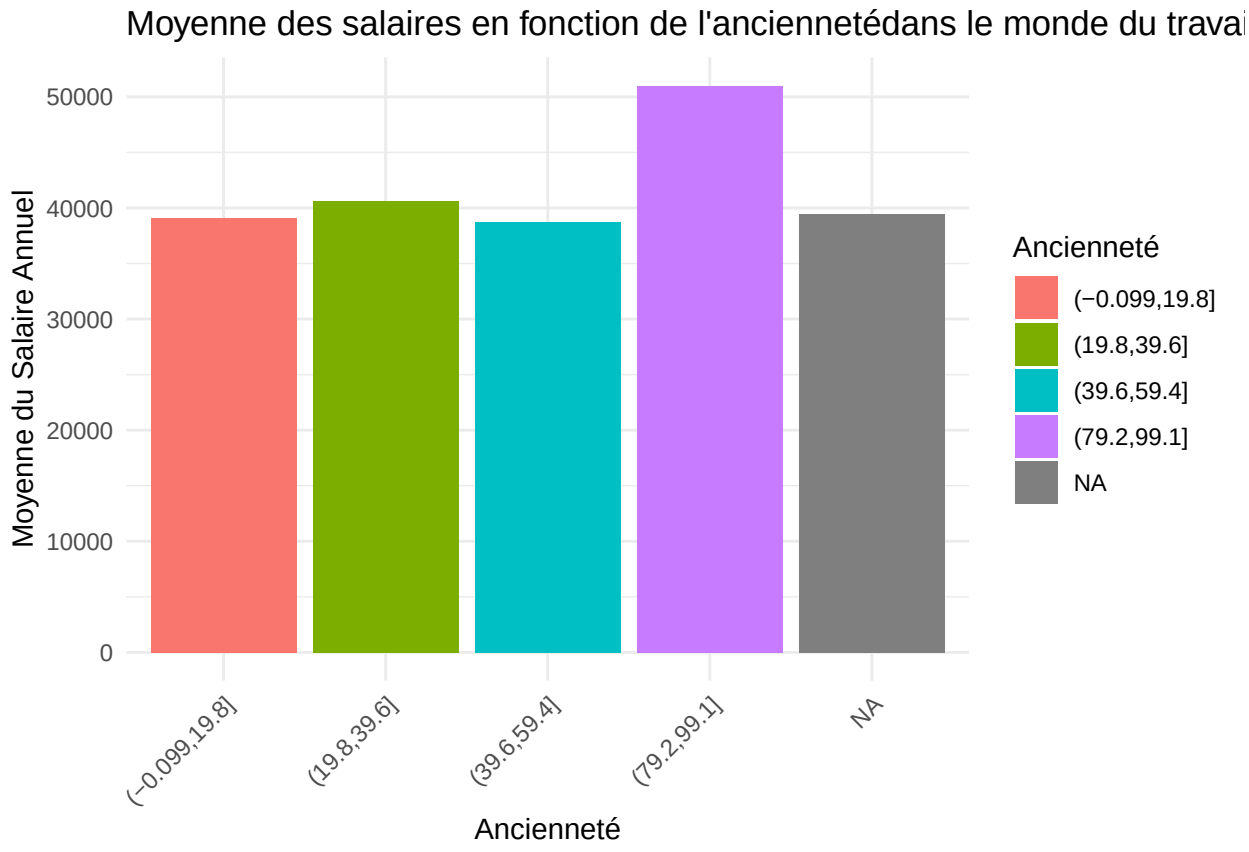
Ce diagramme illustre la diversité des situations salariales en fonction du type de contrat, reflétant les différences en termes de sécurité d'emploi, d'opportunités de croissance de salaire, et de statut professionnel. Il est important de noter que ces données peuvent varier considérablement en fonction du secteur d'activité, de la région, de l'expérience et de la formation des individus.

Distribution du salaire en fonction de l'ancienneté

```
library(ggplot2)
```

```r
ordered_data <- raw_data[order(as.numeric(raw_data$Anciennete)), ]

# Création de catégories d'ancienneté (facultatif)
ordered_data$Anciennete_category <- cut(as.numeric(ordered_data$Anciennete), breaks = 5)

# Diagramme en barres pour la moyenne des salaires en fonction de l'ancienneté
ggplot(ordered_data, aes(x = Anciennete_category, y = Salaire_annuel_brut_primes, fill = Anciennete_cat
  geom_bar(stat = "summary", fun = "mean", position = "dodge") +
  labs(title = "Moyenne des salaires en fonction de l'ancienneté dans le monde du travail",
       x = "Ancienneté",
       y = "Moyenne du Salaire Annuel",
       fill = "Ancienneté") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
## Warning: Removed 202 rows containing non-finite values (`stat_summary()`).
```



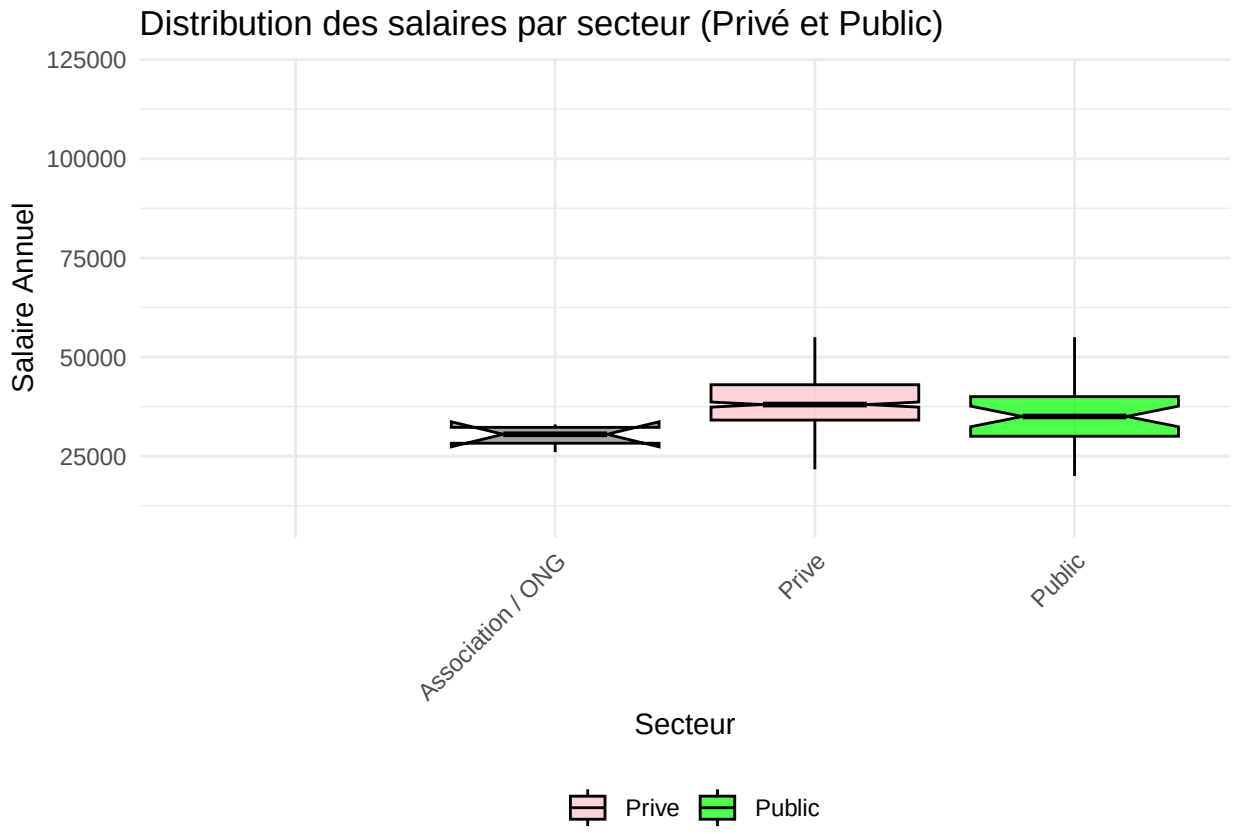
###Distribution du salaire selon le secteur

```r
library(ggplot2)

# Supposons que 'Secteur' est la colonne qui indique si une personne travaille dans le secteur privé ou
# Nous allons exclure les non-salariés qui, dans cet exemple, sont marqués comme 'Non salarie(e)' dans

filtered_data <- raw_data[raw_data$Secteur != 'Non salarie(e)', ]

# Graphique Boîte à moustaches avec deux couleurs pour le secteur privé et le secteur public
```

```r
ggplot(filtered_data, aes(x = Secteur, y = Salaire_annuel_brut_primes, fill = Secteur)) +
  geom_boxplot(
    width = 0.8,
    notch = TRUE,
    outlier.shape = NA,
    color = "black", # La couleur des lignes du boxplot
    alpha = 0.7
  ) +
  scale_fill_manual(values = c("Prive" = "pink", "Public" = "green")) + # Deux couleurs pour privé et p
  labs(
    title = "Distribution des salaires par secteur (Privé et Public)",
    x = "Secteur",
    y = "Salaire Annuel"
  ) +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1),
    legend.position = "bottom",
    legend.title = element_blank() # Enlève le titre de la légende
  )
```

```
## Warning: Removed 202 rows containing non-finite values (`stat_boxplot()`).
```

```
## Notch went outside hinges
## i Do you want `notch = FALSE`?
```



Le diagramme en boîte intitulé “Distribution des salaires par secteur (Privé et Public)” offre une visualisation

des écarts de rémunération entre le secteur privé, le secteur public et les associations/ONG. Chaque boîte dépeint la répartition des salaires annuels au sein de chaque catégorie, avec une médiane qui divise la boîte en deux parties égales, représentant la valeur centrale des données.

Le secteur privé présente une plage salariale plus large que le secteur public, ce qui pourrait indiquer une plus grande diversité dans les types d'emplois et les niveaux de rémunération au sein du privé. La médiane du secteur privé est légèrement supérieure à celle du public, suggérant que les salaires médians sont généralement plus élevés dans le privé.

Le secteur public montre une distribution des salaires plus concentrée avec moins de valeurs extrêmes, ce qui est cohérent avec des grilles salariales plus réglementées et uniformes.

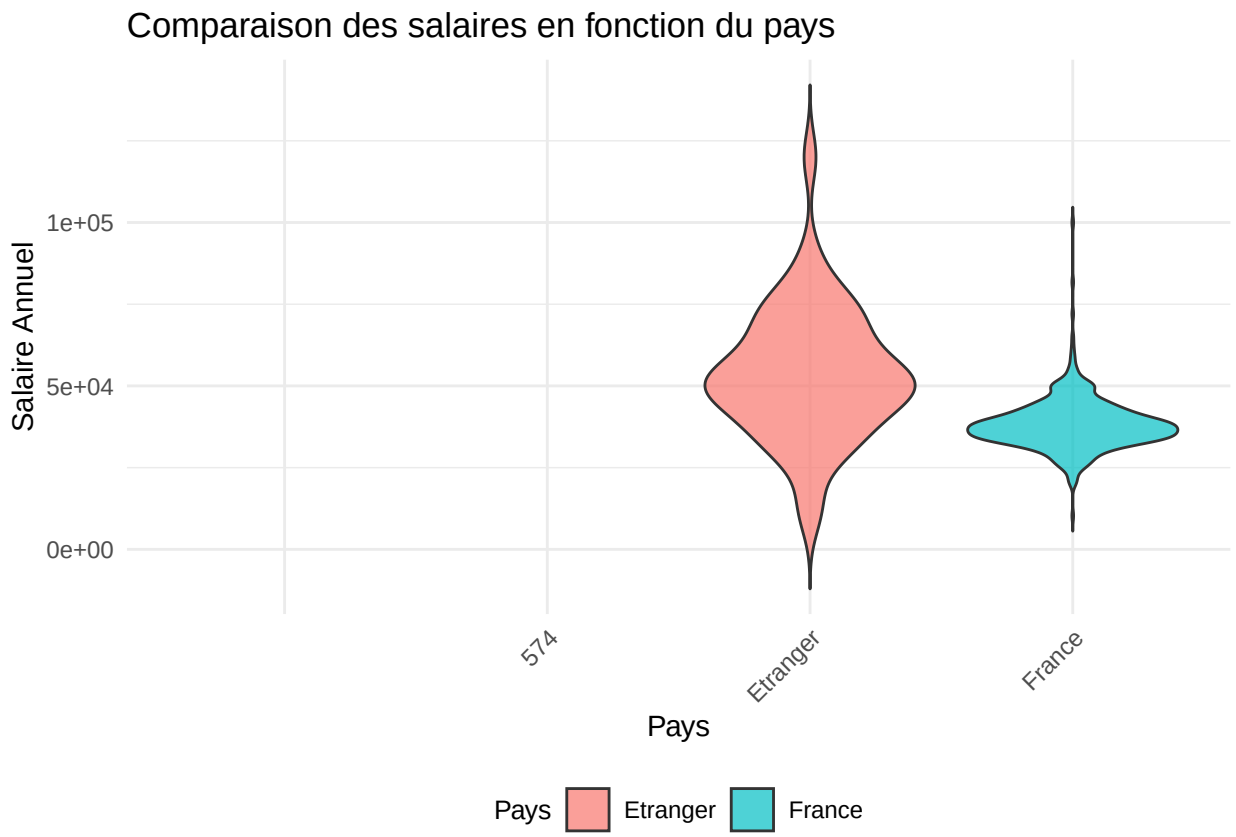
Les associations et ONG présentent la médiane la plus basse, ainsi qu'une gamme de salaires relativement étroite, ce qui est typique pour le secteur non lucratif où les budgets sont souvent limités et les rémunérations moins compétitives par rapport aux secteurs privé et public.

En général, ce diagramme illustre les différences structurelles de rémunération entre les différents secteurs d'activité, reflétant des réalités économiques, des politiques de rémunération et des missions organisationnelles distinctes.

Distribution du salaire en fonction de France/étranger

```
ggplot(raw_data, aes(x = factor(Pays), y = Salaire_annuel_brut_primes, fill = factor(Pays))) +
  geom_violin(trim = FALSE, scale = "width", width = 0.8, alpha = 0.7) +
  labs(
    title = "Comparaison des salaires en fonction du pays",
    x = "Pays",
    y = "Salaire Annuel",
    fill = "Pays"
  ) +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1),
    legend.position = "bottom",
    legend.title = element_text(size = 10)
  )
```

```
## Warning: Removed 202 rows containing non-finite values (`stat_ydensity()`).
```



Le diagramme “Comparaison des salaires en fonction du pays” est un graphique en forme de violon qui illustre la distribution des salaires annuels entre deux catégories : “Étranger” et “France”. Cette forme de représentation permet de visualiser à la fois la répartition de la densité des données (la largeur du violon) et la médiane (la ligne épaisse au centre du violon).

Pour le pays marqué comme “Étranger”, le violon est large en son centre, indiquant une concentration élevée des salaires autour de la médiane, mais avec des pointes s’étendant vers des salaires plus élevés et plus bas, ce qui suggère une variabilité significative des salaires à l’étranger.

La catégorie “France”, en revanche, montre un violon plus étroit et allongé, avec une médiane plus basse que celle de la catégorie “Étranger”. La concentration autour de la médiane est moins prononcée, ce qui implique que les salaires en France sont plus uniformément répartis sur une gamme plus étroite.

Globalement, ce graphique suggère que, bien que la médiane des salaires en France soit inférieure, la distribution des salaires est moins variée qu’à l’étranger. Cela pourrait refléter des différences dans les structures économiques, les niveaux de vie, les politiques de rémunération, ou d’autres facteurs socio-économiques influençant les salaires dans les deux régions.

##Conclusion sur les différentes années

- Le genre : Dans l’ensemble, ces analyses annuelles montrent une tendance persistante d’écart salarial entre les hommes et les femmes, les hommes ayant généralement des salaires plus élevés. Bien que les distributions des salaires puissent varier d’une année à l’autre, la tendance à la disparité salariale entre les genres est cohérente. Cela peut être dû à divers facteurs, tels que la discrimination, les choix de carrière, l’éducation, l’expérience professionnelle, ou d’autres facteurs socio-économiques. Une analyse plus approfondie et des tests statistiques pour évaluer la signification de ces différences seraient nécessaires pour comprendre pleinement ces disparités salariales.
- L’année d’obtention du diplome: L’analyse des données sur plusieurs années met en évidence des

tendances intéressantes. Dans l'ensemble, les données montrent que les salaires des diplômés de Polytech Montpellier varient avec le temps post-diplôme. Les diplômés les plus récents ont tendance à commencer avec des salaires plus bas, ce qui peut refléter des débuts de carrière à des niveaux d'entrée plus modestes. Avec l'expérience, les salaires ont tendance à augmenter, mais il existe également une variabilité croissante des salaires parmi les diplômés, ce qui peut être attribué à des trajectoires professionnelles diverses, des opportunités de marché changeantes et des choix personnels de carrière.

Ces observations soulignent l'importance de prendre en compte l'évolution des conditions du marché de l'emploi et des choix de carrière individuels lors de la planification de carrière et de la gestion des finances personnelles. Il est également essentiel pour les institutions éducatives de fournir des informations pertinentes aux étudiants pour les aider à comprendre les réalités salariales dans leur domaine d'études et à prendre des décisions éclairées sur leur avenir professionnel.

- Le secteur : L'analyse des données sur les salaires des diplômés de Polytech Montpellier montre une tendance générale où les diplômés du secteur privé ont tendance à gagner mieux que ceux du secteur public. Cependant, il existe des variations importantes en fonction de divers facteurs, notamment l'expérience, l'industrie et la spécialisation professionnelle. Le choix entre le secteur public et le secteur privé devrait être basé sur des considérations personnelles, professionnelles et de style de vie, en tenant compte des différences salariales potentielles.
- L'anciennete: une plus grande ancienneté dans le milieu du travail est corrélée à des salaires moyens plus élevés, ce qui suggère que plus vous êtes ancien, plus vous avez de chances d'obtenir un meilleur salaire. Cependant, en 2022, des facteurs économiques ou des choix de carrière différents pourraient avoir contribué à des salaires plus variés pour les diplômés plus récents.
- Le pays: En conclusion, pour toutes les années examinées, il apparaît qu'il existe un meilleur salaire moyen à l'étranger qu'en France pour les diplômés de Polytech Montpellier. Cette tendance suggère une corrélation entre le choix de travailler à l'étranger et des perspectives salariales plus favorables par rapport à une carrière en France.
- Le type de contrat: En général, il est observé que les salariés en CDI ont des salaires moyens plus élevés que ceux en CDD. Cependant, pour d'autres types de contrats, il existe une diversité significative d'une année à l'autre, ce qui rend difficile toute comparaison directe des salaires.
- Le type de formation: Au fil du temps, on observe une évolution dans les salaires en fonction du type de formation. Initialement, la professionnalisation avait le salaire moyen le plus élevé, suivi de près par l'apprentissage, puis la formation initiale. Cependant, cette tendance a changé, avec une augmentation des salaires moyens pour l'apprentissage, suivi par la formation initiale, et enfin la professionnalisation. Cette transformation suggère un changement dans la valorisation des différents types de formation au fil des années.
- La filière: En ce qui concerne les filières, il est clair que le choix de la filière peut avoir un impact significatif sur le salaire. Les filières telles que IG, MSI, et MI ont tendance à présenter des médianes de salaire plus élevées, ainsi que des étendues plus importantes vers des salaires élevés. Cela suggère que ces filières peuvent offrir des opportunités de rémunération plus attractives pour les diplômés.

Etudes approfondies

Analyse plus pousées sur les critères ayant un plus fort impact

Analyse des données des différentes années

En se basant sur l'analyse des différentes années et de façon plus globale, voici ce que l'on peut en tirer.

Critères ayant un faible impact ou aucun impact : - Genre - Année de diplôme - Ancienneté

Critère ayant un impact modéré : - Type de formation (initiale, apprentissage ou contrat de professionnalisation) - Filière - Nature du contrat

Critère ayant un fort impact : - Pays (Etranger/France ainsi qu'à l'étranger)

Les Critères pouvant être sélectionné pour une analyse plus poussée sont : - Pays - Nature du contrat - Filière - Type de formation - Genre (Certes avec un faible impact général, mais pouvant avoir un meilleur impact s'il y a plus de nuances.)

Analyse plus pousées sur les critères ayant un plus fort impact

Vu que l'on a récupérer des données redondantes entre les années, le choix de l'année devrait avoir un faible impact sur les analyses approfondies.

Chargement des données

```
# Trying ISO-8859-1
base_data <- read.csv("data/data_2023.csv", sep = ",", fileEncoding = "UTF-8")

# Lists specifying which columns to convert to factors and numeric
factor_cols <- c("Date", "Identifiant", "Genre", "Annee_diplome","Type_formation","Filiere","Situation_
numeric_cols <- c("Anciennete","Salaire_annuel_brut_primes")

# Convert columns to factors
base_data[factor_cols] <- lapply(base_data[factor_cols], as.factor)

# Convert columns to numeric
base_data[numeric_cols] <- lapply(base_data[numeric_cols], as.numeric)


filtered_data <- remove_rows_by_value(base_data, "Situation_actuelle", "En recherche emploi")
filtered_data <- remove_rows_by_value(filtered_data, "Situation_actuelle", "En poursuite d'études (hors
filtered_data <- remove_rows_by_value(filtered_data, "Situation_actuelle", "These")
filtered_data <- remove_rows_by_value(filtered_data, "Situation_actuelle", "Sans activité")
filtered_data <- remove_rows_by_value(filtered_data, "Situation_actuelle", "Volontariat")
filtered_data <- remove_rows_by_value(filtered_data, "Situation_actuelle", "En création d'entreprise /r
filtered_data <- remove_rows_by_value(filtered_data, "Situation_actuelle", "Etudes")

filtered_data <- remove_rows_by_value(filtered_data, "Nature_contrat", "")

filtered_data <- remove_rows_by_value(filtered_data, "Salaire_annuel_brut_primes", "")

filtered_data <- remove_rows_by_value(filtered_data, "Nature_contrat", "CDD a la cimenterie Lafarge de 




base_data <- filtered_data
```

ANOVA

```
raw_data <- base_data


# ANOVA pour la variable "Pays"
anova_pays <- aov(Salaire_annuel_brut_primes ~ Pays, data = base_data)

# ANOVA pour la variable "Nature_contrat"
```

```r
anova_contrat <- aov(Salaire_annuel_brut_primes ~ Nature_contrat, data = base_data)

# ANOVA pour la variable "Filiere"
anova_filiere <- aov(Salaire_annuel_brut_primes ~ Filiere, data = base_data)

# ANOVA pour la variable "Type_formation"
anova_formation <- aov(Salaire_annuel_brut_primes ~ Type_formation, data = base_data)

# ANOVA pour la variable "Genre"
anova_genre <- aov(Salaire_annuel_brut_primes ~ Genre, data = base_data)

# Afficher les résultats
summary(anova_pays)
```

```
##              Df    Sum Sq   Mean Sq F value Pr(>F)
## Pays          1 9.647e+09 9.647e+09   118.6 <2e-16 ***
## Residuals   513 4.173e+10 8.134e+07
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary(anova_contrat)
```

```
##                 Df    Sum Sq   Mean Sq F value   Pr(>F)
## Nature_contrat   7 2.434e+09 347697616   3.602 0.000848 ***
## Residuals      507 4.894e+10  96530440
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary(anova_filiere)
```

```
##              Df    Sum Sq   Mean Sq F value   Pr(>F)
## Filiere       9 6.163e+09 684756953   7.648 1.46e-10 ***
## Residuals   505 4.521e+10  89528720
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary(anova_formation)
```

```
##                 Df    Sum Sq   Mean Sq F value Pr(>F)
## Type_formation   2 7.561e+08 378058848   3.824 0.0225 *
## Residuals      512 5.062e+10  98864646
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary(anova_genre)
```

```
##              Df    Sum Sq   Mean Sq F value  Pr(>F)
## Genre         2 2.018e+09 1.009e+09   10.47 3.5e-05 ***
## Residuals   512 4.936e+10 9.640e+07
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

-Pays : Degrés de liberté (Df) : 1 pour le groupe, 513 pour les résidus, indiquant probablement une comparaison entre deux pays. Somme des carrés (Sum Sq) : Indique la variabilité totale attribuable au groupe et la variabilité non expliquée (résidus). Moyenne des carrés (Mean Sq) : Variabilité moyenne par degré de liberté. Valeur F : 118.6, qui est le rapport de la variance moyenne due au facteur Pays par rapport à la variance résiduelle. Pr(>F) : La probabilité d'observer une telle valeur F ou plus extrême si l'hypothèse nulle est vraie.

Ici, p < 2.2e-16, indiquant que les différences de salaires entre les pays sont statistiquement significatives.

-Nature_contrat : Df : 7 pour le groupe, 507 pour les résidus, indiquant une comparaison entre huit types de contrats. F value : 3.602, suggérant qu'il y a une différence entre les groupes. Pr(>F) : p < 0.001, indiquant que la nature du contrat a un effet significatif sur les salaires.

-Filiere : Df : 9 pour le groupe, ce qui suggère dix filières différentes. F value : 7.648, indiquant une variation significative entre les filières. Pr(>F) : p < 2.2e-16, confirmant la signification statistique des différences de salaires entre les filières.

-Type_formation : Df : 2 pour le groupe, suggérant trois types de formation. F value : 3.824, qui est à la limite de la significativité. Pr(>F) : p = 0.0225, montrant que les différences de salaires entre les types de formation sont statistiquement significatives, mais moins fortement que pour les pays ou les filières.

- Genre : Df : 2 pour le groupe, indiquant trois catégories de genre. F value : 10.47, qui est élevée, suggérant une différence significative. Pr(>F) : p < 3.5e-05, indiquant que les différences de salaires entre les genres sont statistiquement très significatives.

Les résultats suggèrent donc que tous les facteurs testés (pays, nature du contrat, filière, type de formation, genre) ont un impact significatif sur les salaires, avec des degrés de significativité variables.

Régression multiple

Explication L'analyse de régression multiple indique que le modèle a une certaine significativité globale, mais la significativité individuelle des variables peut varier. Certains pays, types de contrat, filières et genres semblent avoir une influence significative sur le salaire_annuel, tandis que d'autres variables peuvent ne pas être significatives.

```
raw_data <- base_data


# Assuming raw_data is your data frame containing the relevant variables

# Convert categorical variables to factors if not done already
raw_data$Pays <- as.factor(raw_data$Pays)
raw_data$Nature_contrat <- as.factor(raw_data$Nature_contrat)
raw_data$Filiere <- as.factor(raw_data$Filiere)
raw_data$Type_formation <- as.factor(raw_data$Type_formation)
raw_data$Genre <- as.factor(raw_data$Genre)

# Create a linear model
reg_model <- lm(Salaire_annuel_brut_primes ~ Pays + Nature_contrat + Filiere + Type_formation + Genre, d

# Summary of the regression model
summary(reg_model)
```

l'analyse

```
##
## Call:
## lm(formula = Salaire_annuel_brut_primes ~ Pays + Nature_contrat +
##     Filiere + Type_formation + Genre, data = raw_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -35242  -4501   -792   3528  67518
```

```
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         51899.6     2423.4  21.416  < 2e-16 ***
## PaysFrance                         -16909.3     1503.5 -11.247  < 2e-16 ***
## Nature_contratCDI                    4866.8     1281.7   3.797 0.000165 ***
## Nature_contratCDIC                   7035.8     6059.1   1.161 0.246127
## Nature_contratContrat d'expatriation -6286.7     8668.8  -0.725 0.468664
## Nature_contratContrat local         -14375.4     6414.8  -2.241 0.025472 *
## Nature_contratCTT                    -6222.2     6111.5  -1.018 0.309127
## Nature_contratde cdd a cdi            8954.2     8487.9   1.055 0.291972
## Nature_contratInterim                 6125.4     8506.5   0.720 0.471811
## FiliereGBA                          -12327.7     3540.8  -3.482 0.000543 ***
## FiliereIG                            -8311.9     3526.8  -2.357 0.018825 *
## FiliereMAT                          -12375.3     3569.1  -3.467 0.000572 ***
## FiliereMAT                          -17948.9     9031.3  -1.987 0.047430 *
## FiliereMEA                          -10717.7     3500.4  -3.062 0.002320 **
## FiliereMI                           -12686.6     3752.2  -3.381 0.000779 ***
## FiliereMSI                            1338.7     2142.8   0.625 0.532434
## FiliereSE                            -1888.1     1928.2  -0.979 0.327953
## FiliereSTE                          -15059.4     3557.8  -4.233 2.75e-05 ***
## Type_formationInitiale                8043.6     3225.9   2.493 0.012978 *
## Type_formationProfessionalisation     9234.2     3290.4   2.806 0.005208 **
## GenreHomme                            2387.1      932.9   2.559 0.010805 *
## GenreNe souhaite pas repondre         2816.9     8442.1   0.334 0.738768
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8357 on 493 degrees of freedom
## Multiple R-squared:  0.3298, Adjusted R-squared:  0.3012
## F-statistic: 11.55 on 21 and 493 DF,  p-value: < 2.2e-16
```

La sortie de la régression multiple fournit plusieurs informations importantes pour interpréter le modèle. Voici une interprétation détaillée :

Résumé Global du Modèle

- Call: La formule utilisée pour ajuster le modèle.
- Residuals: Statistiques des résidus, y compris le minimum, les quartiles et le maximum.

Coefficients du Modèle Chaque coefficient représente l'effet de la variable correspondante sur le salaire annuel, toutes les autres variables étant maintenues constantes.

- Estimate: La valeur estimée du coefficient.
- Std. Error: L'écart type estimé du coefficient.
- t value: La statistique de test (t-value) pour tester si le coefficient est significativement différent de zéro.
- Pr(>|t|): La valeur p associée à la statistique de test, indiquant la significativité du coefficient.

Intercept Intercept (1 not defined because of singularities): La constante (l'effet lorsque toutes les variables explicatives sont nulles). Cependant, elle peut ne pas être bien définie en raison de singularités dans les données.

Interprétation des données Ici dans les données on peut remarquer que le pays à généralement un impact significatif (en se référant aux étoiles) ainsi que le type de contrat, en revanche la filère, le type de formation et le genre n'ont pas d'impact significatif

Cette régression a été effectuée pour comprendre comment divers facteurs (comme le pays, la nature du contrat, le type de formation et le genre) affectent le salaire annuel brut (plus primes) des individus. Voici une interprétation des résultats clés :

-PaysFrance : La comparaison avec le pays de référence (probablement un autre pays non spécifié) montre une différence négative de 16909.3, ce qui signifie que le salaire en France est en moyenne inférieur à celui du pays de référence, et cette différence est statistiquement significative ($p < 2e\text{-}16$).

-Nature_contratCDI, Nature_contratCDD, etc. : Ces coefficients représentent la différence dans le salaire annuel brut en fonction du type de contrat par rapport à la catégorie de référence (probablement un autre type de contrat non spécifié). Par exemple, être en CDI est associé à une augmentation moyenne de salaire de 4866.8 par rapport au contrat de référence. Les contrats CDI et d'expatriation ne sont pas statistiquement significatifs ($p > 0.05$), tandis que les contrats locaux et les CTT (contrat de travail temporaire) le sont.

-FiliereIG, FiliereMAT, etc. : Ces coefficients indiquent les différences de salaire associées à chaque filière par rapport à la filière de référence. La plupart des filières montrent des différences négatives, indiquant des salaires inférieurs à la filière de référence, avec des niveaux de significativité variés.

-Type_formationInitiale et Type_formationProfessionnalisation : Être dans une formation initiale est associé à une augmentation moyenne de salaire de 9043.6 par rapport à la catégorie de référence (probablement l'apprentissage), tandis que la professionalisation est associée à une augmentation de 8234.2.

-GenreHomme : Être un homme est associé à une augmentation moyenne de salaire de 2387.1 par rapport à la catégorie de référence (probablement les femmes), et cette différence est statistiquement significative ($p < 0.05$).

-Ne souhaite pas répondre : Ceux qui n'ont pas souhaité répondre au sujet de leur genre ont une différence de salaire non significative par rapport à la catégorie de référence.

En conclusion, cette régression montre que plusieurs facteurs influencent le salaire annuel brut des individus. Le pays, certains types de contrat et le genre ont des effets significatifs. Le modèle a un R^2 ajusté de 0.3012, ce qui signifie qu'environ 30.12% de la variabilité du salaire annuel brut est expliquée par les variables incluses dans le modèle. La signification globale du modèle est forte ($p < 2.2e\text{-}16$), indiquant que le modèle est significatif dans la prédiction du salaire annuel brut.

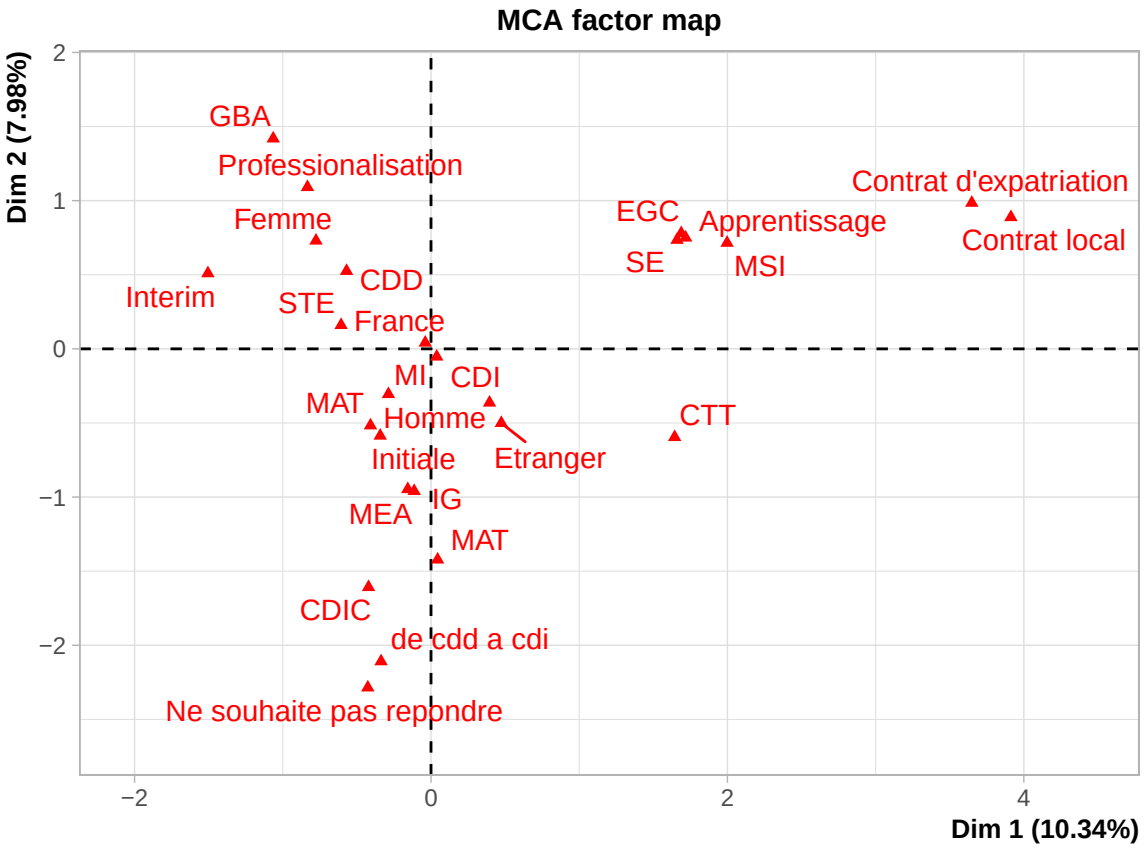
Conclusion En résumé, pour conclure, il apparaît que les filières ont un impact significatif, conforme à nos attentes initiales. De même, le type de contrat et le pays de travail présentent également un impact notable. Bien que les analyses ANOVA aient confirmé l'impact des variables, il semble que le type de formation et le genre affichent une influence légère et trop faible pour être clairement discernée.

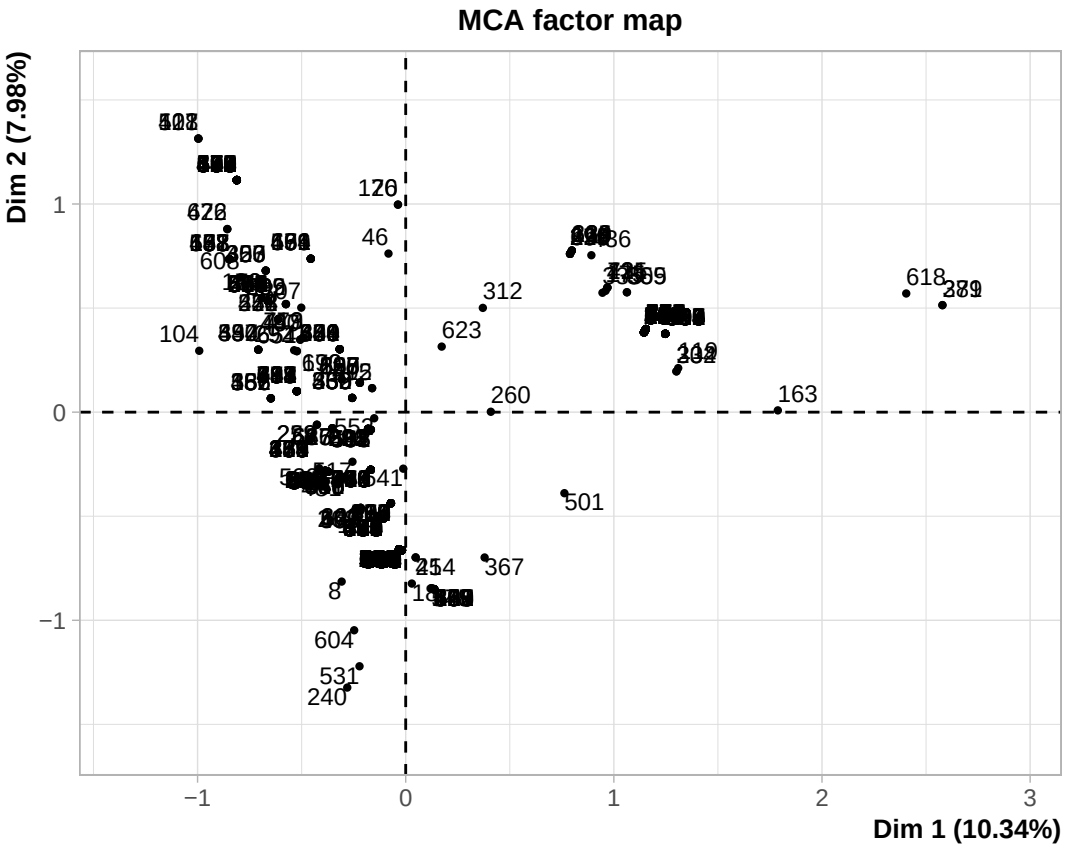
AFC(M)

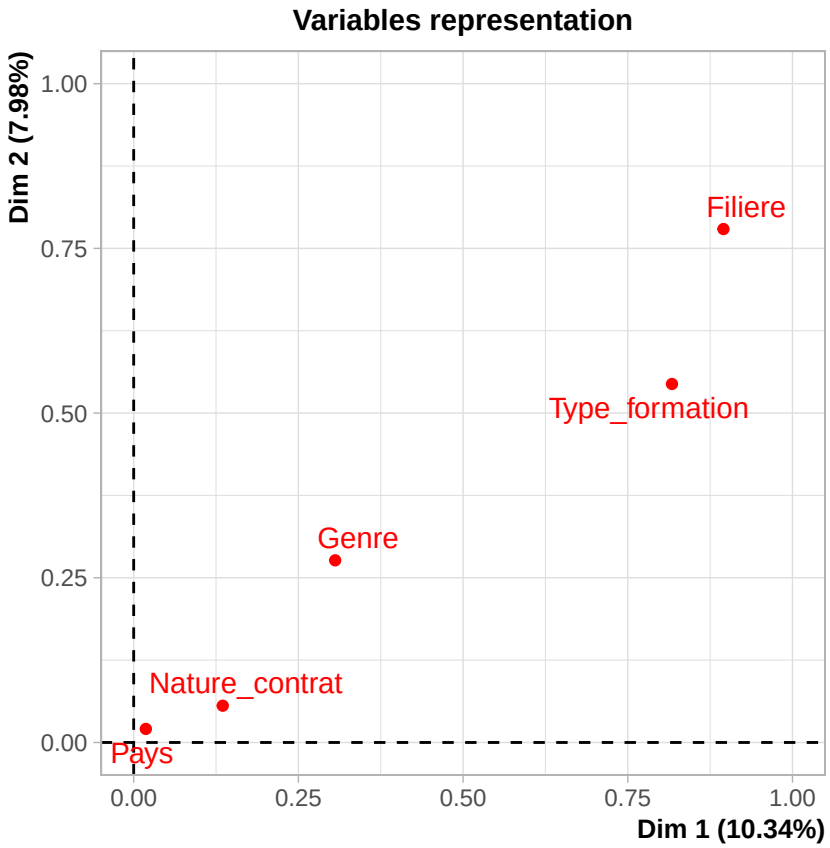
```
library(FactoMineR)
```

```
## Warning: package 'FactoMineR' was built under R version 4.2.3
```

```
# Sélectionner les variables d'intérêt
variables_mca <- c("Pays", "Nature_contrat", "Filiere", "Type_formation", "Genre", "Salaire_annuel_brut_

# Sous-ensemble des données avec les variables sélectionnées
data_mca <- base_data[, variables_mca]
# Supprimer les lignes avec des valeurs manquantes
data_mca <- na.omit(data_mca)

# Effectuer l'Analyse en Correspondances Multiples (MCA)
mca_result <- MCA(data_mca[, -6])
```


MCA factor map
Dim 2 (7.98%)
Dim 1 (10.34%)
2
1
0
−1
−2
−2
0
2
4
GBA
Professionalisation
Femme
Interim
CDD
STE
France
Contrat d'expatriation
Contrat local
EGC
Apprentissage
SE
MSI
MI
CDI
MAT
Homme
Initiale
Etranger
CTT
MEA
IG
MAT
CDIC
de cdd a cdi
Ne souhaite pas repondre


MCA factor map
Dim 2 (7.98%)
Dim 1 (10.34%)
1
0
−1
−1
0
1
2
3
46
312
623
618
104
260
163
501
367
8
604
531
240



Il est difficile de définir des analayse poussées avec cette MCA, il n'y a pas réellement de données sortant du groupe.

Conclusion

L'analyse de la régression linéaire multiple fournit un aperçu significatif de la manière dont divers facteurs contribuent aux différences de salaire annuel brut parmi les individus. En examinant les coefficients et leur pertinence statistique, on observe que le pays de travail, la filière d'études, le type de contrat, le type de formation et le genre ont une influence notable sur les niveaux de rémunération.

Le pays de travail émerge comme le déterminant le plus puissant des salaires, suggérant que les différences économiques internationales jouent un rôle crucial dans la détermination des salaires. Cette distinction peut refléter les disparités en matière de coût de la vie, de politiques fiscales et de demande de compétences spécifiques à chaque pays.

La filière d'études est également un prédicteur significatif des variations de salaire. Les résultats indiquent que certaines filières conduisent à des salaires supérieurs, ce qui peut être dû à la demande sur le marché du travail pour des compétences spécialisées ou à des parcours professionnels plus lucratifs associés à ces filières.

Le type de contrat détient la troisième place en termes d'impact sur les salaires, avec des contrats à durée indéterminée (CDI) qui tendent à offrir de meilleures rémunérations par rapport aux contrats temporaires ou précaires. Cette observation confirme la valeur associée à la sécurité de l'emploi et aux avantages liés aux postes permanents.

En ce qui concerne le type de formation, les formations initiales et de professionalisation semblent être associées à des salaires plus élevés, soulignant l'importance de l'éducation continue et de l'acquisition de compétences en vue d'améliorer les perspectives salariales.

Enfin, le genre est une variable qui continue d'influer sur le salaire, mettant en évidence l'écart de rémunération persistant entre les hommes et les femmes. Les hommes gagnent, en moyenne, plus que les femmes, ce qui appelle à une action continue pour l'équité salariale.

Globalement, ces facteurs ensemble expliquent une portion significative, bien que non exhaustive, de la variabilité des salaires, laissant une marge importante pour d'autres influences non capturées par le modèle. Cette analyse confirme non seulement l'importance des facteurs individuels mais aussi la complexité de la structure salariale qui est affectée par un entrelacement de divers éléments. Cela souligne la nécessité d'approches holistiques pour comprendre pleinement les dynamiques salariales et pour élaborer des politiques visant à promouvoir des opportunités économiques équitables pour tous.

Déroulement du projet

Pour exposer la progression de manière chronologique, initialement, nous avons rencontré des complications liées à la sélection des critères appropriés. Par exemple, nous avions initialement opté pour la sélection des intitulés de postes et des entreprises, mais le problème résidait dans la multitude de valeurs différentes. Il devenait ainsi difficile d'entreprendre une étude sans investir un temps considérable dans la consolidation des données en raison de l'impossibilité de les automatiser.

Par la suite, nous sommes passés à la phase d'harmonisation des données, où de nouveaux problèmes ont émergé. En premier lieu, il était nécessaire de déterminer comment harmoniser les colonnes, puis de normaliser les valeurs. Il fallait veiller à ce que, d'un fichier à l'autre, deux valeurs signifiant la même chose soient strictement équivalentes, afin d'élaborer un script unique pour l'analyse des données.

Enfin, dans la phase de développement du script d'analyse, nous n'avons pas rencontré de problèmes majeurs, mis à part ceux liés à la diversité des sources et à l'adaptation nécessaire pour chaque fichier.

La répartition des tâches est la suivante :

Nicolas : - Création d'un script d'analyse primaire (diagrammes boxplot simples mais variés pour comprendre les données initiales) - Mise en place d'un document partagé pour l'harmonisation, avec explication aux autres participants. - Création d'un fichier partiellement harmonisé pour le script primaire. - Adaptation du script primaire en script final pour l'analyse globale des données. - Sélection des critères pour l'étude approfondie. - Création des scripts pour l'étude approfondie (ANOVA, AFC(M) et Régression Multiple). - Amélioration du site.

Assia : - Tri de la moitié des données. - Harmonisation de la moitié des données. - Création du script final avec tous les diagrammes utilisés. - Harmonisation des scripts selon les fichiers. - Rédaction du rendu. - Analyse de tous les diagrammes. - Création script AFCM. - Sélection des critères pour l'étude approfondie. - Analyse des études approfondies.

Tom : - Harmonisation de la moitié des données. - Création du site.