# proj01

František Kareš

2025-03-28

```
K = 4
L = 5
M = (((K+L)*47)%%(11))+1
print(M)
```

```
## [1] 6
```

```
knitr::opts_chunk$set(warning = FALSE, message = FALSE)
```

```
library(eurostat)
library(dplyr)
library(ggplot2)
library(sf)
library(ggiraph)
library(patchwork)
```

# ŠPANĚLSKO - EMPLOYMENT (NUTS3,NACE)

Analyzuji data o zaměstnanosti v NUTS3 regionech Španělska z roku 2022, rozdělená také podle ekonomických sektorů dle klasifikace NACE.

## Předzpracování dat

```
populatio_data <- populatio_data[populatio_data$TIME_PERIOD == "2023-01-01" & populatio_data$sex == "T"
```

```
spain_population <- populatio_data[grepl("^ES[1-9][0-9][0-9]$",populatio_data$geo),]
head(spain_population)
```

```
## # A tibble: 6 x 7
##   freq  sex   unit  age   geo   TIME_PERIOD  values
##   <chr> <chr> <chr> <chr> <chr> <date>        <dbl>
## 1 A     T     NR    TOTAL ES111 2023-01-01  1123884
## 2 A     T     NR    TOTAL ES112 2023-01-01   324267
## 3 A     T     NR    TOTAL ES113 2023-01-01   304563
## 4 A     T     NR    TOTAL ES114 2023-01-01   946710
## 5 A     T     NR    TOTAL ES120 2023-01-01  1006060
## 6 A     T     NR    TOTAL ES130 2023-01-01   588387
```

```
head(employment_data)
```

```
## # A tibble: 6 x 7
##   freq  unit  wstatus nace_r2 geo   TIME_PERIOD values
##   <chr> <chr> <chr>   <chr>   <chr> <date>       <dbl>
## 1 A     THS   EMP     A       AT    2000-01-01    231.
## 2 A     THS   EMP     A       AT    2001-01-01    230.
## 3 A     THS   EMP     A       AT    2002-01-01    226.
## 4 A     THS   EMP     A       AT    2003-01-01    225.
## 5 A     THS   EMP     A       AT    2004-01-01    218
## 6 A     THS   EMP     A       AT    2005-01-01    215
```

```
total_es <- employment_data[employment_data$geo == 'ES' & employment_data$wstatus == "EMP" & employment_
total_es
```

```
## # A tibble: 1 x 7
##   freq  unit  wstatus nace_r2 geo   TIME_PERIOD values
##   <chr> <chr> <chr>   <chr>   <chr> <date>       <dbl>
## 1 A     THS   EMP     TOTAL   ES    2022-01-01  20828.
```

rozděleni španělska podle NUTS3

```
spain_data <- employment_data[grepl("^ES[1-9][0-9][0-9]$",employment_data$geo)& employment_data$geo != "

spain_data <- spain_data %>%
  mutate(nace_r2 = case_when(
    nace_r2 == "A" ~ "A-F",
    nace_r2 == "B-E" ~ "A-F",
    nace_r2 == "F" ~ "A-F",
    TRUE ~ nace_r2  # Keeps the rest unchanged
  ))

spain_data_total_geo <- employment_data[grepl("^ES[1-9][0-9][0-9]$",employment_data$geo)& employment_da

spain_data_total_geo_self <- employment_data[grepl("^ES[1-9][0-9][0-9]$",employment_data$geo)& employme
head(spain_data)
```

```
## # A tibble: 6 x 7
##   freq  unit  wstatus nace_r2 geo   TIME_PERIOD values
##   <chr> <chr> <chr>   <chr>   <chr> <date>       <dbl>
## 1 A     THS   EMP     A-F     ES111 2022-01-01    22.7
## 2 A     THS   EMP     A-F     ES112 2022-01-01    20.8
## 3 A     THS   EMP     A-F     ES113 2022-01-01     6
## 4 A     THS   EMP     A-F     ES114 2022-01-01    21.5
## 5 A     THS   EMP     A-F     ES120 2022-01-01    13.1
## 6 A     THS   EMP     A-F     ES130 2022-01-01     5.9
```

```
spain_data <- subset(spain_data, select = -c(freq,unit,wstatus) )
```

kontrola NUTS3 a NACE

```r
unique(spain_data$geo)
```

```
##  [1] "ES111" "ES112" "ES113" "ES114" "ES120" "ES130" "ES211" "ES212" "ES213"
## [10] "ES220" "ES230" "ES241" "ES242" "ES243" "ES300" "ES411" "ES412" "ES413"
## [19] "ES414" "ES415" "ES416" "ES417" "ES418" "ES419" "ES421" "ES422" "ES423"
## [28] "ES424" "ES425" "ES431" "ES432" "ES511" "ES512" "ES513" "ES514" "ES521"
## [37] "ES522" "ES523" "ES531" "ES532" "ES533" "ES611" "ES612" "ES613" "ES614"
## [46] "ES615" "ES616" "ES617" "ES618" "ES620" "ES630" "ES640" "ES703" "ES704"
## [55] "ES705" "ES706" "ES707" "ES708" "ES709"
```

```r
unique(spain_data$nace_r2)
```

```
## [1] "A-F" "G-J" "K-N" "O-U"
```

sjednocení canary islands

```r
spain_data <- spain_data %>%
  mutate(geo = ifelse(grepl("^ES7", geo), "Canary Islands", geo)) %>%
  group_by(geo, nace_r2) %>%
  mutate(values = sum(values, na.rm = TRUE)) %>%
  ungroup() %>%
  distinct()
```

rozdělení podle industry

```r
AF_data <- spain_data[spain_data$nace_r2 == "A-F",]
GJ_data <- spain_data[spain_data$nace_r2 == "G-J",]
KN_data <- spain_data[spain_data$nace_r2 == "K-N",]
OU_data <- spain_data[spain_data$nace_r2 == "O-U",]
unique(AF_data$nace_r2)
```

```
## [1] "A-F"
```

```r
unique(GJ_data$nace_r2)
```

```
## [1] "G-J"
```

```r
unique(KN_data$nace_r2)
```

```
## [1] "K-N"
```

```r
unique(OU_data$nace_r2)
```

```
## [1] "O-U"
```

# Úloha 1

```r
spain_summary_geo <- spain_data %>% group_by(geo) %>% summarize(employment = sum(values, na.rm = TRUE))

summary(spain_summary_geo)
```

```
##      geo              employment
##  Length:53          Min.   :  31.4
##  Class :character   1st Qu.: 118.4
##  Mode  :character   Median : 232.0
##                     Mean   : 392.8
##                     3rd Qu.: 397.6
##                     Max.   :3716.2
```

Nejvyšší počet zaměstnanců v jednom regionu je 3705.6, což je velký rozdíl oproti průměru/mediánu.

```r
ggplot(spain_summary_geo,aes(x=reorder(geo, employment),y=employment))+
   geom_col(fill = "steelblue",width=0.8) +
  coord_flip() +
  labs(title = "Employment by NUTS3 Region in Spain",
       x = "NUTS3 Region",
       y = "Employment") +
  theme_minimal() +  theme(
        axis.text.y = element_text(size = 10, face = "bold"),  # Adjust y-axis text size
        axis.text.x = element_text(size = 12),  # Adjust x-axis text size
        plot.title = element_text(hjust = 0.5, size = 23, face = "bold"),  # Center and bold title
        panel.grid.major.y = element_blank(),  # Remove horizontal grid lines
        panel.grid.minor.y = element_blank(),  # Remove minor horizontal grid lines
        axis.line = element_line(color = "black")  # Add axis lines for clarity
    ) +
    scale_y_continuous(expand = expansion(mult = c(0,0.1)))+
    geom_text(aes(label = employment),  # Add value labels
        hjust = -0.2,  # Adjust horizontal position of labels
        size = 4  # Adjust size of labels
    )
```

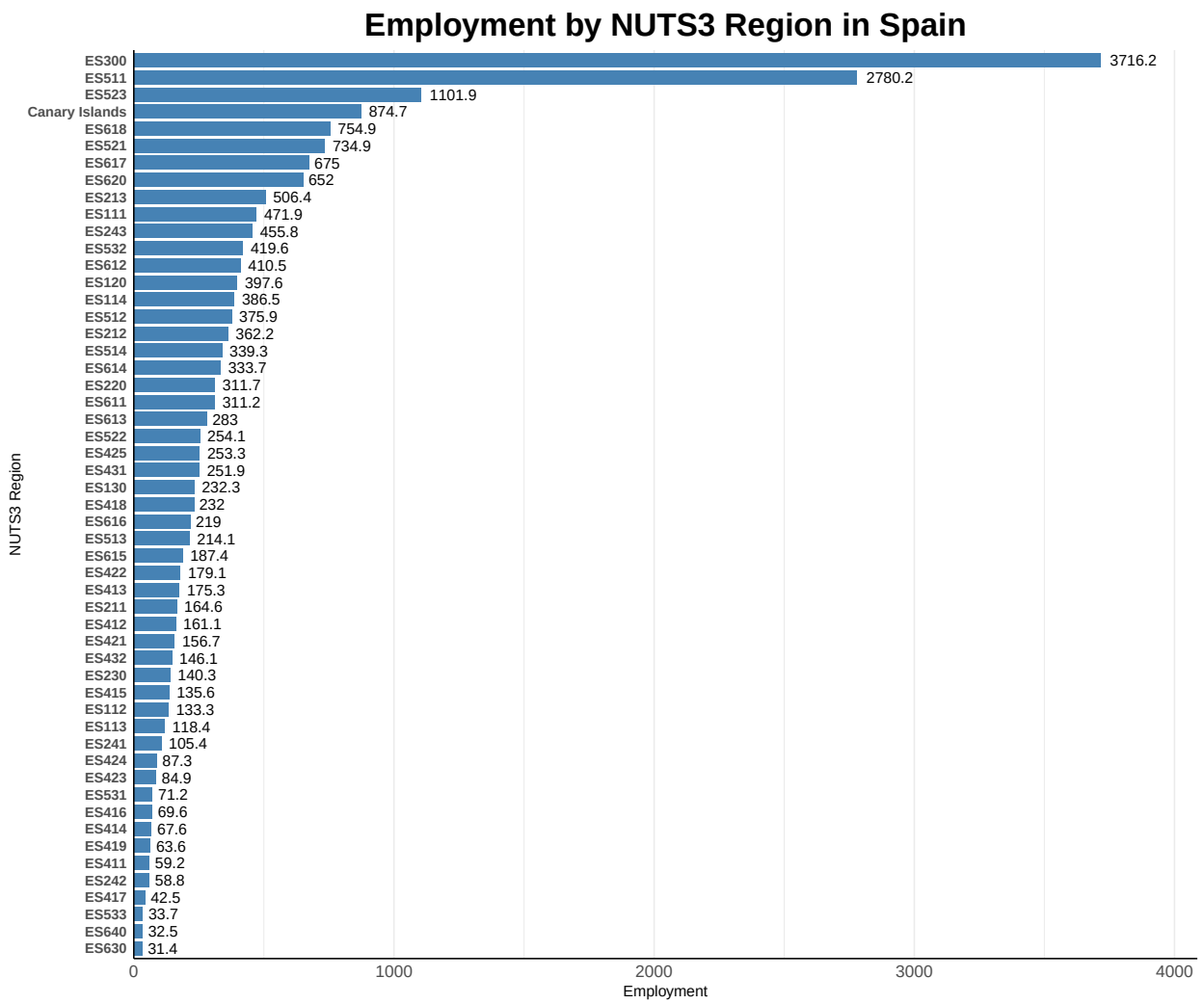



Table 1: Podle grafu nejvíce zaměstnaných lidí najdeme v Madridu, Valencii a Barceloně. Naopak na posledních dvou příčkách se nacházejí autonomní oblasti Melilla a Ceuta, ležící na severu Afriky.

| Umístění | Území | Počet (v tis.) |
|---|---|---|
| 1. | Madrid | 3891 |
| 2. | Barcelona | 3110 |
| 3. | Valencie | 1239 |

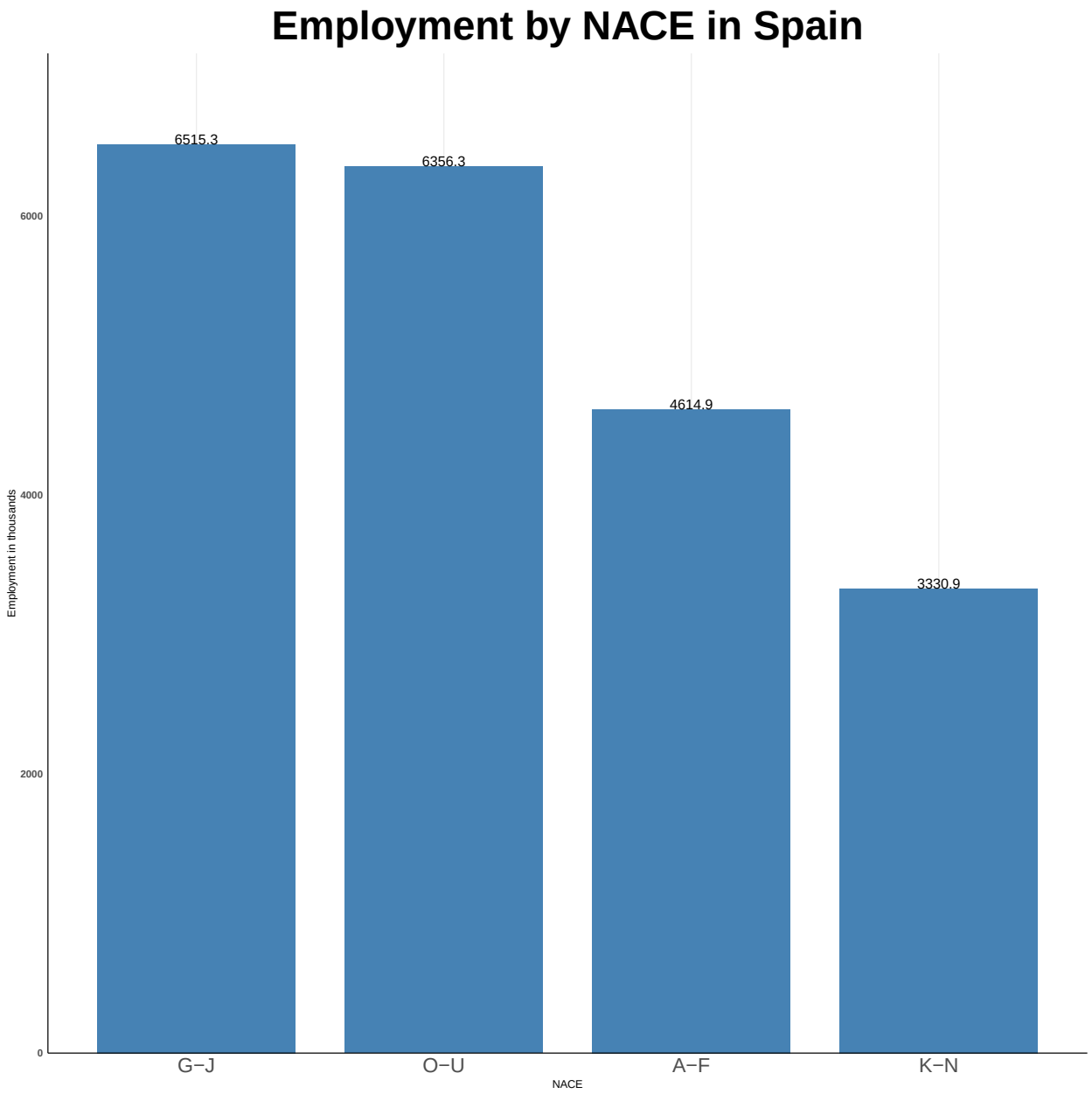
```
spain_summary_nace <- spain_data %>% group_by(nace_r2) %>% summarize(employment=sum(values))
summary(spain_summary_nace)
```

```
##    nace_r2           employment
##  Length:4           Min.   :3331
##  Class :character   1st Qu.:4294
##  Mode  :character   Median :5486
##                     Mean   :5204
##                     3rd Qu.:6396
##                     Max.   :6515
```

```r
unique(spain_data$nace_r2)
```

```
## [1] "A-F" "G-J" "K-N" "O-U"
```

```r
sector_summary <- data.frame(
  nace_r2 = c("A-F", "G-J", "K-N", "O-U"),
  employment = c(sum(AF_data$values),
                 sum(GJ_data$values),
                 sum(KN_data$values),
                 sum(OU_data$values))
)

# Ensure correct data format and sorting
sector_summary <- sector_summary %>%
  arrange(desc(employment))  # Sort sectors by employment

# Create bar plot
ggplot(sector_summary, aes(x = reorder(nace_r2, -employment), y = employment)) +
  geom_col(fill = "steelblue", width = 0.8) +
  labs(title = "Employment by NACE in Spain",
       x = "NACE",
       y = "Employment in thousands") +
  theme_minimal() +
  theme(
        axis.text.y = element_text(size = 10, face = "bold"),
        axis.text.x = element_text(size = 20),
        plot.title = element_text(hjust = 0.5, size = 40, face = "bold"),
        panel.grid.major.y = element_blank(),
        panel.grid.minor.y = element_blank(),
        axis.line = element_line(color = "black")
    ) +
  scale_y_continuous(expand = expansion(mult = c(0,0.1))) +
  geom_text(aes(label = employment),
        vjust = 0,
        size = 5
    )
```



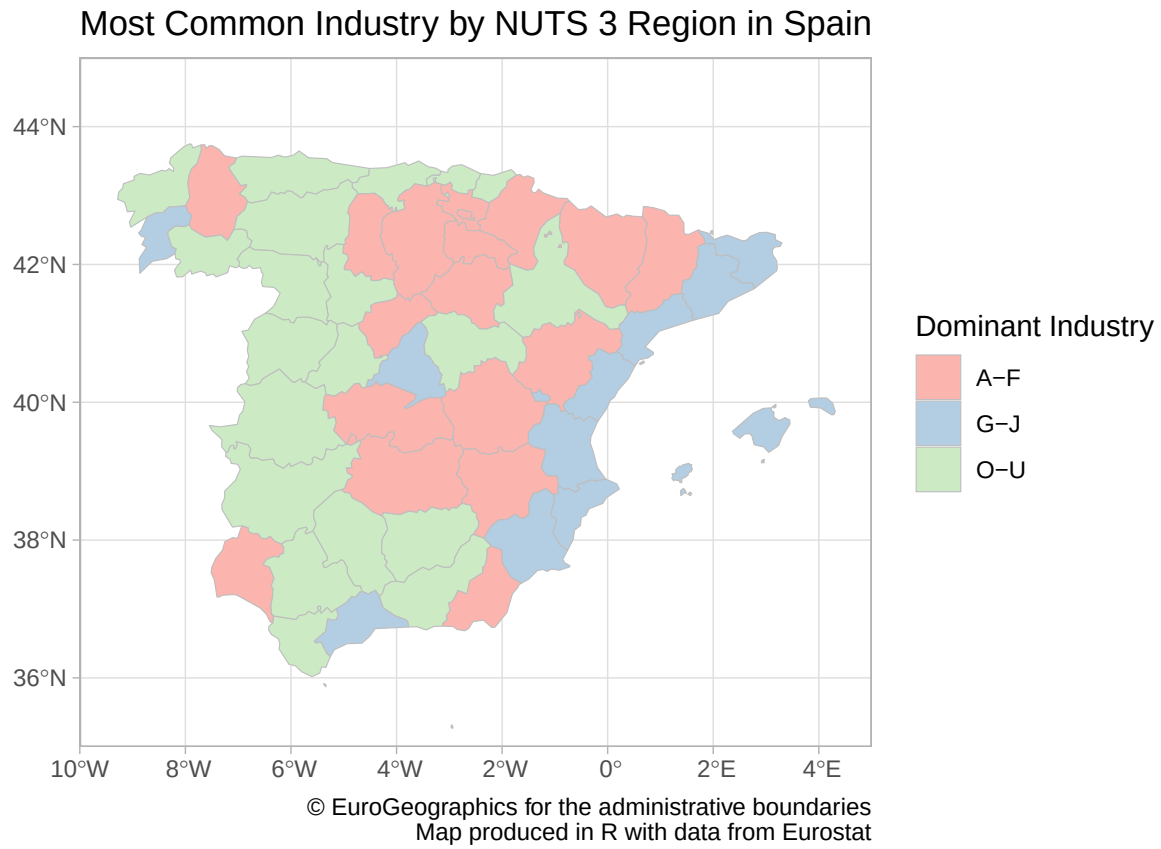

| Umístění | Odvětví | Počet (v tis.) |
|---|---|---|
| 1. | G-J | 6498,4 |
| 2. | O-U | 6347,7 |
| 3. | A-F | 4598,1 |

```r
map_nuts3_base <- get_eurostat_geospatial(nuts_level = 3, resolution = "10", output_class = "sf", counti

# Find the dominant industry in each NUTS3 region

dominant_industry <- spain_data %>%
  group_by(geo) %>%   # Group by region
  slice_max(values, n = 1) %>%  # Select row with max employment per region
```

```r
  ungroup()

head(dominant_industry)
```

```
## # A tibble: 6 x 4
##   nace_r2 geo            TIME_PERIOD values
##   <chr>   <chr>          <date>       <dbl>
## 1 G-J     Canary Islands 2022-01-01   341
## 2 O-U     ES111          2022-01-01   144.
## 3 A-F     ES112          2022-01-01    43.6
## 4 O-U     ES113          2022-01-01    38.2
## 5 G-J     ES114          2022-01-01   118.
## 6 O-U     ES120          2022-01-01   126.
```

```r
# Merge with map data
map_nuts3 <- left_join(map_nuts3_base, dominant_industry, by = "geo")

# Plot the map
ggplot(map_nuts3) +
  # Base layer
  geom_sf(fill = "lightgrey", color = "white") +
  # Choropleth layer (most dominant industry in each region)
  geom_sf(aes(fill = nace_r2), color = "grey", linewidth = 0.2, na.rm = TRUE) +
  scale_fill_brewer(palette = "Pastel1", na.translate = FALSE) +  # Use categorical color palette
  guides(fill = guide_legend(title = "Dominant Industry")) +
  labs(
    title = "Most Common Industry by NUTS 3 Region in Spain",
    caption = "© EuroGeographics for the administrative boundaries
               Map produced in R with data from Eurostat"
  ) +
  theme_light() +
  coord_sf(
    xlim = c(-10, 5),   # Longitude range (Westernmost to Easternmost Spain)
    ylim = c(35, 45),    # Latitude range (Southernmost to Northernmost Spain)
    expand = FALSE
  )
```




```
print(sum(spain_data$values[spain_data$nace_r2 == "O-U"]))
```

```
## [1] 6356.3
```

```
print(sum(spain_data$values[spain_data$nace_r2 == "A-F"]))
```

```
## [1] 4614.9
```

```
print(sum(spain_data$values[spain_data$nace_r2 == "G-J"]))
```

```
## [1] 6515.3
```

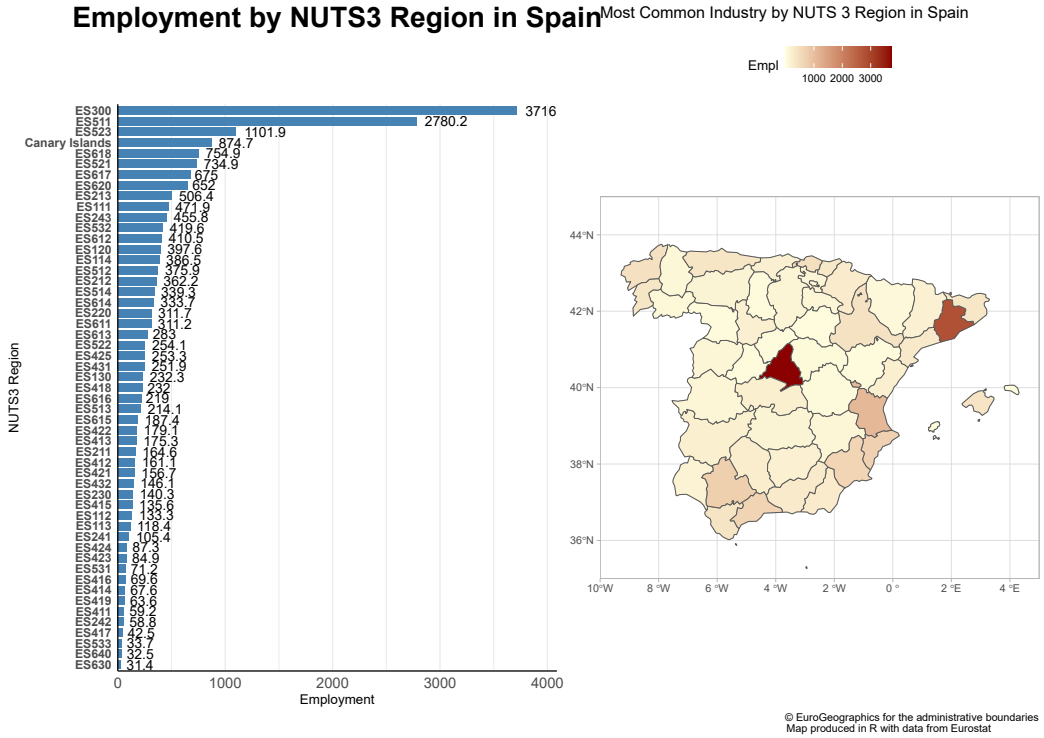
Můžeme si všimnout, že u pobřežních oblastí a Madridu(nejspíše více turistické), převládá Maloobchod, ubytování a stravování a ve zbytku převládá Španělska mix veřejného sektoru a průmyslu/zemědělství/stavebnictví. Co naopak nenalezneme na mapě je skupina K-N, tudíž sektor ICT a finančnictví nebude ve Španělsku nejhlavnějším pilířem španělské ekonomiky.

**INTERAKTIVNÍ!**

```
# Create the second chart (Bar plot)
map_nuts4 <- left_join(map_nuts3_base, spain_data_total_geo, by = "geo")

p2 <- ggplot(spain_summary_geo,aes(x=reorder(geo, employment),y=employment,tooltip = employment, data_i
```

```
   geom_col_interactive(fill = "steelblue",width=0.8) +
  coord_flip() +
  labs(title = "Employment by NUTS3 Region in Spain",
       x = "NUTS3 Region",
       y = "Employment") +
  theme_minimal() +  theme(
        axis.text.y = element_text(size = 10, face = "bold"),  # Adjust y-axis text size
        axis.text.x = element_text(size = 12),  # Adjust x-axis text size
        plot.title = element_text(hjust = 0.5, size = 23, face = "bold"),  # Center and bold title
        panel.grid.major.y = element_blank(),  # Remove horizontal grid lines
        panel.grid.minor.y = element_blank(),  # Remove minor horizontal grid lines
        axis.line = element_line(color = "black")  # Add axis lines for clarity
    ) +
    scale_y_continuous(expand = expansion(mult = c(0,0.1)))+
    geom_text(aes(label = employment),  # Add value labels
        hjust = -0.2,  # Adjust horizontal position of labels
        size = 4  # Adjust size of labels
    )

# Create the third chart (choropleth)

p3 <- ggplot(map_nuts4) +
  # Base layer
  geom_sf(fill = "lightgrey", color = "white") +
  # Choropleth layer (most dominant industry in each region)
  geom_sf(aes(fill = values), color = "grey", linewidth = 0.2, na.rm = TRUE) +
  geom_sf_interactive(
    data = map_nuts4,
    aes(fill = values, tooltip = paste("Value:", values, "<br>Geo:", geo), data_id = geo)
  )+
  scale_fill_gradient(low = "lightyellow", high = "darkred", na.value = "grey",name = "Empl") +  # Use

  labs(
    title = "Most Common Industry by NUTS 3 Region in Spain",
    caption = "© EuroGeographics for the administrative boundaries
                Map produced in R with data from Eurostat"
  ) +
  theme_light() +theme(legend.position = "top")+
  coord_sf(
    xlim = c(-10, 5),    # Longitude range (Westernmost to Easternmost Spain)
    ylim = c(35, 45),     # Latitude range (Southernmost to Northernmost Spain)
    expand = FALSE
  )

# Combine the plots
combined_plot <- p2 + p3 + plot_layout(ncol = 2,heights = c(1, 1))

# Create the interactive plot
interactive_plot <- girafe(ggobj = combined_plot,width_svg = 12, height_svg = 15)
interactive_plot <- girafe_options(
  interactive_plot,
  opts_hover(css = "fill:magenta;stroke:grey;")
)
```

```
htmltools::save_html(interactive_plot, "multiple-ggiraph-4.html")
interactive_plot
```




Největší koncentrace zaměstnanců je v Madridu, Barceloně a Valencii, jak už jsme si popsali o graf výše.

# Souhrn sektorů

## AF GROUP

- A - Agriculture, Forestry and Fishing
- B - Mining and Quarrying
- C - Manufacturing
- D - Electricity, Gas, Steam and Air Conditioning Supply
- E - Water Supply; Sewerage, Waste Management and Remediation Activities
- F - Construction

```
summary(AF_data)
```

```
##    nace_r2              geo             TIME_PERIOD            values
##  Length:53          Length:53          Min.   :2022-01-01   Min.   :  2.20
##  Class :character   Class :character   1st Qu.:2022-01-01   1st Qu.: 32.90
##  Mode  :character   Mode  :character   Median :2022-01-01   Median : 65.10
##                                        Mean   :2022-01-01   Mean   : 87.07
##                                        3rd Qu.:2022-01-01   3rd Qu.:106.40
##                                        Max.   :2022-01-01   Max.   :547.30
```

```
sum(AF_data$values)
```

```
## [1] 4614.9
```

## GJ GROUP

- G - Wholesale and Retail Trade; Repair of Motor Vehicles and Motorcycles
- H - Transportation and Storage
- I - Accommodation and Food Service Activities
- J - Publishing, Broadcasting and Content Production and Distribution Activities

```
summary(GJ_data)
```

```
##    nace_r2              geo             TIME_PERIOD            values
##  Length:53          Length:53          Min.   :2022-01-01   Min.   :   7.4
##  Class :character   Class :character   1st Qu.:2022-01-01   1st Qu.:  35.5
##  Mode  :character   Mode  :character   Median :2022-01-01   Median :  64.3
##                                        Mean   :2022-01-01   Mean   : 122.9
##                                        3rd Qu.:2022-01-01   3rd Qu.: 124.1
##                                        Max.   :2022-01-01   Max.   :1221.7
```

```
sum(GJ_data$values)
```

```
## [1] 6515.3
```

## KN GROUP

- K - Telecommunication, Computer Programming, Consulting, Computing Infrastructure, and other Information Service Activities
- L - Financial and Insurance Activities
- M - Real Estate Activities
- N - Professional, Scientific and Technical Activities

```
summary(KN_data)
```

```
##    nace_r2              geo             TIME_PERIOD            values
##  Length:53          Length:53          Min.   :2022-01-01   Min.   :  2.50
##  Class :character   Class :character   1st Qu.:2022-01-01   1st Qu.: 12.60
##  Mode  :character   Mode  :character   Median :2022-01-01   Median : 25.30
##                                        Mean   :2022-01-01   Mean   : 62.85
##                                        3rd Qu.:2022-01-01   3rd Qu.: 51.90
##                                        Max.   :2022-01-01   Max.   :888.30
```

## OU GROUP

- O - Administrative and Support Service Activities
- P - Public Administration and Defence; Compulsory Social Security
- Q - Education
- R - Human Health and Social Work Activities
- S - Arts, Sports and Recreation
- T - Other Service Activities
- U - Activities of Households as Employers; Undifferentiated Goods and Services Producing Activities of Households for Own Use

```
summary(OU_data)
```

```
##    nace_r2              geo             TIME_PERIOD            values
##  Length:53          Length:53          Min.   :2022-01-01   Min.   :   9.3
##  Class :character   Class :character   1st Qu.:2022-01-01   1st Qu.:  37.4
##  Mode  :character   Mode  :character   Median :2022-01-01   Median :  70.8
##                                        Mean   :2022-01-01   Mean   : 119.9
##                                        3rd Qu.:2022-01-01   3rd Qu.: 126.2
##                                        Max.   :2022-01-01   Max.   :1148.1
```

u všech těchto skupin si můžeme všimnout, že maximum je dost daleko od průměru, nejspíš se bude jednat o outliers v silných městech Španělska.

# Úloha 2

```r
cont_table <- xtabs(values ~ geo + nace_r2, data = spain_data)
addmargins(cont_table)
```

```
##                 nace_r2
## geo                 A-F     G-J     K-N     O-U     Sum
##   Canary Islands  122.7   341.0   122.2   288.8   874.7
##   ES111           114.7   142.4    71.1   143.7   471.9
##   ES112            43.6    39.7    12.6    37.4   133.3
##   ES113            32.6    37.2    10.4    38.2   118.4
##   ES114           111.4   117.5    48.6   109.0   386.5
##   ES120            93.3   122.5    55.6   126.2   397.6
##   ES130            59.2    69.8    28.8    74.5   232.3
##   ES211            55.3    37.1    21.1    51.1   164.6
##   ES212           103.5    92.0    43.9   122.8   362.2
##   ES213           116.5   146.4    82.6   160.9   506.4
##   ES220           108.2    74.5    35.8    93.2   311.7
##   ES230            47.8    35.5    15.5    41.5   140.3
##   ES241            39.4    26.7    10.5    28.8   105.4
##   ES242            20.2    16.2     4.9    17.5    58.8
##   ES243           127.5   124.1    66.0   138.2   455.8
##   ES300           458.1  1221.7   888.3  1148.1  3716.2
##   ES411            16.8    15.2     5.3    21.9    59.2
##   ES412            55.7    41.0    17.6    46.8   161.1
##   ES413            41.0    52.4    20.8    61.1   175.3
##   ES414            23.8    16.2     7.1    20.5    67.6
##   ES415            32.9    41.2    14.9    46.6   135.6
##   ES416            23.6    19.7     6.4    19.9    69.6
##   ES417            16.6    10.1     3.5    12.3    42.5
##   ES418            61.3    60.0    34.4    76.3   232.0
##   ES419            20.2    16.5     5.8    21.1    63.6
##   ES421            49.4    45.4    13.5    48.4   156.7
##   ES422            57.4    49.3    15.7    56.7   179.1
##   ES423            30.7    25.9     6.3    22.0    84.9
##   ES424            21.9    23.9    14.9    26.6    87.3
##   ES425            81.0    68.6    25.3    78.4   253.3
##   ES431            65.1    68.8    29.5    88.5   251.9
##   ES432            39.0    37.3    15.0    54.8   146.1
##   ES511           547.3   873.4   557.5   802.0  2780.2
##   ES512            96.7   132.4    45.9   100.9   375.9
##   ES513            70.0    64.3    21.2    58.6   214.1
##   ES514            91.4   102.8    45.7    99.4   339.3
##   ES521           164.6   261.4    98.4   210.5   734.9
##   ES522            76.0    83.4    29.2    65.5   254.1
##   ES523           257.0   352.1   172.9   319.9  1101.9
##   ES531            12.6    34.4     8.7    15.5    71.2
##   ES532            71.4   152.3    66.2   129.7   419.6
##   ES533             8.1    12.6     3.7     9.3    33.7
##   ES611           106.4   105.7    28.3    70.8   311.2
##   ES612            77.0   136.4    51.9   145.2   410.5
##   ES613            85.1    70.6    30.6    96.7   283.0
##   ES614            66.7   110.2    44.3   112.5   333.7
##   ES615            61.7    52.8    18.6    54.3   187.4
##   ES616            70.5    55.3    20.8    72.4   219.0
```

```
##   ES617            111.4   241.2   117.0   205.4   675.0
##   ES618            155.9   218.4   126.7   253.9   754.9
##   ES620            189.9   204.7    83.4   174.0   652.0
##   ES630              2.6     7.7     2.5    18.6    31.4
##   ES640              2.2     7.4     3.5    19.4    32.5
##   Sum             4614.9  6515.3  3330.9  6356.3 20817.4
```

```r
employment_matrix <- tapply(spain_data$values, list(spain_data$geo, spain_data$nace_r2), sum, na.rm = TR

# Heatmapa tabulky
ggplot(spain_data, aes(x = nace_r2, y = geo, fill = values)) +
   geom_tile(color = "black", size = 0.5) +  # Add borders for visibility
  scale_fill_gradient(low = "white", high = "red") +
  geom_text(aes(label = round(values, 0)), size = 4, color = "black") +  # Add values inside cells
  theme_minimal() +
  theme(
    axis.text.x = element_text(hjust = 1),  # Rotate X labels
    axis.text.y = element_text(size = 12),  # Make Y labels bigger
    legend.position = "bottom",  # Move legend for better space usage
    panel.grid = element_blank()  # Remove grid lines for cleaner look
  ) +
  labs(
    title = "Employment Distribution Heatmap",
    x = "NACE Industry",
    y = "Region",
    fill = "Employment"
  )
```

Employment Distribution Heatmap

| Region | A–F | G–J | K–N | O–U |
|---|---|---|---|---|
| ES640 | 2 | 7 | 4 | 19 |
| ES630 | 3 | 8 | 2 | 19 |
| ES620 | 190 | 205 | 83 | 174 |
| ES618 | 156 | 218 | 127 | 254 |
| ES617 | 111 | 241 | 117 | 205 |
| ES616 | 70 | 55 | 21 | 72 |
| ES615 | 62 | 53 | 19 | 54 |
| ES614 | 67 | 110 | 44 | 112 |
| ES613 | 85 | 71 | 31 | 97 |
| ES612 | 77 | 136 | 52 | 145 |
| ES611 | 106 | 106 | 28 | 71 |
| ES533 | 8 | 13 | 4 | 9 |
| ES532 | 71 | 152 | 66 | 130 |
| ES531 | 13 | 34 | 9 | 16 |
| ES523 | 257 | 352 | 173 | 320 |
| ES522 | 76 | 83 | 29 | 66 |
| ES521 | 165 | 261 | 98 | 210 |
| ES514 | 91 | 103 | 46 | 99 |
| ES513 | 70 | 64 | 21 | 59 |
| ES512 | 97 | 132 | 46 | 101 |
| ES511 | 547 | 873 | 558 | 802 |
| ES432 | 39 | 37 | 15 | 55 |
| ES431 | 65 | 69 | 30 | 88 |
| ES425 | 81 | 69 | 25 | 78 |
| ES424 | 22 | 24 | 15 | 27 |
| ES423 | 31 | 26 | 6 | 22 |
| ES422 | 57 | 49 | 16 | 57 |
| ES421 | 49 | 45 | 14 | 48 |
| ES419 | 20 | 16 | 6 | 21 |
| ES418 | 61 | 60 | 34 | 76 |
| ES417 | 17 | 10 | 4 | 12 |
| ES416 | 24 | 20 | 6 | 20 |
| ES415 | 33 | 41 | 15 | 47 |
| ES414 | 24 | 16 | 7 | 20 |
| ES413 | 41 | 52 | 21 | 61 |
| ES412 | 56 | 41 | 18 | 47 |
| ES411 | 17 | 15 | 5 | 22 |
| ES300 | 458 | 1222 | 888 | 1148 |
| ES243 | 128 | 124 | 66 | 138 |
| ES242 | 20 | 16 | 5 | 18 |
| ES241 | 39 | 27 | 10 | 29 |
| ES230 | 48 | 36 | 16 | 42 |
| ES220 | 108 | 74 | 36 | 93 |
| ES213 | 116 | 146 | 83 | 161 |
| ES212 | 104 | 92 | 44 | 123 |
| ES211 | 55 | 37 | 21 | 51 |
| ES130 | 59 | 70 | 29 | 74 |
| ES120 | 93 | 122 | 56 | 126 |
| ES114 | 111 | 118 | 49 | 109 |
| ES113 | 33 | 37 | 10 | 38 |
| ES112 | 44 | 40 | 13 | 37 |
| ES111 | 115 | 142 | 71 | 144 |
| Canary Islands | 123 | 341 | 122 | 289 |

NACE Industry

Employment
300 600 900 1200

Zaměstnanost není rovnoměrně rozložena mezi regiony NUTS 3. Městské oblasti, jako Madrid a Barcelona, mají vysokou koncentraci zaměstnanosti v sektorech financí, IT a odborných služeb (K-N). Naopak venkovské regiony mají vyšší podíl zaměstnanosti v zemědělství a výrobě (A-F).

### Test hypotézy

$H_0$ : Zaměstnanost napříč všemi kategoriemy NACE je **nezávislá** na NUTS3.

$H_A$ : Zaměstnanost napříč všemi kategoriemy NACE je **závislá** na NUTS3.

```
chi_test <- chisq.test(cont_table)

chi_test
```

```
##
##  Pearson's Chi-squared test
##
## data:  cont_table
## X-squared = 908.44, df = 156, p-value < 2.2e-16
```

Na standardní hladině významnosti 5% **zamítáme** $H_0$ (*p-value* je menší než 0.05). To znamená, že existuje statisticky významný vztah mezi regionem a odvětvím v zaměstnanosti.

## Uloha 3

### Self-employed/Employees v letech 2017 a 2022

#### Hypotéza:

$H_0$ : $\mu_z = 0$ - Střední hodnota rozdílu poměru self/empl v letech 2022 a 2017 **je** rovna nule.

$H_A$ : $\mu_z \neq 0$ - Střední hodnota rozdílu poměru self/empl v letech 2022 a 2017 **není** rovna nule.

Zkoumám zdali se poměr self-employed a employees změnil z roku 2017 na 2022.(před covidem a "po" covidu)

```
spain_data_total_geo_emp_22 <- employment_data[grepl("^ES[1-9][0-9][0-9]$",employment_data$geo)& employm

spain_data_total_geo_self_22 <- employment_data[grepl("^ES[1-9][0-9][0-9]$",employment_data$geo)& employ

spain_data_total_geo_emp_17 <- employment_data[grepl("^ES[1-9][0-9][0-9]$",employment_data$geo)& employm

spain_data_total_geo_self_17 <- employment_data[grepl("^ES[1-9][0-9][0-9]$",employment_data$geo)& employ

spain_data_total_geo_emp_22_t <- employment_data[employment_data$geo == "ES"& employment_data$nace_r2 ==

spain_data_total_geo_self_22_t <- employment_data[employment_data$geo == "ES"& employment_data$nace_r2 =

spain_data_total_geo_emp_17_t <- employment_data[employment_data$geo == "ES"& employment_data$nace_r2 ==

spain_data_total_geo_self_17_t <- employment_data[employment_data$geo == "ES"& employment_data$nace_r2 =

spain_ratio_df_22 <- data.frame(
```

```
  geo = spain_data_total_geo_emp_22$geo,  # Assuming both have the same order of regions
  employed = spain_data_total_geo_emp_22$values,
  self_employed = spain_data_total_geo_self_22$values
)
spain_ratio_df_17 <- data.frame(
  geo = spain_data_total_geo_emp_17$geo,  # Assuming both have the same order of regions
  employed = spain_data_total_geo_emp_17$values,
  self_employed = spain_data_total_geo_self_17$values
)

spain_ratio_df_22$ratio <-  spain_ratio_df_22$self_employed /spain_ratio_df_22$employed


# Compute the ratio of self-employed to employed
spain_ratio_df_17$ratio <-  spain_ratio_df_17$self_employed /spain_ratio_df_17$employed
```

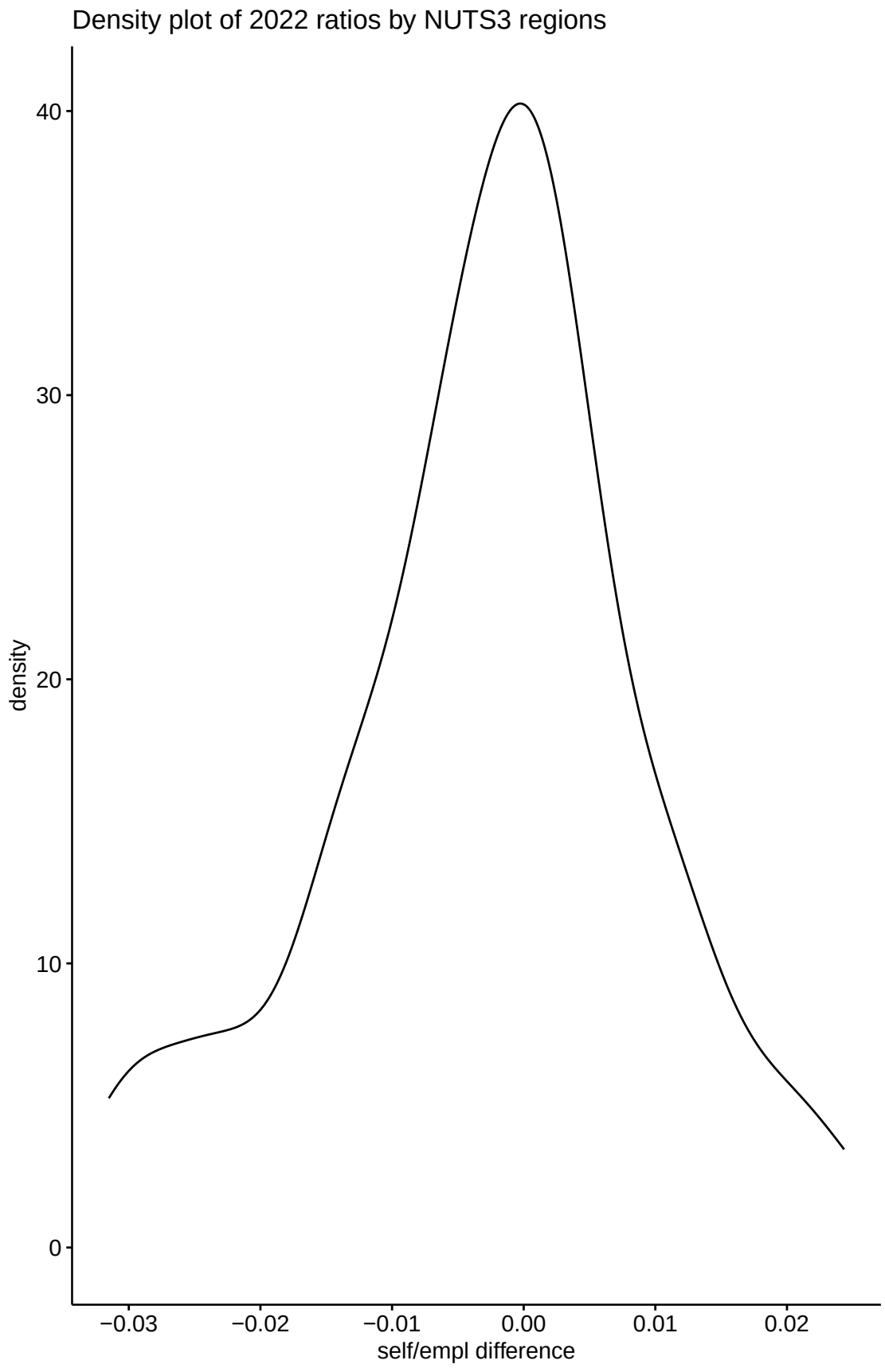
### Test normality

```
shapiro.test(spain_ratio_df_22$ratio - spain_ratio_df_17$ratio)
```

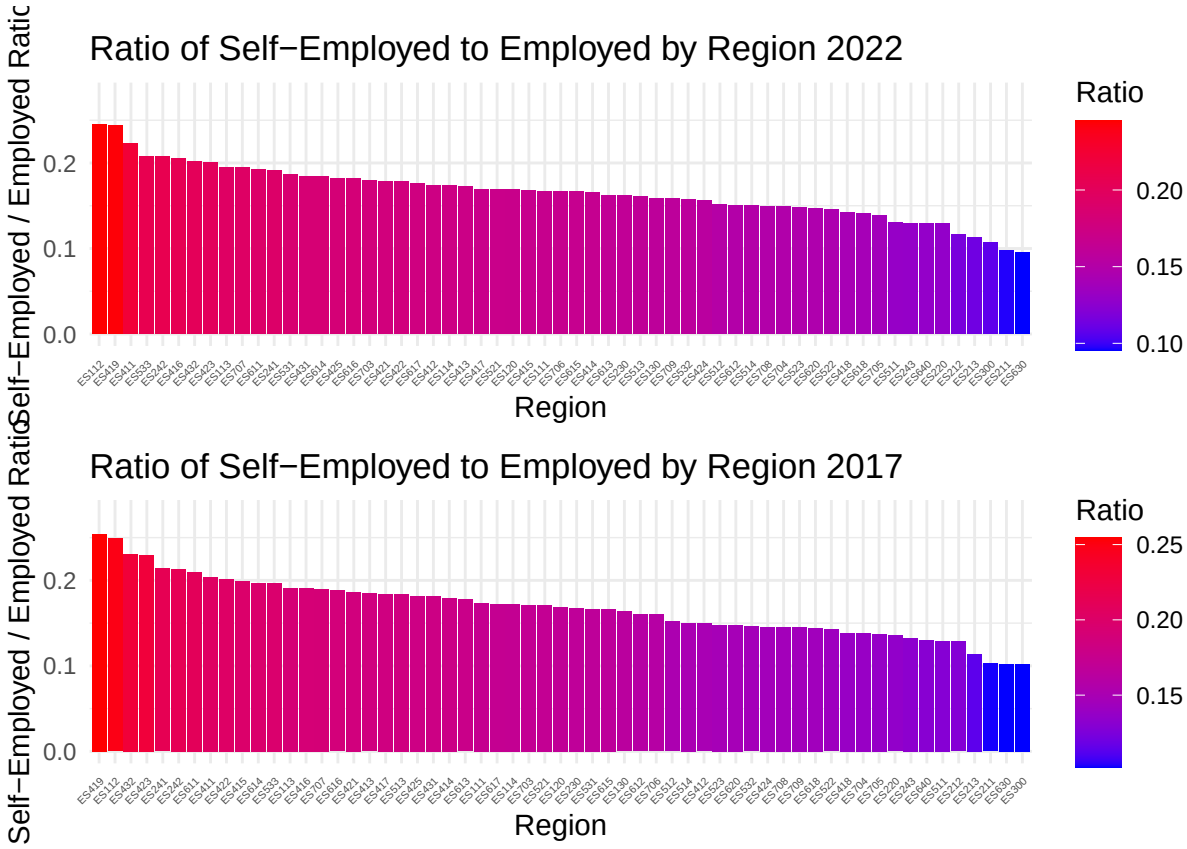
```
##
##  Shapiro-Wilk normality test
##
## data:  spain_ratio_df_22$ratio - spain_ratio_df_17$ratio
## W = 0.978, p-value = 0.361
```

Na standardní hladině významnosti 5% $H_0$ **nezamítáme** (*p-value* je větší než 0.05) a můžeme nyní využít testy, kde je předpokladem **normalní rozdělení** dat.

```
library("ggpubr")

ggdensity(spain_ratio_df_22$ratio - spain_ratio_df_17$ratio,
          main = "Density plot of 2022 ratios by NUTS3 regions",
          xlab = "self/empl difference")
```


Density plot of 2022 ratios by NUTS3 regions
density
40
30
20
10
0
−0.03
−0.02
−0.01
0.00
0.01
0.02
self/empl difference

```r
pr1 <- ggplot(spain_ratio_df_22, aes(x = reorder(geo, -ratio), y = ratio, fill = ratio)) +
  geom_bar(stat = "identity") +
  scale_fill_gradient(low = "blue", high = "red") +  # Color gradient from low to high
  theme_minimal() + ylim(0,0.28)+
  theme(axis.text.x = element_text(angle = 45, hjust = 1,size=4)) +  # Rotate x labels
  labs(title = "Ratio of Self-Employed to Employed by Region 2022",
       x = "Region",
       y = "Self-Employed / Employed Ratio",
       fill = "Ratio")

pr2 <- ggplot(spain_ratio_df_17, aes(x = reorder(geo, -ratio), y = ratio, fill = ratio)) +
  geom_bar(stat = "identity") + ylim(0,0.28)+
  scale_fill_gradient(low = "blue", high = "red") +  # Color gradient from low to high
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1,size=4)) +  # Rotate x labels
  labs(title = "Ratio of Self-Employed to Employed by Region 2017",
       x = "Region",
       y = "Self-Employed / Employed Ratio",
       fill = "Ratio")

combined_plot <- pr1 + pr2 + plot_layout(nrow = 2,heights = c(1, 1))
combined_plot
```



### Párový t-test

párový test jsem si vybral, jelikož mám párové data, která jsou normálně rozdělena s kladným rozptylem a jsou navzájem závislá.

```
t.test(spain_ratio_df_22$ratio, spain_ratio_df_17$ratio, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  spain_ratio_df_22$ratio and spain_ratio_df_17$ratio
## t = -1.7951, df = 58, p-value = 0.07785
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -0.0059086862  0.0003215907
## sample estimates:
## mean difference
##    -0.002793548
```

#### STATISTICKÁ INTERPRETACE

- p-value = 0.09761

Na hladině významnosti 5% **nezamítáme** $H_0$ (*p-value* je větší, než 0,05). To znamená, že na základě našich dat nemáme dostatečné důkazy k zamítnutí nulové hypotézy a nemůžeme tvrdit, že existuje statisticky významný rozdíl mezi průměrnými hodnotami poměru (self-employment ratio) pro rok 2022 a 2017.

#### Motivace

Motivací pro tuto hypotézu, mě zajímalo zda nějaká část populace změnila ze zaměstnaneckého poměru (či přímo začala) do self-emplyoed poměru v době před covidem a "po" covidu. Jelikož častým trendem na internetu bylo, že lidi začínají pracovat sami na sebe. Ukazuje se, že tomu tak nejspíš ve Španělsku nebylo, ale chtělo by to toto téma více rozebrat. Například porovnání jenom self-employed a jejich změny mezi lety.

## 2. Distribuce G-J employees je stejna jak ve Španělsku tak v Česku

Předpokládáme, že oba výběry pocházejí ze spojitých rozdělení s distribučními funkcemi **F** a **G** .

#### Hypotéza:

$H_0 : F = G$ - Rozdělení počtu zaměstnanců ve skupině G–J v Česku a ve Španělsku **jsou** stejné.

$H_A : F \neq G$ - Rozdělení počtu zaměstnanců ve skupině G–J v Česku a ve Španělsku **nejsou** stejné.

```
czech_2022_GJ <- employment_data[grepl("^CZ[0-9][0-9][0-9]$",employment_data$geo)&  employment_data$nace
```

```
shapiro.test(GJ_data$values)
```

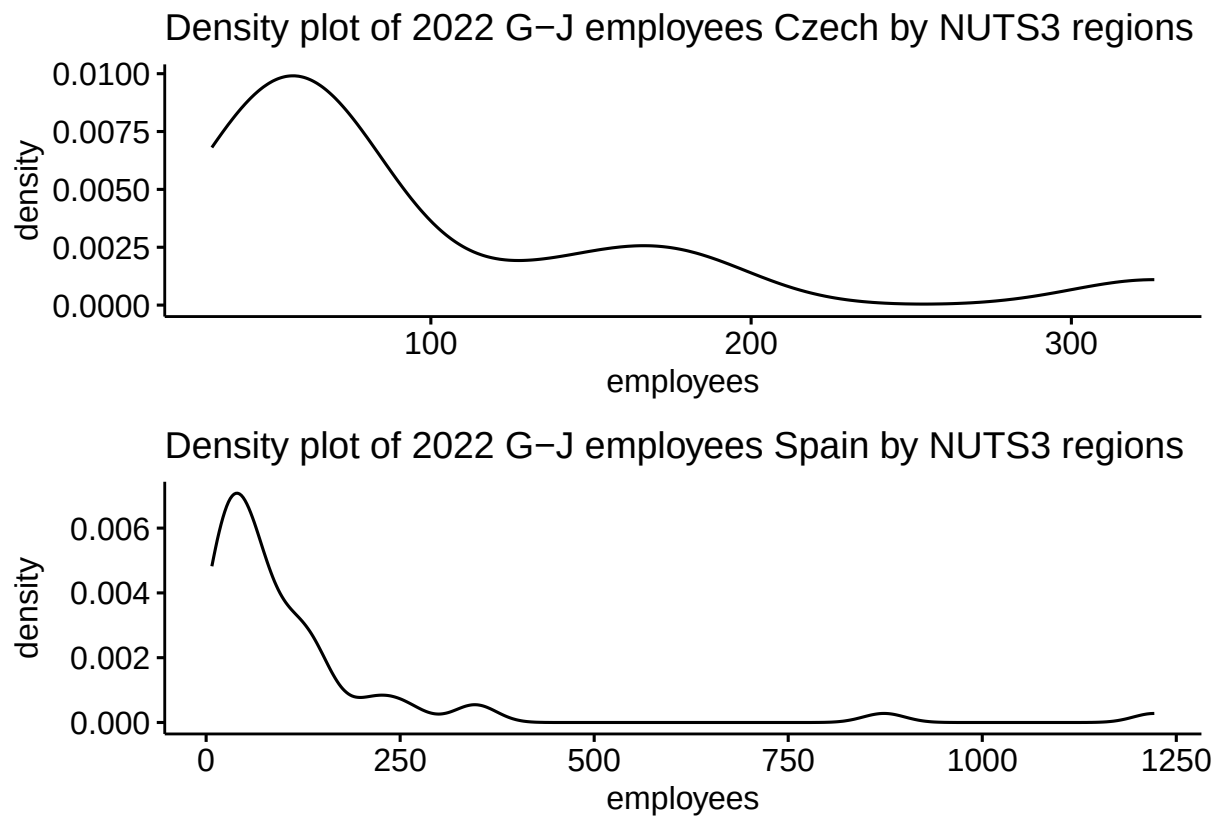
```
##
##  Shapiro-Wilk normality test
##
## data:  GJ_data$values
## W = 0.50122, p-value = 3.682e-12
```

```r
shapiro.test(czech_2022_GJ$values)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  czech_2022_GJ$values
## W = 0.72906, p-value = 0.0007515
```

Na standardní hladině významnosti 5% $H_0$ **zamítáme** (*p-value* je menší než 0.05) ve prospěch alternativní.

```r
af_dens <- ggdensity(GJ_data$values,
          main = "Density plot of 2022 G-J employees Spain by NUTS3 regions",
          xlab = "employees")

cz_dens <- ggdensity(czech_2022_GJ$values,
          main = "Density plot of 2022 G-J employees Czech by NUTS3 regions",
          xlab = "employees")

comb_dens <- cz_dens + af_dens + plot_layout(nrow = 2,heights = c(1, 1))
comb_dens
```




jak je i z vizualizace patrné, data nemají *normální rozdělení.*

### Kolmogorovovův-Smirnovův dvouvýběrový test

Vybral jsem Kolmogorovovův-Smirnovův dvouvýběrový test na základě této informace z přednášky.

Jde tedy o test shody distribucí. Mannův-Whitneyův test, též dvouvýběrový Wilcoxonův test, je neparametrický test. Je poněkud citlivý na posunutí, tj. případ $G(x)=F(x-\delta), \delta > 0, G(x)=F(x-\delta), \delta > 0$ větší odchylky v tvaru či rozptylu. V tom případě je lepší zvolit např. Kolmogorovovův-Smirnovův (KS) test.

```
ks.test(GJ_data$values, czech_2022_GJ$values)
```

```
##
##  Exact two-sample Kolmogorov-Smirnov test
##
## data:  GJ_data$values and czech_2022_GJ$values
## D = 0.26819, p-value = 0.3315
## alternative hypothesis: two-sided
```

#### STATISTICKÁ INTERPRETACE

- **Statistika testu (D)**: 0.268
- **p-hodnota**: 0.335
- **Hladina významnosti**: 0.05

Jelikož je **p-hodnota = 0.335 > 0.05**, tudíž nemáme dostatek důkazů k zamítnutí nulové hypotézy.

Na hladině významnosti 5 % nemůžeme říci, že by se rozdělení počtu zaměstnanců ve skupině G–J v roce 2022 statisticky významně lišilo mezi Českem a Španělskem. Jinými slovy, **nelze vyloučit, že pocházejí ze stejného rozdělení**.

## 3. Nezaměstnanost ES x EU27

**Hypotéza:**

$H_0$ : Střední hodnota míry nezaměstnanosti ve španělských regionech (NUTS2) v roce 2023 **je rovna** váženému průměru míry nezaměstnanosti v zemích EU27.

$H_A$ : Střední hodnota míry nezaměstnanosti ve španělských regionech **se liší** od váženého průměru EU27.

```
unemployment_rate <- get_eurostat("une_rt_a")
unemployment_rate_nuts2 <- get_eurostat("tgs00010")
```

```
unemployed_spain_nuts2 <- unemployment_rate_nuts2[grepl("^ES[1-9][0-9]$",unemployment_rate_nuts2$geo) &

unemployed_spain_nuts2 <- unemployed_spain_nuts2 %>% group_by(geo) %>% summarise(mean_unemp = mean(valu

unemployment_rate <- unemployment_rate[unemployment_rate$TIME_PERIOD == "2023-01-01" & unemployment_rat

eu27_countries <- c(
  "AT", "BE", "BG", "HR", "CY", "CZ", "DK", "EE", "FI", "FR",
  "DE", "GR", "HU", "IE", "IT", "LV", "LT", "LU", "MT", "NL",
  "PL", "PT", "RO", "SK", "SI", "SE"
)
```

```r
merged_df <- left_join(unemployment_rate,populatio_data,by = "geo")
unemployment_data_filtered <- merged_df %>%
  filter(geo %in% eu27_countries)
```

```r
weighted_mean <- unemployment_data_filtered %>%
  summarise(w_mean = sum(values.x * values.y, na.rm = TRUE) / sum(values.y, na.rm = TRUE)) %>%
  pull(w_mean)

# Print the result
print(weighted_mean)
```

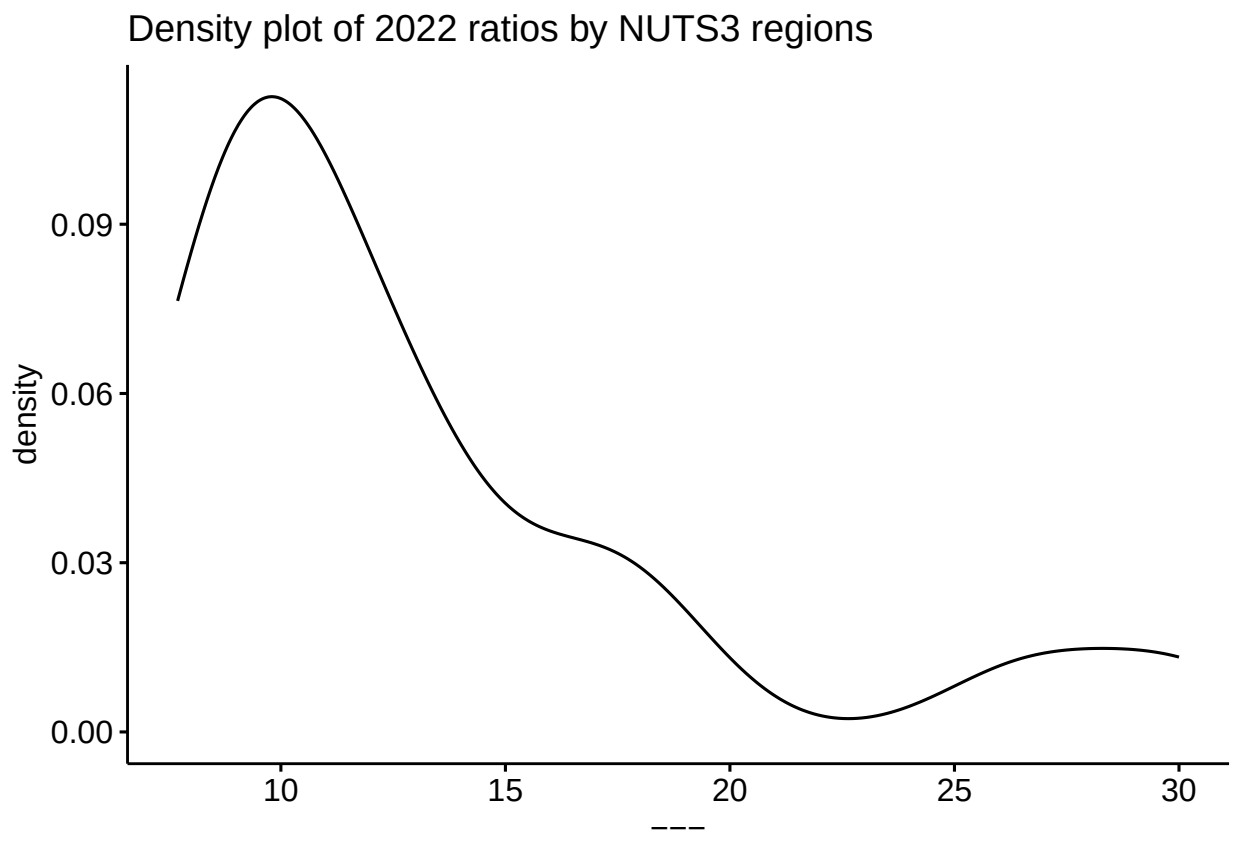
```
## [1] 5.259726
```

Vážený průměr nezaměstnanosti evropských státu je 5.25 %.

```r
shapiro.test(unemployed_spain_nuts2$mean_unemp)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  unemployed_spain_nuts2$mean_unemp
## W = 0.77731, p-value = 0.0005449
```

```r
ggdensity(unemployed_spain_nuts2$mean_unemp,
          main = "Density plot of 2022 ratios by NUTS3 regions",
          xlab = "---")
```




**Wilcoxův test** test jsem si vybral, jelikož rozdělení mých dat není normální.

```
wilcox.test(unemployed_spain_nuts2$mean_unemp, mu = weighted_mean)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  unemployed_spain_nuts2$mean_unemp
## V = 190, p-value = 0.0001426
## alternative hypothesis: true location is not equal to 5.259726
```

- **Testová statistika**: $\mathrm{V} = 190$
- **p-hodnota**: 0.0001426
- **Hladina významnosti**: 5 %

**Statistická Interpretace**:

p-hodnota je výrazně menší než 0.05 → **zamítáme nulovou hypotézu** $H_o$ na standardní hladině významnosti.

To znamená, že **medián míry nezaměstnanosti ve španělských regionech se statisticky významně liší od váženého průměru nezaměstnanosti v EU27**.