

### 3. Forduló

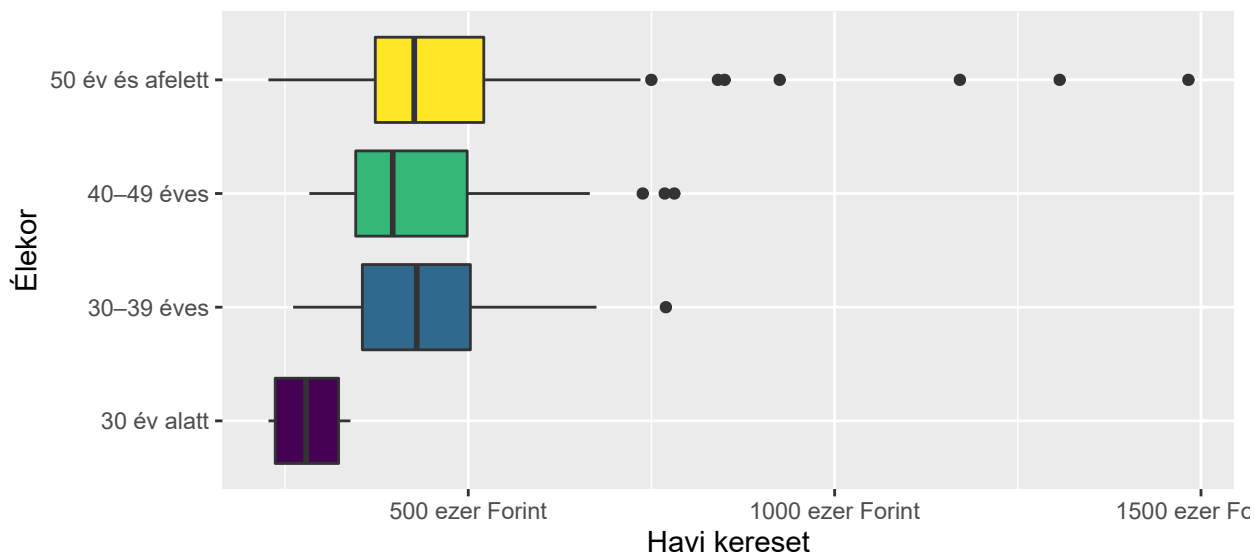
StatWars

Cash Money

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 felettiekénél 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 (lásd: 1. táblázat), míg a többiben nem érte el az 50%-ot.

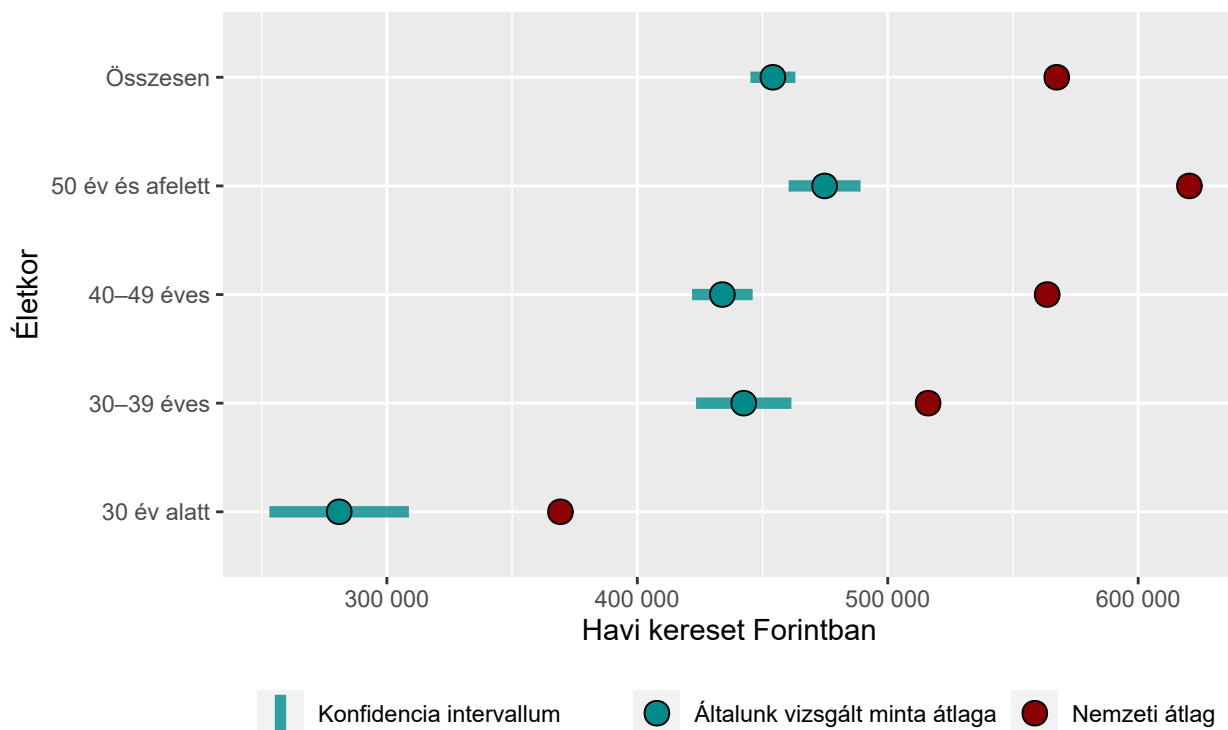


1. ábra. Egyetemi/főiskolai oktatók havi keresetének dobozábrája életkor csoportok szerinti bontásban.

Referenciaértékként a KSH 2020-as 2410-es FEOR '08 kódja (Egyetemi, főiskolai oktató, tanár) alá tartozó értékeket vizsgáltuk<sup>1</sup>. 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

<sup>1</sup>[https://www.ksh.hu/stadat\\_files/mun/hu/mun0059.html](https://www.ksh.hu/stadat_files/mun/hu/mun0059.html)

intézményben, a legnagyobb különbséget azonban az 50 év feletti korosztályban tapasztalhattuk, nagyjából 150 ezer forintos átlagos eltéréssel.



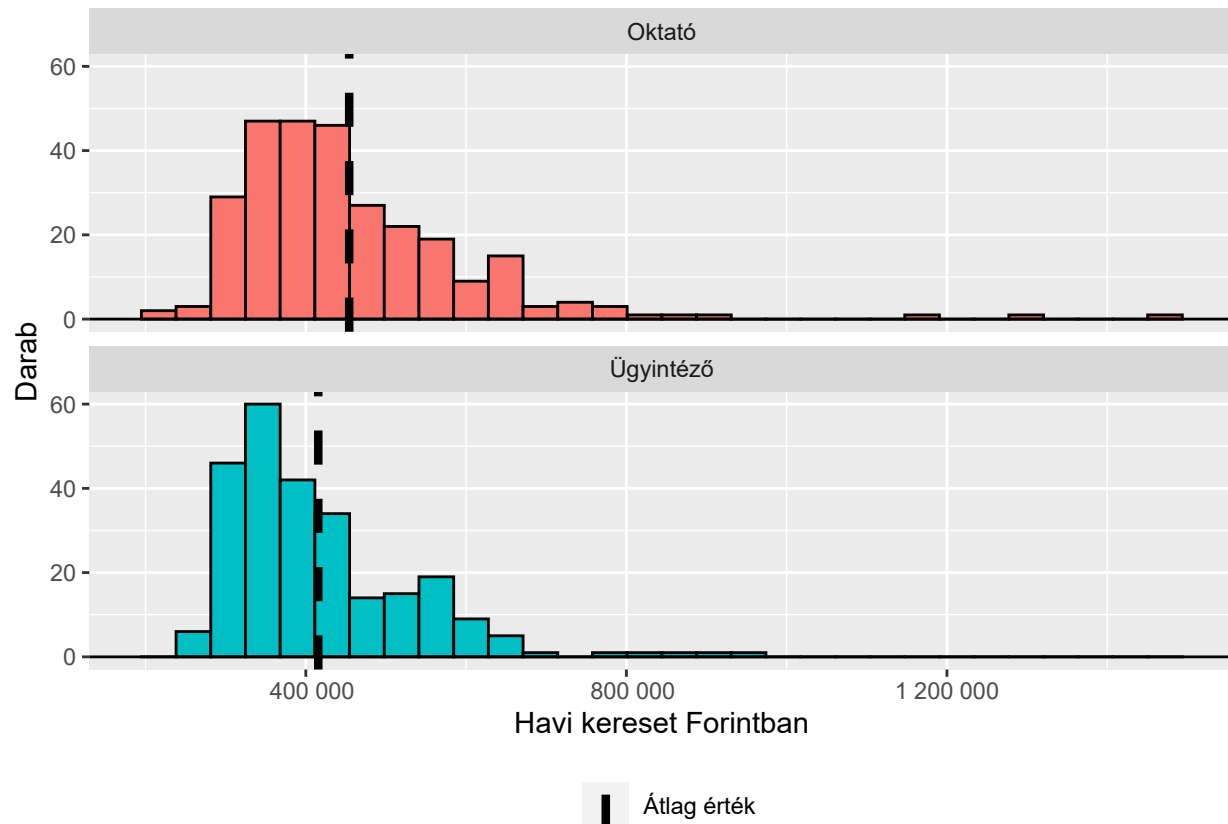
2. ábra. Általunk vizsgált minta és az országos havi átlag keresetek összehasonlítása életkor szerinti bontásban

1. táblázat: Leíró statisztikák a életkor szerinti bontásban

Életkor	Átlag	Medián	Szórás	Relatív szórás	Ferdeség	Csúcsosság	Elemzés
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

Az oktatók és ügyintézők keresetek szerinti eloszlása a 3. ábra két hisztogramon összehasonlítható. A grafikonokról vizuálisan leolvashatjuk, hogy az oktatók bruttó átlagfizetése magasabb, mint az ügyintézőké. Ez a különbség nagyjából 38 600 forintot jelent, melynek szignifikanciáját egy kétmintás t-próba segítségével vizsgáltuk meg. A kétmintás t-próba teszt-statisztikájának értéke 3,3504, ami alapján minden gyakorlatban bevett szignifikancia szinten elvetésre kerül az a nullhipotézis, miszerint a két foglalkoztatási csoportban megegyezne a sokasági átlag. A mediánbér szintén hasonló (39 000 forintos) különbséget mutat.



3. ábra. Fizetések hisztogramja munkakör jellege szerinti bontásban

Az eloszlás magasabb momentumait vizsgáljuk, elmondhatjuk, hogy a két eloszlás szórása közötti különbség ugyan számottevő, a relatív szórásuk hasonló. Így azt mondhatjuk, hogy a magasabb béreket fizető oktatói munkakör is hasonló átlagos eltérést mutat az átlagbérhez képest, hogyha az eltérés mértékét arányosítjuk a bérek nagyságához.

A ferdeség értéke ( $\alpha_3$ ) az eloszlások típusában hasonlóságot mutat: mindkét munkakör fizetései jobbra ferde eloszlást követnek, azonban a ferdeség értéke magasabb az oktatóknál, ez azt jelenti, hogy a felfelé kiugró értékek gyakoribbak az oktatóknál. Kiemelendők még a csúcsossági értékek ( $\alpha_4$ ), mely mindkét munkakör esetében nagyobb, mint 3, így leptokurtikusnak mondhatjuk az eloszlását. Azonban hogyha egymással hasonlítjuk össze a két eloszlást, az oktatók kereseteloszlásának kurtózisa kétszeres értéket vesz fel az ügyintézőkhöz képest, így ebből megállapíthatjuk, hogy az oktatók körében gyakoribbak a kiugró értékek, mint az ügyintézőknél. Ez az eredmény egybevág a korábbi bekezdésben tárgyaltakkal, az eltérések leginkább a pozitív irányban mutatkoznak. Az ábráról is leolvasható, hogy 1 millió forint feletti bruttó fizetések is előfordulnak az oktatók körében.

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

Ahhoz, hogy összehasonlítsuk a két nem közötti keresetbeli különbségeket 3 típusú módszertant használtunk, részben a szakirodalom, részben pedig a saját statisztikai tanulmányaink alapján. Azon különbségeknél, amelyekben a magyarázóváltozó teljes hatását szeretnénk vizsgálni, egy- és kétoldalas t-próbát alkalmaztunk a férfi és női bruttó bérek átlagára, valamint klasszikus legkisebb négyzetek (OLS) regressziót futtattunk a kereset, mint célváltozót magyarázva kizárólag a női nem által alkotott dummy változó bevonásával. Ezzel szemben a közvetlen hatás vizsgálatakor az előbbi regressziót kibővítettük a többi magyarázóváltozóval (életkor, munkakör, iskolai végzettség), így minden változó együtthatója a saját közvetlen hatását mutatja be. (Ilyen módszertannal dolgozott például a nemek közötti kereseti különbségeket kutató cikkben Meara et al. (2017)<sup>2</sup>) Végül pedig a közvetlen hatás második módszereként párosítással is kiszámoltuk a nem más változók hatásától szűrt befolyásának értékét, amely módszertant például Hirsch et al. (2013)<sup>3</sup> is használtak.

3. táblázat: Fizetések eloszlásának jellemzői nemek szerinti bontásban

Nem	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

A férfi és női fizetések kétmintás statisztikai próbájának elvégzésével választ kapunk arra, hogy van-e szignifikáns (nem véletlen mintavételi ingadozásnak betudható) eltérés.

Kétoldali alternatív hipotézis mellett a kétmintás t-próba teszt-statisztikájának értéke 2,2425 (p-érték = 0,0257), ami alapján 5%-os szignifikancia szinten elutasíthatjuk, hogy a férfi és női fizetések sokasági átlaga megegyezne. Egyoldali 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 lenne, mint a nőké.

Az egyszerű statisztikai próba elvégzésén túl azért van szükség további eszközökre, mert a nem számos más a fizetésekről szóló táblázatban megtalálható változóval együtt mozog. Az ismérvek megoszlását nemenkénti bontásban a függelékben található 6. ábra mutatja be. Az ábrán megmutatkozik, hogy mely más magyarázóváltozók alapján különböznek a férfi és női munkavállalók. Megállapíthatjuk, hogy a legalacsonyabb átlagos keresettel rendelkező munkakörbe tartozó betanított/segédmunkát végző munkavállalók túlnyomó többségében férfiak. Azonban a korábban bemutatott ügyintézők ugyanakkor 73%-ban nők. Ez alapján feltételezhetjük, hogy a munkakörök bizonyos dummy változói és a nem dummy változói között erős kapcsolat húzódik. A végzettség szerint minden végzettségi szinten nagyjából hasonló arányban találhatók nők és férfiak, így a kapcsolat szorossága nem mondható magasnak. Az azonban, hogy ezen két változó, valamint az életkor mennyire mutatkozik meg a kereseti különbségekben, szükséges megvizsgálnunk a közvetett hatásokat a regresszióban.

A következőkben arra tér ki elemzésünk, hogy mi magyarázza a jövedelmi eltéréseket: mekkora különbséget okoz önmagában az, hogy valaki nő, és mekkora különbség tulajdonítható a különböző nemeket jellemző eltérő magyarázó változók. A fizetésekben megnyilvánuló nemi diszkrimináció szempontjából ez az igazán fontos kérdés, ugyanis más kontroll változók bevonásával képet kapunk arra, hogy tényleg azért keres-e valaki kevesebbet, mert nő, vagy lehet, hogy harmadik változó áll a háttérben.

Az elemzéshez a klasszikus legkisebb négyzetek módszert (OLS) alkalmaztuk, amelyben először mindösszesen a női nem dummy változóját szerepeltettük, amellyel a többi magyarázóváltozót hatását nem különítettük el,

<sup>2</sup>Meara, Katie and Pastore, Francesco and Pastore, Francesco and Webster, Allan, Is the Gender Pay Gap in the US Just the Result of Gender Segregation at Work?. Available at SSRN: <https://ssrn.com/abstract=2949109> or <http://dx.doi.org/10.2139/ssrn.2949109>

<sup>3</sup>Hirsch, B., König, M., & Möller, J. (2013). Is there a gap in the gap? Regional differences in the gender pay gap. *Scottish Journal of Political Economy*, 60(4), 412-439.

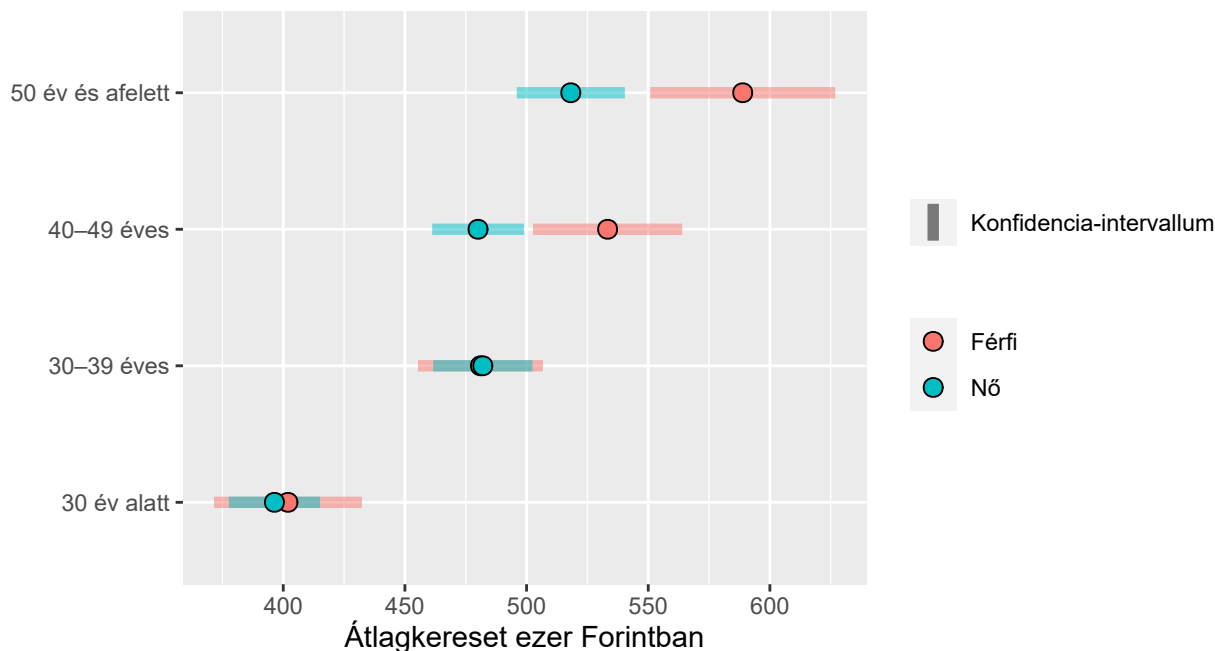
hanem az együttthatóban a teljes hatás szerepelt. Ebben a modellben 5%-os szinten szignifikáns a nem hatása a várható fizetésre, egy nő bruttó bére várhatóan 46680 forinttal alacsonyabb, mint egy férfi. Ez a nemből adódó **teljes hatás**. A modell becsült paramétereinek értékét a függelékben található 5. táblázat mutatja be. A modell jóságát a függelékben található 8. táblázat ismerteti (1. modell).

Következő lépésként olyan regressziót készítettünk, amelyben az összes rendelkezésünkre álló magyarázó változóra kontrolláltunk (2. modell). Ebben a bővebb modellben szintén szignifikáns lett a női nem hatása a keresetre, a koefficiens  $-32000$  Forint. Ezen modell jóságát szintén az 8. táblázat mutatja be.

Mivel a modellben a végzettség keresetre gyakorolt becsült hatása nem volt szignifikánsan különböző 0-tól, így készítettünk olyan modellt is, amelyből azt elhagytuk (3. modell). Ebben a modellben **33 000 Forinttal keres kevesebbet egy nő**, mint férfi társa, aki minden más ismérv szerint ekvivalens. Ez a nemből adódó **közvetlen hatás**. Az  $R^2$ , korrigált  $R^2$ , AIC és más illeszkedés jóságát ismertető mutató nem romlott érdemben az inszignifikáns változó elhagyásával.

A női keresetkülönbség közvetett hatása így tehát a két együtttható különbsége, azaz a szűkített modellt figyelembe véve,  $-16680$  forint. Ez a különbség abból adódik, hogy a nők más ismérvei is olyan eloszlást mutatnak, amely miatt negatív érinti a keresetüket.

Korábbi kutatások eredményei alapján azt igazolták, hogy a férfi-női fizetések különbsége általában az életkorral való interakcióval együtt mutat reális képet<sup>4</sup>. Tehát azt mondhatjuk, hogy empirikusan az életkor növekedésével a fizetési különbségek kitágulnak. A hipotézis relevánsnak tűnt az alapján is, hogy a férfiak és nők átlagfizetésének különbsége a mi mintákban is nagyobb az idősebb korcsoportokban.



4. ábra. Átlagkereset korcsoportok és nemek szerinti bontásban

A 4. ábrán látszik, hogy a 40 évnél fiatalabbak körében nincs szignifikáns eltérés a férfiak és nők között, azonban a 40 év felettiek esetében már igen. Továbbá a 30-39 éves korosztályban még magasabb is a nők átlag fizetése, mint a férfiaké.

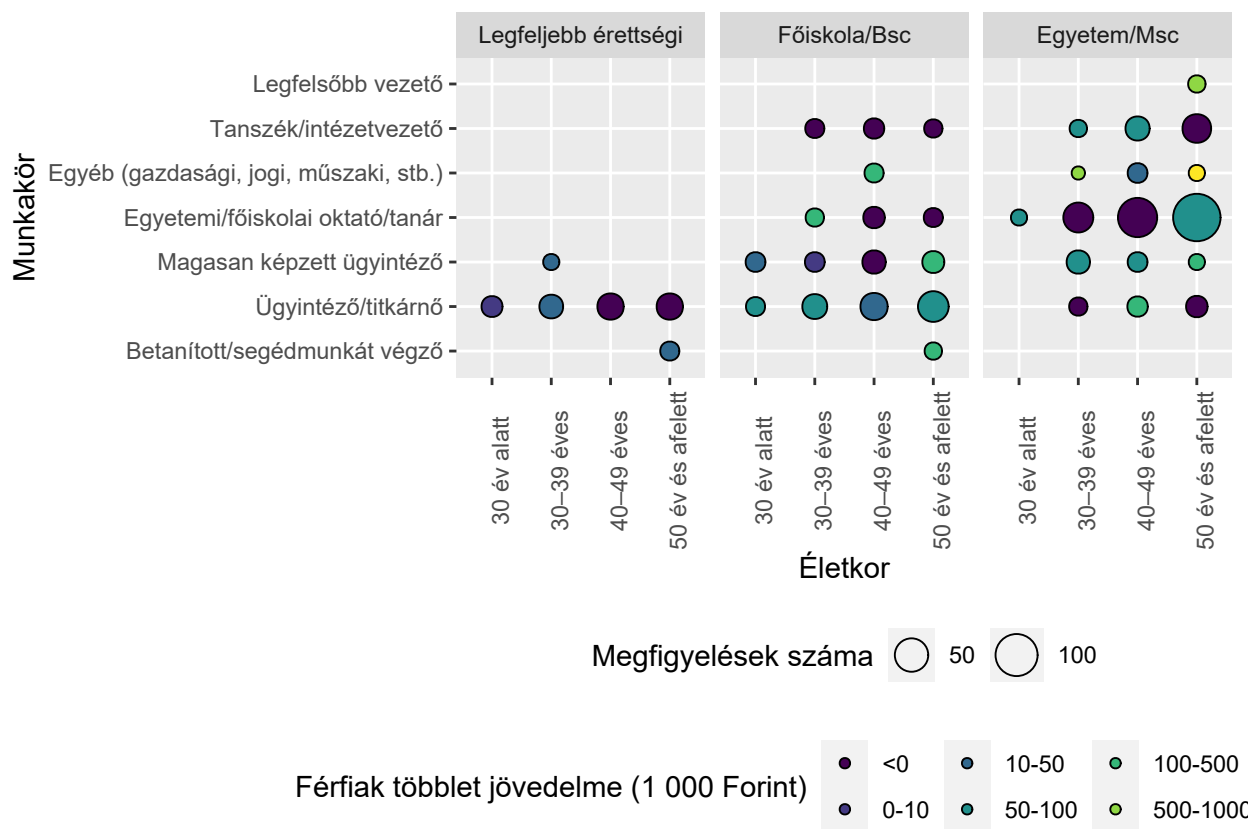
Ez alapján regressziót futtattunk a teljes modellünkön túl a korábban definiált női nem dummy változója és az életkor interakciójának bevonásával együtt is. Azonban az egyetemi dolgozók mintája alapján **az interakció nem bizonyult szignifikánsnak**. A modell koefficienseit a 4. táblázat ismerteti.

<sup>4</sup>Leaker, D. (2008). The gender pay gap in the UK. Economic & Labour Market Review, 2(4), 19-24.

4. táblázat: Kereset regressziós modellje a nem és az életkor interakciójának felhasználásával

Változó	Koefficiens	Standard hiba	T-statisztika	P-érték
Konstans	356,14	76,79	4,64	0,00%
Nő	-13,74	100,12	-0,14	89,09%
Életkor	3,91	1,60	2,45	1,46%
Életkor*Nő	-0,73	2,07	-0,35	72,43%

Másik általunk választott módszertan a többi változó hatásának kiszűrése a nem fizetésre való hatásából a párosítás. Ebben az esetben olyan férfiakat és nőket hasonlítottunk össze, akik más jellemzőik szerint teljesen megegyeznek egymással. Az egyetlen számszerű kontroll változónk a életkor volt, melyet az elemzés során már többször alkalmaztunk korcsoportokra osztást alkalmavá, így a **nyers párosítás** mellett döntöttünk<sup>5</sup>. Első lépésként csoportokba rendeztük a megfigyeléseket korosztály, végzettség és beosztás szerint. Ezt követően kiszámoltuk a férfiak és a nők átlagát, illetve a megfigyelések számát az összes csoportban, majd vettük a férfiak és nők átlag keresetének a különbségét. Ezeket a különbségeket ismerteti a 5. ábra.



5. ábra. Férfiak bértöbblete életkor, munkakör és iskolai végzettség szerinti bontásban

Az kapott különbségeknek a megfigyelések számával vett súlyozott átlagaként számítottuk ki az ATE (Average Treatment Effect - átlagos kezelési hatás) értékét, amely megmutatja, hogy várhatóan hány Forinttal kap alacsonyabb fizetést egy nő, mint minden tulajdonság szerint vele megegyező férfi kollégája. Ez az érték

<sup>5</sup>Másik lehetőségünk a P-score párosítás lett volna

**31 339 Forint**, amely alig tér el a legkisebb négyzetek módszerrel kapott közvetlen hatás értékétől, tehát eredményünk robusztus.

## Konklúzió

Konklúzióként elmondhatjuk, hogy a megfigyelt egyetem oktatói és tanárai között az életkori csoportok szerint az 50 év felettiek rendelkeznek a legmagasabb átlagos fizetéssel, valamint a legtöbb kiugró érték is ebben a csoportban található. Azonban az egyetemi oktatók bruttó bére minden csoportban elmaradt az országos átlagos fizetéstől (összesen átlagosan 120 ezer forinttal), a legnagyobb különbséget pedig éppen a legidősebb korosztály csoportjában találhattuk.

Ezt követően az ügyintézők és az oktatók béreinek eloszlását hasonlítottuk össze. Megfigyeltük, hogy a különböző munkakörök átlagfizetései szignifikánsan különböznek egymástól, az oktatók 38 600 forinttal keresnek többet bruttó értékben az ügyintézőknél. Az eloszlások alakjaira pedig mindkét esetben a jobbra ferde eloszlás jellemző, azonban az oktatók esetében jóval több kiugró értéket láthatunk.

Végül az egyetem összes munkatársának fizetését hasonlítottuk össze, a nemi fizetési különbségekre fókuszálva. Azt mondhatjuk el, hogy a kétoldali t-próba alapján 5%-os szignifikancia szinten a férfi és női fizetések különböznek egymástól, átlagosan 46 690 forinttal. Amennyiben a teljes hatást vizsgáltuk, az egyváltozós regresszió eredménye hasonló eredményt mutat, viszont a többi magyarázóváltozót bevonva a nők átlagos fizetése ceteris paribus 33 000 forinttal volt alacsonyabb a férfiakétól. Végül pedig a nyers párosítás módszer-tanát alkalmazva szintén a közvetlen hatását vizsgálva a nemi változónak 31 339 forintos eltérést kaptunk, amely megerősíti a korábbi elemzésünket.

## Függelék

### Kiegészítő táblázatok és ábrák

5. táblázat: Kereset regressziós modellje a nem változó felhasználásával

Változó	Koefficiens	Standard hiba	T-statisztika	P-érték
Konstans	539,40	17,40	31,01	0,00%
Nem	-46,68	22,12	-2,11	3,52%

6. táblázat: Kereset regressziós modellje az összes regresszor felhasználásával

Változó	Koefficiens	Standard hiba	T-statisztika	P-érték
Konstans	665,17	39,88	16,68	0,00%
Nő	-32,01	15,42	-2,08	3,84%
Életkor	2,04	0,75	2,71	0,68%
Főiskola/Bsc	31,42	21,70	1,45	14,82%
Legfeljebb érettségi	11,03	14,96	0,74	46,12%
Egyéb (gazdasági, jogi, műszaki, stb.)	1068,00	63,66	16,78	0,00%
Egyetemi/főiskolai oktató/tanár	473,25	60,03	7,88	0,00%
Legfelsőbb vezető	187,23	47,97	3,90	0,01%
Magasan képzett ügyintéző	69,51	30,73	2,26	2,40%
Tanszék/intézetvezető	180,96	29,75	6,08	0,00%
Ügyintéző/titkárnő	221,11	25,35	8,72	0,00%

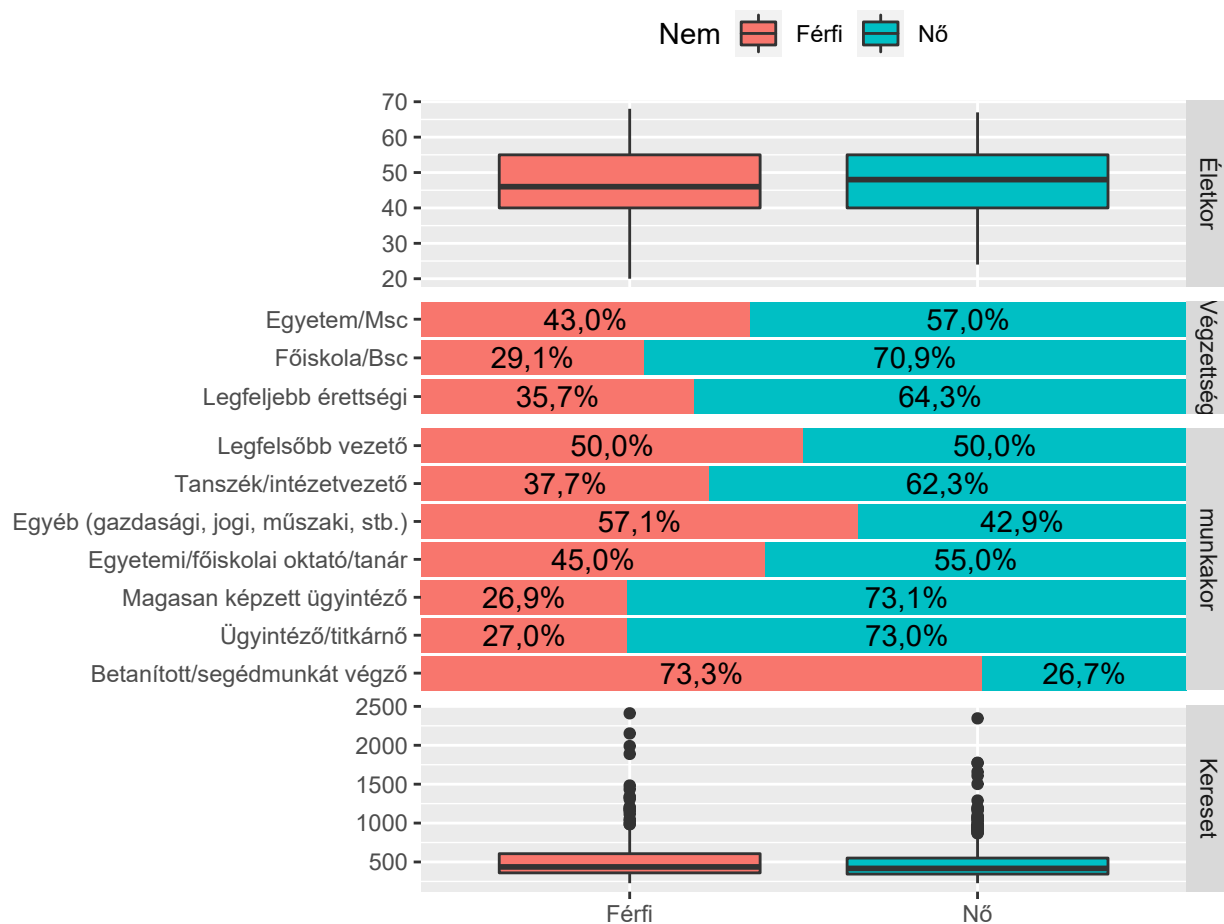
7. táblázat: Kereset regressziós modellje az iskolai végzettség kihagyásával

Változó	Koefficiens	Standard hiba	T-statisztika	P-érték
Konstans	224,55	61,74	3,64	0,03%
Nő	-33,13	15,41	-2,15	3,19%
Életkor	2,17	0,75	2,91	0,37%
Egyéb (gazdasági, jogi, műszaki, stb.)	630,08	62,87	10,02	0,00%
Egyetemi/főiskolai oktató/tanár	139,90	49,33	2,84	0,47%
Legfelsőbb vezető	1444,31	104,29	13,85	0,00%
Magasan képzett ügyintéző	176,62	53,78	3,28	0,11%
Tanszék/intézetvezető	616,72	53,11	11,61	0,00%
Ügyintéző/titkárnő	96,35	50,65	1,90	5,76%

8. táblázat: Modellek jósága

	R-négyzet	Korrigált R-négyzet	Globális F-próba p-értéke	AIC
Modell 1	0,69%	0,53%	3,52%	9100,64
Modell 2	55,14%	54,44%	0,00%	8604,39
Modell 3	54,92%	54,35%	0,00%	8603,64





6. ábra. A vizsgált adattábla változóinak nemenkénti megoszlása

## Alkalmazott R kódok

```

1  # setup -----
2
3  library(tidyverse)
4  library(GGally)
5  options(scipen = 999)
6
7  # data -----
8
9  teacher_df <- readxl::read_excel("3. forduló STAT WARS UNI.xlsx", sheet = 2) %>%
10   mutate(
11     nem = case_when(
12       nem == 1 ~ "Férfi",
13       nem == 2 ~ "Nő"
14     ),
15     életkor = as.integer(életkor),
16     iskveg = factor(iskveg, levels = 1:3, ordered = TRUE),
17     iskveg = fct_relabel(iskveg, function(l) {
18       case_when(
19         l == 1 ~ "Legfeljebb érettségi",
20         l == 2 ~ "Főiskola/Bsc",
21         l == 3 ~ "Egyetem/Msc"
22       )}),
23     munkakor = factor(munkakor, levels = 7:1, ordered = TRUE),
24     munkakor = fct_relabel(munkakor, function(l) {
25       case_when(
26         l == 1 ~ "Legfelsőbb vezető",
27         l == 2 ~ "Tanszék/intézetvezető",
28         l == 3 ~ "Egyéb (gazdasági, jogi, műszaki, stb.)",
29         l == 4 ~ "Egyetemi/főiskolai oktató/tanár",
30         l == 5 ~ "Magasan képzett ügyintéző",
31         l == 6 ~ "Ügyintéző/titkárnő",
32         l == 7 ~ "Betanított/segédmunkát végző"
33       )})
34   )
35
36 teacher_df <- teacher_df %>%
37   mutate(
38     életkor_group = cut(életkor, breaks = c(c(0, 3, 4, 5)*10, Inf), right = FALSE,
39                          labels = FALSE),
40     életkor_group = factor(életkor_group, levels = 1:4, ordered = TRUE),
41     életkor_group = fct_relabel(életkor_group, function(l) {
42       case_when(
43         l == 1 ~ "30 év alatt",
44         l == 2 ~ "30-39 éves",
45         l == 3 ~ "40-49 éves",
46         l == 4 ~ "50 év és afelett"
47       )
48     })
49   )
50
51 # utils -----
52

```

```

53 total_summarise <- function(x, g, ...) {
54   # original summarise function from tidyverse, but contains TOTAL row
55
56   bind_rows(
57     x %>%
58       group_by({{ g }}) %>%
59       summarise(...) %>%
60       ungroup(),
61     x %>%
62       summarise(...) %>%
63       mutate(g = "Összesen") %>%
64       select(g, everything()) %>%
65       rename("{ g }" := 1)
66   )
67
68 }
69
70 print_model <- function(model, label, var_names) {
71   # print the table of coefs in a nice way
72   broom::tidy(model) %>%
73     mutate_at(2:4, ~ format(round(., 2), decimal.mark = ",")) %>%
74     mutate(
75       p.value = scales::percent(p.value, decimal.mark = ",", accuracy = .01),
76       term = c("Konstans", var_names)
77     ) %>%
78     set_names("Változó", "Koefficiens", "Standard hiba", "T-statisztika", "P-érték") %>%
79     knitr::kable(caption = label, align = c("l", rep("c", 4)))
80 }
81
82 teacher_df %>%
83   filter(munkakor == "Egyetemi/főiskolai oktató/tanár") %>%
84   ggplot(aes(kereset, életkor_group, fill = életkor_group)) +
85   geom_boxplot(show.legend = FALSE) +
86   scale_x_continuous(labels = ~ str_c(., " ezer Forint")) +
87   labs(x = "Havi kereset", y = "Élekor")
88
89 national_avg <- rio::import("https://www.ksh.hu/stadat_files/mun/hu/mun0059.csv") %>%
90   # download data from KSH website: https://www.ksh.hu/stadat_files/mun/hu/mun0059.html
91   tibble() %>%
92   janitor::row_to_names(2) %>%
93   select(2, starts_with("2020")) %>%
94   rename_all(str_remove_all, "2020 Korcsoport ") %>%
95   rename_all(str_remove_all, "2020 ") %>%
96   rename(profession = 1, Összesen = Együtt) %>%
97   filter(str_detect(profession, "Egyetemi")) %>%
98   mutate_at(-1, str_remove, " ") %>%
99   mutate_at(-1, as.numeric) %>%
100  pivot_longer(-1, names_to = "életkor_group") %>%
101  select(-profession)
102
103 compare_df <- bind_rows(
104   teacher_df %>%
105   filter(munkakor == "Egyetemi/főiskolai oktató/tanár") %>%

```

```

106   total_summarise(eletkor_group,
107                   value = mean(kereset)*1e3,
108                   s = sd(kereset*1e3),
109                   n = n()
110   ) %>%
111   mutate(type = "Általunk vizsgált minta átlaga"),
112   national_avg %>%
113   mutate(type = "Nemzeti átlag", s = NA, n = NA)
114 )
115
116 compare_df %>%
117   mutate(
118     lb = value - s/(n^.5),
119     ub = value + s/(n^.5),
120   ) %>%
121   mutate(
122     etekor_group = factor(eletkor_group, levels = c("30 év alatt", "30-39 éves",
123                                                    "40-49 éves", "50 év és afelett",
124                                                    "Összesen"), ordered = TRUE)
125   ) %>%
126   ggplot() +
127   geom_linerange(aes(xmin = lb, xmax = ub, y = etekor_group,
128                    color = "Konfidencia intervallum"), size = 2, alpha = .8) +
129   geom_point(aes(value, etekor_group, fill = type), shape = 21, size = 4) +
130   scale_fill_manual(values = c("cyan4", "red4")) +
131   scale_color_manual(values = "cyan4") +
132   scale_x_continuous(labels = ~ format(., big.mark = " ")) +
133   labs(x = "Havi kereset Forintban", y = "Életkor", color = NULL, fill = NULL) +
134   theme(
135     legend.position = "bottom"
136   )
137
138 teacher_df %>%
139   filter(munkakor == "Egyetemi/főiskolai oktató/tanár") %>%
140   select(-etekor, -) %>%
141   GGally::ggpairs(aes(color = etekor_group))
142 total_summarise(teacher_df, etekor_group,
143                 `Átlag` = mean(kereset),
144                 `Medián` = median(kereset),
145                 `Szórás` = sd(kereset),
146                 `Relatív szórás` = sd(kereset) / mean(kereset),
147                 `Ferdesség` = moments::skewness(kereset),
148                 `Csúcsosság` = moments::kurtosis(kereset),
149                 `Elemszám` = n()
150 ) %>%
151   mutate_at(-1, ~ format(round(., 2), decimal.mark = ",")) %>%
152   rename(Életkor = 1) %>%
153   knitr::kable(caption = "Leíró statisztikák a életkor szerinti bontásban",
154               align = c("l", rep("c", 7)))
155 profession_df <- teacher_df %>%
156   filter(
157     munkakor %in% c("Ügyintéző/titkárnő", "Magasan képzett ügyintéző",
158                    "Egyetemi/főiskolai oktató/tanár")

```

```

159 ) %>%
160 mutate(munkakor_group = ifelse(
161   munkakor == "Egyetemi/főiskolai oktató/tanár", "Oktató", "Ügyintéző"
162 ))
163
164 profession_df %>%
165   group_by(munkakor_group) %>%
166   mutate(
167     m = mean(kereset),
168     m = ifelse(!duplicated(m), m, NA)
169   ) %>%
170   ggplot(aes(kereset, fill = munkakor_group)) +
171   geom_histogram(color = "black", show.legend = FALSE) +
172   geom_hline(yintercept = 0) +
173   geom_vline(aes(xintercept = m, lty = "Átlag érték"), size = 1.5) +
174   facet_wrap(~ munkakor_group, ncol = 1) +
175   scale_linetype_manual(values = 2, name = NULL) +
176   scale_x_continuous(labels = ~ format(.*1e3, big.mark = " ")) +
177   theme(
178     legend.position = "bottom"
179   ) +
180   labs(x = "Havi kereset Forintban", y = "Darab")
181
182 t.test(kereset ~ munkakor_group, data = profession_df)
183
184 profession_df %>%
185   GGally::ggpairs(aes(color = munkakor_group))
186
187 profession_df %>%
188   total_summarise(g = munkakor_group,
189     `Átlag` = mean(kereset),
190     `Medián` = median(kereset),
191     `Szórás` = sd(kereset),
192     `Relatív szórás` = sd(kereset) / mean(kereset),
193     `Ferdeség` = moments::skewness(kereset),
194     `Csúcsosság` = moments::kurtosis(kereset),
195   ) %>%
196   mutate_at(-1, ~ format(round(., 2), decimal.mark = ",")) %>%
197   rename(`Munkakör jellege` = 1) %>%
198   knitr::kable(caption =
199     "Fizetések eloszlásának jellemzői munkakör jellege szerinti bontásban",
200     align = c("l", rep("c", 7)))
201
202 t.test(kereset ~ nem, data = profession_df, alternative = "two.sided")
203
204 t.test(kereset ~ nem, data = profession_df, alternative = "greater")
205
206 teacher_df %>%
207   total_summarise(g = nem,
208     `Átlag` = mean(kereset),
209     `Medián` = median(kereset),
210     `Szórás` = sd(kereset),
211     `Relatív szórás` = sd(kereset) / mean(kereset),

```

```

212         `Ferdesség` = moments::skewness(kereset),
213         `Csúcsosság` = moments::kurtosis(kereset),
214     ) %>%
215     mutate_at(-1, ~ format(round(., 2), decimal.mark = ",")) %>%
216     rename(`Nem` = 1) %>%
217     knitr::kable(caption =
218         "Fizetések eloszlásának jellemzői nemek szerinti bontásban",
219         align = c("l", rep("c", 7)))
220
221 teacher_df %>%
222     group_by(eletkor_group, nem) %>%
223     summarise(m = mean(kereset), s = sd(kereset), n = n()) %>%
224     mutate(
225         cl = m - s/(n^.5),
226         ch = m + s/(n^.5)
227     ) %>%
228     ggplot() +
229     aes(m, életkor_group) +
230     geom_linerange(aes(xmin = cl, xmax = ch, color = nem,
231         linetype = "Konfidencia-intervallum"),
232         size = 2, alpha = .5) +
233     geom_point(aes(fill = nem), shape = 21, size = 3) +
234     guides(color = guide_none()) +
235     labs(
236         y = NULL, x = "Átlagkereset ezer Forintban",
237         fill = NULL, linetype = NULL
238     )
239
240 teacher_df %>%
241     mutate(
242         nem = nem == "Nő",
243         int = nem*életkor
244     ) %>%
245     lm(formula = kereset ~ nem + életkor + int) %>%
246     print_model(
247         label =
248             "Kereset regressziós modellje a nem és az életkor interakciójának felhasználásával",
249         var_names = c("Nő", "Életkor", "Életkor*Nő")
250     )
251
252 matching_df <- teacher_df %>%
253     group_by(nem, iskvegz, munkakor, életkor_group) %>%
254     summarise(kereset = mean(kereset), n = n()) %>%
255     pivot_wider(names_from = nem, values_from = c(kereset, n)) %>%
256     janitor::clean_names() %>%
257     mutate(
258         d = kereset_ferfi - kereset_no,
259         n = n_ferfi + n_no
260     )
261
262 matching_df %>%
263     filter(!is.na(d)) %>%
264     arrange(d) %>%

```

```

265 mutate(
266   d_cat = cut(d, breaks = c(-Inf, 0, 10, 50, 100, 500, 1000, Inf),
267             right = FALSE, labels = FALSE),
268   d_cat = factor(d_cat, levels = 1:7, ordered = TRUE),
269   d_cat = fct_relabel(d_cat, function(l) {
270     case_when(
271       l == 1 ~ "<0",
272       l == 2 ~ "0-10",
273       l == 3 ~ "10-50",
274       l == 4 ~ "50-100",
275       l == 5 ~ "100-500",
276       l == 6 ~ "500-1000",
277       l == 7 ~ ">1000"
278     )
279   })
280 ) %>%
281 ggplot(aes(eletkor_group, munkakor, fill = d_cat, size = n)) +
282 geom_point(shape = 21, color = "black") +
283 facet_wrap(~ iskvegzs) +
284 scale_size(range = c(2, 8))+
285 theme(
286   axis.text.x = element_text(angle = 90),
287   legend.position = "bottom",
288   legend.box = "vertical"
289 ) +
290 labs(
291   x = "Életkor",
292   y = "Munkakör",
293   fill = "Férfiak többlet jövedelme (1 000 Forint)",
294   size = "Megfigyelések száma"
295 )
296
297 matching_df %>%
298 ungroup() %>%
299 summarise(ate = weighted.mean(d, n, na.rm = T),
300           atet = weighted.mean(d, n_ferfi, na.rm = TRUE),
301           atet_no = weighted.mean(d, n_no, na.rm = TRUE)
302 )
303
304 fit1 <- teacher_df %>%
305   lm(formula = kereset ~ nem)
306
307 fit1 %>%
308   print_model(
309     label = "Kereset regressziós modellje a nem változó felhasználásával",
310     var_names = "Nem"
311   )
312
313 fit2 <- teacher_df %>%
314   lm(formula = kereset ~ .-eletkor_group)
315
316 fit2 %>%
317   print_model(

```

```

318   label = "Kereset regressziós modellje az összes regresszor felhasználásával",
319   var_names = c("Nő", "Életkor", "Főiskola/Bsc", "Legfeljebb érettségi",
320                 "Egyéb (gazdasági, jogi, műszaki, stb.)",
321                 "Egyetemi/főiskolai oktató/tanár", "Legfelsőbb vezető",
322                 "Magasan képzett ügyintéző",
323                 "Tanszék/intézetvezető",
324                 "Ügyintéző/titkárnő")
325 )
326
327 fit3 <- teacher_df %>%
328   mutate_if(is.factor, as.character) %>%
329   lm(formula = kereset ~ .-eletkor_group - iskvegz)
330
331 fit3 %>%
332   print_model(
333     label = "Kereset regressziós modellje az iskolai végzettség kihagyásával",
334     var_names = c("Nő", "Életkor",
335                   "Egyéb (gazdasági, jogi, műszaki, stb.)",
336                   "Egyetemi/főiskolai oktató/tanár", "Legfelsőbb vezető",
337                   "Magasan képzett ügyintéző",
338                   "Tanszék/intézetvezető",
339                   "Ügyintéző/titkárnő")
340   )
341
342 tibble(fit = list(fit1, fit2, fit3)) %>%
343   transmute(
344     model = str_c("Modell ", row_number()),
345     map_df(fit, broom::glance)
346   ) %>%
347   select(
348     model,
349     `R-négyzet` = r.squared,
350     `Korrigált R-négyzet` = adj.r.squared,
351     `Globális F-próba p-értéke` = p.value,
352     AIC
353   ) %>%
354   column_to_rownames("model") %>%
355   mutate_at(1:3, ~ scales::percent(., accuracy = .01, decimal.mark = ",")) %>%
356   mutate_at(4, ~ format(round(., 2), decimal.mark = ",")) %>%
357   knitr::kable(caption = "Modellek jósága", align = c("l", rep("c", 4)))
358 teacher_df %>%
359   select(-eletkor_group) %>%
360   rename(
361     `Életkor` = életkor,
362     `Végzettség` = iskvegz,
363     `Munkakör` = ,
364     `Kereset` = kereset,
365     `Nem` = nem
366   ) %>%
367   GGally::ggbivariate(outcome = "Nem",
368                       rowbar_args = list(
369                         label_format = scales::label_percent(decimal.mark = ",",
370                                                                 accuracy = .1)

```



371 )  
372  
373 )