3. Feladat

StatWars

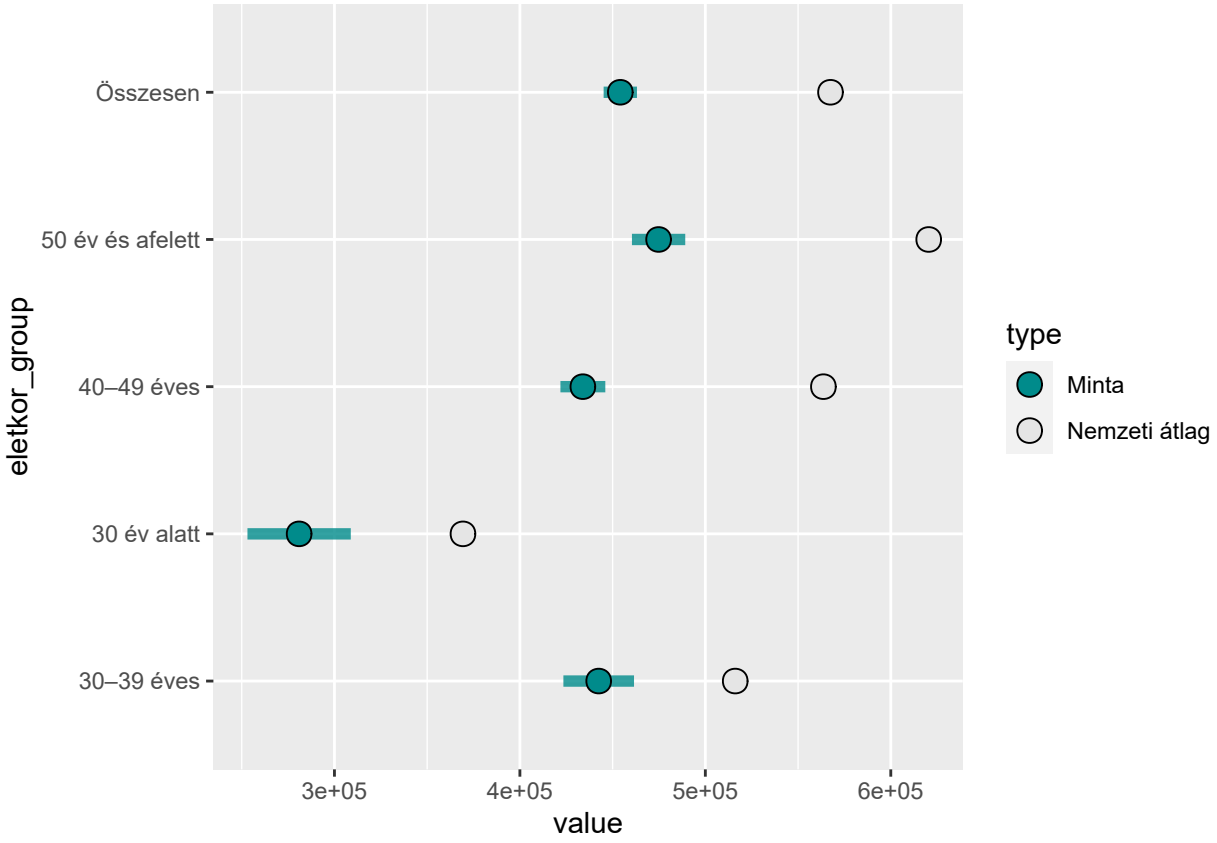
2021. november 16.

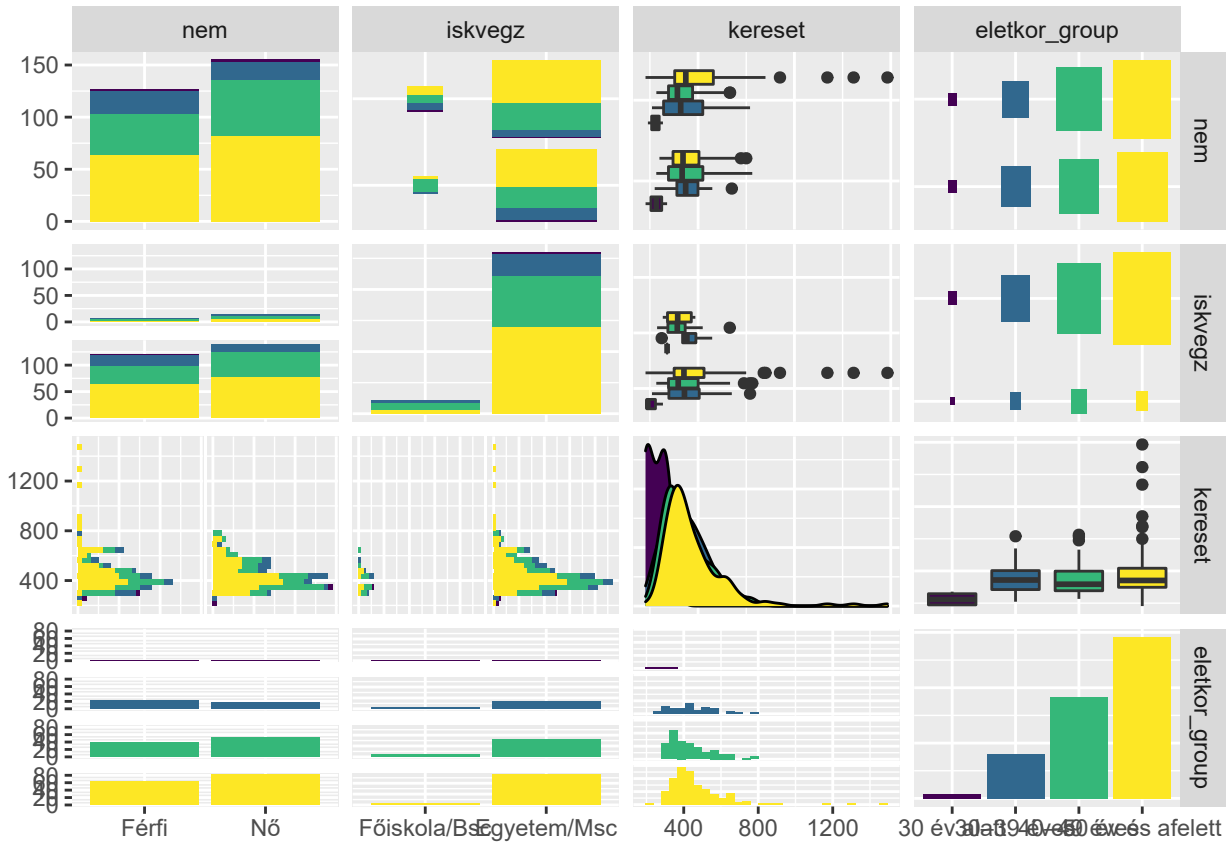
Tartalomjegyzék

1. Az egyetemi főiskolai oktatók/tanárok fizetése életkori csoportonként hogyan különbözik egymástól? Hasonlítsák össze valamilyen benchmark adattal, ezen az egyetemen mennyivel keresnek jobban/rosszabbul az egyetemi/főiskolai oktatók/tanárokéletkori csoportonkénti bontásban, mint az országos átlag?

☒ ábra a ksh összehasonlításról
☐ kereset vs. eletkor boxplot

```
## Warning: Removed 5 rows containing missing values (geom_segment).
```

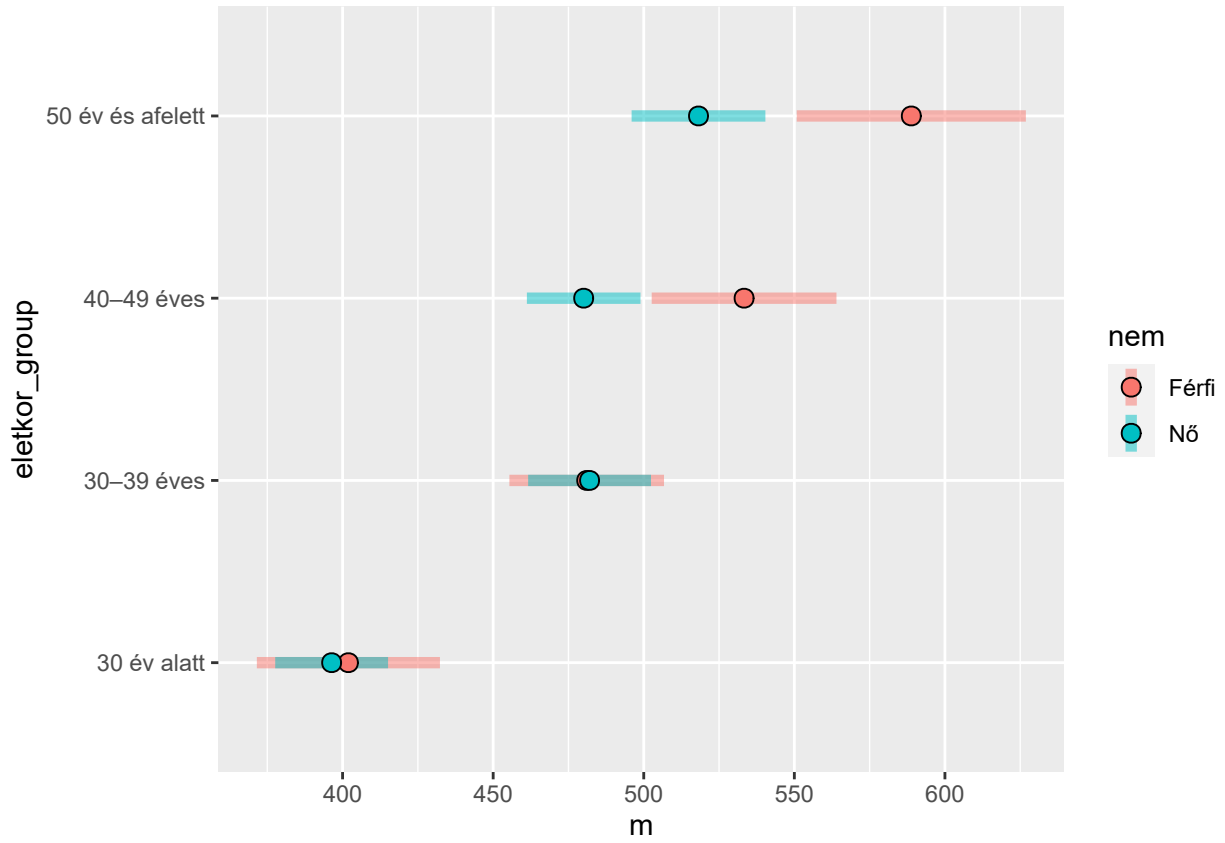





```
## # A tibble: 5 x 7
##   eletkor_group     mean median    sd alpha3 kurtosis     n
##   <chr>            <dbl>  <dbl> <dbl>  <dbl>    <dbl> <int>
## 1 30 év alatt       398.   372.  96.9  0.636     2.50    37
## 2 30-39 éves        482.   443. 171.   1.28      4.24   116
## 3 40-49 éves        501.   419. 242.   2.14      8.46   209
## 4 50 év és afelett  544.   426. 335.   2.90     13.0    285
## 5 Összesen          511.   425. 274.   3.02     15.6    647
```

2. Hasonlítsák össze az oktatók (4-es csoport) és az ügyintézők (5 és 6-os csoport együtt) keresetek szerinti eloszlását a lehető legteljesebben!

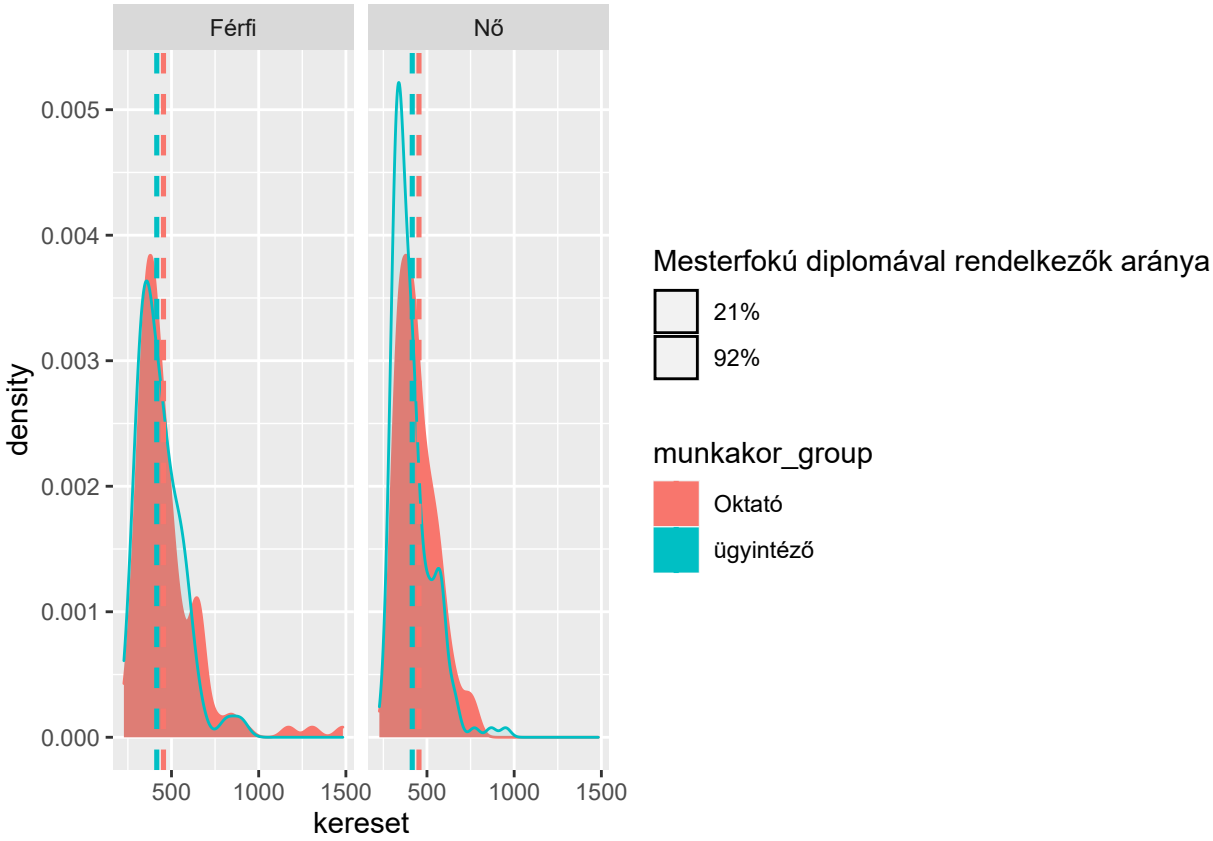
Az eredményeket foglalják össze, ahol annak helye van érzékeltessék ábrákkal! Igyekezzenek tömören, lényegretörően végezni a számításokat! Kérjük, egy word vagy pdf fájlban legyenek az eredmények, elemzések! Excelt, vagy más szoftvert természetesen használhatnak, de azok outputja ha feltétlenül kell, függelékként lehet az elemzésükben.

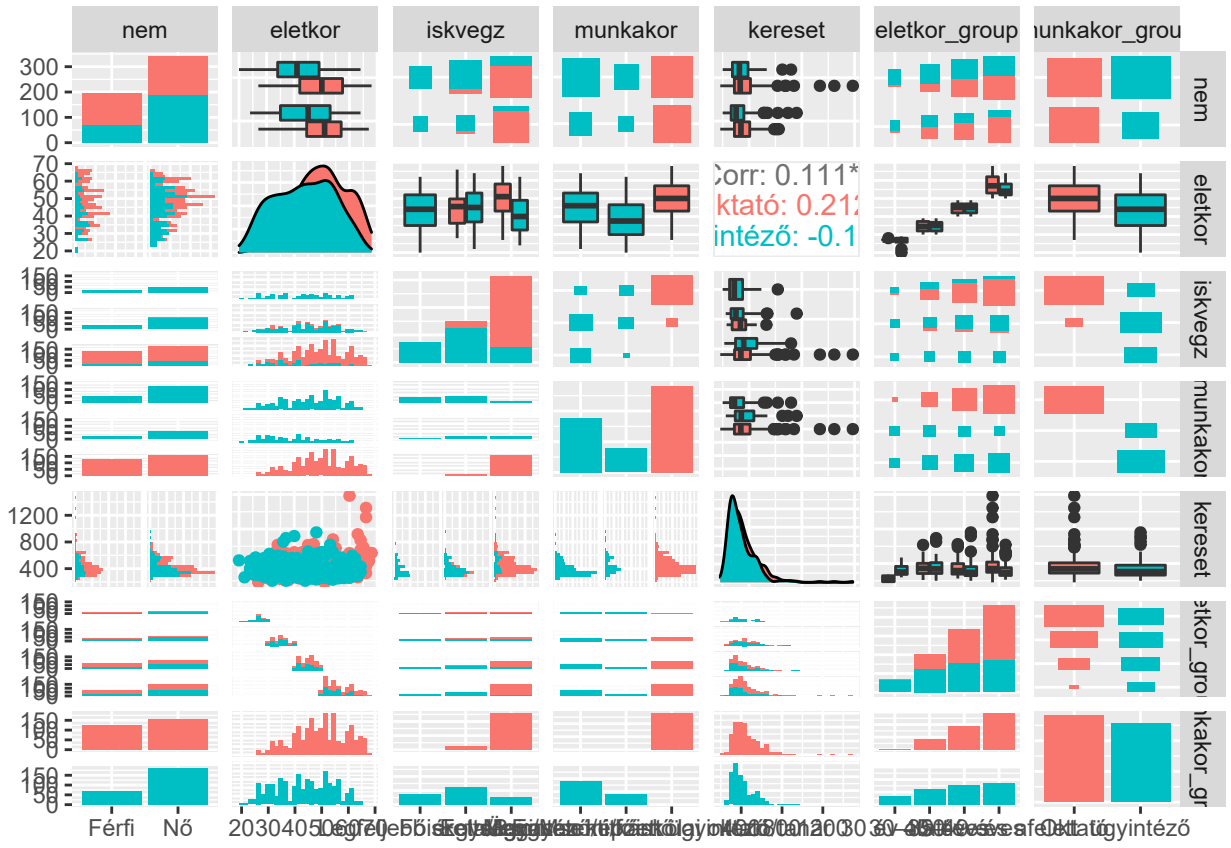

50 év és afelett
40–49 éves
30–39 éves
30 év alatt
eletkor_group
400
450
500
550
600
m
nem
Férfi
Nő

Életkor alapján való eloszlása a tanári fizetéseknek

Oktatük és ügyintézők kereseti eloszlása

```
## Warning: Using alpha for a discrete variable is not advised.
```

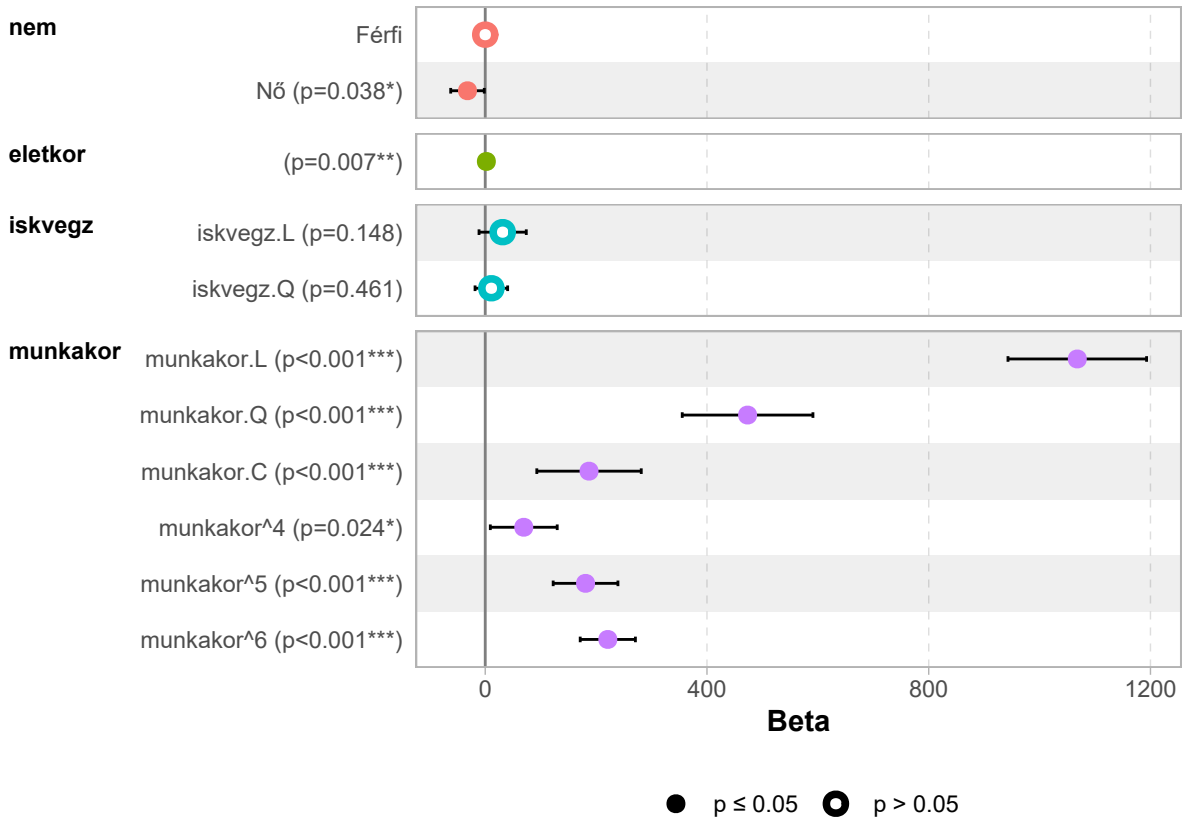

Férfi
Nő
0.005
0.004
0.003
0.002
0.001
0.000
density
500
1000
1500
kereset
Mesterfokú diplomával rendelkezők aránya
21%
92%
munkakor_group
Oktató
ügyintéző



3. Készítsenek elemzést arról, hogy nem (férfi-nő) szerint a havi átlagos bruttó keresetekben mekkora az átlagos különbség összességében és az egyéb ismérvek hatását kiszűrve, illetve azokkal összekapcsolódva! Használjanak az elemzéshez kétféle módszertant/modellt és hasonlítsák össze a kétféle módszerrel kapott eredmény(eke)t! Írjanak egy összefoglalást is az elemzések tapasztalatairól!

p-score, ols, fa, ..

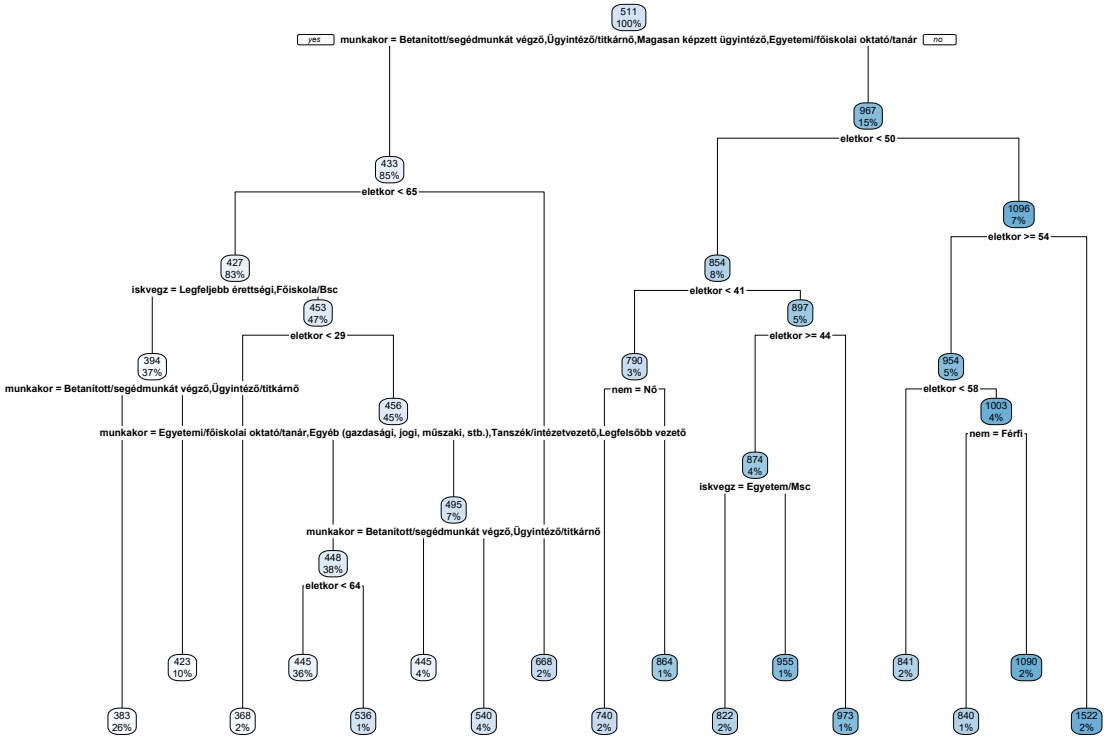
```
## # A tibble: 2 x 2
##   nem   `mean(kereset)`
##   <chr>           <dbl>
## 1 Férfi            539.
## 2 Nő               493.
```




```
## # A tibble: 647 x 8
##        z nem    eletkor iskvegz      munkakor        kereset eletkor_group     id
##    <dbl> <chr>    <int> <ord>        <ord>             <dbl> <ord>          <int>
##  1 0.521 Férfi       50 Egyetem/Msc  Legfelsőbb vez~   2411. 50 év és afel~     1
##  2 0.516 Nő          51 Egyetem/Msc  Legfelsőbb vez~   1073. 50 év és afel~     2
##  3 0.507 Férfi       53 Egyetem/Msc  Legfelsőbb vez~   1990. 50 év és afel~     3
##  4 0.455 Nő          64 Egyetem/Msc  Legfelsőbb vez~   1609. 50 év és afel~     4
##  5 0.393 Férfi       32 Főiskola/Bsc Tanszék/intéze~    706. 30-39 éves         5
##  6 0.477 Nő          33 Egyetem/Msc  Tanszék/intéze~    994. 30-39 éves         6
##  7 0.384 Nő          34 Főiskola/Bsc Tanszék/intéze~    632. 30-39 éves         7
##  8 0.380 Férfi       35 Főiskola/Bsc Tanszék/intéze~    512. 30-39 éves         8
##  9 0.468 Férfi       35 Egyetem/Msc  Tanszék/intéze~    987. 30-39 éves         9
## 10 0.468 Nő          35 Egyetem/Msc  Tanszék/intéze~    880. 30-39 éves        10
## # ... with 637 more rows
```

```
## # A tibble: 1 x 3
##     ate  atet atet_no
##   <dbl> <dbl>   <dbl>
## 1  31.3  33.9    29.7
```

```
## Warning: Cannot retrieve the data used to build the model (so cannot determine roundint and is.binary
## To silence this warning:
##     Call rpart.plot with roundint=FALSE,
##     or rebuild the rpart model with model=TRUE.
```


511
100%
yes
munkakor = Betanitott/segédmunkát végző,Ügyintéző/titkárnő,Magasan képzett ügyintéző,Egyetemi/főiskolai oktató/tanár
no
967
15%
eletkor < 50
433
85%
eletkor < 65
1096
7%
eletkor >= 54
427
83%
iskvegz = Legfeljebb érettségi,Főiskola/Bsc
854
8%
eletkor < 41
897
5%
eletkor >= 44
453
47%
eletkor < 29
790
3%
nem = Nő
954
5%
eletkor < 58
1003
4%
394
37%
munkakor = Betanitott/segédmunkát végző,Ügyintéző/titkárnő
456
45%
munkakor = Egyetemi/főiskolai oktató/tanár,Egyéb (gazdasági, jogi, műszaki, stb.),Tanszék/intézetvezető,Legfelsőbb vezető
nem = Férfi
874
4%
iskvegz = Egyetem/Msc
495
7%
munkakor = Betanitott/segédmunkát végző,Ügyintéző/titkárnő
448
38%
eletkor < 64
423
10%
445
36%
445
4%
668
2%
864
1%
955
1%
841
2%
1090
2%
383
26%
368
2%
536
1%
540
4%
740
2%
822
2%
973
1%
840
1%
1522
2%

Függelék: R kódok

```r
# setup ------------------------------------------------------------------------

library(tidyverse)
# data -------------------------------------------------------------------------

teacher_df <- readxl::read_excel("3. forduló STAT WARS UNI.xlsx", sheet = 2) %>%
  mutate(
    nem = case_when(
      nem == 1 ~ "Férfi",
      nem == 2 ~ "Nő"
    ),
    eletkor = as.integer(eletkor),
    iskvegz = factor(iskvegz, levels = 1:3, ordered = TRUE),
    iskvegz = fct_relabel(iskvegz, function(l) {
      case_when(
        l == 1 ~ "Legfeljebb érettségi",
        l == 2 ~ "Főiskola/Bsc",
        l == 3 ~ "Egyetem/Msc"
      )}),
    munkakor = factor(munkakor, levels = 7:1, ordered = TRUE),
    munkakor = fct_relabel(munkakor, function(l) {
      case_when(
        l == 1 ~ "Legfelsőbb vezető",
        l == 2 ~ "Tanszék/intézetvezető",
        l == 3 ~ "Egyéb (gazdasági, jogi, műszaki, stb.)",
        l == 4 ~ "Egyetemi/főiskolai oktató/tanár",
        l == 5 ~ "Magasan képzett ügyintéző",
        l == 6 ~ "Ügyintéző/titkárnő",
        l == 7 ~ "Betanított/segédmunkát végző"
      )})
  )
# utils ------------------------------------------------------------------------

total_summarise <- function(x, g, ...) {
  # original summarise function from tidyverse, but contains TOTAL row
  bind_rows(
    x %>%
      group_by({{ g }}) %>%
      summarise(...) %>%
      ungroup(),
    x %>%
      summarise(...) %>%
      mutate(g = "Összesen") %>%
      select(g, everything()) %>%
      rename("{{ g }}" := 1)
  )
}
national_avg <- rio::import("https://www.ksh.hu/stadat_files/mun/hu/mun0059.csv") %>%
  # download data from KSH website: https://www.ksh.hu/stadat_files/mun/hu/mun0059.html
  tibble() %>%
  janitor::row_to_names(2) %>%
  select(2, starts_with("2020")) %>%
```

```r
  rename_all(str_remove_all, "2020 Korcsoport ") %>%
  rename_all(str_remove_all, "2020 ") %>%
  rename(profession = 1, Összesen = Együtt) %>%
  filter(str_detect(profession, "Egyetemi")) %>%
  mutate_at(-1, str_remove, " ") %>%
  mutate_at(-1, as.numeric) %>%
  pivot_longer(-1, names_to = "eletkor_group") %>%
  select(-profession)
teacher_df <- teacher_df %>%
  mutate(
    eletkor_group = cut(eletkor, breaks = c(c(0, 3, 4, 5)*10, Inf), right = FALSE,
                        labels = FALSE),
    eletkor_group = factor(eletkor_group, levels = 1:4, ordered = TRUE),
    eletkor_group = fct_relabel(eletkor_group, function(l) {
      case_when(
        l == 1 ~ "30 év alatt",
        l == 2 ~ "30-39 éves",
        l == 3 ~ "40-49 éves",
        l == 4 ~ "50 év és afelett"
      )
    })
  )
compare_df <- bind_rows(
teacher_df %>%
  filter(munkakor == "Egyetemi/főiskolai oktató/tanár") %>%
  total_summarise(eletkor_group,
                  value = mean(kereset)*1e3,
                  s = sd(kereset*1e3),
                  n = n()
                  ) %>%
  mutate(type = "Minta"), # TODO név
national_avg %>%
  mutate(type = "Nemzeti átlag", s = NA, n = NA) # TODO teljes munkaidő hipotézise
)
compare_df %>%
  mutate(
    lb = value - s/(n^.5),
    ub = value + s/(n^.5),
  ) %>%
  ggplot() +
  geom_linerange(aes(xmin = lb, xmax = ub, y = eletkor_group),
                 color = "cyan4", size = 2, alpha = .8) +
  geom_point(aes(value, eletkor_group, fill = type), shape = 21, size = 4) +
  scale_fill_manual(values = c("cyan4", "grey90"))
teacher_df %>%
  filter(munkakor == "Egyetemi/főiskolai oktató/tanár") %>%
  select(-eletkor, -munkakor) %>%
  GGally::ggpairs(aes(color = eletkor_group))
total_summarise(teacher_df, eletkor_group,
                mean = mean(kereset),
                median = median(kereset),
                sd = sd(kereset),
                alpha3 = moments::skewness(kereset),
```

```
                kurtosis = moments::kurtosis(kereset),
                n = n()
)
teacher_df %>%
  group_by(eletkor_group, nem) %>%
  summarise(m = mean(kereset), s = sd(kereset), n = n()) %>%
  mutate(
    cl = m - s/(n^.5),
    ch = m + s/(n^.5)
  ) %>%
  ggplot() +
  aes(m, eletkor_group) +
  geom_linerange(aes(xmin = cl, xmax = ch, color = nem), size = 2, alpha = .5) +
  geom_point(aes(fill = nem), shape = 21, size = 3)

GGally::ggpairs(teacher_df, ggplot2::aes(colour=iskvegz))

teacher_df %>%
  filter(
    munkakor %in% c("Ügyintéző/titkárnő", "Magasan képzett ügyintéző", "Egyetemi/főiskolai oktató/tanár
  ) %>%
  mutate(munkakor_group = ifelse(
    munkakor == "Egyetemi/főiskolai oktató/tanár", "Oktató", "ügyintéző"
  )) %>%
  group_by(munkakor_group) %>%
  mutate(iskvegz_ratio = round(sum(iskvegz == "Egyetem/Msc")/n(), 2),
         iskvegz_ratio = scales::percent(iskvegz_ratio),
         mean = mean(kereset)) %>%
  ungroup() %>%
  ggplot(aes(x = kereset, group = munkakor_group, color = munkakor_group,
             fill = munkakor_group, alpha = iskvegz_ratio)) +
  geom_density() +
  geom_vline(aes(xintercept = mean, color = munkakor_group), linetype = "dashed", size = 0.9) +
  facet_wrap(~nem)+
  labs(alpha = "Mesterfokú diplomával rendelkezők aránya")
teacher_df %>%
  filter(
    munkakor %in% c("Ügyintéző/titkárnő", "Magasan képzett ügyintéző", "Egyetemi/főiskolai oktató/tanár
  ) %>%
  mutate(munkakor_group = ifelse(
    munkakor == "Egyetemi/főiskolai oktató/tanár", "Oktató", "ügyintéző"
  )) %>%
  GGally::ggpairs(aes(color = munkakor_group))
teacher_df %>%
  group_by(nem) %>%
  summarise(mean(kereset)) # TODO
teacher_df %>%
  lm(formula = kereset ~ .-eletkor_group) %>%
  GGally::ggcoef_model()
teacher_df %>%
  select(- eletkor_group) %>%
  mutate(nem = nem == "Férfi") %>%
  glm(formula = nem ~ eletkor + iskvegz + munkakor, family = "binomial") %>%
```

```r
  predict( type = "response") %>%
  cbind(teacher_df) %>%
  rename(z = 1) %>%
  tibble() %>%
  mutate(id = row_number())
teacher_df %>%
  group_by(nem, iskvegz, munkakor, eletkor_group) %>%
  summarise(kereset = mean(kereset), n = n()) %>%
  pivot_wider(names_from = nem, values_from = c(kereset, n)) %>%
  mutate(
    d = kereset_Férfi - kereset_Nő,
    n = n_Férfi + n_Nő
  ) %>%
  ungroup() %>%
  summarise(ate = weighted.mean(d, n, na.rm = T),
            atet = weighted.mean(d, n_Férfi, na.rm = TRUE),
            atet_no = weighted.mean(d, n_Nő, na.rm = TRUE))
teacher_df %>%
  select(- eletkor_group) %>%
  rpart::rpart(formula = kereset ~ ., cp = .001) %>%
  rpart.plot::rpart.plot()
```