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

Semestrální projekt 3

František Kareš, Šimon Kubeš

```
library(eurostat)
library(ggplot2)
library(lmtest)
library(car)
library(rnaturalearth)
library(rnaturalearthdata)
library(giscoR)
library(sf)
library(dplyr)
```

```
K <- 6
L <- 5
M <- (((K + L) * 47) %% 11) + 1
sprintf("M = %s", M)
```

```
## [1] "M = 1"
```

M je v našem případě 1, proto budeme uvažovat data z roku 2013.

Příprava dat

```
data <- get_eurostat("nama_10_fte")
head(data)
```

```
## # A tibble: 6 x 5
##   freq  unit  geo   TIME_PERIOD values
##   <chr> <chr> <chr> <date>       <dbl>
## 1 A     EUR   AT    1995-01-01   26543
## 2 A     EUR   AT    1996-01-01   26522
## 3 A     EUR   AT    1997-01-01   25943
## 4 A     EUR   AT    1998-01-01   26863
## 5 A     EUR   AT    1999-01-01   27714
## 6 A     EUR   AT    2000-01-01   28415
```

```
unique(data$freq)
```

```
## [1] "A"
```

```r
unique(data$unit)
```

```
## [1] "EUR" "NAC"
```

```r
unique(data$geo)
```

```
##  [1] "AT"        "BE"        "BG"        "CY"        "CZ"        "DE"
##  [7] "DK"        "EA20"      "EE"        "EL"        "ES"        "EU27_2020"
## [13] "FI"        "FR"        "HR"        "HU"        "IE"        "IT"
## [19] "LT"        "LU"        "LV"        "MT"        "PL"        "PT"
## [25] "RO"        "SE"        "SI"        "SK"
```

```r
unique(data$TIME_PERIOD)
```

```
##  [1] "1995-01-01" "1996-01-01" "1997-01-01" "1998-01-01" "1999-01-01"
##  [6] "2000-01-01" "2001-01-01" "2002-01-01" "2003-01-01" "2004-01-01"
## [11] "2005-01-01" "2006-01-01" "2007-01-01" "2008-01-01" "2009-01-01"
## [16] "2010-01-01" "2011-01-01" "2012-01-01" "2013-01-01" "2014-01-01"
## [21] "2015-01-01" "2016-01-01" "2017-01-01" "2018-01-01" "2019-01-01"
## [26] "2020-01-01" "2021-01-01" "2022-01-01" "2023-01-01" "2024-01-01"
```

V datasetu jsou záznamy o průměrném platu pro období jednoho roku. Vyfiltrujeme si záznamy z roku 2013, kde jednotka je euro (druhá možnost je national currency, to se nehodí pro porovnávání zemí mezi sebou).

```r
df <- data[data$TIME_PERIOD == "2013-01-01" & data$unit == "EUR",]
head(df)
```

```
## # A tibble: 6 x 5
##   freq  unit  geo   TIME_PERIOD values
##   <chr> <chr> <chr> <date>       <dbl>
## 1 A     EUR   AT    2013-01-01   40037
## 2 A     EUR   BE    2013-01-01   43435
## 3 A     EUR   BG    2013-01-01    5704
## 4 A     EUR   CY    2013-01-01   21594
## 5 A     EUR   CZ    2013-01-01   12260
## 6 A     EUR   DE    2013-01-01   37739
```

Zbavíme se nepotřebných sloupců a pak odebereme záznamy, kde *geo* je EA20 (země, které používají euro) nebo EU27_2020 (členské země Evropské unie).

```r
df <- df[ ! df$geo %in% c("EA20", "EU27_2020"), c("geo", "values")]
df
```

```
## # A tibble: 26 x 2
##    geo   values
##    <chr>  <dbl>
##  1 AT     40037
##  2 BE     43435
##  3 BG      5704
##  4 CY     21594
```

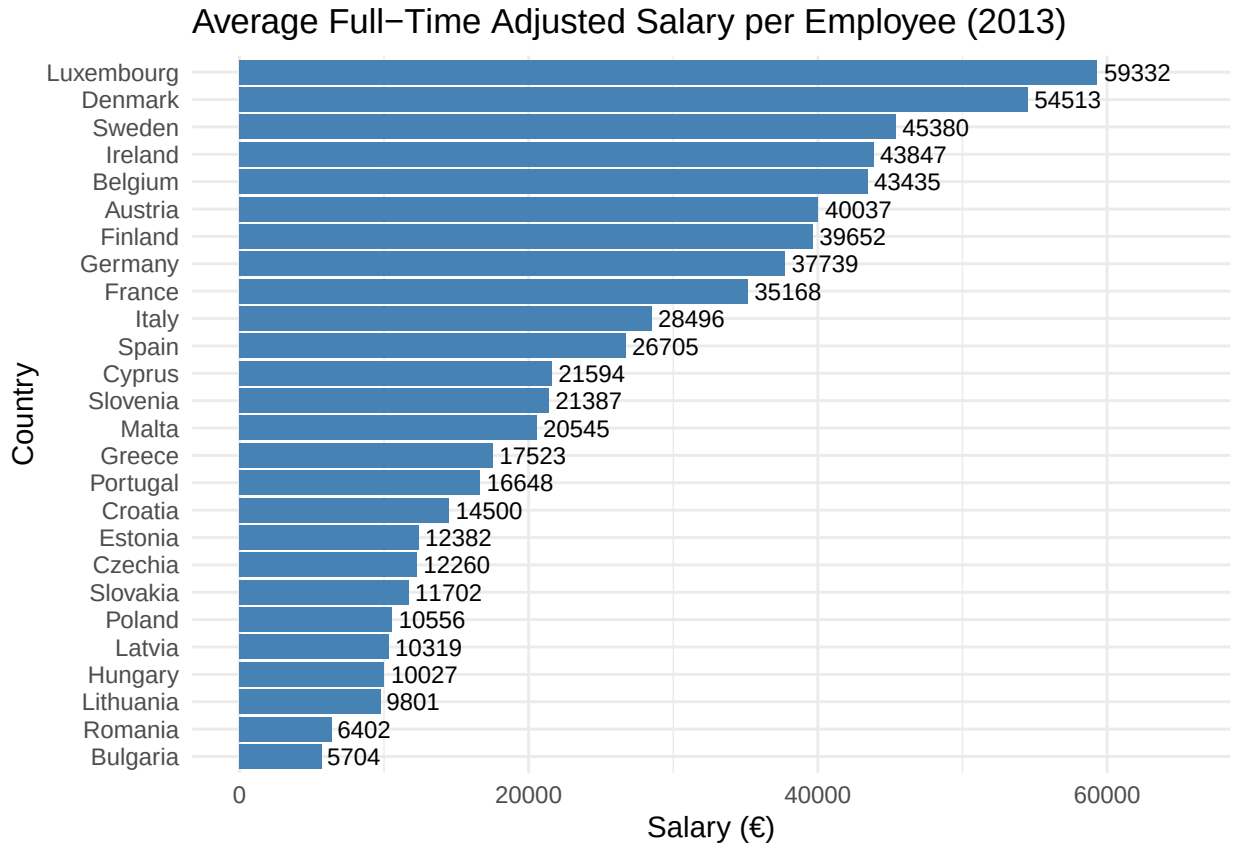
```
##  5 CZ     12260
##  6 DE     37739
##  7 DK     54513
##  8 EE     12382
##  9 EL     17523
## 10 ES     26705
## # i 16 more rows
```

Základní statistická šetření

```r
summary(df$values)
```

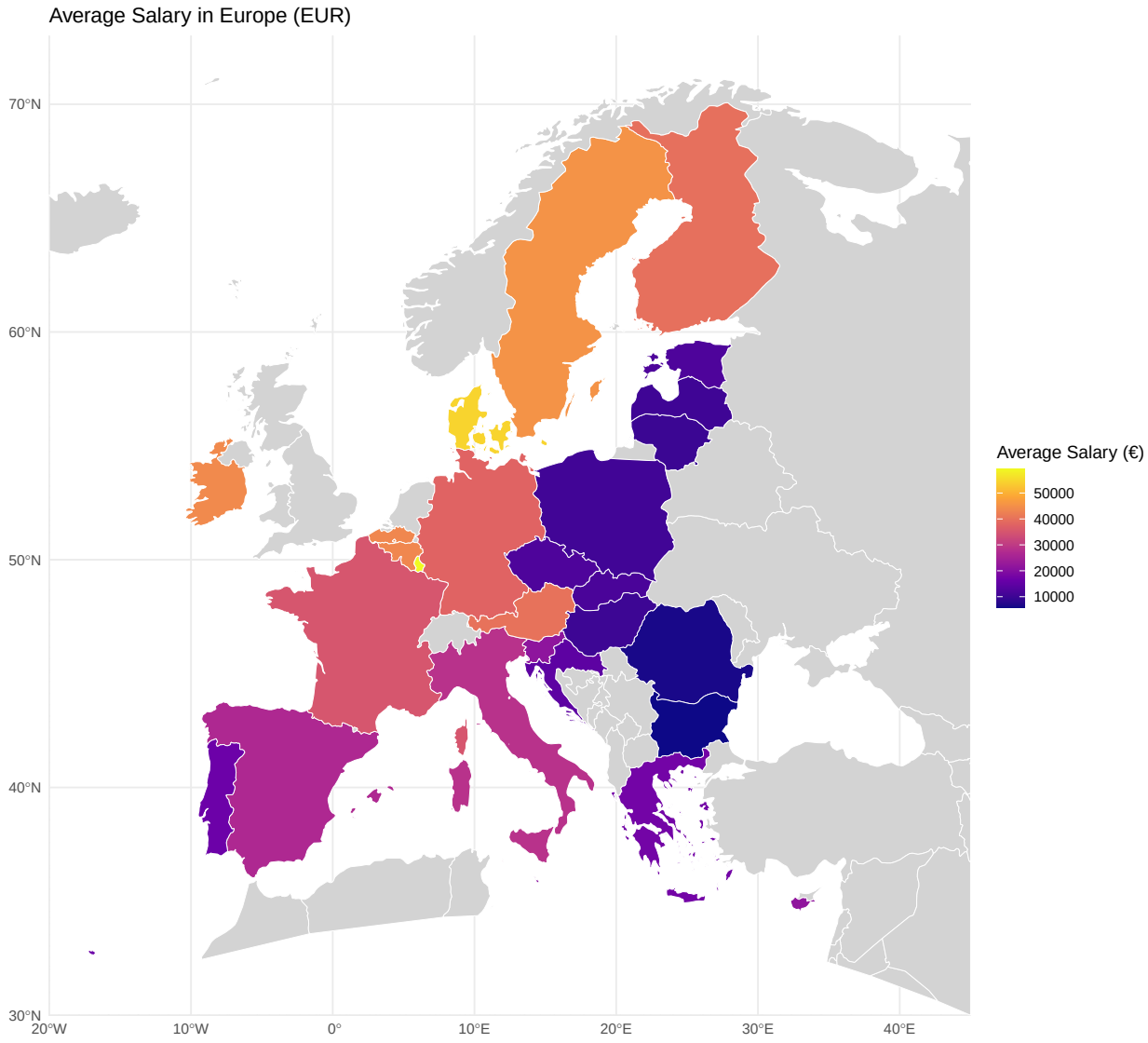
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5704   11842   20966   25217   39174   59332
```

```r
df_labeled <- df
df_labeled$geo <- label_eurostat(df$geo, dic = "geo")
ggplot(df_labeled, aes(x = reorder(geo, values), y = values)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  geom_text(aes(label = round(values, 0)), hjust = -0.1, size = 3) +
  coord_flip() +

  labs(
    title = "Average Full-Time Adjusted Salary per Employee (2013)",
    x = "Country",
    y = "Salary (€)"
  ) +
  theme_minimal() +
  expand_limits(y = max(df_labeled$values) * 1.1)
```



Nejmenší mzda byla v Bulharsku, pak v Rumunsku a v Litvě. Naopak největší mzda byla v roce 2013 v Lucembursku, Dánsku a Švédsku. Lucembursko by se vzhledem ke své malé rozloze a počtu obyvatel (663 430) dalo považovat spíše za outlier. I přesto je však patrný výrazný rozdíl mezi 1. a 3. kvartilem.

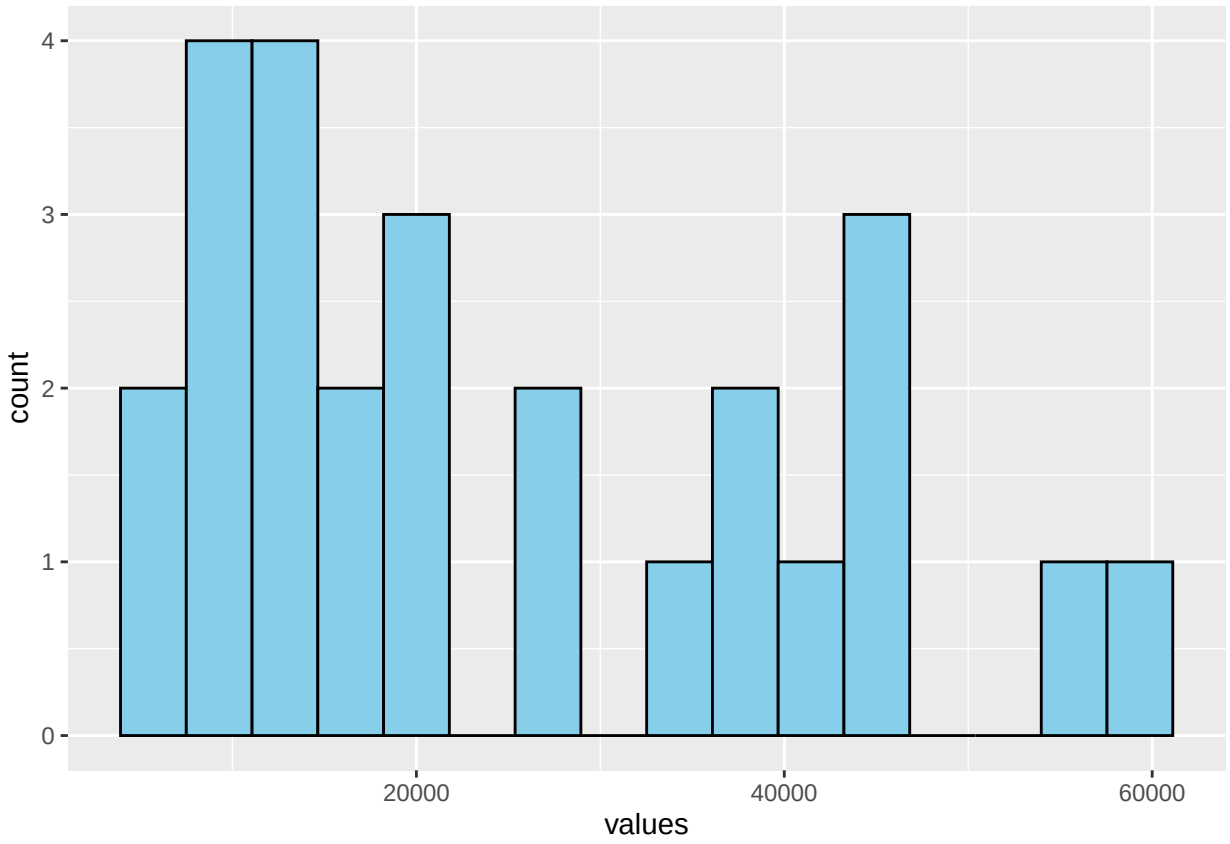
```
df$geo[df$geo == "EL"] <- "GR"
map_nuts0_base <- get_eurostat_geospatial(nuts_level = 0, resolution = "10", output_class = "sf",)

worldmap <- ne_countries(scale = 'medium', type = 'map_units',
                         returnclass = 'sf')
europe_cropped <- st_crop(worldmap, xmin = -20, xmax = 45,
                                    ymin = 30, ymax = 73)

map_nuts0 <- left_join(europe_cropped, df, by = c("iso_a2_eh" = "geo"))


ggplot(map_nuts0) +
  geom_sf(aes(fill = values), color = "white") +
  scale_fill_viridis_c(option = "plasma", na.value = "lightgrey") +
  labs(
    title = "Average Salary in Europe (EUR)",
    fill = "Average Salary (€)"
  ) +
  theme_minimal() +
  coord_sf(xlim = c(-20, 45), ylim = c(30, 73), expand = FALSE)
```




```
df$geo[df$geo == "GR"] <- "EL"
```

Z mapy je patrný rozdíl mezi východní Evropou a západní Evropou.

```r
ggplot(df, aes(x=values)) +
    geom_histogram(bins=16, colour="black", fill="skyblue")
```



Histogram naznačuje, že data nejspíše nebudou normálně rozdělena. Normalitu rovnou otestujeme.

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

```
##
##  Shapiro-Wilk normality test
##
## data:  df$values
## W = 0.9069, p-value = 0.0224
```

Zamítáme nulovou hypotézu ve prospěch alternativy (**p-value** < 0.05), data nejsou normálně rozdělena.

Volba regresorů

S průměrnou mzdou by mohlo souviset HDP, které slouží k měření ekonomického růstu. Nabízí se také vzdělání. Očekávali bychom, že lidé s vyšším vzděláním budou mít vyšší mzdy. Dále třeba sektor, ve kterém je nejvíce zaměstnanců dané země (zemědělství, služby, průmysl). My jsme si vybrali následující regresory:

První regresor - HDP na osobu

Prvním regresorem bude HDP na osobu v eurech z datasetu nama_10_pc.

```
tmp <- get_eurostat("nama_10_pc")
tmp <- tmp[tmp$TIME_PERIOD == "2013-01-01",]
tmp <- tmp[tmp$na_item == "B1GQ",]  # HDP v trznich cenach
tmp <- tmp[tmp$unit == "CP_EUR_HAB",]  # HDP v eurech
tmp$gdp <- tmp$values
tmp$values <- NULL
tmp
```

```
## # A tibble: 43 x 6
##    freq  unit       na_item geo   TIME_PERIOD   gdp
##    <chr> <chr>      <chr>   <chr> <date>      <dbl>
##  1 A     CP_EUR_HAB B1GQ    AL    2013-01-01   3330
##  2 A     CP_EUR_HAB B1GQ    AT    2013-01-01  37890
##  3 A     CP_EUR_HAB B1GQ    BE    2013-01-01  35360
##  4 A     CP_EUR_HAB B1GQ    BG    2013-01-01   5870
##  5 A     CP_EUR_HAB B1GQ    CH    2013-01-01  66900
##  6 A     CP_EUR_HAB B1GQ    CY    2013-01-01  20940
##  7 A     CP_EUR_HAB B1GQ    CZ    2013-01-01  15280
##  8 A     CP_EUR_HAB B1GQ    DE    2013-01-01  35600
##  9 A     CP_EUR_HAB B1GQ    DK    2013-01-01  46240
## 10 A     CP_EUR_HAB B1GQ    EA    2013-01-01  30020
## # i 33 more rows
```

```
df <- merge(df, tmp[,c("geo", "gdp")], by="geo", all.x=TRUE)
df
```

```
##    geo values   gdp
## 1   AT  40037 37890
## 2   BE  43435 35360
## 3   BG   5704  5870
## 4   CY  21594 20940
## 5   CZ  12260 15280
## 6   DE  37739 35600
## 7   DK  54513 46240
## 8   EE  12382 14520
## 9   EL  17523 16240
## 10  ES  26705 22020
## 11  FI  39652 37410
## 12  FR  35168 32280
## 13  HR  14500 10590
## 14  HU  10027 10350
## 15  IE  43847 39590
## 16  IT  28496 26880
## 17  LT   9801 11780
## 18  LU  59332 90030
## 19  LV  10319 10930
## 20  MT  20545 19120
## 21  PL  10556 10260
## 22  PT  16648 16300
## 23  RO   6402  7150
## 24  SE  45380 45820
## 25  SI  21387 17500
## 26  SK  11702 13790
```

Vztah HDP a průměrného platu

Protože jsme **zamítnuli** normalitu rozdělení průměrných mezd, použijeme Spearmanův korelační koeficient k ověření vztahu mezi HDP a průměrným platem.

```
cor(df$values, df$gdp, method="spearman")
```

```
## [1] 0.9753846
```

Korelační koeficient je 0.975, veličiny **jsou** silně **kladně korelované**.

```
summary(df$gdp)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5870   12282   18310   24990   35540   90030
```

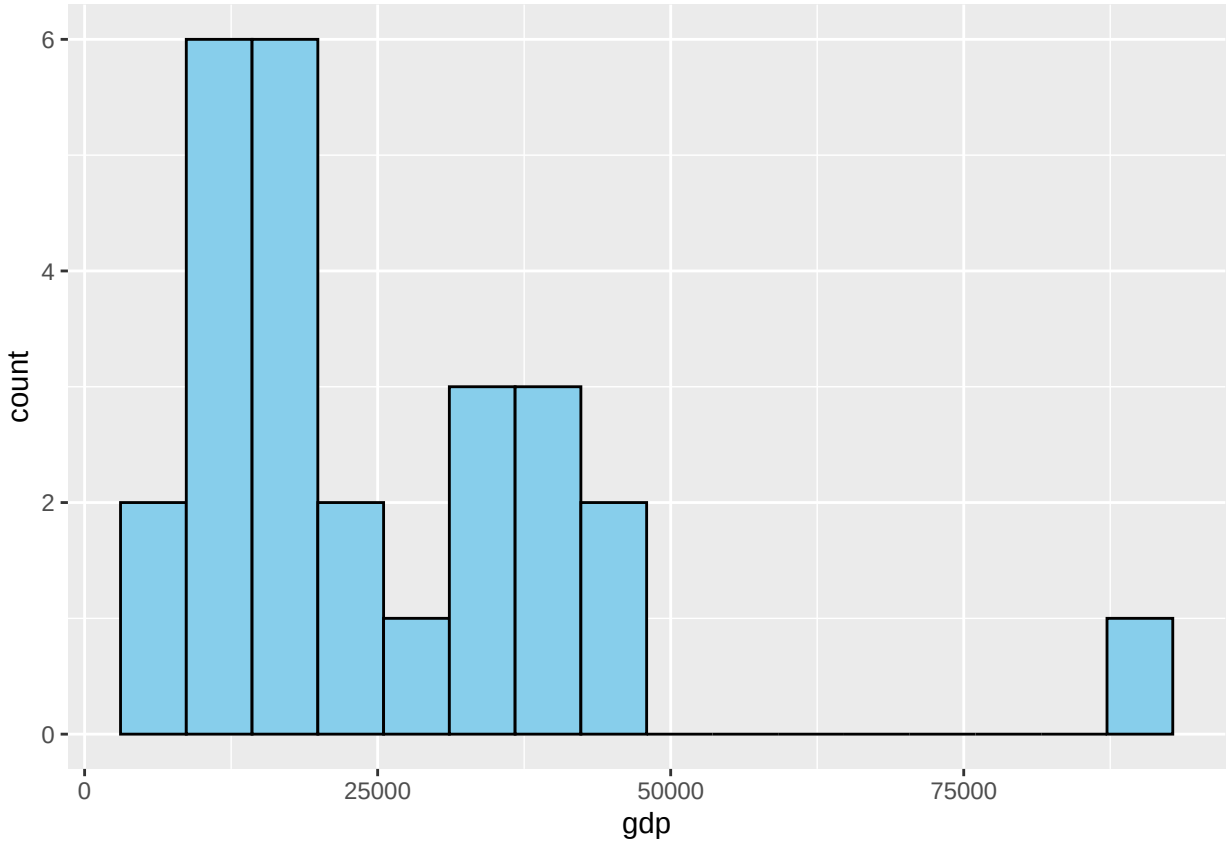
Průměrné HDP na obyvatele je 24988 EUR, ale medián 18310 EUR, maximum je 90030 EUR a minimum 5870 EUR.

```
rbind(
    head(df[order(-df$gdp),], 3),
    tail(df[order(-df$gdp),], 3)
)
```

```
##    geo values   gdp
## 18  LU  59332 90030
## 7   DK  54513 46240
## 24  SE  45380 45820
## 21  PL  10556 10260
## 23  RO   6402  7150
## 3   BG   5704  5870
```

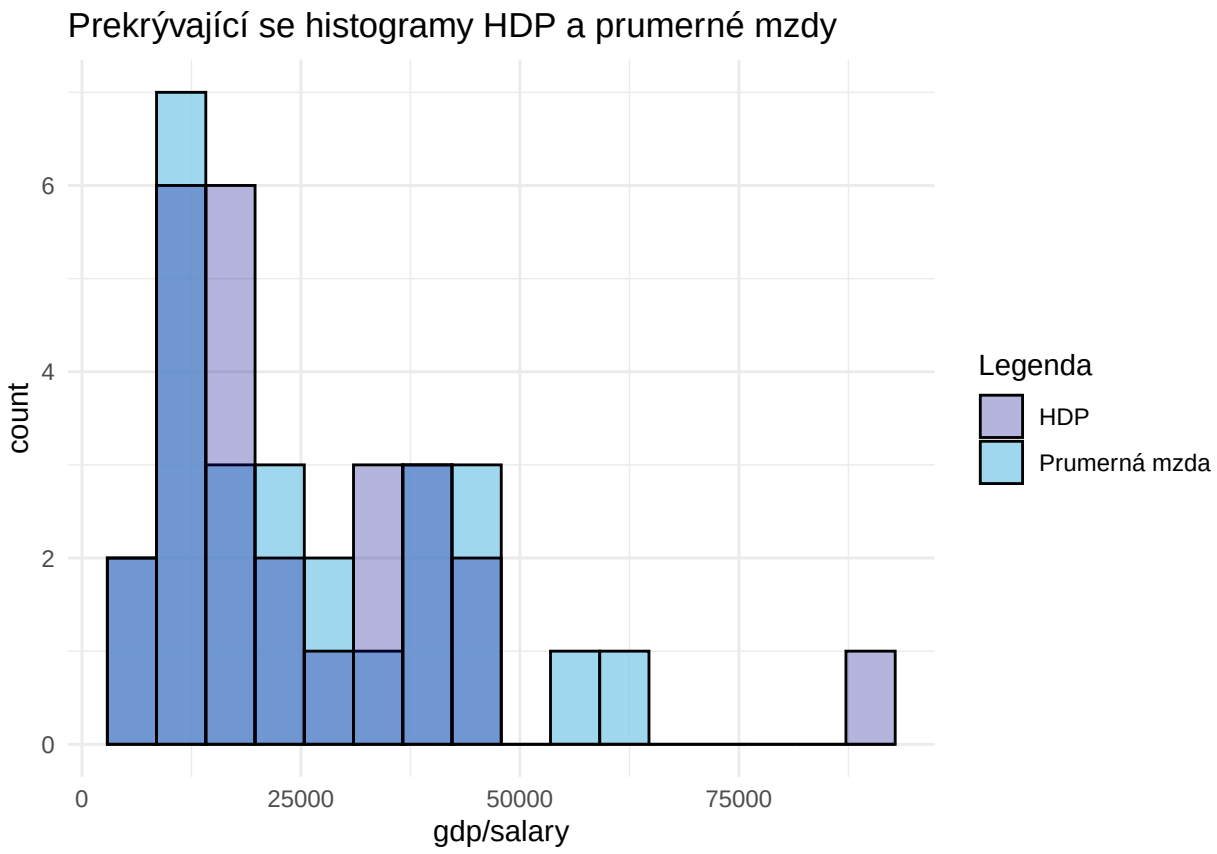
Největší HDP na obyvatele bylo v roce 2013 v Lucembursku, Dánsku a Švédsku nejmenší naopak v Bulharsku, Rumunsku a Polsku. To je velmi podobné jako u průměrné mzdy. Ještě se podívejme na histogram.

```
ggplot(df, aes(x=gdp)) +
    geom_histogram(bins=16, colour="black", fill="skyblue")
```



Většina zemí měla v roce 2013 HDP na obyvatele menší než 25000 EUR, nějaké země měly mezi 25000 a 50000 EUR a pak můžeme vidět odlehlé Lucembursko s 90030 EUR.

```
ggplot() +
  geom_histogram(data = df, aes(x = values, fill = "Průměrná mzda"),
                 bins = 16, colour = "black", alpha = 0.8) +

  geom_histogram(data = df, aes(x = gdp, fill = "HDP"),
                 bins = 16, colour = "black", alpha = 0.3) +

  scale_fill_manual(values = c("Průměrná mzda" = "skyblue", "HDP" = "darkblue")) +
  labs(title = "Překrývající se histogramy HDP a průměrné mzdy",
       x = "gdp/salary", y = "count", fill = "Legenda") +
  theme_minimal()
```



Z grafů je patrné, že rozdělení obou veličin jsou podobná a téměř se překrývají, což odpovídá vysoké hodnotě Spearmanova korelačního koeficientu.

Druhý regresor - míra nezaměstnanosti

Druhý regresor bude míra nezaměstnanosti lidí ve věku 15-74 let v procentech z datasetu lfsa_urgan.

```
tmp <- get_eurostat("lfsa_urgan")
tmp <- tmp[tmp$TIME_PERIOD == "2013-01-01",]
tmp <- tmp[tmp$sex == "T",]  # Total
tmp <- tmp[tmp$age == "Y15-74",]
tmp <- tmp[tmp$citizen == "TOTAL",]
tmp$unempl <- tmp$values
tmp$values <- NULL
tmp
```

```
## # A tibble: 37 x 8
##    freq  unit  sex   age    citizen geo   TIME_PERIOD unempl
##    <chr> <chr> <chr> <chr>  <chr>   <chr> <date>       <dbl>
##  1 A     PC    T     Y15-74 TOTAL   AT    2013-01-01     5.4
##  2 A     PC    T     Y15-74 TOTAL   BE    2013-01-01     8.4
##  3 A     PC    T     Y15-74 TOTAL   BG    2013-01-01    13
##  4 A     PC    T     Y15-74 TOTAL   CH    2013-01-01     4.8
##  5 A     PC    T     Y15-74 TOTAL   CY    2013-01-01    15.9
##  6 A     PC    T     Y15-74 TOTAL   CZ    2013-01-01     7
```

```
##  7 A     PC    T     Y15-74 TOTAL   DE    2013-01-01     5.2
##  8 A     PC    T     Y15-74 TOTAL   DK    2013-01-01     7.4
##  9 A     PC    T     Y15-74 TOTAL   EA21  2013-01-01    12
## 10 A     PC    T     Y15-74 TOTAL   EE    2013-01-01     8.6
## # i 27 more rows
```

```
df <- merge(df, tmp[,c("geo", "unempl")], by="geo", all.x=TRUE)
df
```

```
##    geo values   gdp unempl
## 1   AT  40037 37890    5.4
## 2   BE  43435 35360    8.4
## 3   BG   5704  5870   13.0
## 4   CY  21594 20940   15.9
## 5   CZ  12260 15280    7.0
## 6   DE  37739 35600    5.2
## 7   DK  54513 46240    7.4
## 8   EE  12382 14520    8.6
## 9   EL  17523 16240   27.5
## 10  ES  26705 22020   26.1
## 11  FI  39652 37410    8.2
## 12  FR  35168 32280    9.9
## 13  HR  14500 10590   17.3
## 14  HU  10027 10350   10.2
## 15  IE  43847 39590   13.8
## 16  IT  28496 26880   12.2
## 17  LT   9801 11780   11.8
## 18  LU  59332 90030    5.9
## 19  LV  10319 10930   11.9
## 20  MT  20545 19120    6.1
## 21  PL  10556 10260   10.3
## 22  PT  16648 16300   16.5
## 23  RO   6402  7150    7.1
## 24  SE  45380 45820    8.1
## 25  SI  21387 17500   10.1
## 26  SK  11702 13790   14.2
```

Vztah míry nezaměstnanosti a průměrného platu

```
cor(df$values, df$unempl, method="spearman")
```

```
## [1] -0.3319658
```

U nezaměstnanosti je korelační koeficient s průměrnou mzdou **záporný**, to dává smysl. Čím více nezaměstnanosti, tím nižší platy.

```
summary(df$unempl)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   5.200   7.575  10.150  11.465  13.600  27.500
```

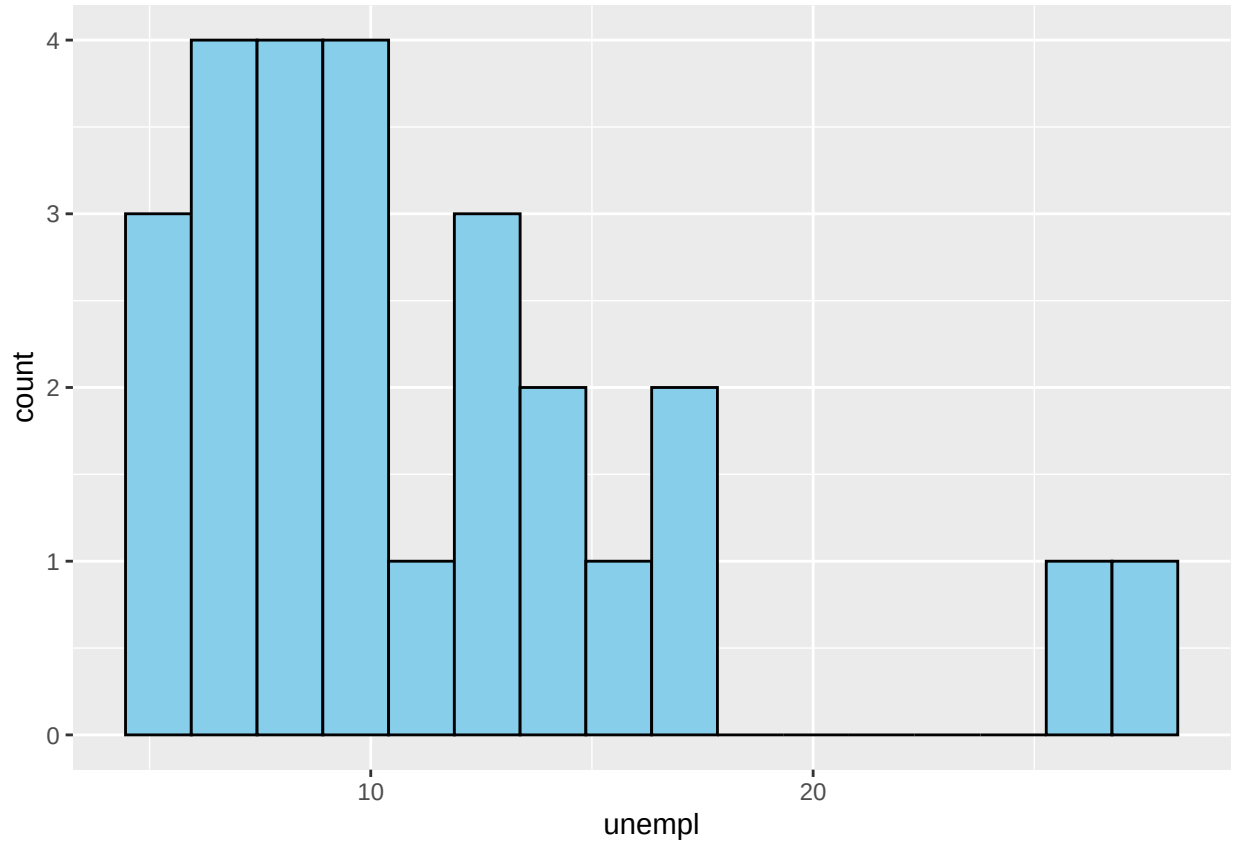
Průměrná nezaměstnanost je 11.465 %, medián je 10.15 %, minimum 5.2 % a maximum 27.5 %.

```
rbind(
    head(df[order(-df$unempl),], 3),
    tail(df[order(-df$unempl),], 3)
)
```

```
##    geo values   gdp unempl
## 9   EL  17523 16240   27.5
## 10  ES  26705 22020   26.1
## 13  HR  14500 10590   17.3
## 18  LU  59332 90030    5.9
## 1   AT  40037 37890    5.4
## 6   DE  37739 35600    5.2
```

Největší nezaměstnanost byla v roce 2013 v Řecku, nejspíše kvůli dopadům způsobeným dluhovou krizí, kde její vrchol byl v letech 2011-2012. Nejmenší naopak v Německu, Rakousku a Lucembursku.

```
ggplot(df, aes(x=unempl)) +
    geom_histogram(bins=16, colour="black", fill="skyblue")
```



Třetí regresor - míra populace s terciárním vzděláním

Třetí regresor bude míra populace ve věku 25-64 let s terciárním vzděláním (VOŠ, bakalářské, magisterské, doktorské studium VŠ) v procentech z datasetu edat_lfse_03.

```
tmp <- get_eurostat("edat_lfse_03")
tmp <- tmp[tmp$TIME_PERIOD == "2013-01-01",]
tmp <- tmp[tmp$sex == "T",]
tmp <- tmp[tmp$age == "Y25-64",]
tmp <- tmp[tmp$isced11 == "ED5-8",]
tmp$edu <- tmp$values
tmp$values <- NULL
tmp
```

```
## # A tibble: 38 x 8
##    freq  sex   age    unit  isced11 geo   TIME_PERIOD   edu
##    <chr> <chr> <chr>  <chr> <chr>   <chr> <date>      <dbl>
##  1 A     T     Y25-64 PC    ED5-8   AT    2013-01-01   20.6
##  2 A     T     Y25-64 PC    ED5-8   BE    2013-01-01   35.5
##  3 A     T     Y25-64 PC    ED5-8   BG    2013-01-01   25.6
##  4 A     T     Y25-64 PC    ED5-8   CH    2013-01-01   37.4
##  5 A     T     Y25-64 PC    ED5-8   CY    2013-01-01   39.3
##  6 A     T     Y25-64 PC    ED5-8   CZ    2013-01-01   20.5
##  7 A     T     Y25-64 PC    ED5-8   DE    2013-01-01   28.6
##  8 A     T     Y25-64 PC    ED5-8   DK    2013-01-01   35.2
##  9 A     T     Y25-64 PC    ED5-8   EA20  2013-01-01   27.9
## 10 A     T     Y25-64 PC    ED5-8   EA21  2013-01-01   27.9
## # i 28 more rows
```

```
df <- merge(df, tmp[,c("geo", "edu")], by="geo", all.x=TRUE)
df
```

```
##    geo values   gdp unempl  edu
## 1   AT  40037 37890    5.4 20.6
## 2   BE  43435 35360    8.4 35.5
## 3   BG   5704  5870   13.0 25.6
## 4   CY  21594 20940   15.9 39.3
## 5   CZ  12260 15280    7.0 20.5
## 6   DE  37739 35600    5.2 28.6
## 7   DK  54513 46240    7.4 35.2
## 8   EE  12382 14520    8.6 37.4
## 9   EL  17523 16240   27.5 27.4
## 10  ES  26705 22020   26.1 33.7
## 11  FI  39652 37410    8.2 40.5
## 12  FR  35168 32280    9.9 32.1
## 13  HR  14500 10590   17.3 20.9
## 14  HU  10027 10350   10.2 22.6
## 15  IE  43847 39590   13.8 42.6
## 16  IT  28496 26880   12.2 16.4
## 17  LT   9801 11780   11.8 35.2
## 18  LU  59332 90030    5.9 40.7
## 19  LV  10319 10930   11.9 31.0
## 20  MT  20545 19120    6.1 19.6
## 21  PL  10556 10260   10.3 25.8
## 22  PT  16648 16300   16.5 19.3
## 23  RO   6402  7150    7.1 15.6
## 24  SE  45380 45820    8.1 37.0
## 25  SI  21387 17500   10.1 27.9
```

```
## 26  SK  11702 13790   14.2 19.9
```

Vztah míry populace s terciárním vzděláním a průměrného platu

```
cor(df$values, df$edu, method="spearman")
```

```
## [1] 0.5228244
```

Korelační koeficient průměrných mezd a procenta populace s terciárním vzděláním je **kladný**.

```
summary(df$edu)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   15.60   20.68   28.25   28.88   35.42   42.60
```

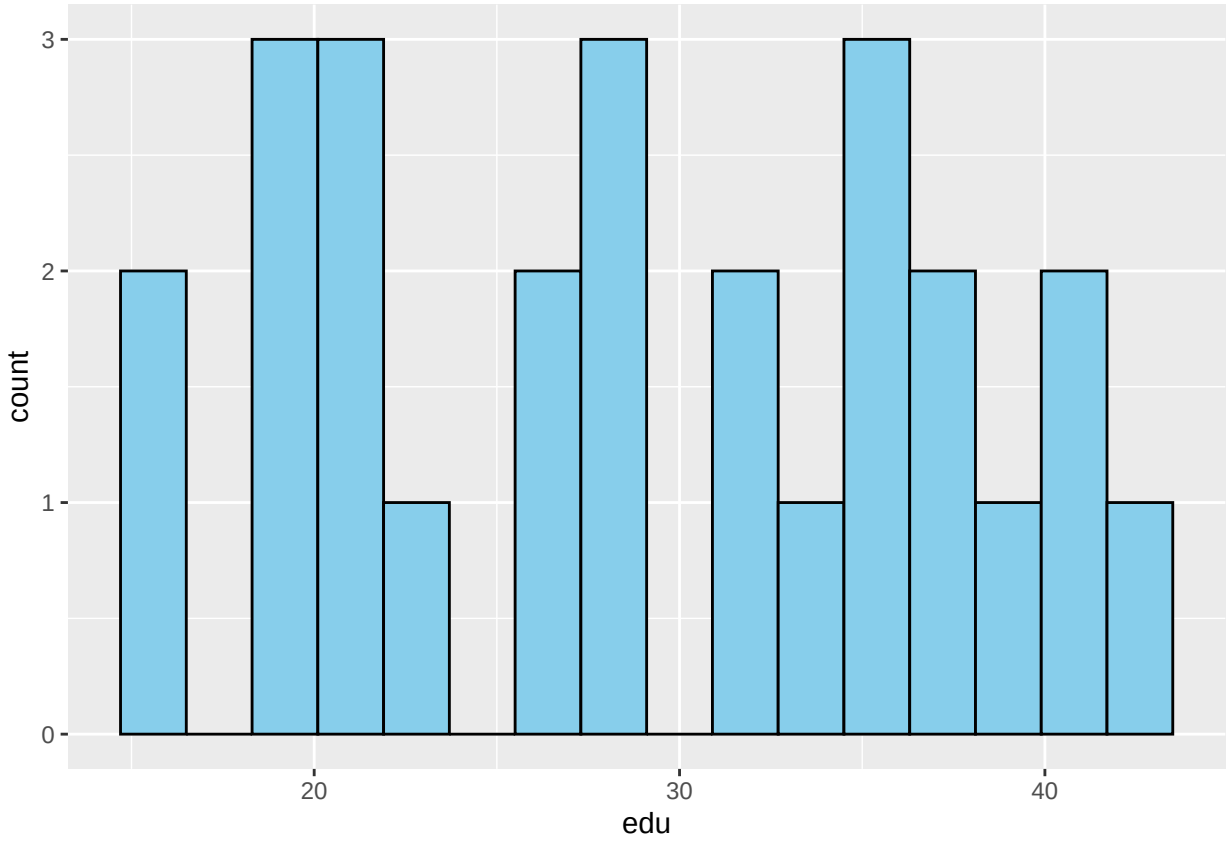
Průměrně má terciární vzdělání 28.88 % populace, medián je 28.25 %, minimum 15.6 % a maximum 42.6 %.

```
rbind(
    head(df[order(-df$edu),], 3),
    tail(df[order(-df$edu),], 3)
)
```

```
##    geo values   gdp unempl  edu
## 15  IE  43847 39590   13.8 42.6
## 18  LU  59332 90030    5.9 40.7
## 11  FI  39652 37410    8.2 40.5
## 22  PT  16648 16300   16.5 19.3
## 16  IT  28496 26880   12.2 16.4
## 23  RO   6402  7150    7.1 15.6
```

Největší podíl populace s terciárním vzděláním byl v roce 2013 v Irsku, Lucembursku a Finsku, nejmenší naopak v Rumunsku, Itálii a Portugalsku.

```
ggplot(df, aes(x=edu)) +
    geom_histogram(bins=16, colour="black", fill="skyblue")
```



Čtvrtý regresor - before/after (2000) EU membership

Čtvrtý regresor bude, zda se země stala členem EU před rokem 2000, nebo po něm.
Zdroj

```
eu_membership <- list()
eu_membership$AT <- "Before"
eu_membership$BE <- "Before"
eu_membership$BG <- "After"
eu_membership$CY <- "After"
eu_membership$CZ <- "After"
eu_membership$DE <- "Before"
eu_membership$DK <- "Before"
eu_membership$EE <- "After"
eu_membership$EL <- "Before"
eu_membership$ES <- "Before"
eu_membership$FI <- "Before"
eu_membership$FR <- "Before"
eu_membership$HR <- "After"
eu_membership$HU <- "After"
eu_membership$IE <- "Before"
eu_membership$IT <- "Before"
eu_membership$LT <- "After"
eu_membership$LU <- "Before"
eu_membership$LV <- "After"
```

```r
eu_membership$MT <- "After"
eu_membership$PL <- "After"
eu_membership$PT <- "Before"
eu_membership$RO <- "After"
eu_membership$SE <- "Before"
eu_membership$SI <- "After"
eu_membership$SK <- "After"
eu_membership <- unlist(eu_membership)
```

```r
df$eu_join <- eu_membership[df$geo]
df$eu_join <- as.factor(df$eu_join)
df
```

```
##    geo values   gdp unempl  edu eu_join
## 1   AT  40037 37890    5.4 20.6  Before
## 2   BE  43435 35360    8.4 35.5  Before
## 3   BG   5704  5870   13.0 25.6   After
## 4   CY  21594 20940   15.9 39.3   After
## 5   CZ  12260 15280    7.0 20.5   After
## 6   DE  37739 35600    5.2 28.6  Before
## 7   DK  54513 46240    7.4 35.2  Before
## 8   EE  12382 14520    8.6 37.4   After
## 9   EL  17523 16240   27.5 27.4  Before
## 10  ES  26705 22020   26.1 33.7  Before
## 11  FI  39652 37410    8.2 40.5  Before
## 12  FR  35168 32280    9.9 32.1  Before
## 13  HR  14500 10590   17.3 20.9   After
## 14  HU  10027 10350   10.2 22.6   After
## 15  IE  43847 39590   13.8 42.6  Before
## 16  IT  28496 26880   12.2 16.4  Before
## 17  LT   9801 11780   11.8 35.2   After
## 18  LU  59332 90030    5.9 40.7  Before
## 19  LV  10319 10930   11.9 31.0   After
## 20  MT  20545 19120    6.1 19.6   After
## 21  PL  10556 10260   10.3 25.8   After
## 22  PT  16648 16300   16.5 19.3  Before
## 23  RO   6402  7150    7.1 15.6   After
## 24  SE  45380 45820    8.1 37.0  Before
## 25  SI  21387 17500   10.1 27.9   After
## 26  SK  11702 13790   14.2 19.9   After
```

```r
summary(df$eu_join)
```

```
##  After Before
##     13     13
```

Ze zemí v datasetu se jich třináct stalo členem EU před rokem 2000 a třináct po něm.

Vztah roku připojení k EU a průměrného platu

```r
summary(df[df$eu_join == "Before",]$values)
```

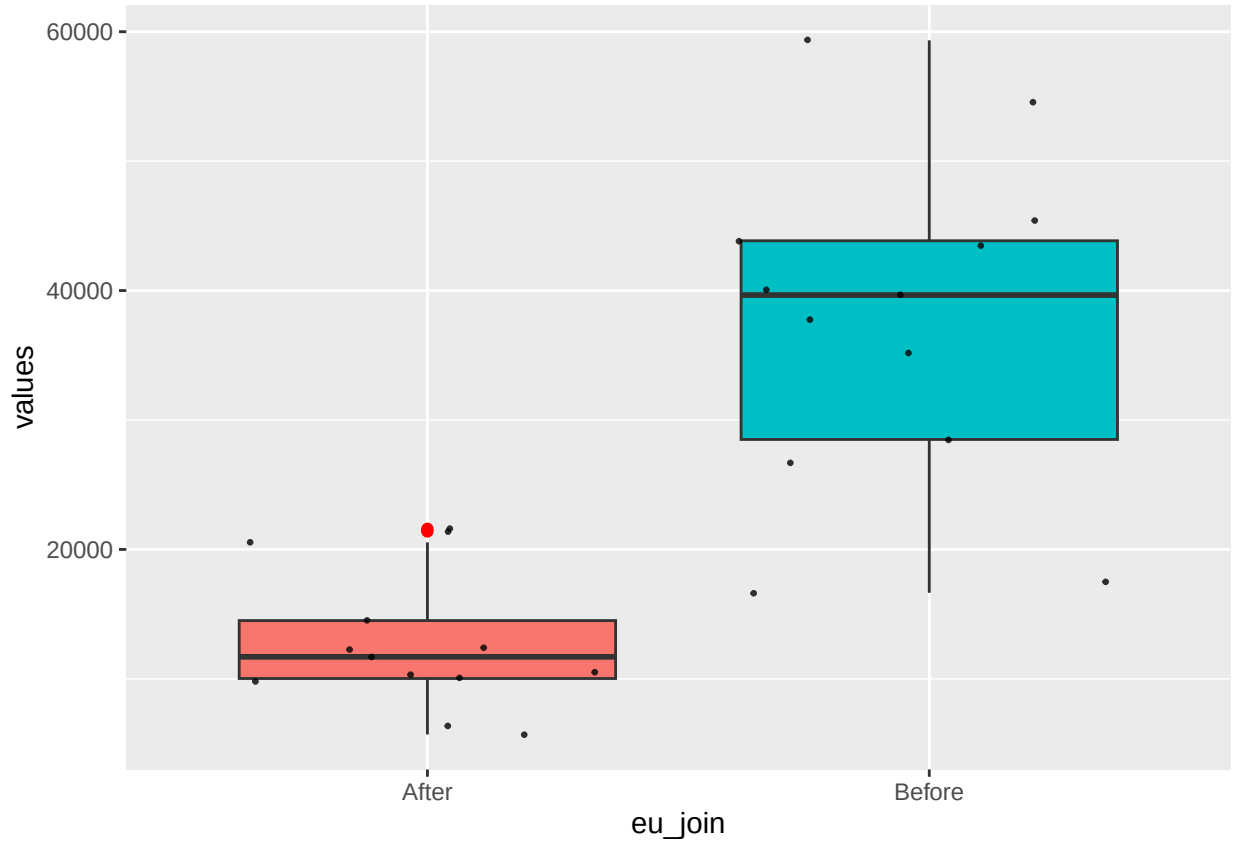
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   16648   28496   39652   37575   43847   59332
```

```r
summary(df[df$eu_join == "After",]$values)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5704   10027   11702   12860   14500   21594
```

Vidíme, že se hodnoty mezi skupinami dost liší. V roce 2013 byl průměrný plat v zemi, která se do EU přidala před rokem 2000, větší než maximální plat zemí, které se přidaly do EU po roce 2000.

```r
ggplot(df, aes(x=eu_join, y=values, fill=eu_join)) +
    geom_boxplot(outlier.color="red", show.legend=F) +
    geom_jitter(size=0.5, alpha=0.8, show.legend=F)
```



Na boxplotech je rozdíl opravdu patrný.

Ještě se podívejme na korelační matici a variance inflation factor.

```r
cor(df[,c(-1, -6)], method="spearman")
```

```
##            values        gdp      unempl         edu
## values  1.0000000  0.9753846 -0.33196581  0.52282442
## gdp     0.9753846  1.0000000 -0.38871795  0.52282442
## unempl -0.3319658 -0.3887179  1.00000000 -0.04787143
## edu     0.5228244  0.5228244 -0.04787143  1.00000000
```

Mezi žádnou dvojicí regresorů není příliš velká korelace.

```
vif(lm(values ~ gdp + unempl + edu + eu_join, data=df))
```

```
##      gdp   unempl      edu  eu_join
## 3.578598 1.547398 1.528866 2.406740
```

Variance inflation factor pro žádný regresor není větší než 5, nejsou na sobě závislé.

Lineární regresní model

```
model <- lm(values ~ gdp + unempl + edu + eu_join, data=df)
summary(model)
```

```
##
## Call:
## lm(formula = values ~ gdp + unempl + edu + eu_join, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -7306  -2670  -1155   2807   9864
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.825e+03  3.817e+03   1.264 0.220033
## gdp            4.545e-01  9.439e-02   4.815 9.29e-05 ***
## unempl        -4.291e+02  1.984e+02  -2.164 0.042176 *
## edu            2.627e+02  1.345e+02   1.952 0.064351 .
## eu_joinBefore  1.274e+04  2.752e+03   4.629 0.000145 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4523 on 21 degrees of freedom
## Multiple R-squared:  0.9313, Adjusted R-squared:  0.9183
## F-statistic: 71.21 on 4 and 21 DF,  p-value: 6.584e-12
```

Regresní model

$$\text{Salary} = 4835 + 0.4547 \cdot \text{gdp} - 428.5 \cdot \text{unempl} + 262.1 \cdot \text{edu} + \begin{cases} 0 & \text{pro eu-join} = After, \\ 12730 & \text{pro eu-join} = Before \end{cases}$$

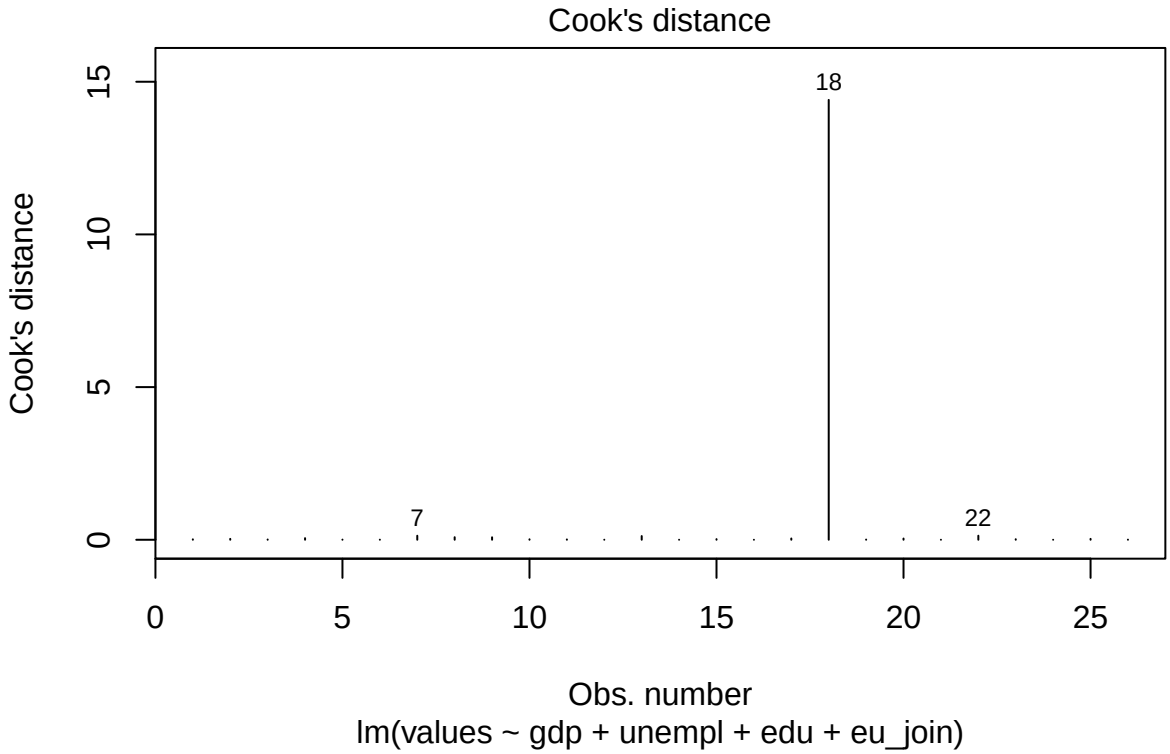
Vysvětlení koeficientů

- **Intercept (4835):** Hodnota průměrného platu, když jsou ostatní regresory nulové a *eu_join* je referenční kategorie (After).
- **Koeficient u gdp (0.4547):** S každým zvýšením HDP na obyvatele o jedno euro se očekává zvýšení průměrného platu o 0.4547 eura.
- **Koeficient u unempl (-428.5):** Tento záporný koeficient naznačuje, že s růstem míry nezaměstnanosti o jeden procentní bod, se očekává pokles průměrného platu o 428.5 eura.
- **Koeficient u edu (262.1):** S každým zvýšením podílu lidí ve věku 25–64 let s terciárním vzděláním o jeden procentní bod se očekává zvýšení průměrného platu o 262.1 eura.
- **Koeficient u eu_joinBefore (12730):** Pokud se země připojila k EU před rokem 2000, očekává se, že průměrný plat bude o 12 730 eur vyšší ve srovnání se zeměmi, které se připojily po roce 2000.

Koeficient determinace R^2 je 0.9311, model vysvětluje 93.11 % variability. Adjustovaný koeficient je 0.918, moc se neliší. Regresory *gdp*, *unempl* a *eu_joinBefore* jsou statisticky významné. Regresor *edu* a intercept nejsou statisticky významné.

Multikolinearitu už jsme zkoumali výše - mezi regresory není silná korelace. Pomocí Cookovy vzdálenosti identifikujeme odlehlá pozorování.

```
plot(model, which=4)
```



```
df[18,]
```

```
##    geo values   gdp unempl  edu eu_join
## 18  LU  59332 90030    5.9 40.7  Before
```

Bod s číslem 18 má velkou Cookovu vzdálenost, jde o Lucembursko, které má *gdp* téměř dvojnásobné oproti druhému největšímu Dánsku, jak jsme viděli výše.

Předpoklady

Otestujeme předpoklady, nejprve nezávislost a homoskedasticitu reziduí.

```
dwtest(model)
```

```
##
##  Durbin-Watson test
##
## data:  model
## DW = 2.0802, p-value = 0.5837
## alternative hypothesis: true autocorrelation is greater than 0
```

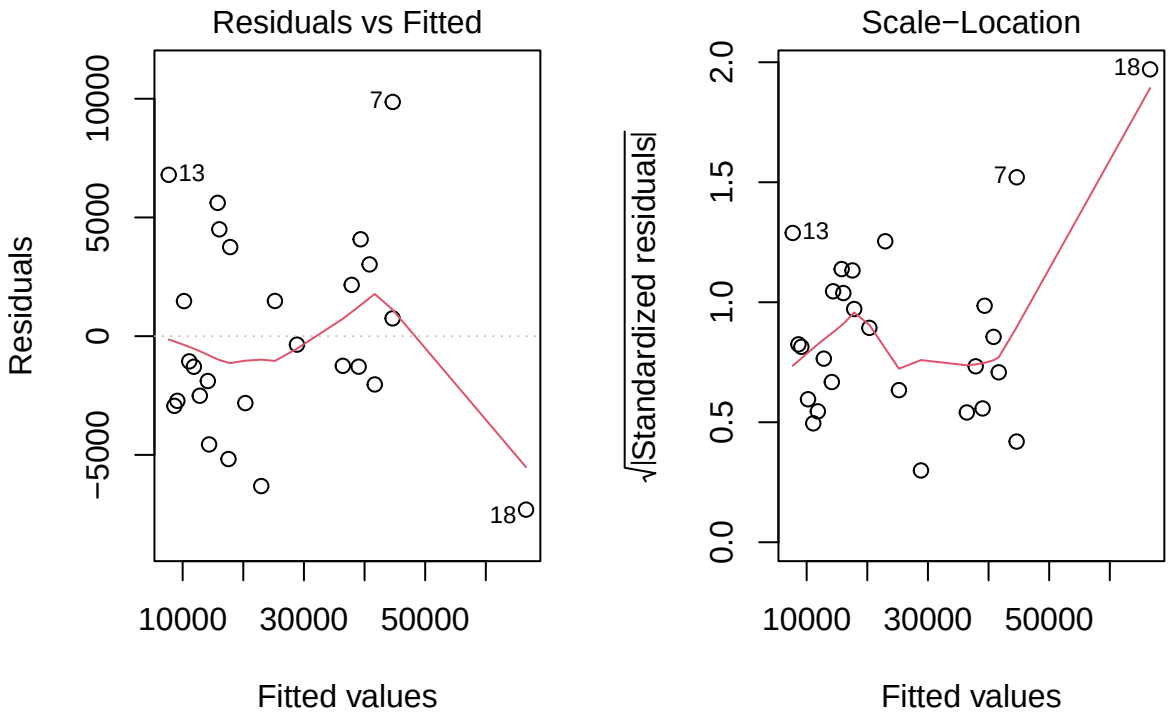
Jelikož **p-hodnota** > 0.05, **nezamítáme** nezávislost reziduí.

```
bptest(model)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model
## BP = 4.482, df = 4, p-value = 0.3447
```

Nezamítáme ani homoskedasticitu, i když grafy níže ukazují mírné zvýšení rozptylu pro vyšší hodnoty predikce. Zároveň je na nich vidět, že residuum pro Lucembursko je větší.

```
par(mfrow=c(1, 2))
plot(model, which=1)
plot(model, which=3)
```

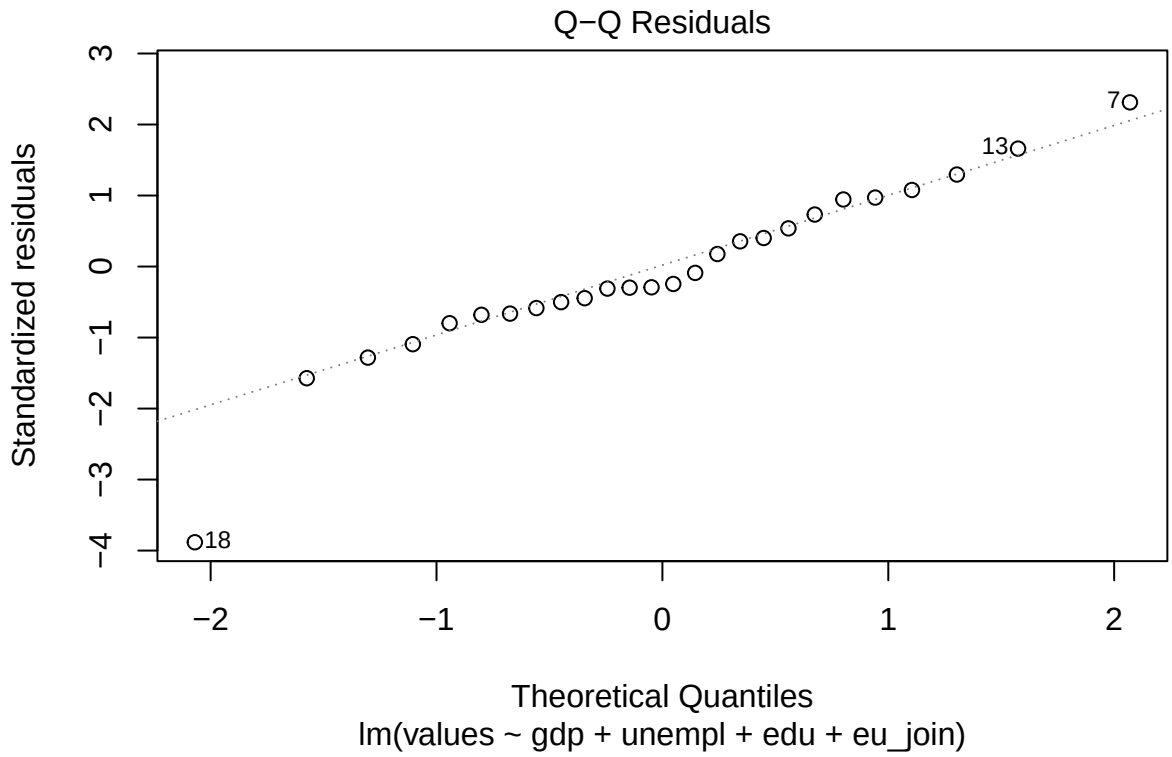


Ještě otestujeme normalitu reziduí.

```
shapiro.test(residuals(model))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(model)
## W = 0.97671, p-value = 0.7975
```

```
plot(model, which=2)
```



Ani normalitu reziduí nezamítáme, předpoklady **jsou splněny**.

Výběr podmodelu

Viděli jsme, že model měl nevýznamné komponenty, zkusíme ho **redukovat** na podmodel.

```
summary(model)
```

```
##
## Call:
## lm(formula = values ~ gdp + unempl + edu + eu_join, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -7306  -2670  -1155   2807   9864
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.825e+03  3.817e+03   1.264 0.220033
## gdp            4.545e-01  9.439e-02   4.815 9.29e-05 ***
## unempl        -4.291e+02  1.984e+02  -2.164 0.042176 *
## edu            2.627e+02  1.345e+02   1.952 0.064351 .
## eu_joinBefore  1.274e+04  2.752e+03   4.629 0.000145 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 4523 on 21 degrees of freedom
## Multiple R-squared:  0.9313, Adjusted R-squared:  0.9183
## F-statistic: 71.21 on 4 and 21 DF,  p-value: 6.584e-12
```

Nejprve odebereme *edu*.

```
model2 <- lm(values ~ gdp + unempl + eu_join, data=df)
summary(model2)
```

```
##
## Call:
## lm(formula = values ~ gdp + unempl + eu_join, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9292.3 -2533.1  -848.5  2511.0 10455.3
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    9.255e+03  3.259e+03   2.839 0.009540 **
## gdp            5.501e-01  8.567e-02   6.422 1.84e-06 ***
## unempl        -3.178e+02  2.017e+02  -1.575 0.129522
## eu_joinBefore  1.172e+04  2.869e+03   4.083 0.000492 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4804 on 22 degrees of freedom
## Multiple R-squared:  0.9189, Adjusted R-squared:  0.9078
## F-statistic: 83.05 on 3 and 22 DF,  p-value: 3.731e-12
```

Nyní je nevýznamné *unempl*, ale intercept už je významný.

```
anova(model2, model)
```

```
## Analysis of Variance Table
##
## Model 1: values ~ gdp + unempl + eu_join
## Model 2: values ~ gdp + unempl + edu + eu_join
##   Res.Df       RSS Df Sum of Sq      F  Pr(>F)
## 1     22 507628465
## 2     21 429641667  1  77986798 3.8118 0.06435 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
AIC(model)
```

```
## [1] 517.9143
```

```
AIC(model2)
```

```
## [1] 520.2511
```

P-hodnota testu model-podmodel je **0.06529**, mezi modely není statisticky významný rozdíl. Jednodušší model má ale o něco vyšší AIC.

Zkusíme odebrat *unempl*.

```
model3 <- lm(values ~ gdp + eu_join, data=df)
summary(model3)
```

```
##
## Call:
## lm(formula = values ~ gdp + eu_join, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11246.3  -2273.3   -933.7   2908.5  11213.6
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.806e+03  1.678e+03   2.864  0.00877 **
## gdp           6.230e-01  7.441e-02   8.372 1.95e-08 ***
## eu_joinBefore 9.689e+03  2.646e+03   3.662  0.00130 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4956 on 23 degrees of freedom
## Multiple R-squared:  0.9097, Adjusted R-squared:  0.9019
## F-statistic: 115.9 on 2 and 23 DF,  p-value: 9.757e-13
```

Všechny komponenty už jsou statisticky významné.

```
anova(model3, model2)
```

```
## Analysis of Variance Table
##
## Model 1: values ~ gdp + eu_join
## Model 2: values ~ gdp + unempl + eu_join
##   Res.Df       RSS Df Sum of Sq      F Pr(>F)
## 1     23 564868462
## 2     22 507628465  1  57239997 2.4807 0.1295
```

```
AIC(model2)
```

```
## [1] 520.2511
```

```
AIC(model3)
```

```
## [1] 521.029
```

Mezi modely opět není statisticky významný rozdíl, je možné ho znovu redukovat, ale AIC je zase o trochu vyšší.

Náš finální model je následující:

$$\text{Salary} = 4812 + 0.623 \cdot \text{gdp} + \begin{cases} 0 & \text{pro eu-join} = After, \\ 9678 & \text{pro eu-join} = Before \end{cases}$$

- **Intercept (4812):** Hodnota průměrného platu, když *gdp* je 0 a *eu_join* je After.
- **Koeficient u gdp (0.623):** S každým zvýšením HDP na obyvatele o jedno euro se očekává zvýšení průměrného platu o 0.623 eura.
- **Koeficient u eu_joinBefore (9678):** Pokud se země připojila k EU před rokem 2000, očekává se, že průměrný plat bude o 9678 eur vyšší ve srovnání se zeměmi, které se připojily po roce 2000.

Koeficient determinace R^2 je 0.9096, adjustovaný je 0.9017. Model vysvětluje necelých 91 % variability a nemá zbytečně moc regresorů.

```
df <- cbind(df, predict(model3, interval="prediction"))
ggplot(df, aes(x=gdp, y=values, group=eu_join, color=eu_join)) +
    geom_point() +
    geom_line(aes(y=fit))
```