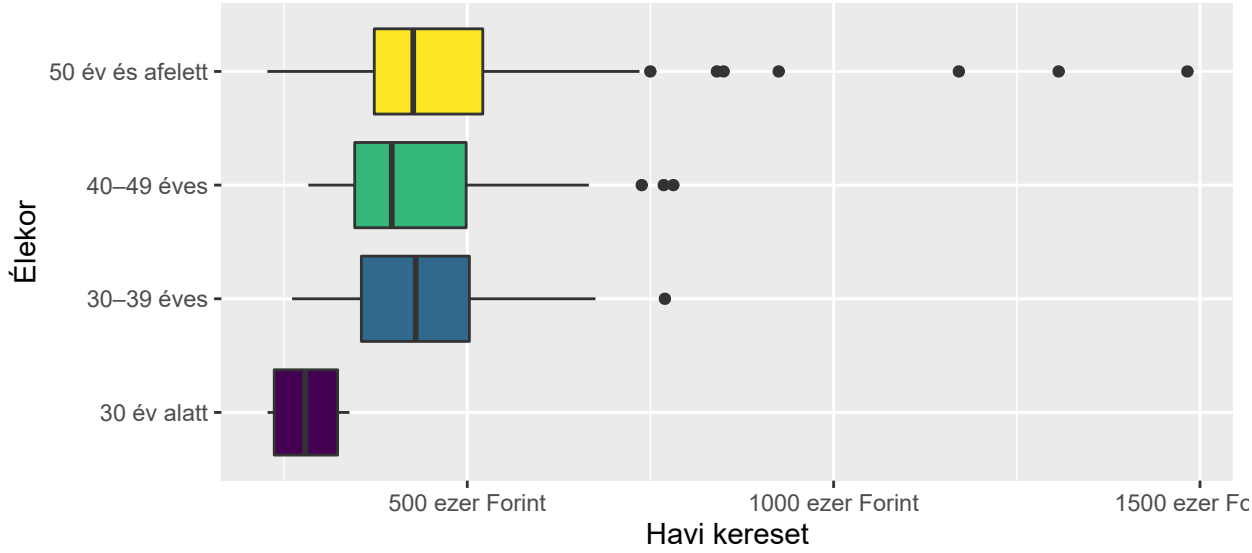
3. Feladat

StatWars

2021. november 18.

1. Feladat

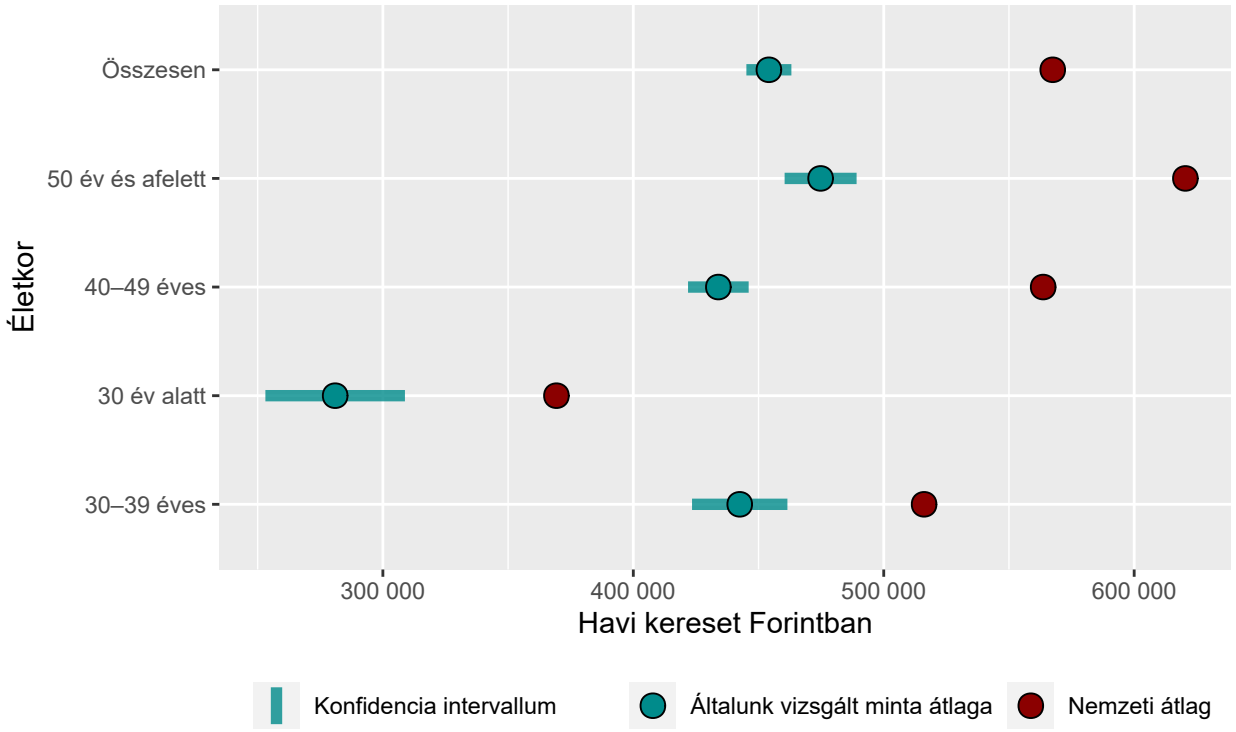
A korosztályok felbontásakor figyelembe vettük a KSH módszertanát, így 4 korcsoportot alkottunk a megfigyelésekből, a 30 év alatti, 30-39 év közötti, 40-49 év közötti, és 50 év felettiek csoportját. Ez alapján elmondhatjuk, hogy a kereset átlagos értéke a legalacsonyabb a 30 év alatti korosztálynál, míg korcsoportonként fokozatosan növekszik. Azonban fontos kiemelni, hogy a fizetések mediánértéke a korosztályokon belül a 30-39 évesek között a legmagasabb, így a másik két korosztálynál a kiugró értékek jobbra ferde eloszlást implikálnak. Láthatjuk, hogy a legtöbb kiugró értéket az 50 év felettieknél találjuk (közülük is a férfiaknál), ahol akár 1 milliós bruttó fizetéssel is rendelkező oktatókat találhatunk, így a relatív szórás értéke ebben a csoportban a 60%-ot is meghaladta, míg a többiben nem érte el az 50%-ot.



1. ábra. Egyetemi/főiskolai oktatók havi keresetének dobozábrája életkor csoportok szerinti bontásban.

Referenciaértékként a KSH 2020-as 2410-es FEOR '08 kódja (Egyetemi, főiskolai oktató, tanár) alá tartozó értékeket vizsgáltuk[1]. Mivel a nemzeti bruttó átlagbér a teljes munkaidőben dolgozó oktatókra vonatkozik, így ezt az összehasonlítást megtehetjük, mivel a mintában szereplő munkavállalók is főállású alkalmazottak voltak. Összehasonlítva a mintában szereplő életkori csoportokat a KSH módszertanában megadott referenciacsoportokkal azt láthatjuk, hogy az 5%-os szignifikancia szinten vizsgált kétoldalas t-próba alapján a 2020-as országos bruttó átlagfizetések mind az 5 korcsoportban meghaladják az általunk vizsgált egyetem oktatóinak fizetéseit. Az összesen vizsgált bruttó átlagfizetés nagyjából 120 ezer forinttal volt alacsonyabb az intézményben, a legnagyobb különbséget azonban az 50 év feletti korosztályban tapasztalhattuk, nagyjából 150 ezer forintos átlagos eltéréssel.

[1] https://www.ksh.hu/stadat_files/mun/hu/mun0059.html



2. ábra. Általunk vizsgált minta és az országos havi átlag keresetek összehasonlítása életkor szerinti bontásban

1. táblázat: Leíró statiszikák a életkor szerinti bontásban

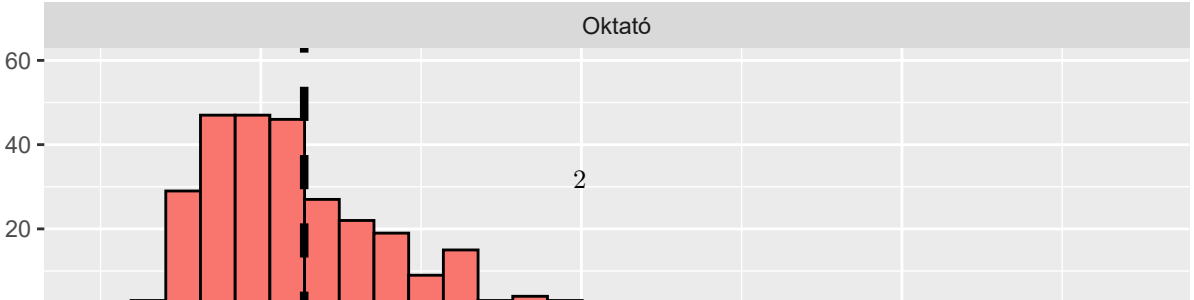
Életkor	Átlag	Medián	Szórás	Relatív szórás	Ferdeség	Csúcsosság	Elemszám
30 év alatt	398,32	371,7	96,91	0,24	0,64	2,50	37
30–39 éves	481,62	443,2	171,49	0,36	1,28	4,24	116
40–49 éves	501,20	418,9	241,66	0,48	2,14	8,46	209
50 év és afelett	543,71	426,5	334,52	0,62	2,90	13,05	285
Összesen	510,53	424,9	274,13	0,54	3,02	15,56	647

2. Feladat

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.

- ☐ t-próba
- ☐ > 1m
- ☐ ezer HUF



kancia szinten elvetésre kerül az a nullhipotézis, miszerint a két foglalkoztatási csoportban megegyezne a sokassági átlag.

2. táblázat: Fizetések eloszlásának jellemzői munkakör jellege szerinti bontásban

Munkakör jellege	Átlag	Medián	Szórás	Relatív szórás	Ferdeség	Csúcsosság
Oktató	454,14	422,30	150,90	0,33	2,57	14,72
Ügyintéző	415,54	383,30	115,38	0,28	1,50	6,14
Összesen	435,77	404,15	136,42	0,31	2,36	13,84

3. Feladat

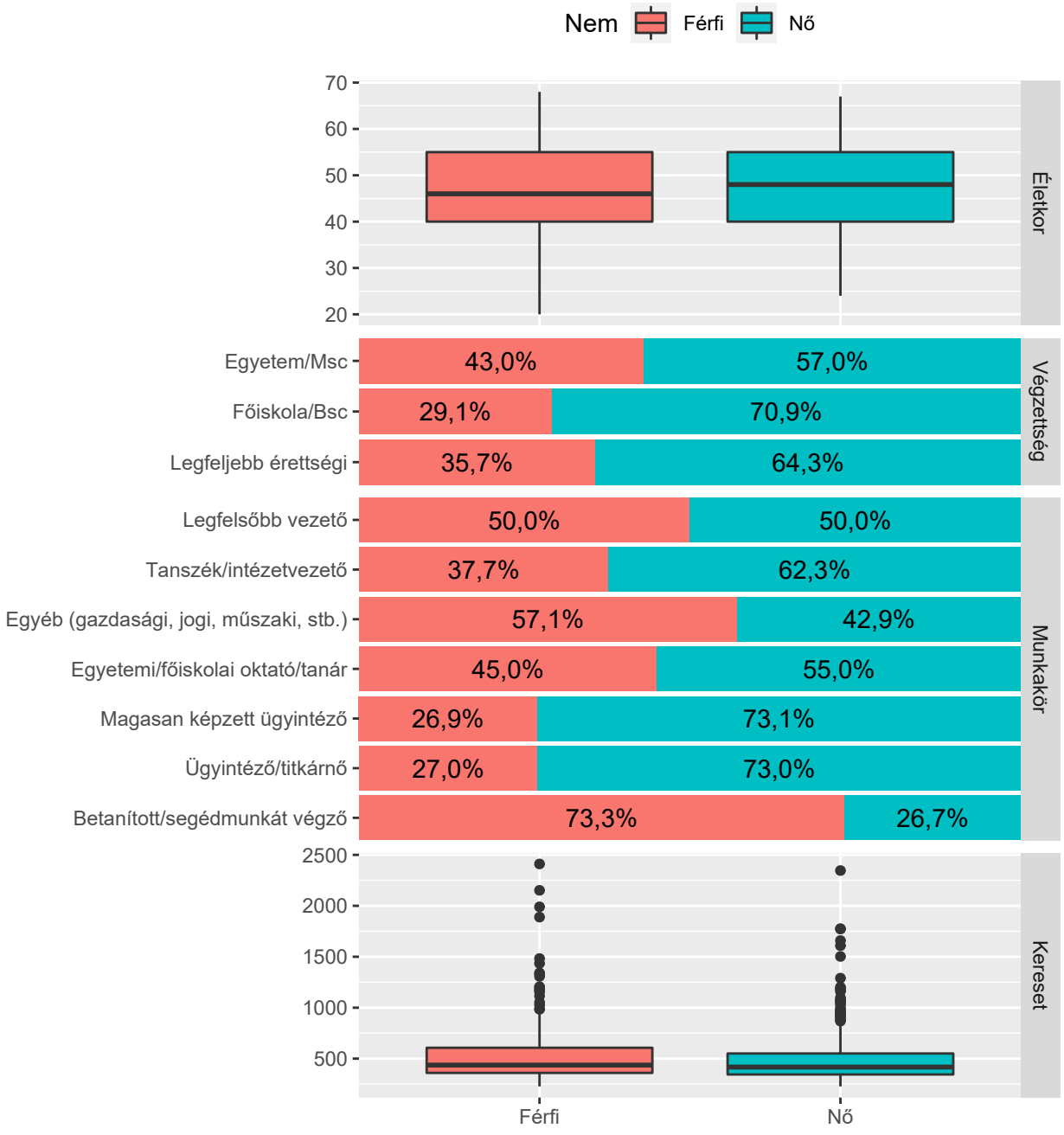
Készítsenek elemzést arról, hogy nem (férfi-nő) szerint a havi átlagos bruttó keresetekben mekkora az átlagos különbség összességében és az egyéb ismérvek hatását kiszűrve, illetve azokkal összekapcsolódva! Használjanak az elemzéshez kétféle módszertant/modellt és hasonlítsák össze a kétféle módszerrel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

p-score, ols, fa, ..

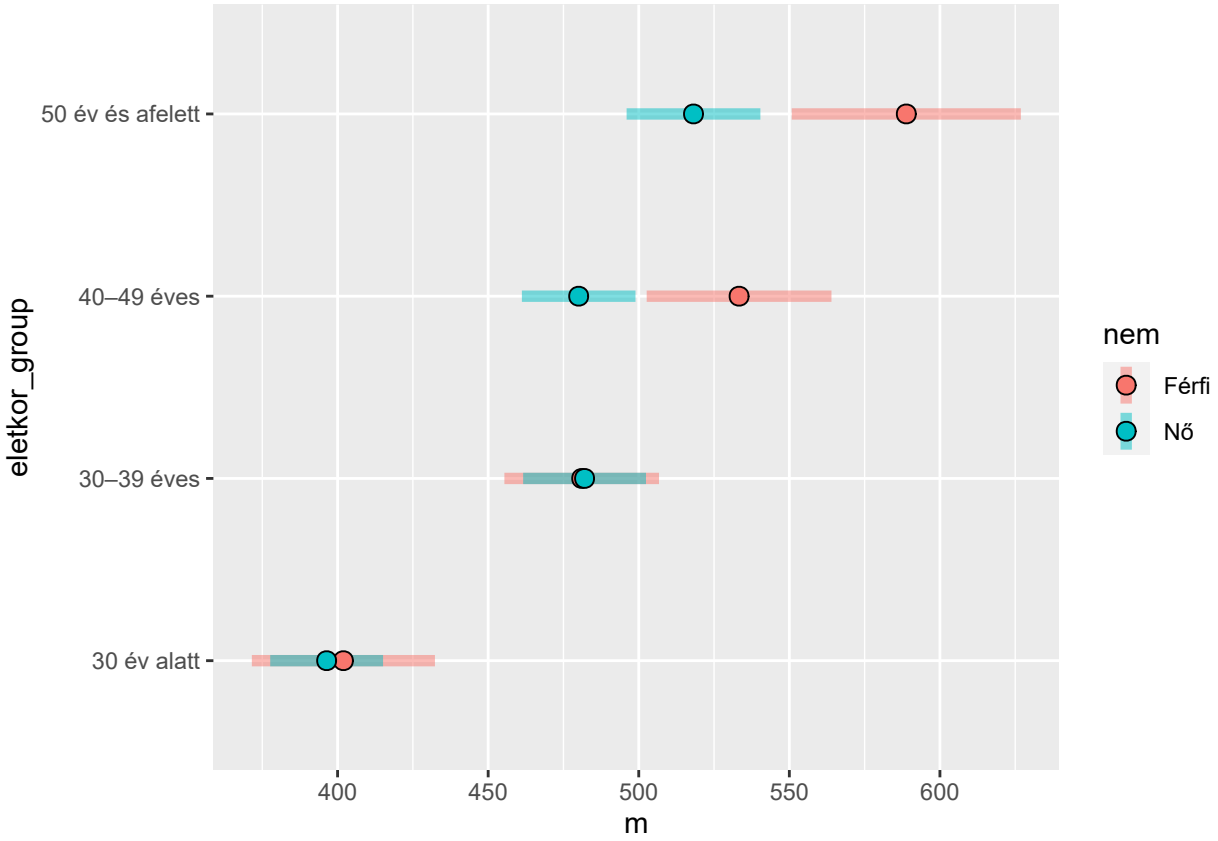
Kétoldalú alternatív hipotézis mellet a kétmintás t-próba teszt-statisztikájának értéke $2,2425$ (p-érték $= 0,0257$), ami alapján 5%-os szignifikanci szinten elutasíthatjuk, hogy a férfi és női fizetések sokassági átlaga megegyezne. Egyoldalú nullhipotézis mellett (H_0 : Férfiak fizetése $\leq$ Nők fizetése), a p-érték $0,0128$, ami mellett 1%-os szignifikancia szinten továbbra sem tudjuk elvetni a nullhipotézist, miszerint a Férfiak fizetése szignifikánsan magasabb, mint a nőké.

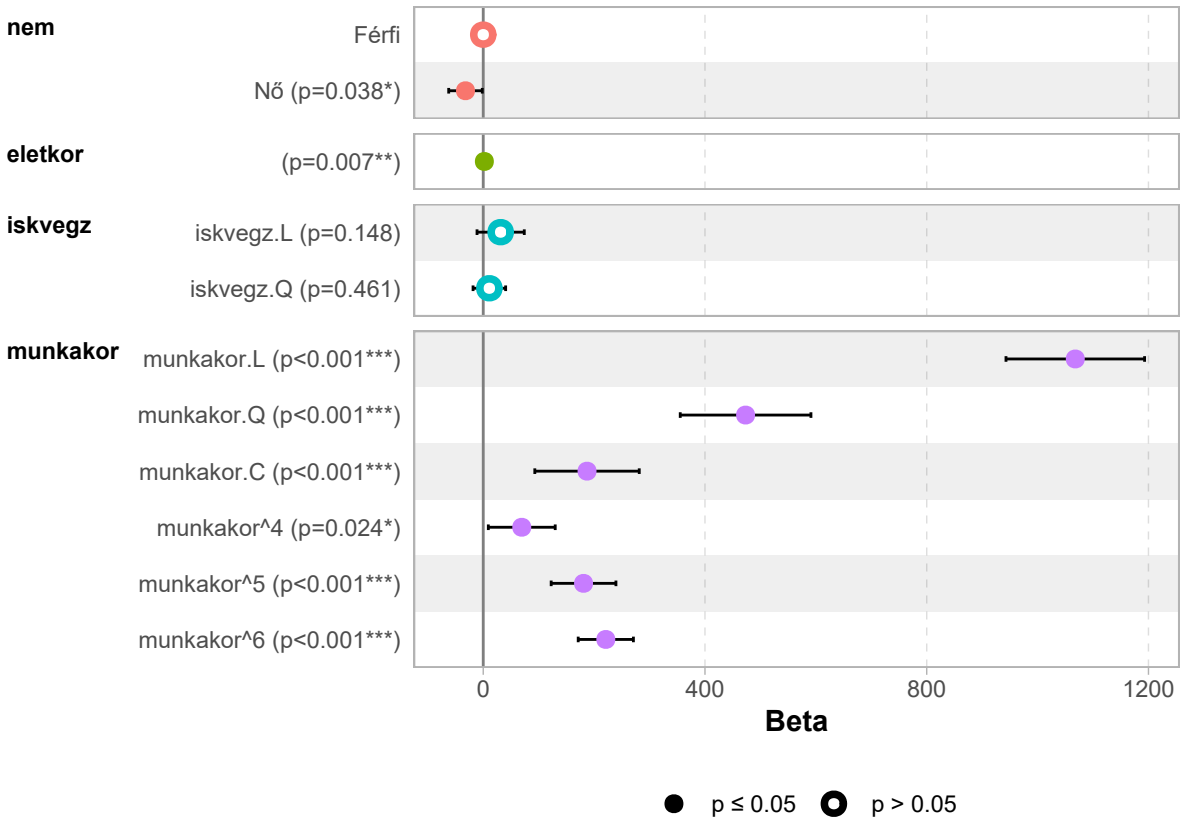
3. táblázat: Fizetések eloszlásának jellemzői nemek szerinti bontásban

Munkakör jellege	Átlag	Medián	Szórás	Relatív szórás	Ferdeség	Csúcsosság
Férfi	539,40	436,4	311,71	0,58	2,82	13,40
Nő	492,71	417,2	246,83	0,50	3,08	16,51
Összesen	510,53	424,9	274,13	0,54	3,02	15,56



4. ábra. A vizsgált adattábla változóinak nemenkénti megoszlása


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




```
## # 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
```

Függelék: R kódok

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

library(tidyverse)
library(GGally)
options(scipen = 999)

# 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"
      )})
  )

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"
      )
    })
  )

# utils ------------------------------------------------------------------------

```

```r
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)
  )
}

teacher_df %>%
  filter(munkakor == "Egyetemi/főiskolai oktató/tanár") %>%
  ggplot(aes(kereset, eletkor_group, fill = eletkor_group)) +
  geom_boxplot(show.legend = FALSE) +
  scale_x_continuous(labels = ~ str_c(., " ezer Forint")) +
  labs(x = "Havi kereset", y = "Élekor")

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")) %>%
  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)

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 = "Általunk vizsgált minta átlaga"),
  national_avg %>%
    mutate(type = "Nemzeti átlag", s = NA, n = NA)
)

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 = "Konfidencia intervallum"), size = 2, alpha = .8) +
  geom_point(aes(value, eletkor_group, fill = type), shape = 21, size = 4) +
  scale_fill_manual(values = c("cyan4", "red4")) +
  scale_color_manual(values = "cyan4") +
  scale_x_continuous(labels = ~ format(., big.mark = " ")) +
  labs(x = "Havi kereset Forintban", y = "Életkor", color = NULL, fill = NULL) +
  theme(
    legend.position = "bottom"
  )

teacher_df %>%
  filter(munkakor == "Egyetemi/főiskolai oktató/tanár") %>%
  select(-eletkor, -munkakor) %>%
  GGally::ggpairs(aes(color = eletkor_group))
total_summarise(teacher_df, eletkor_group,
                `Átlag` = mean(kereset),
                `Medián` = median(kereset),
                `Szórás` = sd(kereset),
                `Relatív szórás` = sd(kereset) / mean(kereset),
                `Ferdeség` = moments::skewness(kereset),
                `Csúcsosság` = moments::kurtosis(kereset),
                `Elemszám` = n()
) %>%
  mutate_at(-1, ~ format(round(., 2), decimal.mark = ",")) %>%
  rename(Életkor = 1) %>%
  knitr::kable(caption = "Leíró statiszikák a életkor szerinti bontásban",
               align = c("l", rep("c", 7)))
profession_df <- 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"
  ))

profession_df %>%
  group_by(munkakor_group) %>%
  mutate(
    m = mean(kereset),
    m = ifelse(!duplicated(m), m, NA)
  ) %>%
  ggplot(aes(kereset, fill = munkakor_group)) +
  geom_histogram(color = "black", show.legend = FALSE) +
  geom_hline(yintercept = 0) +
  geom_vline(aes(xintercept = m, lty = "Átlag érték"), size = 1.5) +
  facet_wrap(~ munkakor_group, ncol = 1) +
  scale_linetype_manual(values = 2, name = NULL) +
  scale_x_continuous(labels = ~ format(.*1e3, big.mark = " ")) +
  theme(
```

```r
    legend.position = "bottom"
  ) +
  labs(x = "Havi kereset Forintban", y = "Darab")

t.test(kereset ~ munkakor_group, data = profession_df)

profession_df %>%
  GGally::ggpairs(aes(color = munkakor_group))

profession_df %>%
  total_summarise(g = munkakor_group,
                  `Átlag` = mean(kereset),
                  `Medián` = median(kereset),
                  `Szórás` = sd(kereset),
                  `Relatív szórás` = sd(kereset) / mean(kereset),
                  `Ferdeség` = moments::skewness(kereset),
                  `Csúcsosság` = moments::kurtosis(kereset),
  ) %>%
  mutate_at(-1, ~ format(round(., 2), decimal.mark = ",")) %>%
  rename(`Munkakör jellege` = 1) %>%
  knitr::kable(caption =
                 "Fizetések eloszlásának jellemzői munkakör jellege szerinti bontásban",
               align = c("l", rep("c", )))

t.test(kereset ~ nem, data = profession_df, alternative = "two.sided")

t.test(kereset ~ nem, data = profession_df, alternative = "greater")

teacher_df %>%
  total_summarise(g = nem,
                  `Átlag` = mean(kereset),
                  `Medián` = median(kereset),
                  `Szórás` = sd(kereset),
                  `Relatív szórás` = sd(kereset) / mean(kereset),
                  `Ferdeség` = moments::skewness(kereset),
                  `Csúcsosság` = moments::kurtosis(kereset),
  ) %>%
  mutate_at(-1, ~ format(round(., 2), decimal.mark = ",")) %>%
  rename(`Munkakör jellege` = 1) %>%
  knitr::kable(caption =
                 "Fizetések eloszlásának jellemzői nemek szerinti bontásban",
               align = c("l", rep("c", )))
teacher_df %>%
  select(-eletkor_group) %>%
  rename(
    `Életkor` = eletkor,
    `Végzettség` = iskvegz,
    `Munkakör` = munkakor,
    `Kereset` = kereset,
    `Nem` = nem
  ) %>%
  GGally::ggbivariate(outcome = "Nem",
                      rowbar_args = list(
```

```r
                        label_format = scales::label_percent(decimal.mark = ",", accuracy = .1)
                      )

  )

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)
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") %>%
  predict( type = "response") %>%
  cbind(teacher_df) %>%
  rename(z = 1) %>%
  tibble() %>%
  mutate(id = row_number())
matching_df <- 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)) %>%
  janitor::clean_names() %>%
  mutate(
    d = kereset_ferfi - kereset_no,
    n = n_ferfi + n_no
  )
matching_df %>%
  ungroup() %>%
  summarise(ate = weighted.mean(d, n, na.rm = T),
            atet = weighted.mean(d, n_ferfi, na.rm = TRUE),
            atet_no = weighted.mean(d, n_no, na.rm = TRUE))
```