Analyse des Variations de l'Emploi en Europe (1990-2023)

```
Loading required package: stringr
```

M1 ISIFAR

Université Paris Cité

Academic Year 2024-2025

Project Members : Hengze WANG ; Ruoqi ZHU

Github Page

```
        to_be_loaded <- c("skimr",
                          "ggplot2",
                          "dplyr",
                          "glue",
                          "DT",
                          "knitr",
                          "unilur",
                          "restatapi",
                          "stats")

        for (pck in to_be_loaded) {
          if (!require(pck, character.only = TRUE)) {
            install.packages(pck)
            stopifnot(require(pck, character.only = T))
          }
        }
```

```
Loading required package: skimr

Loading required package: ggplot2

Loading required package: dplyr


Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```
Loading required package: glue

Loading required package: DT

Loading required package: knitr

Loading required package: unilur

Loading required package: restatapi

restatapi: - version 0.22.1
           - config file with the API version 2 loaded from GitHub (the 'current'
API version number is 2).
           - 2 from the 8 cores are used for parallel computing, can be changed
with 'options(restatapi_cores=...)'
           - 'auto' method will be used for file download, can be changed with
'options(restatapi_dmethod=...)'
           - the Table of contents (TOC) was not pre-loaded into the deafult cache
('.restatapi_env').
```

0. Importation des données d'Eurostat

Data code	liens et nom du data
lfsi_emp_a_h	Employment and activity by sex and age (1992-2020) - annual data
lfsi_educ_a_h	Employment by educational attainment level (1998-2020) - annual data
earn_ses18_30	Mean annual earnings by sex, economic activity and educational attainment

Définition de la structure des données :

DSD of Employment and activity by sex and age (1992-2020) - annual data :

```
datatable(get_eurostat_dsd("lfsi_emp_a_h"))
```

Show 10 entries Search:

	concept	code	name
1	freq	A	Annual
2	age	Y15-24	From 15 to 24 years
3	age	Y15-64	From 15 to 64 years
4	age	Y20-64	From 20 to 64 years

5	age	Y25-54	From 25 to 54 years
6	age	Y55-64	From 55 to 64 years
7	unit	THS_PER	Thousand persons
8	unit	PC_POP	Percentage of total population
9	sex	T	Total
10	sex	M	Males

Showing 1 to 10 of 56 entries Previous 1 2 3 4 5 6 Next

DSD of Employment by educational attainment level (1998-2020) - annual data :

```
datatable(get_eurostat_dsd("lfsi_educ_a_h"))
```

Show 10 entries Search:

	concept	code	name
1	freq	A	Annual
2	age	Y15-24	From 15 to 24 years
3	age	Y15-64	From 15 to 64 years
4	age	Y20-64	From 20 to 64 years
5	age	Y25-54	From 25 to 54 years
6	age	Y55-64	From 55 to 64 years
7	unit	THS_PER	Thousand persons
8	unit	PC_POP	Percentage of total population
9	unit	PC_EMP	Percentage of total employment
10	sex	T	Total

Showing 1 to 10 of 56 entries Previous 1 2 3 4 5 6 Next

DSD of Mean annual earnings by sex, economic activity and educational attainment :

```
datatable(get_eurostat_dsd("earn_ses18_30"))
```

Show 10 entries Search:

	concept	code	name
1	freq	A	Annual
2	indic_se	ERN	Gross earnings
3	indic_se	BNS	Annual bonuses
4	isced11	TOTAL	All ISCED 2011 levels
5	isced11	ED0-2	Less than primary, primary and lower secondary education (levels 0-2)
6	isced11	ED3_4	Upper secondary and post-secondary non-tertiary education (levels 3 and 4)
7	isced11	ED5-8	Tertiary education (levels 5-8)
8	nace_r2	B-S	Industry, construction and services (except activities of households as employers and extra-territorial organisations and bodies)
9	nace_r2	B-S_X_O	Industry, construction and services (except public administration, defense, compulsory social security)
10	nace_r2	B-N	Business economy

Showing 1 to 10 of 84 entries Previous 1 2 3 4 5 ... 9 Next

I. Exploration des données

```
# Importer des données d'Eurostat
emp_sex_age <- get_eurostat_data("lfsi_emp_a_h")
emp_educ <- get_eurostat_data("lfsi_educ_a_h")
earn <- get_eurostat_data("earn_ses18_30")
# Visualiser la structure des données
glimpse(emp_sex_age)
```

```
Rows: 56,928
Columns: 7
$ age      <fct> Y15-24, Y15-24, Y15-24, Y15-24, Y15-24, Y15-24, Y15-24, Y15-2…
$ unit     <fct> PC_POP, PC_POP, PC_POP, PC_POP, PC_POP, PC_POP, PC_POP, PC_PO…
$ sex      <fct> F, F, F, F, F, F, F, F, F, F, F, F, F, F, F, F, F, F, F, F, F…
$ indic_em <fct> ACT, ACT, ACT, ACT, ACT, ACT, ACT, ACT, ACT, ACT, ACT, ACT, A…
$ geo      <fct> AT, AT, AT, AT, AT, AT, AT, AT, AT, AT, AT, AT, AT, AT, AT, A…
$ time     <fct> 1995, 1996, 1997, 1998, 1999, 2000, 2001, 2002, 2003, 2004, 2…
$ values   <dbl> 57.6, 55.8, 55.1, 54.7, 54.7, 50.5, 49.7, 50.3, 49.8, 53.0, 5…
```

```
glimpse(emp_educ)
```

```
Rows: 101,703
Columns: 7
$ age     <fct> Y15-24, Y15-24, Y15-24, Y15-24, Y15-24, Y15-24, Y15-24, Y15-24…
$ unit    <fct> PC_EMP, PC_EMP, PC_EMP, PC_EMP, PC_EMP, PC_EMP, PC_EMP, PC_EMP…
$ sex     <fct> F, F, F, F, F, F, F, F, F, F, F, F, F, F, F, F, F, F, F, F,…
$ isced11 <fct> ED0-2, ED0-2, ED0-2, ED0-2, ED0-2, ED0-2, ED0-2, ED0-2, ED0-2,…
$ geo     <fct> AT, AT, AT, AT, AT, AT, AT, AT, AT, AT, AT, AT, AT, AT, AT, AT…
$ time    <fct> 1999, 2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 20…
$ values  <dbl> 30.5, 29.8, 29.8, 29.9, 30.1, 22.6, 24.0, 28.0, 31.6, 31.9, 29…
```

```
glimpse(earn)
```

```
Rows: 111,072
Columns: 9
$ indic_se <fct> BNS, BNS, BNS, BNS, BNS, BNS, BNS, BNS, BNS, BNS, BNS, BNS, B…
$ isced11  <fct> ED0-2, ED0-2, ED0-2, ED0-2, ED0-2, ED0-2, ED0-2, ED0-2, ED0-2…
$ nace_r2  <fct> B, B, B, B, B, B, B, B, B, B, B, B, B, B, B, B, B, B, B, B, B…
$ sex      <fct> F, F, F, F, F, F, F, F, F, F, F, F, F, F, F, F, F, F, F, F, F…
$ sizeclas <fct> GE10, GE10, GE10, GE10, GE10, GE10, GE10, GE10, GE10, GE10, G…
$ unit     <fct> EUR, EUR, EUR, EUR, EUR, EUR, EUR, EUR, EUR, EUR, EUR, EUR, E…
$ geo      <fct> AT, CH, CZ, DE, DK, EA17, EA18, EA19, EE, ES, EU27_2007, EU27…
$ time     <fct> 2018, 2018, 2018, 2018, 2018, 2018, 2018, 2018, 2018, 2018, 2…
$ values   <dbl> 6273, 2232, 867, 1846, 1500, 3179, 2979, 2963, 214, 2734, 300…
```

Nous avons sélectionné la dataframe “Employment and activity by sex and age (1992-2020) - annual data” à partir d’Eurostat, identifiée par le code “lfsi_emp_a_h”, que nous utilisons dans notre code. Un examen rapide avec la fonction `glimpse` révèle la présence de 7 variables, dont 6 sont qualitatives : age, unit, sex, indic_em, geo, time, et 1 variable quantitative : values.

De même, nous avons opté pour la dataframe “Employment by educational attainment level (1998-2020) - annual data” provenant d’Eurostat, identifiée par le code “lfsi_educ_a_h”, que nous utilisons également dans notre code. Un aperçu avec la fonction `glimpse` indique la présence de 7 variables, dont 6 sont qualitatives : age, unit, sex, isced11, geo, time, et 1 variable quantitative : values.

Enfin, nous avons retenu la dataframe “Mean annual earnings by sex, economic activity and educational attainment” d’Eurostat, identifiée par le code “earn_ses18_30”, que nous utilisons dans notre code. Un examen rapide avec la fonction `glimpse` montre la présence de 9 variables, dont 8 sont qualitatives : indic_se, isced11, nace_r2, sex, sizeclas, unit, geo, time, et 1 variable quantitative : values.

1. Filtration des données et visualiser la structure et les statistiques sommaires des données qu’on a besion et vérifier les valeurs manquantes

```r
emp_sex_age_filtered <- get_eurostat_data("lfsi_emp_a_h",
                                         filters = c("PC_POP","T","Y15-64","EMP_LFS
# On filtre les données qui correspondent :
# unit = PC_POP = Percentage of total population,
# sex = T = Total, age = Y15-64 = From 15 to 64 years,
# indic_em = EMP_LFS = Total employment (resident population concept -LFS);

emp_sex_age_filtered_EU27_2020 <- get_eurostat_data("lfsi_emp_a_h",
                                         filters = c("PC_POP","T","Y15-64","EU27_20
# On filtre les données qui correspondent :
# unit = PC_POP = Percentage of total population,
# sex = T = Total, age = Y15-64 = From 15 to 64 years,
# geo = EU27_2020 = European Union-27 countries (from 2020),
# indic_em = EMP_LFS = Total employment (resident population concept -LFS);


# Visualiser la structure des données
glimpse(emp_sex_age_filtered)
```

```
Rows: 943
Columns: 7
$ geo      <fct> AT, AT, AT, AT, AT, AT, AT, AT, AT, AT, AT, AT, AT, AT, AT, A…
$ unit     <fct> PC_POP, PC_POP, PC_POP, PC_POP, PC_POP, PC_POP, PC_POP, PC_PO…
$ sex      <fct> T, T, T, T, T, T, T, T, T, T, T, T, T, T, T, T, T, T, T, T, T…
$ indic_em <fct> EMP_LFS, EMP_LFS, EMP_LFS, EMP_LFS, EMP_LFS, EMP_LFS, EMP_LFS…
$ age      <fct> Y15-64, Y15-64, Y15-64, Y15-64, Y15-64, Y15-64, Y15-64, Y15-6…
$ time     <fct> 1995, 1996, 1997, 1998, 1999, 2000, 2001, 2002, 2003, 2004, 2…
$ values   <dbl> 68.6, 67.9, 67.9, 68.2, 68.5, 68.5, 68.4, 68.7, 68.9, 66.5, 6…
```

```r
glimpse(emp_sex_age_filtered_EU27_2020)
```

```
Rows: 21
Columns: 7
$ geo      <fct> EU27_2020, EU27_2020, EU27_2020, EU27_2020, EU27_2020, EU27_2…
$ unit     <fct> PC_POP, PC_POP, PC_POP, PC_POP, PC_POP, PC_POP, PC_POP, PC_PO…
$ sex      <fct> T, T, T, T, T, T, T, T, T, T, T, T, T, T, T, T, T, T, T, T, T
$ indic_em <fct> EMP_LFS, EMP_LFS, EMP_LFS, EMP_LFS, EMP_LFS, EMP_LFS, EMP_LFS…
$ age      <fct> Y15-64, Y15-64, Y15-64, Y15-64, Y15-64, Y15-64, Y15-64, Y15-6…
$ time     <fct> 2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2…
$ values   <dbl> 61.0, 61.3, 61.2, 61.4, 61.7, 62.2, 63.2, 64.3, 64.8, 63.6, 6…
```

```r
# Statistiques de synthèse Vérification des valeurs manquantes
skimr :: skim(emp_sex_age_filtered_EU27_2020)
```

Name	emp_sex_age_filtered_EU27...
Number of rows	21
Number of columns	7

Key	NULL
_____	
Column type frequency:	
factor	6
numeric	1
_____	
Group variables	None

Data summary

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
geo	0	1	FALSE	1	EU2: 21
unit	0	1	FALSE	1	PC_: 21
sex	0	1	FALSE	1	T: 21
indic_em	0	1	FALSE	1	EMP: 21
age	0	1	FALSE	1	Y15: 21
time	0	1	FALSE	21	200: 1, 200: 1, 200: 1, 200: 1

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
values	0	1	63.92	2.25	61	62.2	63.4	64.8	68.5	▇▇▂▁▂

II. Visualisation des données (Études univariées et bivariées)

```
# La seule variable quantitative est values(taux d'emploi), on commence ave
emp_sex_age_filtered$values %>%
  skimr::skim() %>%
  select(1:11, -starts_with('skim')) %>%
  knitr::kable(caption = "Summary statistics pour taux d'emploi")
```

n_missing	complete_rate	numeric.mean	numeric.sd	numeric.p0	numeric.p25	numeric.p50	numer
0	1	64.35408	7.645043	39.6	59.75	64.3	

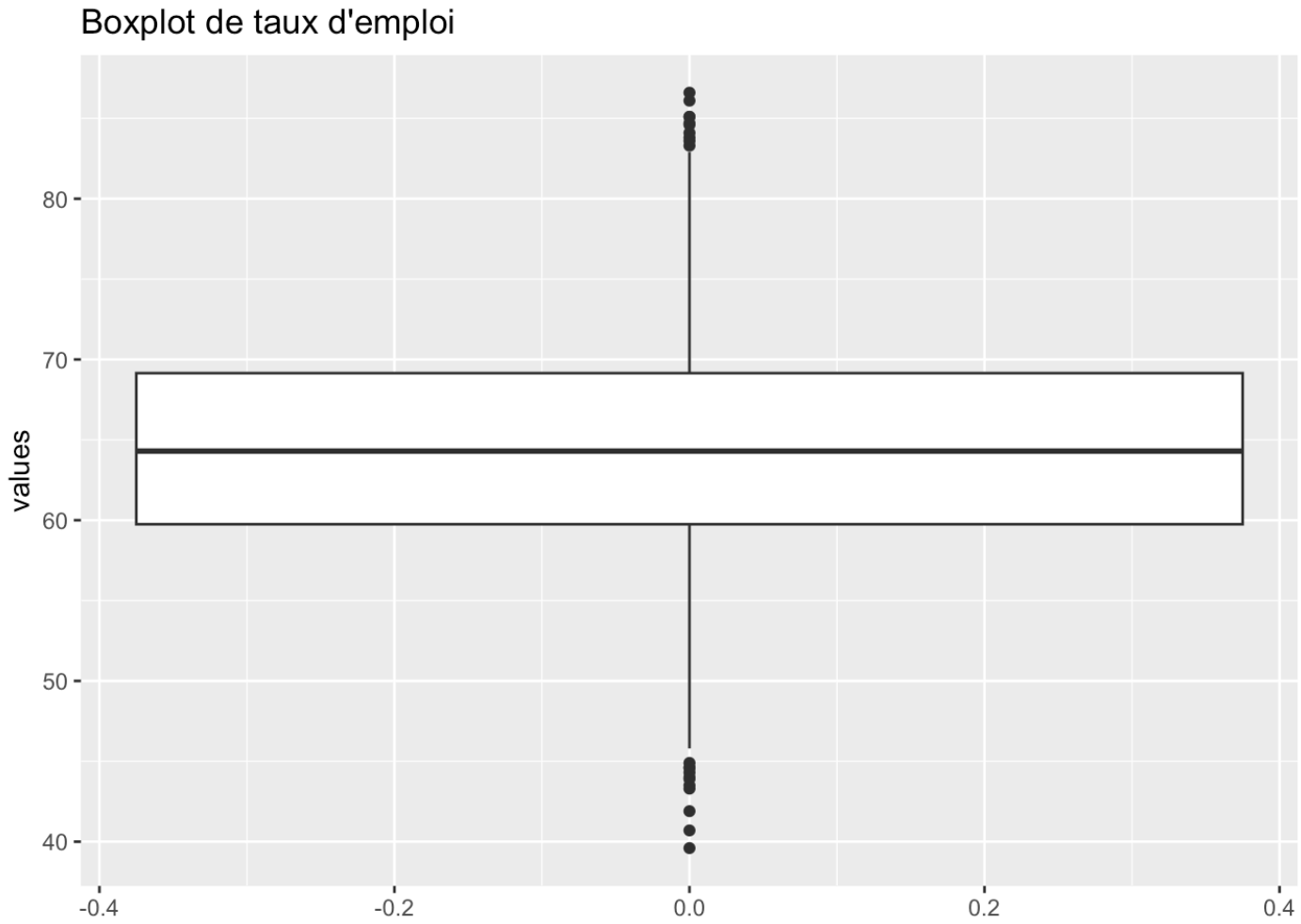
Summary statistics pour taux d'emploi

Nous constatons que le taux d'emploi moyen est 63.92%, l'écart-type est 2.25, le minimal taux

est 61, le médian est 63.4, le maximal est 68.5;

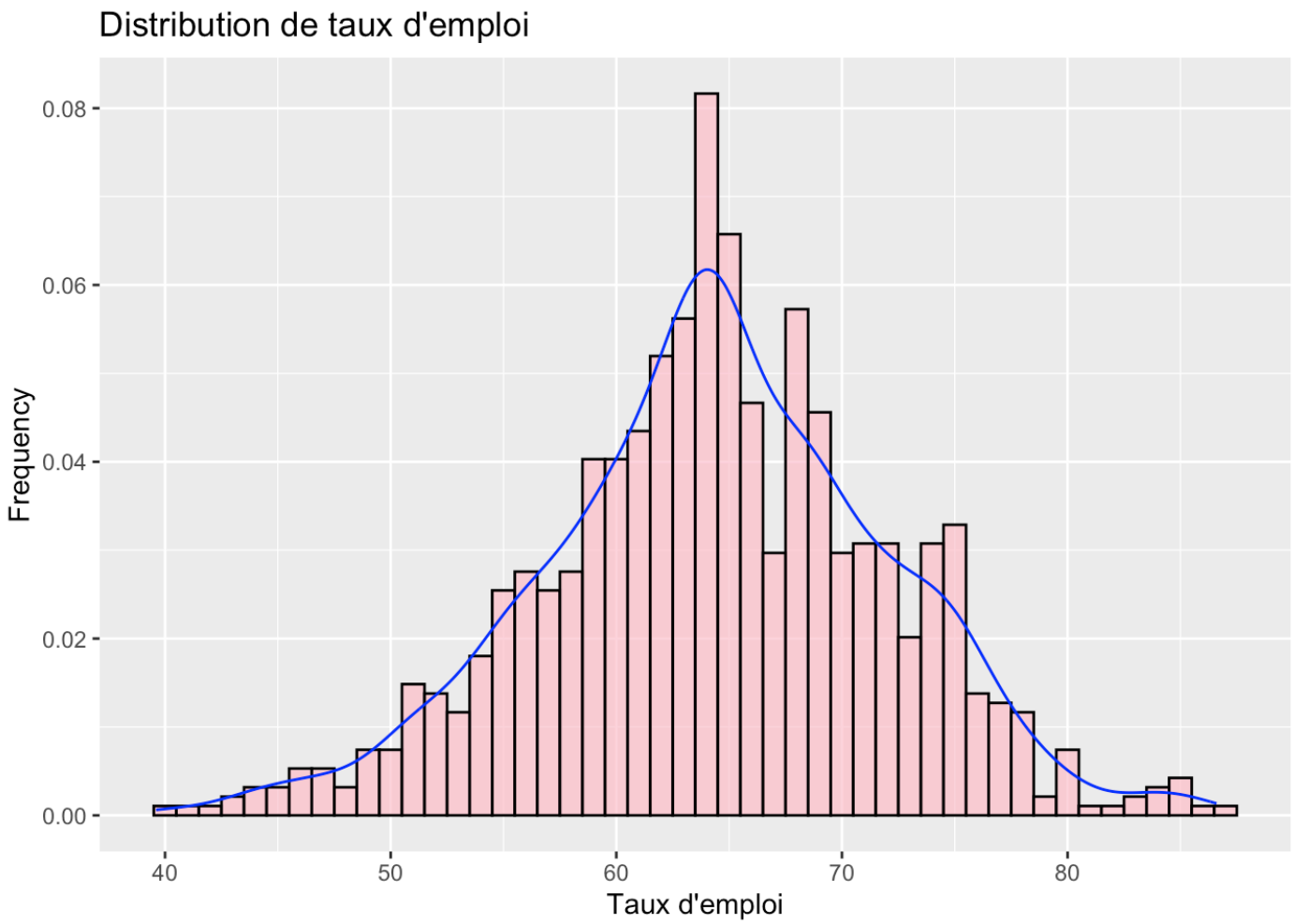
1. Visualisation de la distribution des taux d'emploi

```
# Boxplot
emp_sex_age_filtered %>%
  ggplot(aes(y = values)) +
  geom_boxplot() +
  labs(title = "Boxplot de taux d'emploi")
```



Par le boxplot, nous constatons que le minimum taux d'emploi est de 61%, le premier quantile est environ 62.2%, le médian est environ 63.4%, le troisième quantile est environ 64.8%, le maximum est environ 68.6%, ces valeurs là sont à peu près égales aux valeurs qu'on a obtenu précédent par skimr::skim;

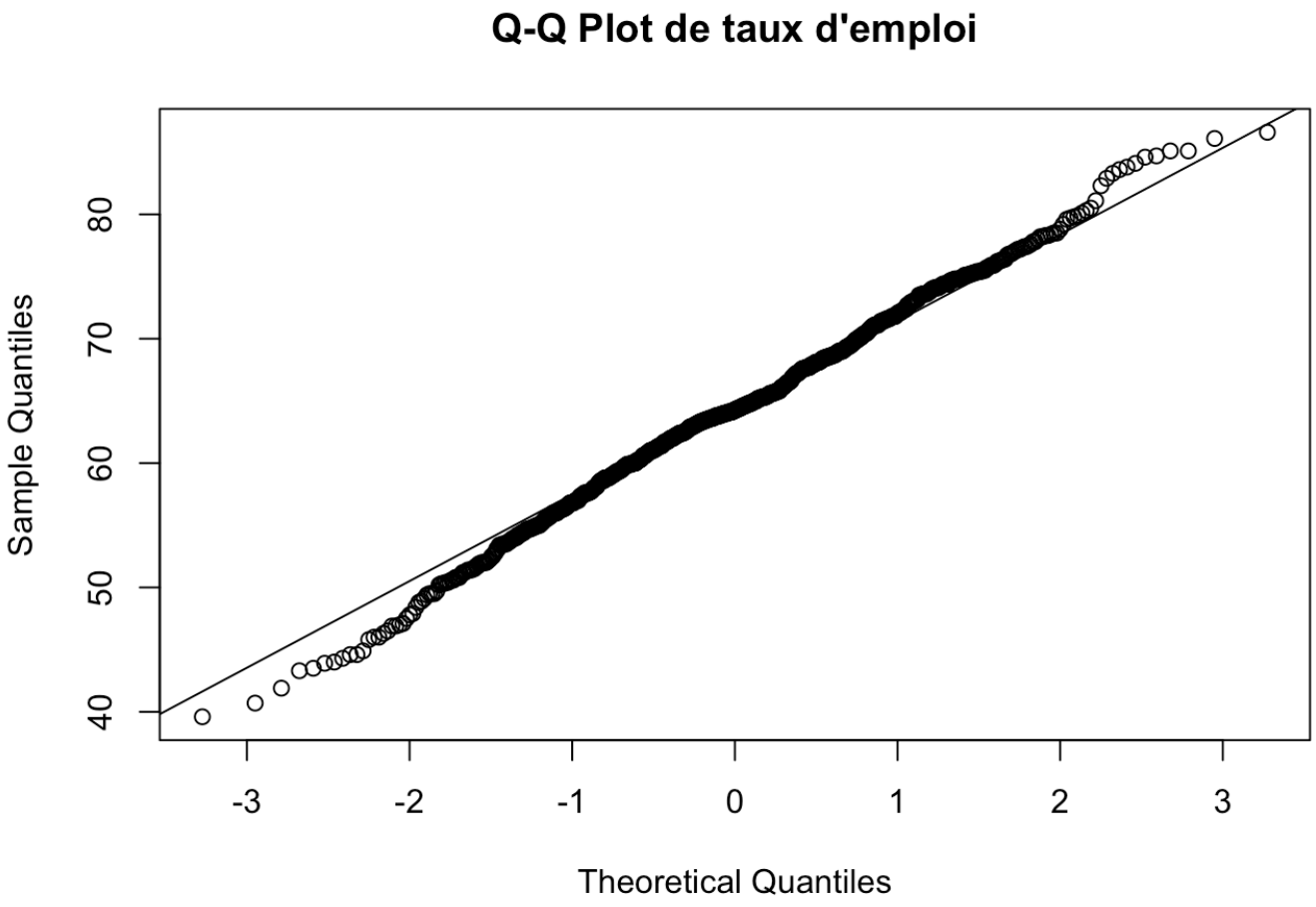
```
# Visualisation de la distribution des taux d'emploi
ggplot(emp_sex_age_filtered, aes(x = values)) +
  geom_histogram(aes(y = after_stat(density)), binwidth = 1, fill = "pink",
  labs(title = "Distribution de taux d'emploi",
       x = "Taux d'emploi",
       y = "Frequency")+
  geom_density(color = "blue")
```



En examinant la distribution du taux d'emploi global de 1992 à 2020, nous constatons que le taux d'emploi est principalement réparti entre 60% et 70%, et que le taux d'emploi se rapproche de la symétrie autour de 65 %.

Il semble que le taux d'emploi est normalement distribué, mais nous ne pouvons pas en être sûrs. Nous allons donc utiliser le graphique qq plot et calculer skewness et kurtosis pour déterminer si le taux d'emploi suit réellement une distribution normale.

```
# QQ plot
qqnorm(emp_sex_age_filtered$values, main = "Q-Q Plot de taux d'emploi")
qqline(emp_sex_age_filtered$values)
```




```
# calcul de skewness et de kurtosis
library(e1071)
skewness(emp_sex_age_filtered$values)
```

```
[1] -0.1186479
```

```
kurtosis(emp_sex_age_filtered$values)
```

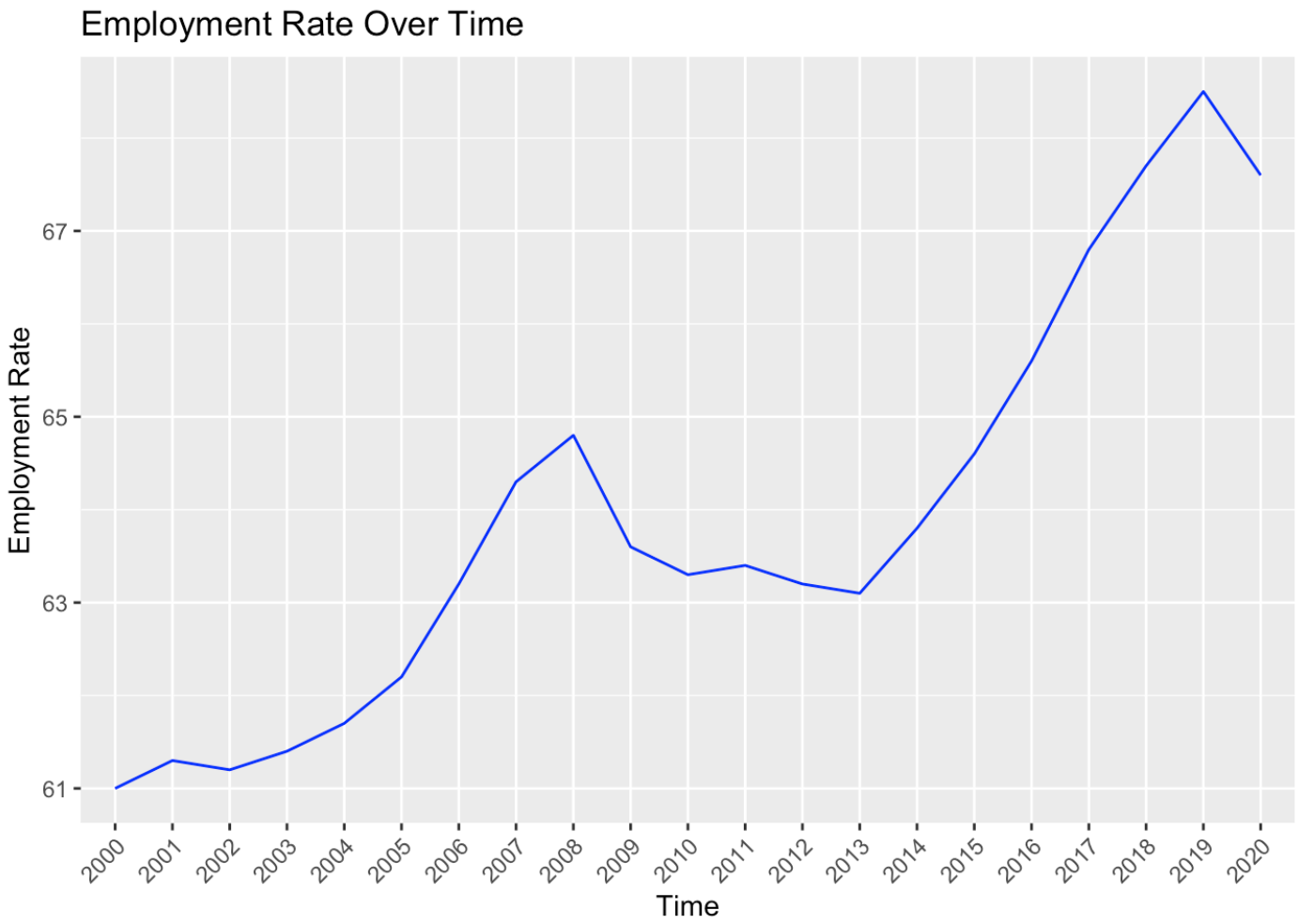
```
[1] 0.1695281
```

Dans le graphique qq plot, la grande majorité des points se trouvent le long de la ligne droite. De cela, nous pouvons conclure que le taux d'emploi suit une distribution normale, avec une skewness proche de zéro (-0.1182076) et une kurtosis proche de zéro (0.1653475) également. Cela vérifie une fois de plus que le taux d'emploi suit une distribution normale.

2. Visualisation de l'évolution des taux d'emploi dans le temps

```
ggplot(emp_sex_age_filtered_EU27_2020, aes(x = time, y = values, group = 1)
  geom_line(color = "blue") +
  labs(title = "Employment Rate Over Time",
       x = "Time",
```

```
       y = "Employment Rate")+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Nous pouvons constater que le taux d'emploi dans les 27 pays de l'Union européenne varie entre 61% et 69% au fil du temps. Entre 2000 et 2008, le taux d'emploi a augmenté de manière constante, tandis qu'entre 2008 et 2013, le taux d'emploi a diminué chaque année, probablement en raison de la crise financière de 2008. De 2013 à 2019, le taux d'emploi a progressivement augmenté, puis après 2019, en raison de la pandémie de la COVID-19, le taux d'emploi a de nouveau diminué

3. Visualisation de la comparaison des taux d'emploi par sexe

```
# Filtration des données et comparaison des taux d'emploi par sexe
emp_data_by_sex <- get_eurostat_data("lfsi_emp_a_h",
                                     filters = c("PC_POP","Y15-64","EU27_20
# Convert the time column to Date format with a fixed month and day
emp_data_by_sex$time <- as.Date(paste(emp_data_by_sex$time, "-01-01", sep =
ggplot(emp_data_by_sex, aes(x = time, y = values, color = sex)) +
  geom_line() +
  labs(title = "Employment Rate Comparison by Gender",
       x = "Time",
       y = "Employment Rate",
```

```
                color = "Gender")
```

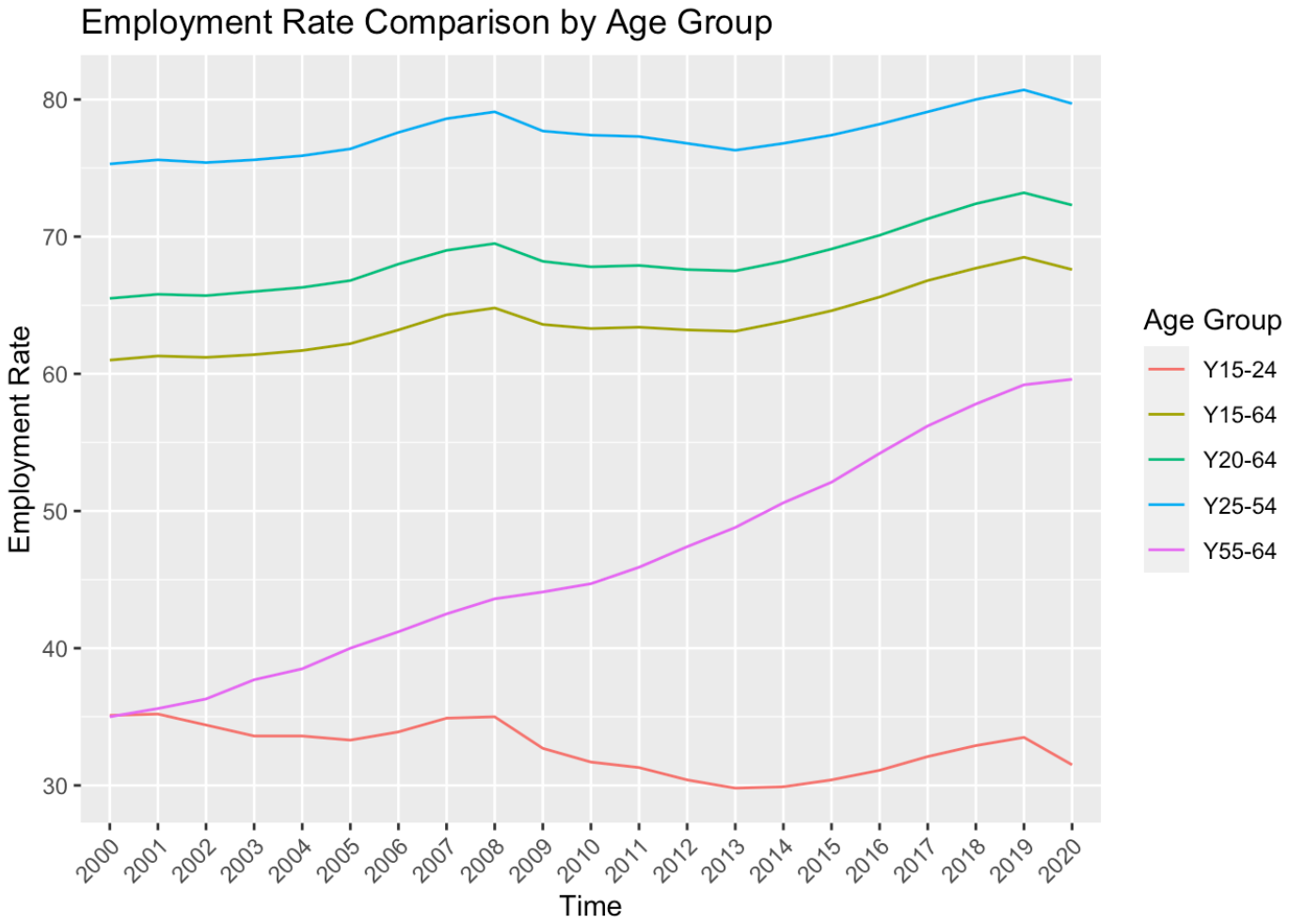
Lors de la comparaison des taux d'emploi entre les sexes de 2000 à 2020, nous constatons que le taux d'emploi des hommes a toujours été plus élevé que celui des femmes. Le taux d'emploi des hommes fluctue autour de 70%, tandis que celui des femmes, qui était d'environ 50% en 2000, a augmenté pour atteindre un peu plus de 60%. Après 2008, le taux d'emploi des hommes a été davantage impacté par la crise économique, subissant une baisse plus significative, tandis que le taux d'emploi des femmes semble avoir été moins affecté.

Nous pouvons observer que, bien que le taux d'emploi global des femmes soit inférieur à celui des hommes, il diminue progressivement, réduisant ainsi l'écart avec les hommes. Cette tendance suggère que, au fil du temps, les femmes ont vu une amélioration relative de leurs taux d'emploi par rapport aux hommes.

4. Visualisation de la comparaison des taux d'emploi par groupe d'âge

```
emp_data_by_age <- get_eurostat_data("lfsi_emp_a_h",
                                     filters = c("T","PC_POP","EU27_2020","EMP_LFS
emp_data_by_age <- emp_data_by_age %>% arrange(time)
ggplot(emp_data_by_age, aes(x = time, y = values, color = age, group = age)
```

```
        geom_line() +
        labs(title = "Employment Rate Comparison by Age Group",
             x = "Time",
             y = "Employment Rate",
             color = "Age Group")+
        theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Lors de la visualisation comparative des taux d'emploi dans différentes tranches d'âge de 15 à 24 ans, de 15 à 64 ans, de 20 à 64 ans, de 25 à 54 ans et de 55 à 64 ans de 2000 à 2020, nous avons observé deux groupes d'âge particulièrement intéressants : les jeunes de 15 à 24 ans et les personnes âgées de 55 à 64 ans.

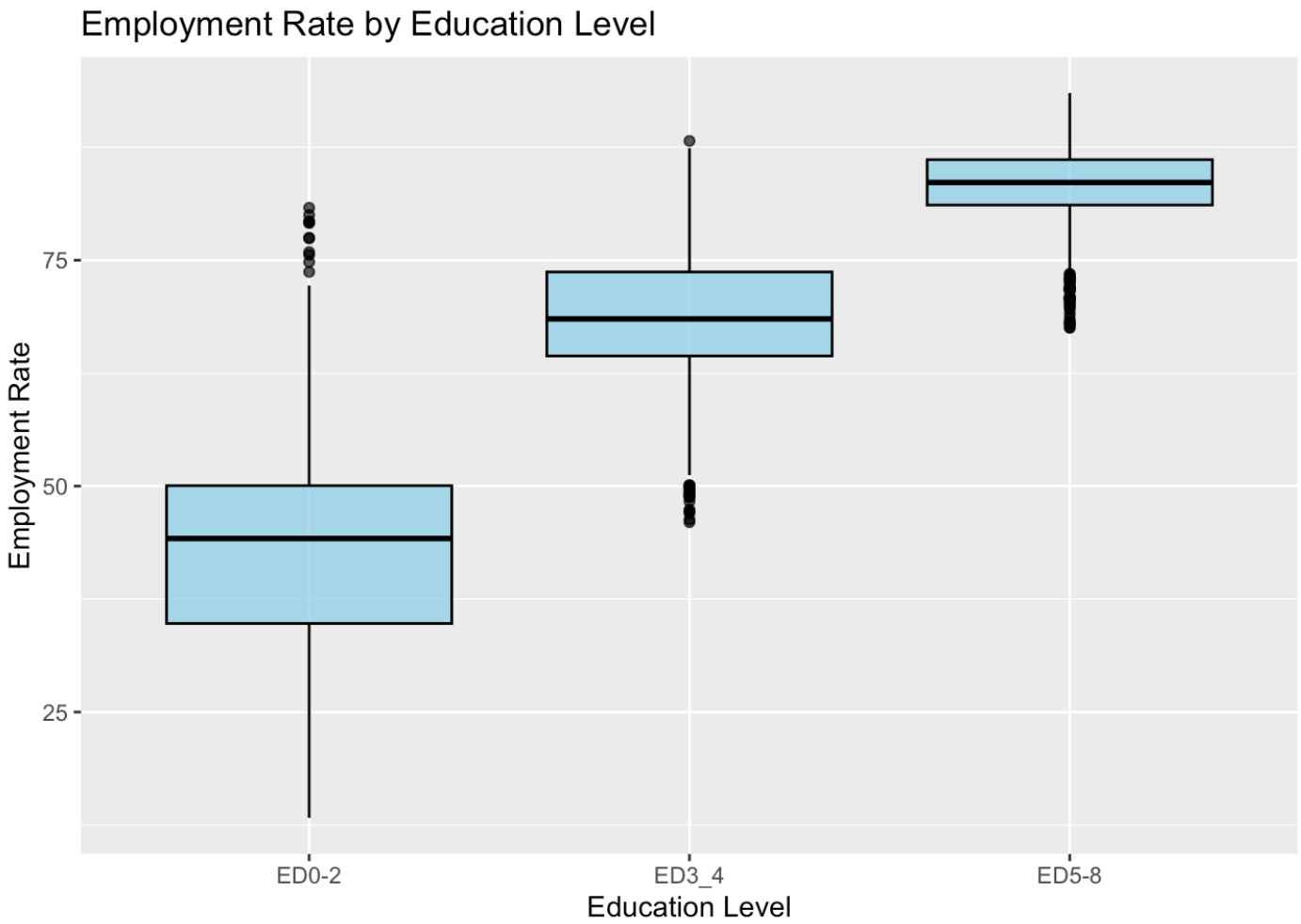
Nous avons constaté que le taux d'emploi des jeunes de 15 à 24 ans et des personnes âgées de 55 à 64 ans était d'environ 35% en 2000. Cela signifie qu'en 2000, les taux d'emploi des jeunes et des personnes âgées étaient relativement bas, mais au fil du temps, leurs données ont divergé, suivant des trajectoires différentes. Le taux d'emploi des personnes âgées de 55 à 64 ans n'a cessé d'augmenter de 2000 à 2020, atteignant même 60% en 2020. En revanche, le taux d'emploi des jeunes de 15 à 24 ans a connu une baisse, atteignant son point le plus bas à 30%.

Nous émettons l'hypothèse que cela pourrait être dû à l'élévation constante de l'âge de la retraite, obligeant les personnes âgées à travailler pendant quelques années supplémentaires. En revanche, les jeunes pourraient nécessiter de plus en plus de temps d'éducation, passant plus de temps à étudier et à obtenir davantage de diplômes, retardant ainsi leur entrée sur le marché du

travail.

5. Visualisation de la relation entre le niveau d'éducation et le taux d'emploi

```
emp_by_edu_filtered <- get_eurostat_data("lfsi_educ_a_h", filters = c("PC_P
EU27_2020"))
# boxplot
ggplot(emp_by_edu_filtered, aes(x = isced11, y = values)) +
  geom_boxplot(fill = "skyblue", color = "black", alpha = 0.7) +
  labs(title = "Employment Rate by Education Level",
       x = "Education Level",
       y = "Employment Rate")
```



Nous observe une tendance où un niveau d'éducation plus élevé est généralement lié à des taux d'emploi plus élevés. De plus, les données relatives aux taux d'emploi pour les individus ayant un niveau d'éducation supérieur montrent une dispersion moindre, concentrée généralement à des niveaux élevés. En d'autres termes, un niveau d'éducation supérieur est généralement corrélé à des taux d'emploi plus élevés, et cette relation semble relativement stable.

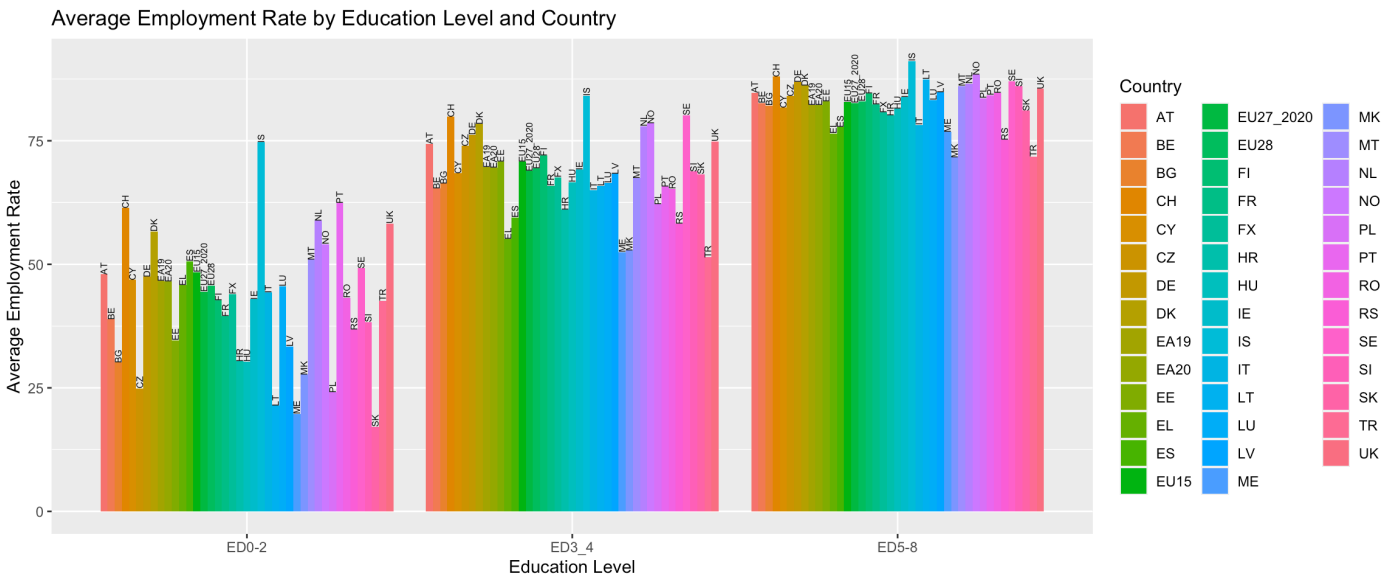
Cependant, lorsqu'on se concentre sur les niveaux d'éducation plus bas, on observe une dispersion plus importante entre le premier quartile et le troisième quartile. De plus, il existe de

nombreux points de données aberrants. Cela suggère que, bien que globalement, un niveau d'éducation plus bas soit associé à un taux d'emploi plus bas, cette relation n'est pas aussi stable. Il existe des situations où, malgré un niveau d'éducation bas, le taux d'emploi peut être élevé. En résumé, la corrélation entre le niveau d'éducation bas et le taux d'emploi bas n'est pas aussi certaine, et des variations importantes peuvent être observées en fonction du lieu et du moment.

Taux d'emploi moyens en fonction des niveaux d'éducation et des pays

```
# Calcul des moyennes par sous-groupe de pays et de niveau d'éducation
summary_emp_by_edu_filtered <- emp_by_edu_filtered %>%
  group_by(geo, isced11) %>%
  summarize(mean_values = mean(values, na.rm = TRUE),.groups = "drop")

# Graphique des comparaisons entre groupes
ggplot(summary_emp_by_edu_filtered, aes(x = isced11, y = mean_values, fill
  geom_bar(stat = "identity", position = "dodge") +
  geom_text(aes(label = geo),
            position = position_dodge(width = 0.9),
            vjust = 0.5,
            size =2,
            angle = 90,
            hjust = 0)  +
  labs(title = "Average Employment Rate by Education Level and Country",
       x = "Education Level",
       y = "Average Employment Rate",
       fill = "Country")
```



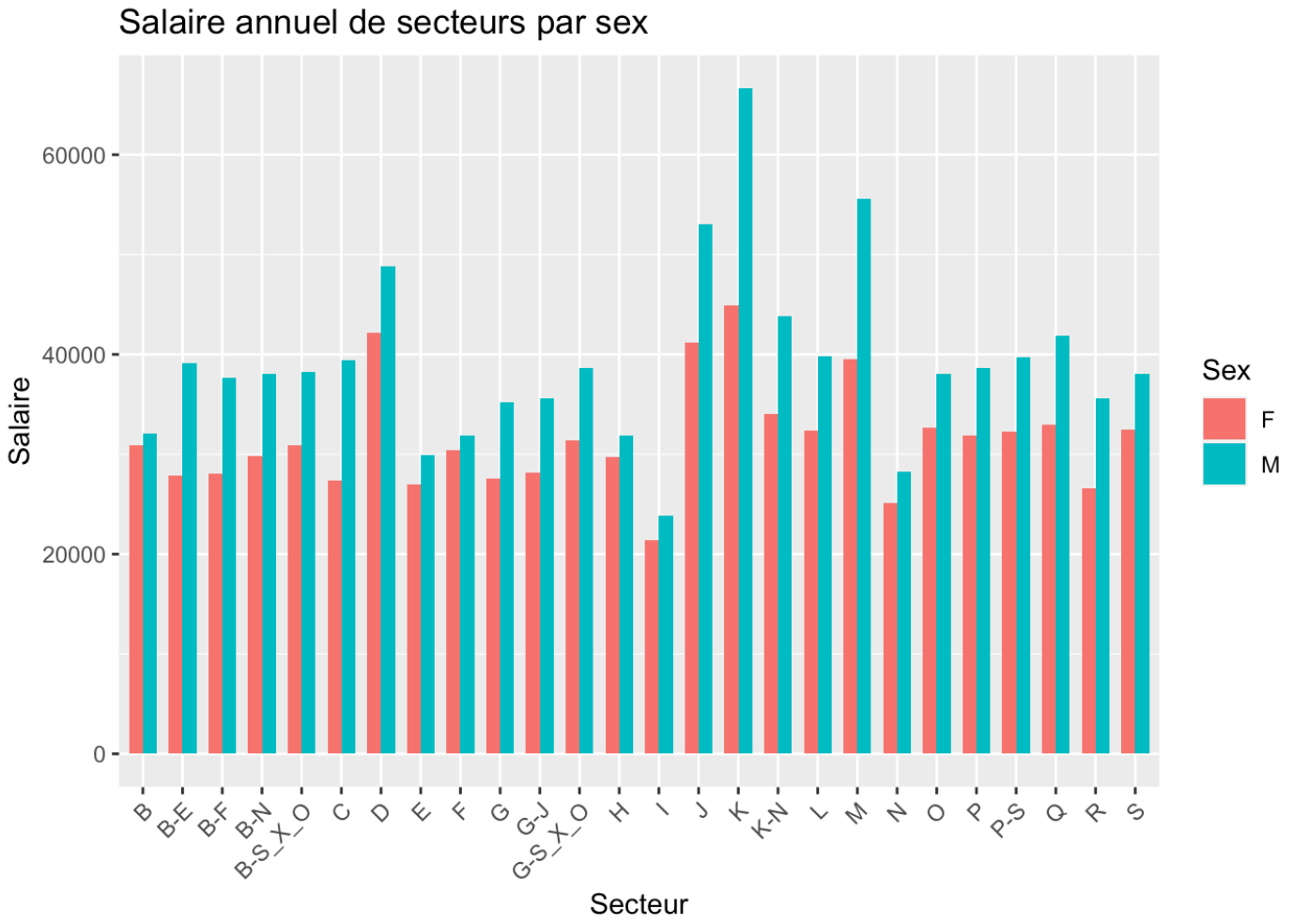
Lors de notre étude sur les taux d'emploi moyens selon différents niveaux d'éducation et pays, nous pouvons clairement observer que les données de certains pays confirment nos hypothèses précédentes. Prenons l'exemple de l'Islande. Bien que, comme dans d'autres pays, un niveau d'éducation plus bas soit associé à un taux d'emploi plus bas, en comparaison avec d'autres

nations, l'Islande présente une particularité. Les personnes ayant un niveau d'éducation relativement bas en Islande affichent des taux d'emploi élevés, parfois même supérieurs à ceux de certains pays où le niveau d'éducation est plus élevé. Examinons maintenant un autre exemple avec la Slovaquie ("SK"). Dans ce pays,les faibles niveaux d'éducation se traduisent par de faibles taux d'emploi, mais à mesure que le niveau d'éducation augmente, les taux d'emploi en Slovaquie deviennent également élevés. Cela signifie que la corrélation entre le niveau d'éducation et les taux d'emploi est très forte.

6. Étude sur la relation entre le revenu, le sexe et l'activité économique

```
salaire_dsd <- get_eurostat_dsd("earn_ses18_30")
salaire <- get_eurostat_data("earn_ses18_30",
                             filters = c("F,M","ERN", "TOTAL", "GE10", "EUR
salaire_sex_ecoact <- salaire %>%
filter(sex != "T")

# Visualisation
ggplot(salaire_sex_ecoact, aes(x = nace_r2, y = values, fill = sex)) +
  geom_bar(stat = "identity", position = "dodge", width = 0.7) +
  labs(title = "Salaire annuel de secteurs par sex",
       x = "Secteur",
       y = "Salaire",
       fill = "Sex")+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Dans tous les secteurs de l'UE, les salaires des hommes sont systématiquement supérieurs à ceux des femmes, avec des écarts plus importants dans les secteurs K, M, J, C, B_E (Activités financières et d'assurance; Activités scientifiques et techniques; Information et communication; Fabrication, Industrie (à l'exclusion de la construction)). En revanche, les secteurs B, E, F, H, I (Extraction de minerais et de carrières, Approvisionnement en eau; gestion des eaux usées et des déchets; Construction; Transport et entreposage; Hébergement et restauration) présentent des disparités moins marquées entre les revenus des hommes et des femmes.

Les secteurs D, J, K, M (Production et distribution d'électricité, de gaz, de vapeur et d'air conditionné; Information et communication; Activités financières et d'assurance; Activités scientifiques et techniques) affichent des niveaux de revenus élevés pour les hommes et les femmes (supérieurs à 40000); tandis que les secteurs E, I, N (Approvisionnement en eau; Gestion des eaux usées et des déchets; Hébergement et restauration; Activités de services administratifs et de soutien) présentent des niveaux de revenus plus bas (inférieurs à 30000).

III. Régression linéaire

Régression linéaire de salaire en fonction du taux d'emploi

```
salaire_dsd <- get_eurostat_dsd("earn_ses18_30")
salaire_total <- get_eurostat_data("earn_ses18_30",
```

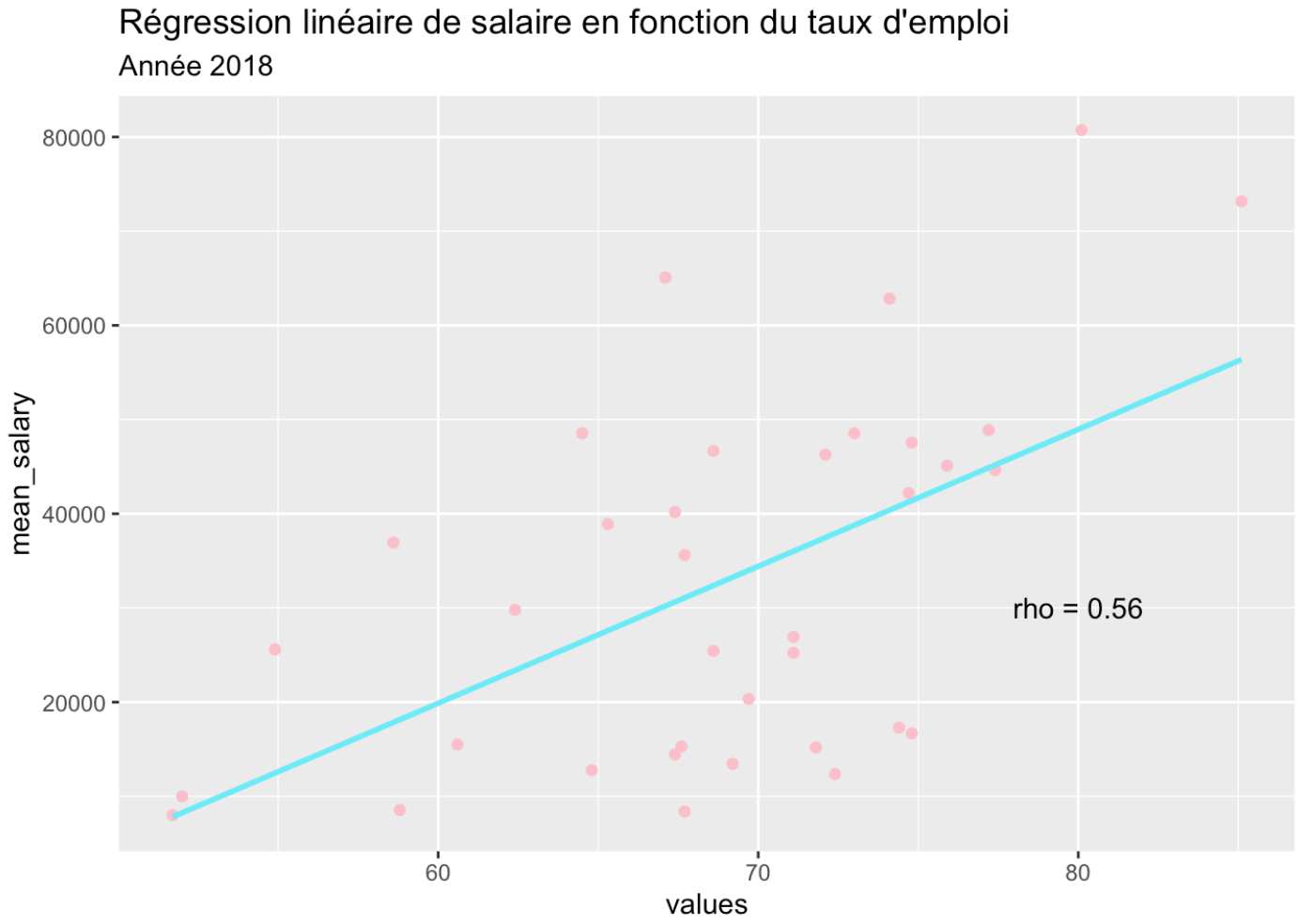
```
                                filters = c("T","ERN", "TOTAL", "GE10", "EUR")
salaire_sex_ecoact_total <- salaire_total %>%
  group_by(geo)%>%
  summarise(mean_salary = mean(values))

employment_geo_2018 <- get_eurostat_data("lfsi_emp_a_h",
                                  filters = c("PC_POP","T","Y15-64","EMP_LFS
  filter( time == "2018")

emp_sal <- employment_geo_2018 %>%
  inner_join(salaire_sex_ecoact_total, by = join_by(geo))

#scatterplot
p <- emp_sal %>%
  ggplot(aes(x=values, y=mean_salary)) +
  geom_point(color='pink') +
  labs(title = "Régression linéaire de salaire en fonction du taux d'emploi
       subtitle = "Année 2018")

rho <- cor(emp_sal$mean_salary, emp_sal$values)
#régression linéaire
p + geom_smooth(method="lm",
                se=FALSE,
                color='#71EAF8',
                formula="y~x") +
  annotate(geom = "text",
           x = 80, y = 30000,
           label = glue("rho = {round(rho, 2)}"))
```

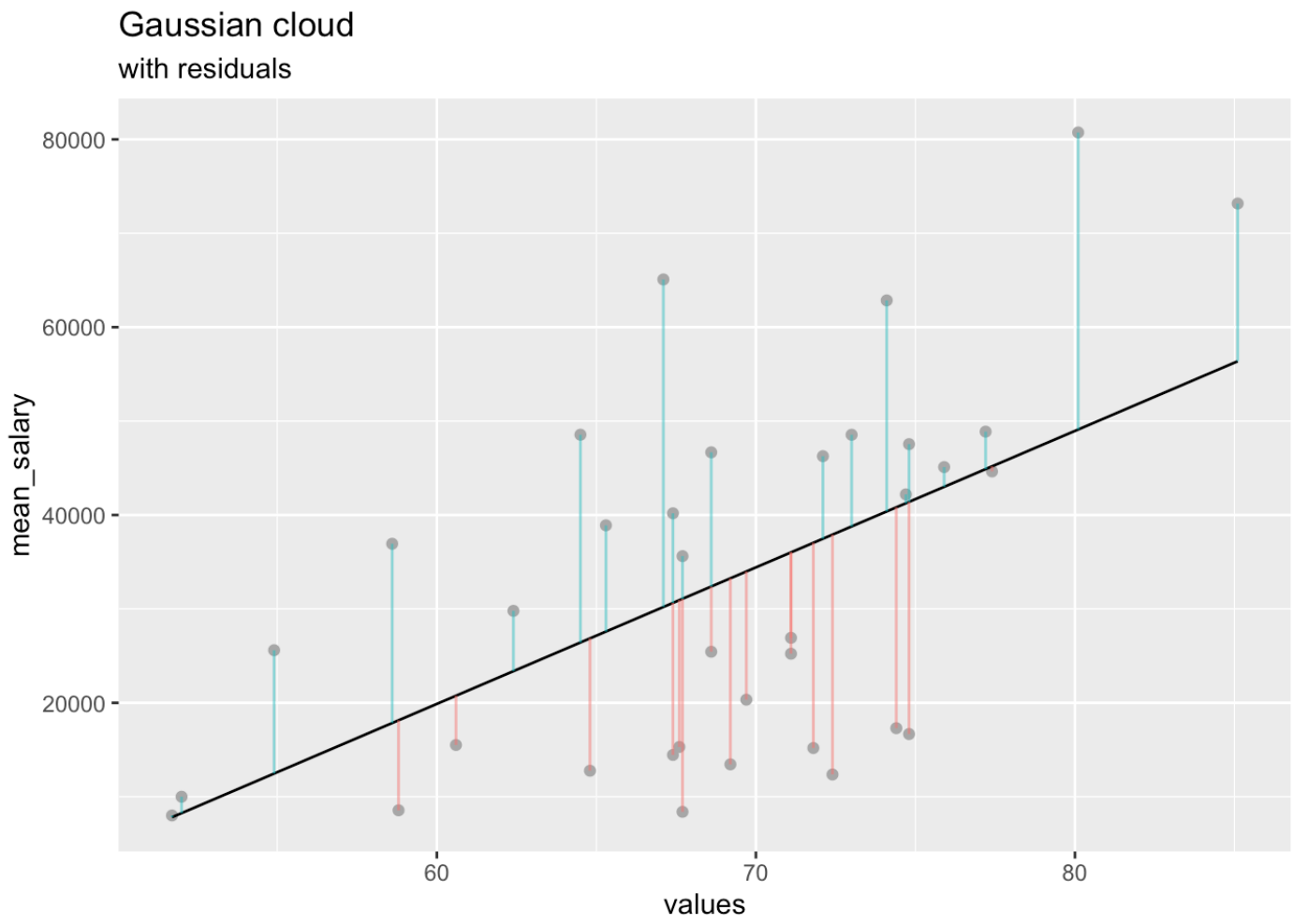


En calculant le coefficient de corrélation de Pearson rho, nous obtenons que rho est égal à 0.55 > 0, ce qui signifie que le salaire moyen a tendance à augmenter lorsque le taux d'emploi augmente. Dans le graphique, bien que la régression linéaire soit positivement corrélée, de nombreux points se trouvent à une grande distance de la droite, nous devrons donc analyser l'écart entre les résidus et le salaire moyen.

```
# Ajustement d'un modèle linéaire
lm1 <- lm(mean_salary~values, data=emp_sal)

# Utilisation d'augment sur le modèle linéaire
emp_sal_augmented <- broom::augment(lm1)

# Création de plot
ggplot(emp_sal_augmented) +
  geom_point(aes(x = values, y = mean_salary),
             color = "darkgrey") +
  geom_line(aes(x = values, y = .fitted)) +
  geom_segment(aes(x = values, xend = values, y = .fitted, yend = mean_sala
                   color = forcats::as_factor(sign(.resid))),
               alpha = 0.5) +
  theme(legend.position = "None") +
  ggtitle("Gaussian cloud", subtitle = "with residuals")
```



Dans le graphique de Gaussien cloud, les segments bleues représentent les résidus positifs, les segments rouges représentent les résidus négatifs, le nombre de points de résidus positifs et négatifs est approximativement égal, plus le milieu de la droite d'ajustement est proche, plus il y a de points de données, plus les résidus des données sont importants.

IV. Évaluation de la qualité des dépendances linéaires

```
lm1 <- lm(mean_salary~values, data=emp_sal)
summary(lm1)
```

```
Call:
lm(formula = mean_salary ~ values, data = emp_sal)

Residuals:
   Min     1Q Median     3Q    Max
-25542 -13762   1343  10150  34878

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -67339.9    25739.9  -2.616 0.013169 *
values        1453.7      372.3   3.904 0.000425 ***
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16460 on 34 degrees of freedom
Multiple R-squared:  0.3096,    Adjusted R-squared:  0.2893
F-statistic: 15.25 on 1 and 34 DF,  p-value: 0.0004254
```

```r
coeff <- lm1$coefficients # â, b_chapeau
kable(coeff)
```

	x
(Intercept)	-67339.889
values	1453.744

```r
#Summary statistique du modèle lm1
kable(broom::tidy(lm1))
```

term	estimate	std.error	statistic	p.value
(Intercept)	-67339.889	25739.8632	-2.616171	0.0131691
values	1453.744	372.3261	3.904493	0.0004254

```r
#Informations sur le diagnostic du modèle lm1
kable(broom::glance(lm1))
```

r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC	BIC	deviance
0.3095755	0.2892689	16458.6	15.24506	0.0004254	1	-399.5627	805.1253	809.8759	9210111993

Selon le modèle de régression linéaire établi, nous avons étudié la relation entre le salaire moyen (‘mean_salary’) et le taux d’emploi(‘values’). Les résultats du modèle indiquent que cette relation linéaire est statistiquement significative et offre une certaine explication des variations du salaire moyen.

Tout d’abord, l’Intercept du modèle est de -66865.4, ce qui signifie que lorsque le taux d’emploi est nulle, la valeur estimée du salaire moyen est négative. Cependant, il est important de noter que, compte tenu du contexte actuel, l’Intercept pourrait ne pas avoir de signification pratique.

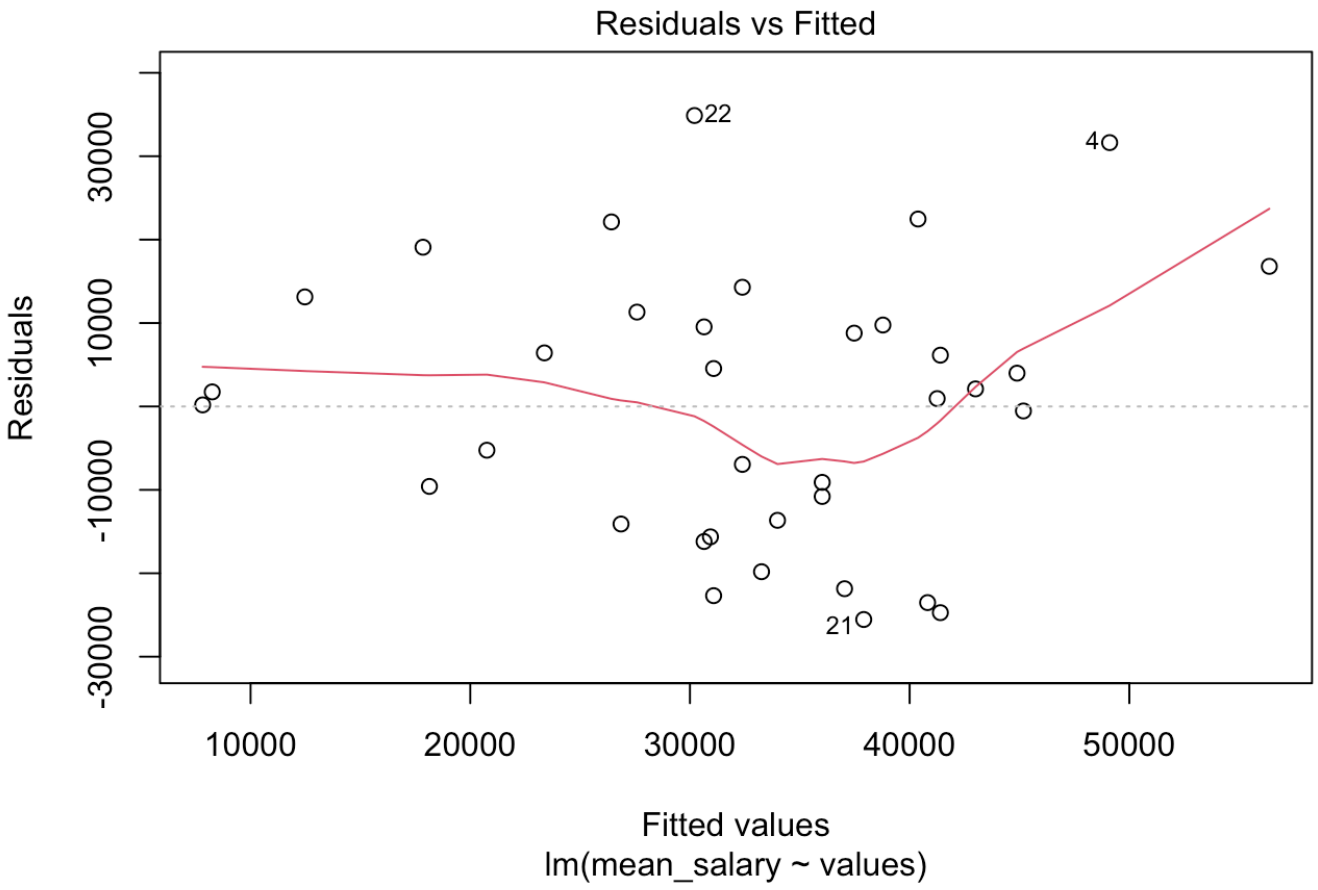
Ensuite, le coefficient de la variable le taux d’emploi est de 1446.4, ce qui signifie qu’une augmentation d’une unité du taux d’emploi est associée à une augmentation prévue de 1446.4 pour le salaire moyen. La significativité de ce coefficient, avec une valeur p de 0.000449, indique que l’impact de taux d’emploi sur le salaire moyen est statistiquement significatif.

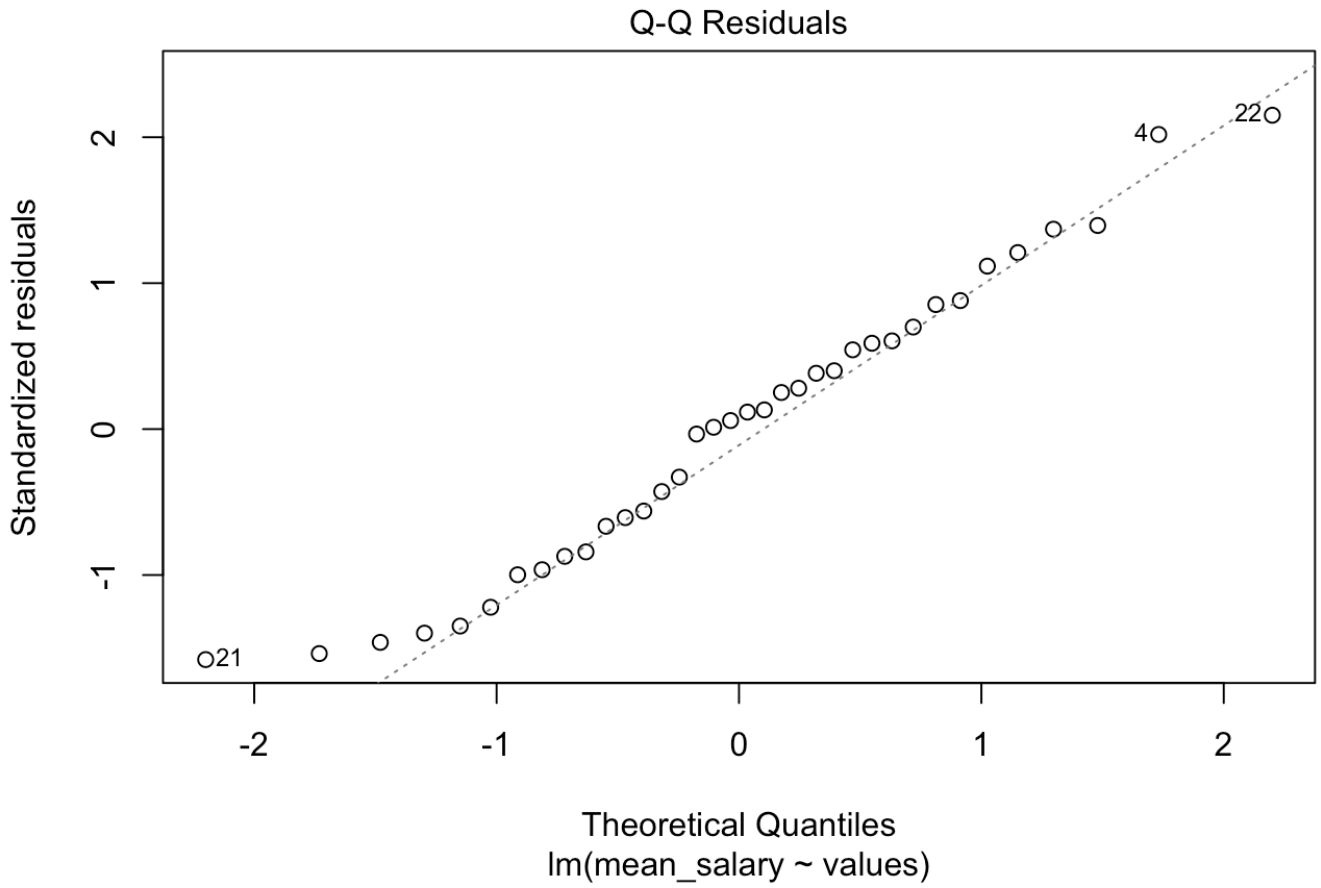
En ce qui concerne l’ajustement global du modèle, le R^2 est de 0.3075, ce qui signifie que le modèle peut expliquer environ 30.75% des variations du salaire moyen. Cela suggère que notre modèle offre une explication partielle, mais pas complète, des fluctuations du salaire moyen.

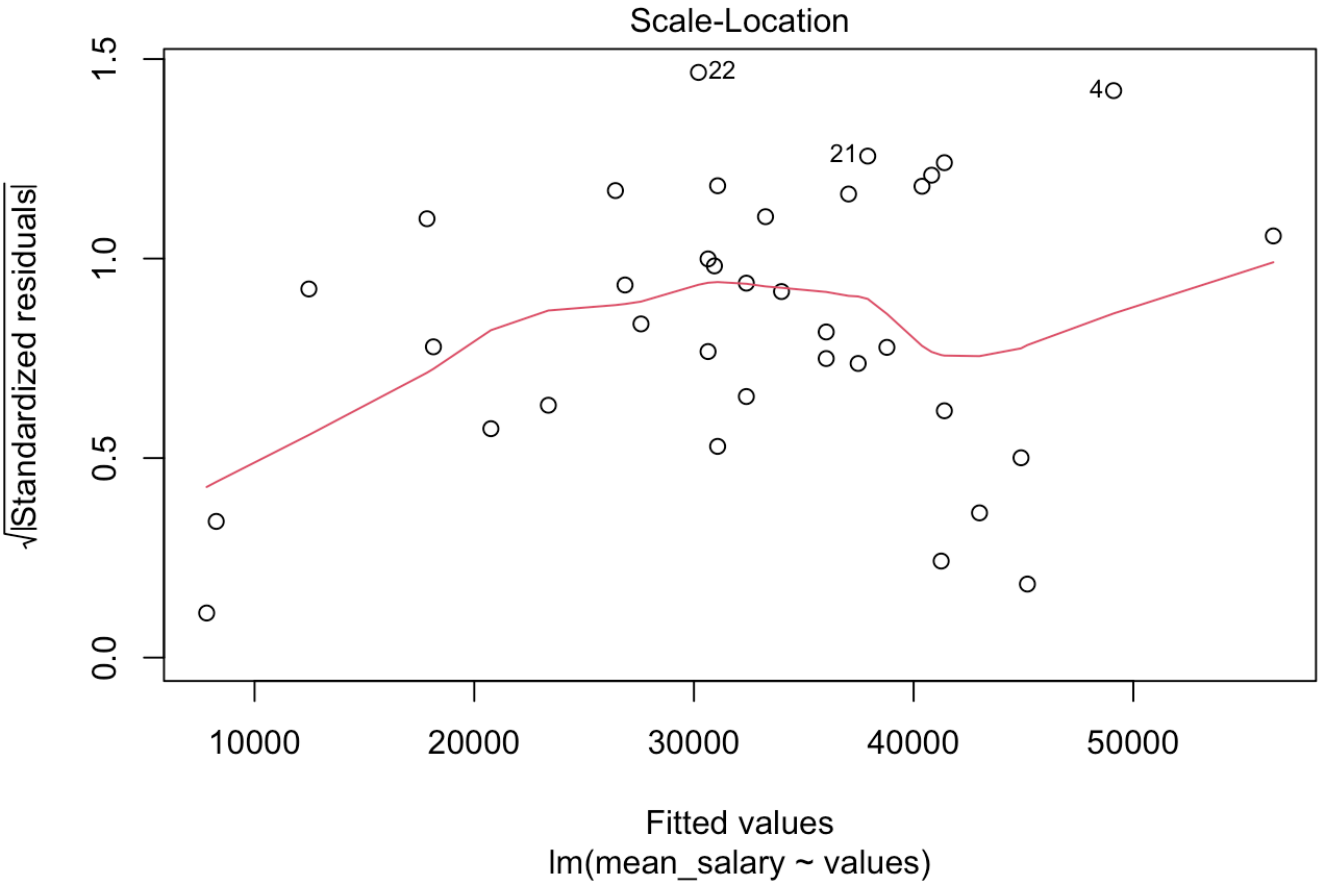
Enfin, le F-statistique de 15.1, avec la p-value correspondante de 0.0004492, confirme que le modèle dans son ensemble est significatif. Cela indique que notre modèle est significativement meilleur que le modèle nul qui ne contient aucune variable prédictive.

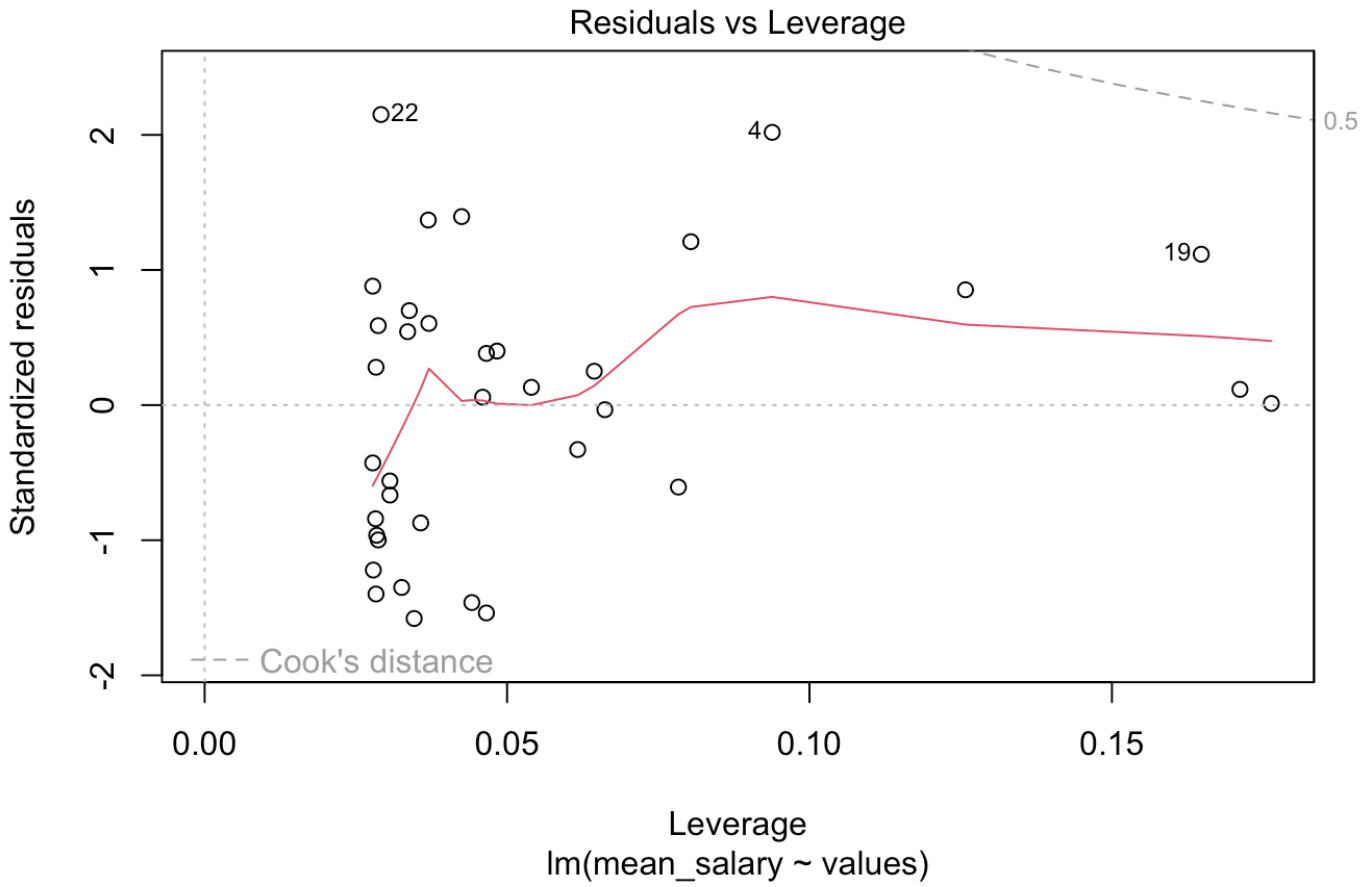
En conclusion, d'après les résultats de ce modèle de régression linéaire, nous pouvons affirmer que le taux d'emploi a un impact significatif sur le salaire moyen. Cependant, la capacité explicative du modèle est limitée, suggérant la nécessité d'examiner d'autres facteurs potentiels.

```
#diagnostic plots
plot(lm1)
```




Q-Q Residuals
Standardized residuals
Theoretical Quantiles
lm(mean_salary ~ values)
21
4
22


Scale-Location
√|Standardized residuals|
Fitted values
lm(mean_salary ~ values)
22
4
21



Le premier plot est Residuals vs Fitted Values Plot, il est pour vérifier la variance constante des résidus: la courbe rouge est la tendance des résidus (la moyenne de les residus), par observation de la courbe, on en déduit que la variance des résidus ne sont pas presque constantes, donc $\hat{y}$ ne satisfait pas $e \sim N(0, \sigma^2)$.

Le deuxième plot est Quantile-Quantile (Q-Q) Plot, il est pour vérifier la normalité des résidus: dans le graphique, nous observons que tous les points forment une ligne droite et que seul un très petit nombre de points s'écarte de la ligne, ce qui nous permet de conclure que les résidus du modèle de régression sont normalement distribués.

Le troisième plot est Scale-Location Plot, il est pour vérifier l'homoscédasticité: nous constatons qu'il n'y a pas de tendance ou de variation claire dans la dispersion des points, ce qui peut indiquer la présence d'homoscédasticité.

Le quatrième plot est Residuals vs Leverage Plot, il est pour identifier les points de données influents: dans ce plot, il n'y a pas de points se trouvent en dehors de la ligne pointillée (distance de Cook), donc il n'y a aucun point influent dans notre modèle de régression.

V. Motivation du choix de l'ensemble de données et conclusion

Avant tout, l'étude de l'emploi revêt une importance capitale dans la compréhension des dynamiques socio-économiques. L'emploi constitue souvent la pierre angulaire de la subsistance

individuelle, garantissant des revenus stables et contribuant à la stabilité financière des individus. Par ailleurs, il est étroitement lié à la croissance économique, étant un indicateur crucial d'une économie active.

Tout d'abord, l'analyse de l'emploi selon le sexe se révèle essentielle pour identifier et remédier aux disparités de genre sur le marché du travail. Cette approche contribue à promouvoir l'équité salariale, à assurer un accès équitable aux opportunités professionnelles et à lutter contre la discrimination.

Ensuite, en examinant l'emploi selon l'âge, on peut comprendre les défis spécifiques auxquels font face différentes générations. Mettre en lumière les obstacles à l'employabilité des jeunes et favoriser un environnement inclusif pour les travailleurs plus âgés devient impératif.

Par la suite, la considération de l'emploi en fonction du niveau d'éducation offre des indications sur l'équité des opportunités d'emploi en lien avec l'accès à l'éducation. Identifier les lacunes dans cet accès. Cette approche contribue à assurer que l'éducation conduit effectivement à des opportunités professionnelles équitables.

En plus, La recherche sur la relation entre le revenu, le sexe et l'activité se révèle essentielle pour comprendre les inégalités économiques et sociales. En révélant les écarts de revenu entre hommes et femmes dans tous les secteurs d'activité, cette étude offre des bases solides pour réduire l'inégalité entre les sexes dans divers domaines professionnels.

Enfin, l'étude de la dépendance linéaire entre le taux d'emploi et le salaire fournit des informations cruciales pour élaborer des politiques économiques éclairées. La conclusion d'une dépendance linéaire significative, où une augmentation du taux d'emploi est corrélée à une hausse des salaires, encourage la création d'initiatives visant à stimuler l'emploi. Cette conclusion guide également l'identification des tendances sur le marché du travail et la promotion de conditions de travail et de salaires plus équitables.

En résumé, l'étude de l'emploi sous diverses dimensions offre une perspective complète des enjeux socio-économiques. Les conclusions de ces analyses sont cruciales pour formuler des politiques inclusives, promouvoir l'équité et travailler vers des conditions de travail et des salaires plus équitables.

Annexe

Analyse des Données avec `skimr`

Visualisation des Données avec `ggplot2`

Manipulation de Données avec `dplyr`

Utilisation de `glue` pour l'Interpolation de Texte

Création de Tableaux Interactifs avec `DT`

Utilisation de `knitr` pour Intégrer le Code dans le Rapport

Utilisation de `restatapi` pour télécharger des données Eurostat