Correction Devoir Maison

Analyse du *gender wage gap* en France

Florentine Oliveira
2025-02-18

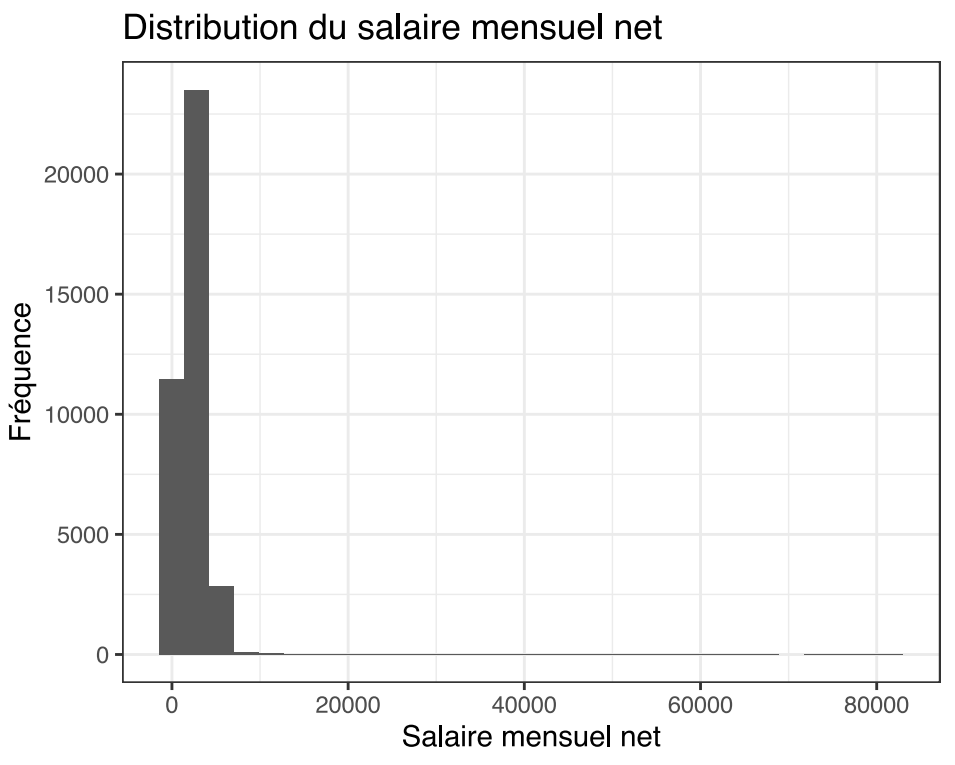
Partie 1: Traitement de la variable de salaire

Cette première partie vise à examiner et préparer la variable de salaire mensuel net pour l'analyse de l'écart de salaire H/F ultérieure. L'objectif est de comprendre la distribution des salaires, d'identifier et de traiter les éventuels problèmes engendrés par l'utilisation de données brutes.

Question 1 (1pt)

Représenter graphiquement la distribution du salaire net mensuel des individus sous la forme d'un histogramme, en ajoutant un titre et le nom des axes. Commentez.

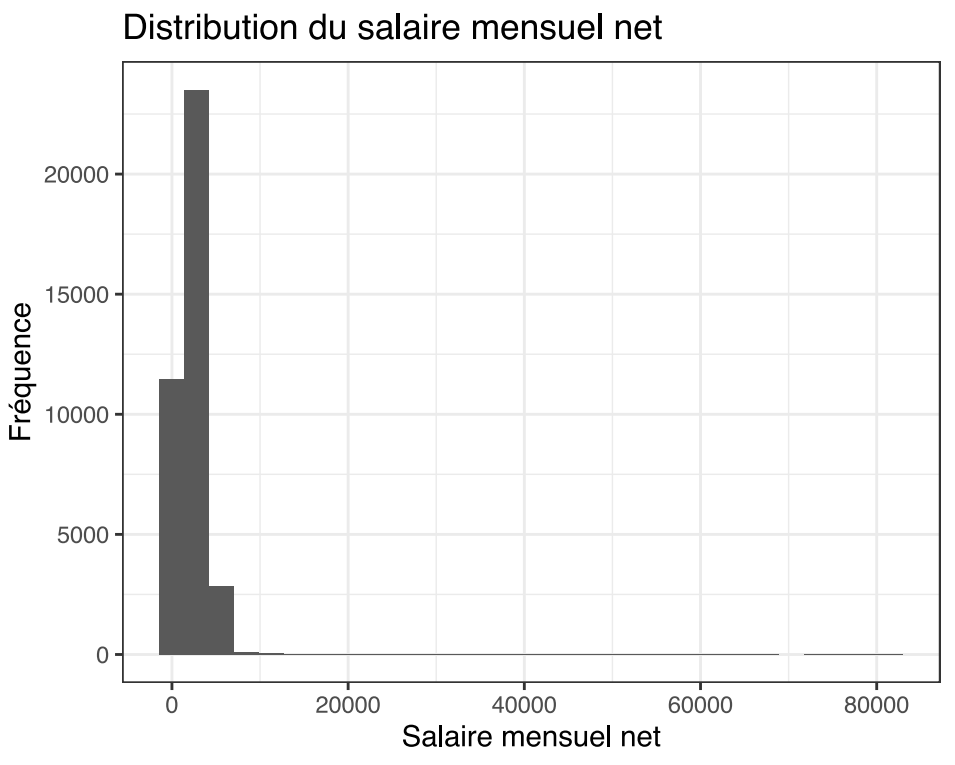
```
ggplot(data, aes(x = wage)) +
  geom_histogram() +
  labs(title="Distribution du salaire mensuel net",
       x = "Salaire mensuel net",
       y = "Fréquence") +
  theme_bw()
```



Question 1 (1pt)

Représenter graphiquement la distribution du salaire net mensuel des individus sous la forme d'un histogramme, en ajoutant un titre et le nom des axes. **Commentez**.

```
ggplot(data, aes(x = wage)) +
  geom_histogram() +
  labs(title="Distribution du salaire mensuel net",
       x = "Salaire mensuel net",
       y = "Fréquence") +
  theme_bw()
```



- Distribution *right-skewed*
- Longue queue de distribution à droite
- Suspicion de valeurs extrêmes

Question 2 (2pts)

Quel peut être le problème induit par l'utilisation de ces données "brutes" de salaire dans un modèle économétrique? Comment peut-on traiter les données en conséquence?

Problèmes:

- Présence d'outliers peut tirer les coefficients estimés vers le haut (la moyenne est sensible aux valeurs extrêmes, pas le cas de la variance)

Soultions:

- Winsoring
- Log

Question 3 (1pt)

Calculez les percentiles du salaire mensuel net (*hint: utiliser la fonction* `quantile()`). Stockez la valeur du 99ème percentile dans `q_99`. Indiquez la valeur du 99ème percentile et interprétez.

```
q_99 = quantile(data$wage, seq(0,1,0.01), na.rm = T)[100]
q_99
```

```
##      99%
## 7485.299
```

99% des individus de l'échantillon ont un salaire inférieur à 7485,299€, 1% ont un salaire supérieur.

Question 4 (1pt)

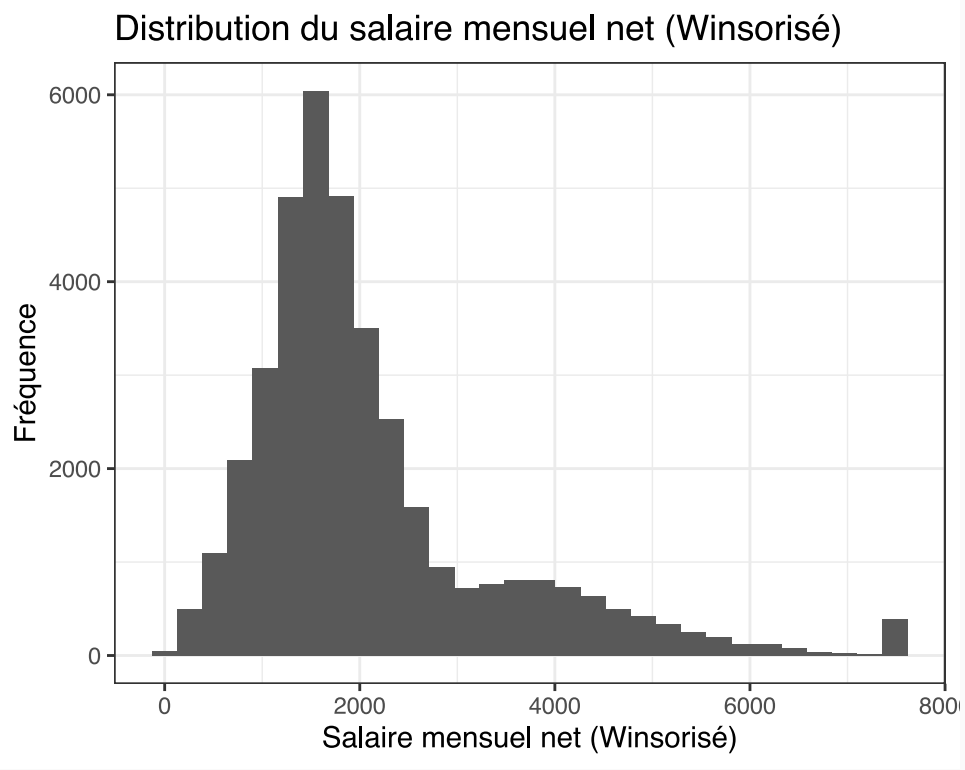
- **Créez la variable** `wage_winsor` qui correspond au salaire mensuel net, `wage`, où les 1% des valeurs les plus élevées sont remplacées par la valeur du 99ème percentile du salaire mensuel net
- Représentez graphiquement la distribution de `wage_winsor`. Commentez.

```
data$wage_winsor=ifelse(data$wage>q_99,q_99,data$wage)
```

Question 4 (1pt)

- Créez la variable `wage_winsor` qui correspond au salaire mensuel net, `wage`, où les 1% des valeurs les plus élevées sont remplacées par la valeur du 99ème percentile du salaire mensuel net
- **Représentez graphiquement la distribution de** `wage_winsor`. Commentez.

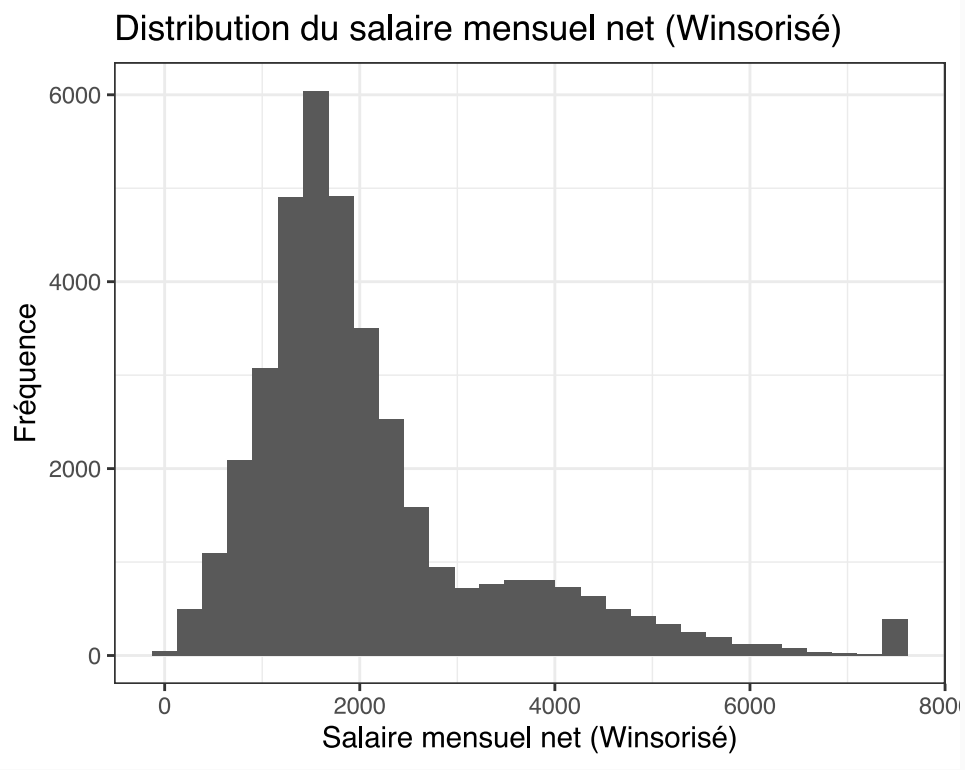
```
data$wage_winsor=ifelse(data$wage>q_99,q_99,data$wage)

ggplot(data, aes(x = wage_winsor)) +
  geom_histogram() +
  labs(title="Distribution du salaire mensuel net (Winsc
       x = "Salaire mensuel net (Winsorisé)",
       y = "Fréquence") +
  theme_bw()
```



Question 4 (1pt)

- Créez la variable `wage_winsor` qui correspond au salaire mensuel net, `wage`, où les 1% des valeurs les plus élevées sont remplacées par la valeur du 99ème percentile du salaire mensuel net
- Représentez graphiquement la distribution de `wage_winsor`. **Commentez**.

```
data$wage_winsor=ifelse(data$wage>q_99,q_99,data$wage)

ggplot(data, aes(x = wage_winsor)) +
  geom_histogram() +
  labs(title="Distribution du salaire mensuel net (Winsc
       x = "Salaire mensuel net (Winsorisé)",
       y = "Fréquence") +
  theme_bw()
```



- Excess-mass à q_99 due au winsoring
- Distribution plus "lisible"
- Encore right-skewed

Question 5 (1pt)

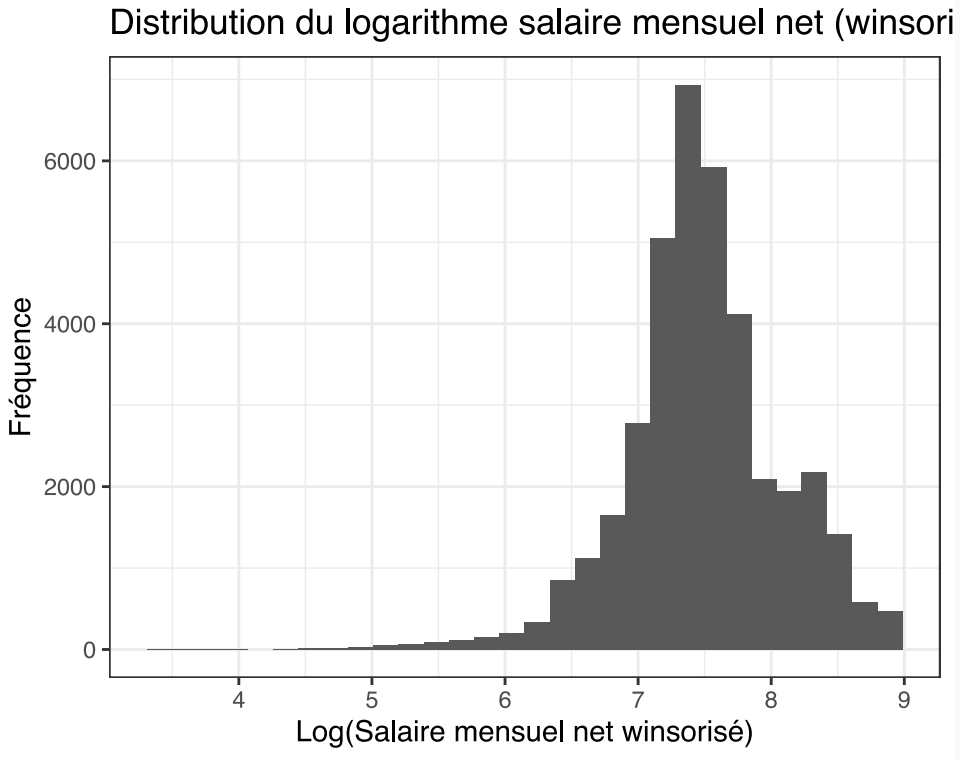
- **Créez la variable** `log_wage_winsor`, le logarithme de `wage_winsor`
- Représentez graphiquement la distribution de `log_wage_winsor`. Commentez.

```
data$log_wage_winsor = log(data$wage_winsor)
```

Question 5 (1pt)

- Créez la variable `log_wage_winsor`, le logarithme de `wage_winsor`
- **Représentez graphiquement la distribution de** `log_wage_winsor`. Commentez.

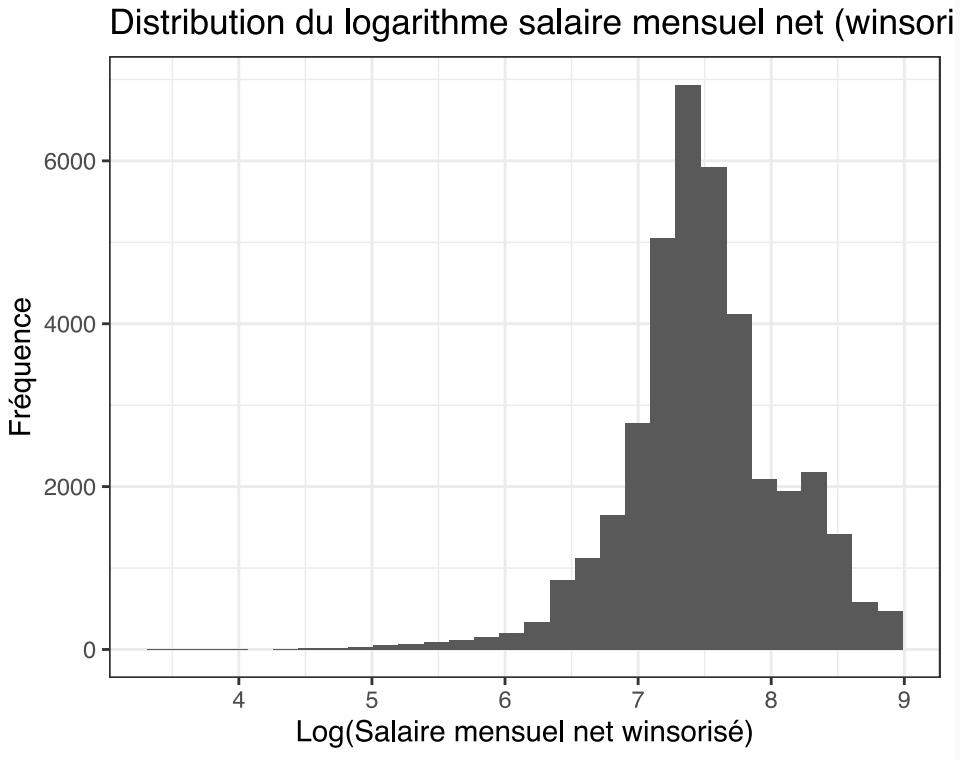
```
ggplot(data, aes(x = log_wage_winsor)) +
  geom_histogram() +
  labs(title="Distribution du logarithme salaire mensuel
       x = "Log(Salaire mensuel net winsorisé)",
       y = "Fréquence") +
  theme_bw()
```



Question 5 (1pt)

- Créez la variable `log_wage_winsor`, le logarithme de `wage_winsor`
- Représentez graphiquement la distribution de `log_wage_winsor`. **Commentez**.

```
ggplot(data, aes(x = log_wage_winsor)) +
  geom_histogram() +
  labs(title="Distribution du logarithme salaire mensuel
       x = "Log(Salaire mensuel net winsorisé)",
       y = "Fréquence") +
  theme_bw()
```



- Log permet de recentrer davantage les observations
- Apparence qui se rapproche d'une loi normale

Partie 2: Statistiques Descriptives

Cette partie a pour objectif d'explorer les données de l'enquête emploi afin de mieux appréhender les écarts entre femmes et hommes en termes de niveau de diplôme, d'activité, d'occupation et de rémunération.

Partie A: Diplôme

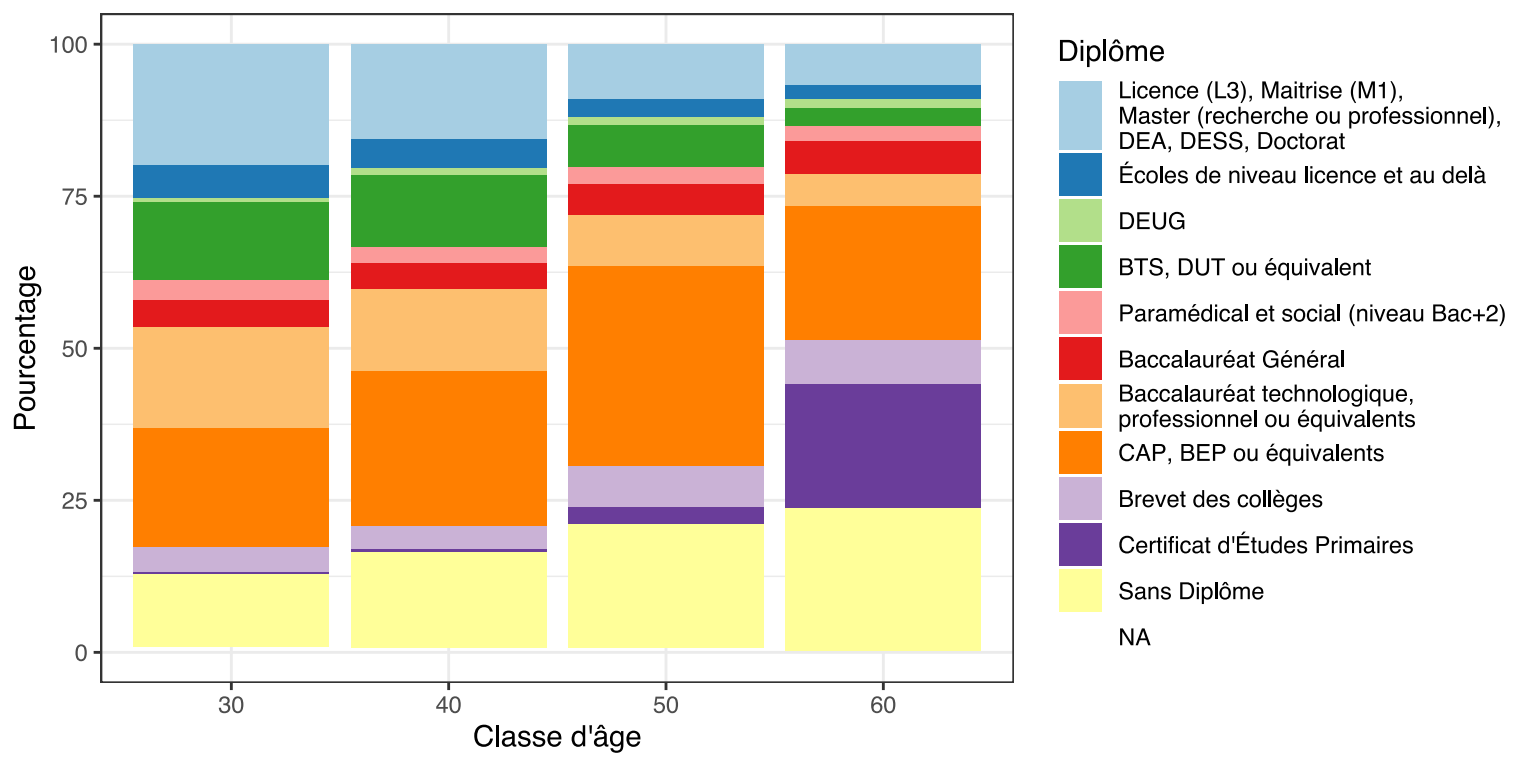
Question 6 (1pt)

Représenter graphiquement la distribution du niveau de diplôme selon la classe d'âge. Conservez uniquement les individus âgés de 30 ans ou plus (*hint 1: utiliser la variable* `AGE5`; *hint 2: on pourra par exemple faire un barplot*). Veillez à ce que la légende indique le libellé du niveau de diplôme et non la modalité correspondante. Commentez.

```
library(RColorBrewer) #<< Palettes color-blind friendly

data %>%
  filter(AGE5 =/= 15) %>%
  group_by(AGE5, DIP11) %>%
  summarise(count = n()) %>%
  mutate(perc = count/sum(count)) %>%
  ggplot(aes(x = AGE5, y = perc*100, fill = DIP11)) +
  geom_bar(stat="identity") +
  labs(x = "Classe d'âge", y = "Pourcentage") +
  scale_fill_brewer(palette = "Paired", name = "Diplôme",
                    labels = c("10" = "Licence (L3), Maitrise (M1),\nMaster (recherche ou professionnel),\nDEA, DESS,
                               "11" = "Écoles de niveau licence et au delà",
                               "30" = "DEUG", "31" = "BTS, DUT ou équivalent",
                               "33" = "Paramédical et social (niveau Bac+2)", "41" = "Baccalauréat Général",
                               "42" = "Baccalauréat technologique,\nprofessionnel ou équivalents",
                               "50" = "CAP, BEP ou équivalents","60" = "Brevet des collèges",
                               "70" = "Certificat d'Études Primaires", "71" = "Sans Diplôme")) +
  theme_bw()
```

Question 6 (1pt)

Représenter graphiquement la distribution du niveau de diplôme selon la classe d'âge. Conservez uniquement les individus âgés de 30 ans ou plus (*hint 1: utiliser la variable* `AGE5`; *hint 2: on pourra par exemple faire un barplot*). Veillez à ce que la légende indique le libellé du niveau de diplôme et non la modalité correspondante. Commentez.



Question 6 (1pt)

Représenter graphiquement la distribution du niveau de diplôme selon la classe d'âge. Conservez uniquement les individus âgés de 30 ans ou plus (*hint 1: utiliser la variable* `AGE5`; *hint 2: on pourra par exemple faire un barplot*). Veillez à ce que la légende indique le libellé du niveau de diplôme et non la modalité correspondante. **Commentez**.

Ce que l'on observe: augmentation de la durée de scolarisation moyenne au fil des cohortes: les enfants sont moins nombreux à ne pas avoir de diplôme et atteignent de niveaux de diplômes plus élevés

- **Massification scolaire**: allongement de la durée des études et accès d'une large partie de la population à un niveau de qualification élevé, qui était auparavant réservé à une minorité d'élèves généralement issus des catégories sociales les plus privilégiées
    - Allongement de la durée de scolarité obligatoire (à 14 ans en 1936, 16 ans en 1967)
    - Loi Berthoin de 1959
    - Loi Haby de 1975 et mise en œuvre à la rentrée 1977 qui instaure le collège unique (tous les enfants entrent en sixième dans des classes indifférenciées)
- **Stratification scolaire**: multiplication des possibilités de diplôme, notamment avec l'apparition des filières professionnelles dans les années

Question 7 (1pt):

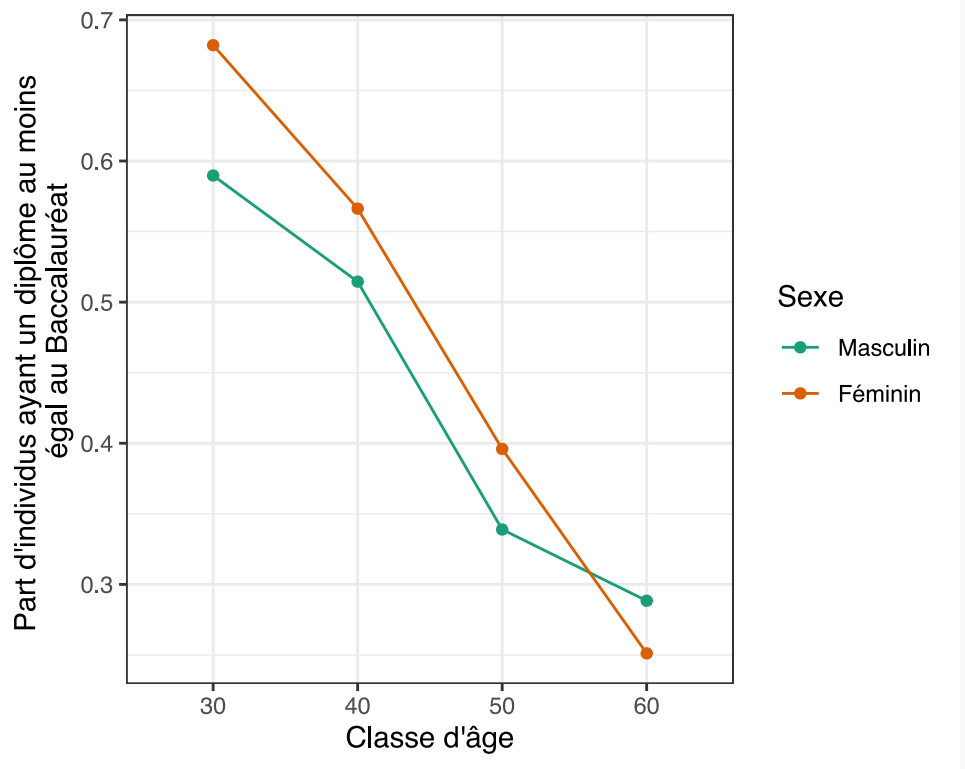
- **Créez une nouvelle variable `at_least_bac` qui vaut 1 si l'individu a un diplôme au moins égal au Bac, 0 sinon.**
- Comparez graphiquement la proportion d'individus ayant un diplôme au moins égal au Bac (`at_least_bac`) entre les hommes et les femmes pour chaque tranche d'âge. Représentez vos résultats graphiquement. Veillez à conserver uniquement les individus de 30 ans ou plus. Commentez.

```
data = data %>%
  mutate(at_least_bac = ifelse(DIP11 < 50, 1, 0))

data %>%
  filter(AGE5 =/= 15) %>%
  group_by(SEXE, AGE5) %>%
  summarise(mean_by_sexe = mean(at_least_bac, na.rm = T)) %>%
  ggplot(aes(x = AGE5, y = mean_by_sexe,
             group = SEXE, color = SEXE )) +
  geom_point() +
  geom_line() +
  labs(x = "Classe d'âge",
       y = "Part d'individus ayant un diplôme au moins égal au Baccalauréat") +
  scale_color_brewer(palette = "Dark2",
                     name = "Sexe",
                     labels = c("1" = "Masculin", "2" = "Féminin")) +
  theme_bw()
```

Question 7 (1pt)

- Créez une nouvelle variable `at_least_bac` qui vaut 1 si l'individu a un diplôme au moins égal au Bac, 0 sinon.
- **Comparez graphiquement la proportion d'individus ayant un diplôme au moins égal au Bac (`at_least_bac`) entre les hommes et les femmes pour chaque tranche d'âge**. Représentez vos résultats graphiquement. Veillez à conserver uniquement les individus de 30 ans ou plus. Commentez.

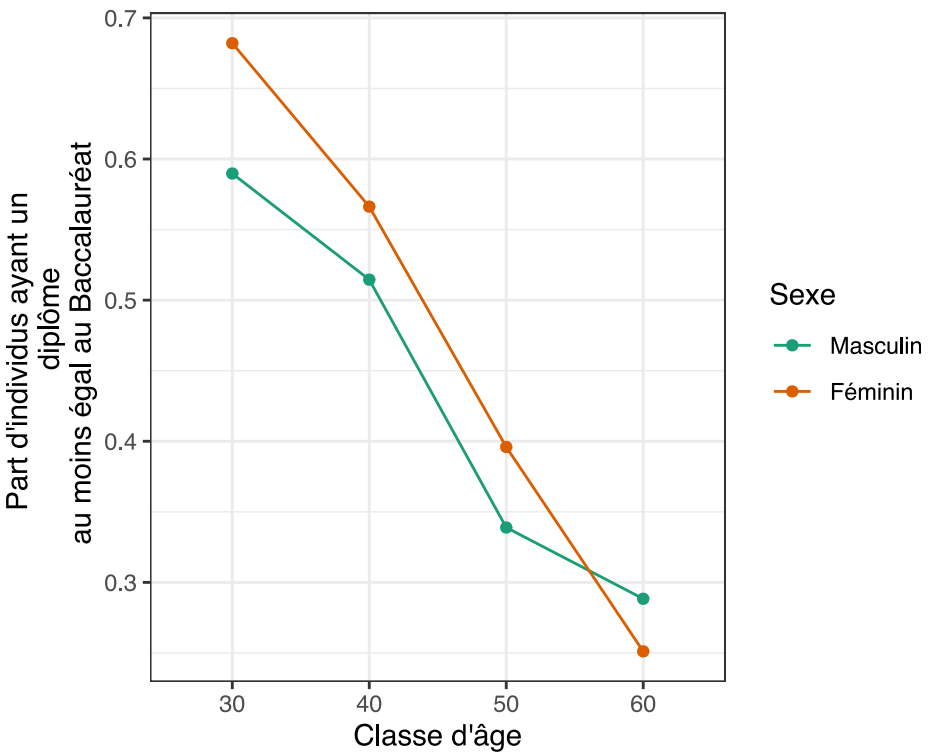
```
data = data %>%
  mutate(at_least_bac = ifelse(DIP11 < 50, 1, 0))

data %>%
  filter(AGE5 =/= 15) %>%
  group_by(SEXE, AGE5) %>%
  summarise(mean_by_sexe = mean(at_least_bac, na.rm = T)
  ggplot(aes(x = AGE5, y = mean_by_sexe,
             group = SEXE, color = SEXE )) +
  geom_point() +
  geom_line() +
  labs(x = "Classe d'âge",
       y = "Part d'individus ayant un diplôme au moins
       égal au Baccalauréat") +
  scale_color_brewer(palette = "Dark2",
                     name = "Sexe",
                     labels = c("1" = "Masculin", "2" = "
  theme_bw()
```



Question 7 (1pt)

- Créez une nouvelle variable `at_least_bac` qui vaut 1 si l'individu a un diplôme au moins égal au Bac, 0 sinon.
- Comparez graphiquement la proportion d'individus ayant un diplôme au moins égal au Bac (`at_least_bac`) entre les hommes et les femmes pour chaque tranche d'âge. Représentez vos résultats graphiquement. Veillez à conserver uniquement les individus de 30 ans ou plus. **Commentez**.

- Augmentation de la part d'individus ayant un diplôme au moins égal au baccalauréat au fil des cohortes
- Rattrapage des filles
  - Auparavant écartées de l'enseignement secondaire et supérieur (autorisées à passer le bac pour la première fois en 1924)



Partie B: Participation au marché du travail

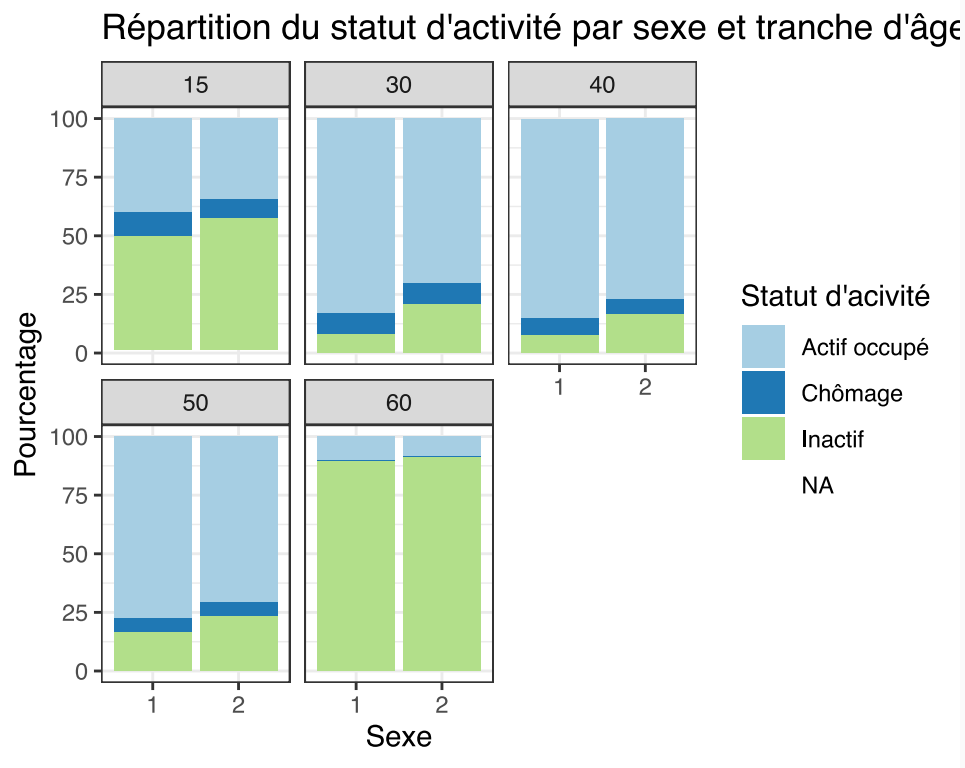
Question 8 (1pt)

Calculez la distribution de la variable `ACTEU`, par genre, par tranche d'âge, et par genre × tranche d'âge. Proposez une représentation graphique. Que peut-on dire de la participation au marché du travail des femmes?

```
data %>%
  group_by(AGE5, SEXE, ACTEU) %>%  # Groupement par tranche d'âge, sexe, statut
  summarise(count = n()) %>%
  group_by(AGE5, SEXE) %>%  # Regrouper uniquement par tranche d'âge et sexe pour recalculer les pourcentages
  mutate(perc = count / sum(count)) %>%
  ggplot(aes(x = as.factor(SEXE), y = perc * 100, fill = as.factor(ACTEU))) +
  geom_bar(stat = "identity", position = "stack") +
  labs(
    x = "Sexe",
    y = "Pourcentage",
    fill = "Statut d'activité",
    title = "Répartition du statut d'activité par sexe et tranche d'âge"
  ) +
  scale_fill_brewer(palette = "Paired",
                    name = "Statut d'acivité",
                    labels = c("1" = "Actif occupé", "2" = "Chômage", "3" = "Inactif")) +
  theme_bw() +
  facet_wrap(~ AGE5)
```

Question 8 (1pt)

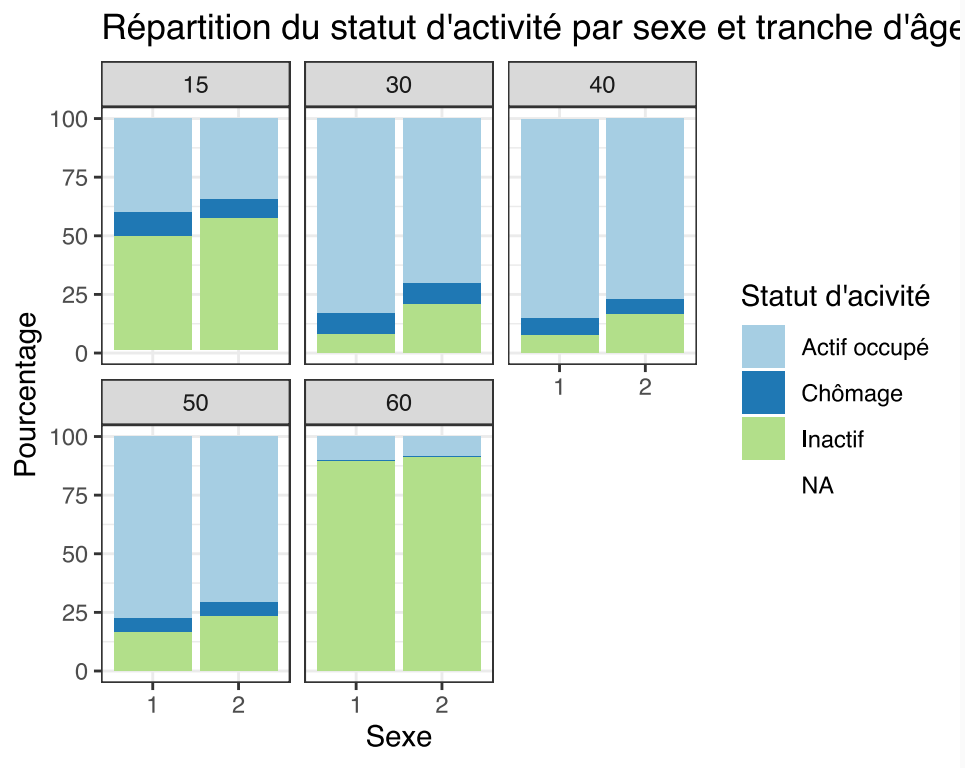
Calculez la distribution de la variable `ACTEU`, par genre, par tranche d'âge, et par genre × tranche d'âge. Proposez une représentation graphique. Que peut-on dire de la participation au marché du travail des femmes?



Question 8 (1pt)

Calculez la distribution de la variable `ACTEU`, par genre, par tranche d'âge, et par genre × tranche d'âge. Proposez une représentation graphique. Que peut-on dire de la participation au marché du travail des femmes?

- Les plus jeunes et plus âgés sont plus fréquemment inactifs que les autres tranches d'âge (parce qu'en études et à la retraite)
- Les femmes plus inactives que les hommes, tout au long de leur vie
  - Intégration du marché du travail différée liée à une grossesse?
  - Discimination?
  - Préférences?



Question 9 (2pts)

Selon vous, quels sont les facteurs qui déterminent la participation au marché du travail des femmes? Expliquez pourquoi.

- **Facteurs institutionnels**:
  - Existence, durée, et montant des congés parentaux, modes de garde
  - Quotas
  - Spécialisations dans le diplômes
- **Facteurs économiques**:
  - Besoins financiers (familles monoparentales ou précaires)
- **Facteurs culturels**:
  - Statut matrimonial (**SVP** attention à la formulation de cet argument dans vos DMs!!!!!!)
  - Stéréotypes de genre
- **Facteurs contextuels/environnementaux**:
  - Nombre d'enfants: *child penalty*
- **Discrimination**: les femmes sont moins embauchées parce qu'on anticipe qu'elles devront quitter temporairement le

Partie C: Rémunération

Question 10 (1pt)

Créez la variable *quotité* qui vaut (*hint: utilisez les variables* `TPPRED` *et* `TXTPPRED`):

- 1 si l'individu travaille à temps complet
- 2 si l'individu travaille plus de 80%
- 3 si l'individu travaille à 80%
- 4 si l'individu travaille à temps partiel entre 50 et 80%
- 5 si l'individu travaille à mi-temps (50%)
- 6 si l'individu travaille moins d'un mi-temps

```
data$quotité = case_when(data$TPPRED == 1 ~ 1,
                         data$TXTPPRED == 5 ~ 2,
                         data$TXTPPRED == 4 ~ 3,
                         data$TXTPPRED == 3 ~ 4,
                         data$TXTPPRED == 2 ~ 5,
                         data$TXTPPRED == 1 ~ 6)
```

Question 11 (1pt)

Étudiez la distribution de la quotité de temps de travail (`quotité`) des individus en fonction de la présence d'au moins un enfant dans le ménage (`ENFRED`), **sur l'échantillon global**, puis séparément sur celui des hommes et des femmes. Commentez.

```
#ENFRED: individu avec au moins 1 enfant dans le ménage:; 1 = Oui, 2 = Non
round(prop.table(table(data$quotité, data$ENFRED), margin = 2), 2)
#       1    2
#  1 0.80 0.81
#  2 0.02 0.02
#  3 0.05 0.03
#  4 0.06 0.06
#  5 0.03 0.03
#  6 0.03 0.06
```

Question 11 (1pt)

Étudiez la distribution de la quotité de temps de travail (`quotité`) des individus en fonction de la présence d'au moins un enfant dans le ménage (`ENFRED`), sur l'échantillon global, puis **séparément sur celui des hommes et des femmes**. Commentez.

```
#ENFRED: individu avec au moins 1 enfant dans le ménage:; 1 = Oui, 2 = Non
round(prop.table(table(data$quotité[data$SEXE == 1], data$ENFRED[data$SEXE == 1]), margin = 2), 2)
#       1    2
#  1 0.94 0.88
#  2 0.01 0.01
#  3 0.01 0.01
#  4 0.02 0.04
#  5 0.01 0.02
#  6 0.01 0.04

round(prop.table(table(data$quotité[data$SEXE == 2], data$ENFRED[data$SEXE == 2]), margin = 2), 2)
#       1    2
#  1 0.67 0.72
#  2 0.04 0.02
#  3 0.10 0.04
#  4 0.09 0.08
#  5 0.05 0.04
#  6 0.06 0.09
```

Question 11 (1pt)

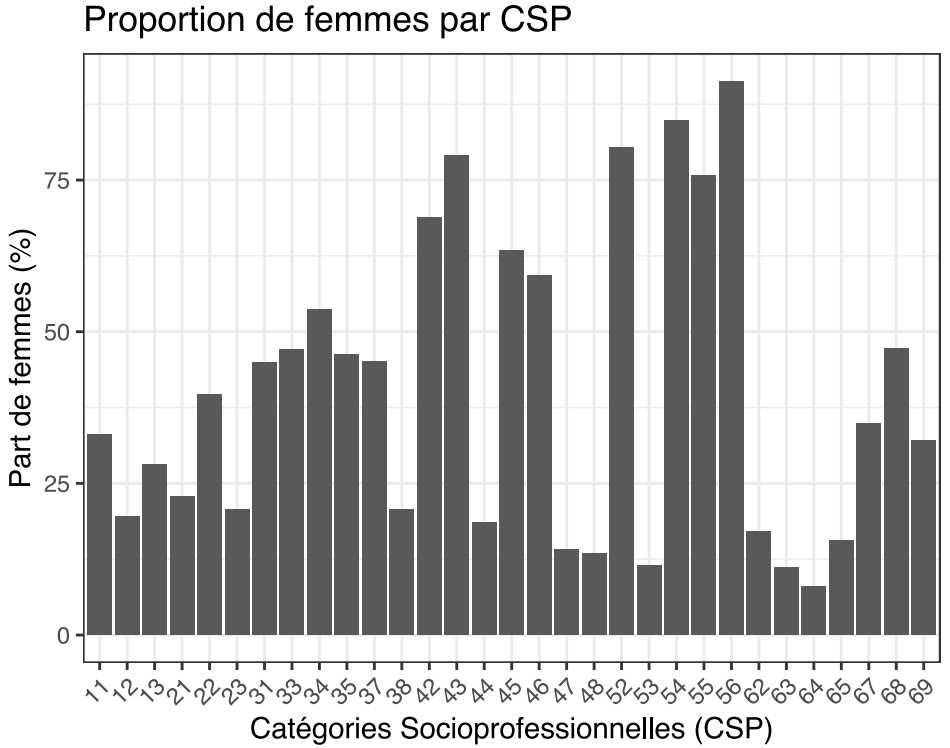
Étudiez la distribution de la quotité de temps de travail (`quotité`) des individus en fonction de la présence d'au moins un enfant dans le ménage (`ENFRED`), sur l'échantillon global, puis séparément sur celui des hommes et des femmes. **Commentez**.

- Sur l'échantillon global, participation au marché du travail à temps plein ne semble pas être affectée par la présence d'un enfant ou non dans le ménage
- Lorsqu'on regarde en détail comment cela affecte la quotité de temps de travail selon le genre:
  - Les hommes ayant au moins 1 enfant dans le ménage sont plus fréquemment à temps plein et à une quotité de temps partiel élevée que ceux sans enfant
  - Au contraite, les femmes ayant au moins un enfant dans le ménage sont plus fréquemment à temps partiel

Question 12 (1pt)

Étudiez la répartition par genre dans chacune des CSP (variable `CSP`) en ne conservant que les individus âgés d'au moins 30 ans et dont la CSP est bien définie et non nulle. Commentez.

```
data %>%
  filter(AGE5 =/= 15, !(CSP %in% c("00", NA))) %>%
  group_by(CSP) %>%
  summarise(n = n(),
            share_women = sum(SEXE == 2)/n) %>%
  ungroup() %>%
  arrange(share_women) %>%
  ggplot(aes(x = CSP, y = share_women* 100)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  labs(
    title = "Proportion de femmes par CSP",
    x = "Catégories Socioprofessionnelles (CSP)",
    y = "Part de femmes (%)"
  ) +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1
```



- CSP très masculines et d'autres très féminines

Question 13 (1pt)

Créez un dataframe qui comprend, pour chaque CSP, la moyenne du log du salaire mensuel net winsorisé et la proportion de femmes en vous restreignant aux individus de 30 ans ou plus et aux CSP connues et non nulles. Qu'observez-vous?

```
data_cor = data %>%
  filter(AGE5 =/= 15, !(CSP %in% c("00", NA))) %>%
  group_by(CSP) %>%
  summarise(mean_wage = mean(wage_winsor, na.rm = T),
            share_women = sum(SEXE == 2)/n()) %>%
  ungroup() %>%
  filter(!is.na(mean_wage))

head(data_cor)
```

```
## # A tibble: 6 × 3
##   CSP   mean_wage share_women
##   <fct>     <dbl>       <dbl>
## 1 21        2333.       0.229
## 2 22        2077.       0.397
## 3 23        4293.       0.207
## 4 31        2896.       0.449
## 5 33        3803.       0.471
## 6 34        3854.       0.536
```

Question 13 (1pt)

Créez un dataframe qui comprend, pour chaque CSP, la moyenne du log du salaire mensuel net winsorisé et la proportion de femmes en vous restreignant aux individus de 30 ans ou plus et aux CSP connues et non nulles. **Qu'observez-vous?**

```
data_cor = data %>%
  filter(AGE5 =/= 15, !(CSP %in% c("00", NA))) %>%
  group_by(CSP) %>%
  summarise(mean_wage = mean(wage_winsor, na.rm = T),
            share_women = sum(SEXE == 2)/n()) %>%
  ungroup() %>%
  filter(!is.na(mean_wage))

head(data_cor)
```

```
## # A tibble: 6 × 3
##   CSP   mean_wage share_women
##   <fct>     <dbl>       <dbl>
## 1 21        2333.       0.229
## 2 22        2077.       0.397
## 3 23        4293.       0.207
## 4 31        2896.       0.449
## 5 33        3803.       0.471
## 6 34        3854.       0.536
```

- Femme sous-représentées dans les CSP où le salaire moyen est le plus élevé
  - Ingénieur (gender-gap dans les STEM)
  - Chef d'entreprise
- et sur-représentées dans les CSP où le salaire moyen est le plus faible
  - métiers du *care*

Partie 3: Analyse Économétrique

Cette analyse économétrique explore les déterminants du salaire horaire et évalue l'effet de leur prise en compte sur l'écart salarial moyen entre femmes et hommes.

Question 14 (0.5pt)

- **Créez la variable** `femme` qui vaut 1 si l'individu est une femme, 0 sinon. Assurez-vous que la variable soit de type `factor`
- Créez la variable `CSP_r` qui est égale au premier chiffre de la CSP (chiffre des dizaines) si la CSP est différente de 23 (Chefs d'entreprises de 10 salariés ou plus); si la CSP est égale à 23, alors attribuer à `CSP_r` la valeur 3 (i.e. on considère que leur situation se rapproche davantage de celle des Cadres et Professions intermédiaires que de celle des Artisans et Commerçants).

```
data = data %>%
  mutate(femme = factor(ifelse(data$SEXE == 2, 1, 0)),
         CSP_r = ifelse(data$CSP == 23, 3, substr(data$CSP, 1, 1)))
```

Question 14 (0.5pt)

- Créez la variable `femme` qui vaut 1 si l'individu est une femme, 0 sinon. Assurez-vous que la variable soit de type `factor`
- **Créez la variable `CSP_r`** qui est égale au premier chiffre de la CSP (chiffre des dizaines) si la CSP est différente de 23 (Chefs d'entreprises de 10 salariés ou plus); si la CSP est égale à 23, alors attribuer à `CSP_r` la valeur 3 (i.e. on considère que leur situation se rapproche davantage de celle des Cadres et Professions intermédiaires que de celle des Artisans et Commerçants).

```
data = data %>%
  mutate(femme = factor(ifelse(data$SEXE == 2, 1, 0)),
         CSP_r = ifelse(data$CSP == 23, 3, substr(data$CSP, 1, 1)))
```

Question 15 (0.5pt)

Créez le dataframe `data_reg` à partir de `data` qui comprend uniquement les observations:

- des individus âgés de 30 ans ou plus
- qui sont en CDD ou CDI
- qui appartiennent à une `CSP_r` différente de 0
- qui appartiennent à une `CSP` comprend au moins 30 individus
- des individus qui travaillent au moins une heure par semaine

```
data_reg = data %>%
  filter(AGE5 =≠ 15,
         STATUTR %in% c(4,5),
         ! (CSP_r %in% c(0,1)),
         HHCE > 0) %>%
  group_by(CSP) %>%
  mutate(n = n()) %>%
  ungroup() %>%
  filter(n>30)
```

```
data_reg = data %>%
  filter(AGE5 =≠ 15,
         CONTRA %in% c(1,2),
         ! (CSP_r %in% c(0,1)),
         HHCE > 0) %>%
  group_by(CSP) %>%
  mutate(n = n()) %>%
  ungroup() %>%
  filter(n>30)
```

Question 16 (1pt)

Calculez l'écart du salaire moyen entre hommes et femmes (en niveau et en pourcentage, en utilisant le salaire mensuel net winsorisé).

```
(mean(data_reg$wage_winsor[data_reg$femme == 1], na.rm = T) - mean(data_reg$wage_winsor[data_reg$femme == 0], na.rm =
```

```
## [1] -581.7032
```

```
(mean(data_reg$wage_winsor[data_reg$femme == 1], na.rm = T)-mean(data_reg$wage_winsor[data_reg$femme == 0], na.rm = T
```

```
## [1] -0.2311041
```

- Les femmes gagnent en moyenne 581,7€ euros de moins que les hommes.
- Le salaire moyen des femmes est inférieur de 23,1% à celui des hommes.

Question 17 (2pt)

Estimez le modèle $log(\text{Salaire mensuel Winsorisé}) = \alpha + \beta \text{Femme} + \varepsilon$. Que représentent α et β? Peut-on dire que β est causal?

```
reg = lm(log_wage_winsor ~ femme, data = data_reg)
#summary(reg)
coeftest(reg) # utiliser cette commande pour avoir des std errors robustes à l'hétéroscédasticité
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error  t value  Pr(>|t|)
## (Intercept)  7.6970322  0.0043897 1753.413 < 2.2e-16 ***
## femme1      -0.2910540  0.0062270  -46.741 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question 17 (2pt)

Estimez le modèle $log(\text{Salaire mensuel Winsorisé}) = \alpha + \beta \text{Femme} + \varepsilon$. **Que représentent α et β?** Peut-on dire que β est causal?

Ici, c'est un modèle **Log-level**, i.e. l'outcome est exprimé en log et la variable explicative en niveau.

- α: moyene du log du salaire des hommes
- β:
  - **Interprétation rigoureuse**: en moyenne, être une femme est associé à une diminution de salaire de $(e^{\beta} - 1) * 100 = -25.25247$ %
  - **Approximation** possible 🚨 lorsque β est **petit** 🚨: en moyenne, être une femme est associé à une diminution de salaire de $\beta * 100 = -29.10540\%$

Question 17 (2pt)

Estimez le modèle $log(\text{Salaire mensuel Winsorisé}) = \alpha + \beta \text{Femme} + \varepsilon$. Que représentent α et β? **Peut-on dire que β est causal?**

Ici, c'est un modèle **Log-level**, i.e. l'outcome est exprimé en log et la variable explicative en niveau.

- α: moyene du log du salaire des hommes
- β:
  - **Interprétation rigoureuse**: en moyenne, être une femme est associé à une diminution de salaire de $(e^{\beta} - 1) * 100 = -25.25247$ %
  - **Approximation** possible 🚨 lorsque β est **petit** 🚨: en moyenne, être une femme est associé à une diminution de salaire de $\beta * 100 = -29.10540\%$

On ne peut pas dire que β est causal: comme on l'a vu au fil des précédentes questions, le diplôme, la CSP, le nombre d'enfants sont **corrélés avec le fait d'être une femme** et **influencent également le salaire**.

Question 18 (1.5pt)

- **Créez la variable salaire (winsorisé)** horaire **et son log**, que vous nommerez `hwage_winsor` et `log_hwage_winsor` (*hint: utiliser la variable `HHCE`, le nombre d'heures travaillées en moyenne **par semaine** dans l'emploi principal*)
- Quel est maintenant l'écart de salaire moyen entre femmes et hommes? Comment se compare-t-il à l'écart de salaire calculé en question 16 et comment expliquez-vous cette différence?

```
data_reg = data_reg %>%
  mutate(hwage_winsor = wage_winsor/(data_reg$HHCE*151.67/35),
         log_hwage_winsor = log(hwage_winsor))

(mean(data_reg$hwage_winsor[data_reg$femme == 1], na.rm = T) - mean(data_reg$hwage_winsor[data_reg$femme == 0], na.rm
```

```
## [1] -0.1309692
```

Question 18 (1.5pt)

- Créez la variable salaire (winsorisé) **horaire** et son log, que vous nommerez `hwage_winsor` et `log_hwage_winsor` (*hint: utiliser la variable `HHCE`, le nombre d'heures travaillées en moyenne **par semaine** dans l'emploi principal*)
- **Quel est maintenant l'écart de salaire moyen entre femmes et hommes**? **Comment se compare-t-il à l'écart de salaire calculé en question 16 et comment expliquez-vous cette différence?**

```
data_reg = data_reg %>%
  mutate(hwage_winsor = wage_winsor/(data_reg$HHCE*151.67/35),
         log_hwage_winsor = log(hwage_winsor))

(mean(data_reg$hwage_winsor[data_reg$femme == 1], na.rm = T) - mean(data_reg$hwage_winsor[data_reg$femme == 0], na.rm
```

```
## [1] -0.1309692
```

- L'écart de salaire moyen entre hommes et femmes, une fois pris en compte la quotité de temps de travail, est de 13,09692%, soit presque la moitié du gap calculé à la question 16
- Suggère qu'une partie on négligeable du gap de salaire H/F est dû à des différences dans le temps travaillé (= *effet de composition*)

Question 19 (2pt)

Proposez un modèle de régression linéaire qui permet d'atténuer les problèmes d'identification soulevés à la question 17. Commentez les résultats. En particulier, comment varie β? Qu'est-ce que cela indique de l'écart de salaire moyen entre hommes et femmes calculé en question 17?

Par exemple, $log(\text{Salaire horraire Winsorisé}) = \alpha + \beta \text{Femme} + \gamma\ 1(\geq Bac) + \delta \text{CSP} + \varepsilon$

Quelque soit le modèle que vous avez estimé, l'inclusion de ces variables de contrôle entraîne une diminution de β (en valeur absolue).

Question 20 (2pt)

Selon vous, peut-on dire que β estimé à la question 19 est causal? Expliquez.

Effectivement, à force d'inclure des variables de contrôle, on se rapproche d'un estimateur de β qui capture uniquement l'effet du genre sur le salaire.

MAIS, il existe toujours des **facteurs inobservés et inobservables**, tels que:

- la motivation
- la discrimination sur le marché du travail
- salaire de réserve
- confiance en soi/capacité de négociation (*gender ask gap*)

qui nous empêchent de dire que β est causal.

Partie 4: Décomposition d'Oaxaca-Blinder

Décomposition d'Oaxaca-Blinder

La méthode de décomposition d'Oaxaca-Blinder permet de mesurer la part de l'écart entre le salaire moyen des femmes et celui des hommes qui est due à des différences dans les caractéristiques moyennes et celle due à des différences d'effets de ces caractéristiques sur le salaire moyen (dit autrement à des différences dans les coefficients de régression). On considère deux groupes, celui des femmes *F* et celui des hommes *H*. L'écart de salaire moyen entre les deux groupes s'écrit:

$$\Delta\overline{Y} = \overline{Y}_H - \overline{Y}_F$$

où $\overline{Y}_i, i \in \{H, F\}$ est le salaire horaire moyen du groupe i.

Décomposition du *gender wage gap* en deux parties

Cette décomposition permet d'écrire l'écart moyen de salaire H/F comme la somme d'une partie expliquée par des différences de caractéristiques et d'une partie inexpliquée.

$$\Delta\overline{Y} = \underbrace{\left(\overline{X}'_H - \overline{X}'_F\right)'\hat{\beta}_H}_{\text{Expliquée}} + \underbrace{\overline{X}'_F\left(\hat{\beta}_F - \hat{\beta}_H\right)}_{\text{Inexpliquée}}$$

Question 21 (1pt)

Décrivez ce que mesure la seconde partie de l'équation (2)? Pourquoi parle t-on de composante inexpliquée de l'écart de salaire hommes/femmes?

La seconde partie de l'équation mesure, à caractéristiques moyennes des femmes données, l'écart de rendement de ces caractéristiques entre femmes et hommes. Dit autrement, cet indicateur mesure la part de l'écart du salaire moyen entre hommes et femmes qui est dû à des différences dans les rendements des caractéristiques et non à des différences de caractéristiques.

On parle de composante non expliquée car elle ne relève pas d'un *effet de composition* mais s'apparente plutôt à de la discrimination.

Question 22 (2pts)

Installez et chargez le package `oaxaca`. Utilisez la fonction `oaxaca` pour réaliser une décomposition d'Oaxaca-Blinder en deux parties. Que peut-on conclure sur les écarts de rémunération salariale entre femmes et hommes? Vous pourrez vous référer à la documentation du package disponible en ligne et vous aider d'un graphique pour interpréter les résultats.

```
library(oaxaca)
oaxaca = oaxaca(log_hwage_winsor ~ as.factor(CSP_r) + as.factor(at_least_bac)  + as.factor(ANCENTR4)  | femme , data

# oaxaca$y # y.diff= 0.1768753

oaxaca$twofold$overall[,c(1,2,4)] # coefficients de référence estimés chez les hommes
```

```
##      group.weight coef(explained) coef(unexplained)
## [1,]    0.0000000      0.03842147        0.10548794
## [2,]    1.0000000      0.04887939        0.09503002
## [3,]    0.5000000      0.04365043        0.10025898
## [4,]    0.5030341      0.04368217        0.10022725
## [5,]   -1.0000000      0.06429823        0.07961118
## [6,]   -2.0000000      0.04782188        0.09608753
```

```
# 0.077419846: différence de salaire moyen de h et femmes si les caractéristiques des femmes étaient valorisées comme
# 0.099455490: différence entre le salaire que toucheraient les femmes si leurs caractéristiques étaient valorisées c
```

Question 22 (2pts)

Explained: contribution de chaque variable à l'écart expliqué

Unexplained: contribution de chaque variable à l'écart inexpliqué

Signe du coefficient:

- les coefficients positifs sont ceux qui contribuent à augmenter l'écart les coefficients négatifs sont ceux qui contribuent à diminuer l'écart

Magnitude du coefficient: le rapporter à l'écart de salaire moyen expliqué et inexpliqué calculé plus haut

```
plot(oaxaca, decomposition = "twofold", group.weight = -
```