Nagybeadandó

Szendrei József

2021 12 12

# Matematikai statisztika beadandó

## Adathalmaz betöltése

Első lépésben betöltöm az adatsetet.

Most hogy betöltöttük nézzük meg alaposabban, mit is tartalmaz ez a file.

## 'data.frame': 500 obs. of 6 variables:  
## $ nem : chr "férfi" "férfi" "nő" "férfi" ...  
## $ kor : num 38 39 46 27 47 46 36 37 33 55 ...  
## $ város : chr "Budapest" "Budapest" "Békés" "Budapest" ...  
## $ foglalkozás : chr "Pincér" "Pincér" "Pincér" "Szabó" ...  
## $ testmagasság: num 172 175 174 169 165 163 159 167 179 168 ...  
## $ jövedelem : num 249 120 488 274 122 ...

Az adatsetünk sorainak és oszlopainak száma: 500, 6  
Az első pár sora:

## nem kor város foglalkozás testmagasság jövedelem  
## 1 férfi 38 Budapest Pincér 172 249  
## 2 férfi 39 Budapest Pincér 175 120  
## 3 nő 46 Békés Pincér 174 488  
## 4 férfi 27 Budapest Szabó 169 274  
## 5 nő 47 Békés Szabó 165 122

Minden sor egy ember adatait tartalmazza, az adatok pedig a következők: nem, kor, város, foglalkozás, testmagasság, jövedelem

**Azaz összességében 500 ember fent említett adatait látjuk ebben az adathalmazban**

# Az adathalmaz vizsgálata

## Hibák keresése

Nézzük meg először, hogy vannak-e nullértékek. A táblázatban összesen 0 nullérték található, Szerencsénkre nincs benne.

Nézzük meg, hogy vannak-e üres értékek benne. A táblázatban összesen 1 üres érték található.

*Mivel ez a későbbi elemzés során hibát okozna, ezért ezt az értéket javítani kell.*

## Hiba kijavítása

Első lépésben megkeressük az üres értéket. Az érték a 2500 . helyen található az adathalmazban ami pont az 50. sorban van.

## nem kor város foglalkozás testmagasság jövedelem  
## 498 férfi 49 Békés Szabó 193 228  
## 499 nő 36 Budapest Szabó 163 334  
## 500 nő 37 Békés Pincér NA 209

Az üres adat helyére az oszlopnak az átlagát fogom tenni.

A javítás sikerességét gyorsan ellenőrizhetjük is.

## nem kor város foglalkozás testmagasság jövedelem  
## 498 férfi 49 Békés Szabó 193 228  
## 499 nő 36 Budapest Szabó 163 334  
## 500 nő 37 Békés Pincér 173 209

## Kiugró értékek ellenőrzése

A 3 számot tartalmazó oszlopra érdemes megnézni a kiugró értékeket, azaz a maximumot és a minimumot.

Az életkorok minimuma: 22év, maximuma: 68év, amik teljesen normális értékek.

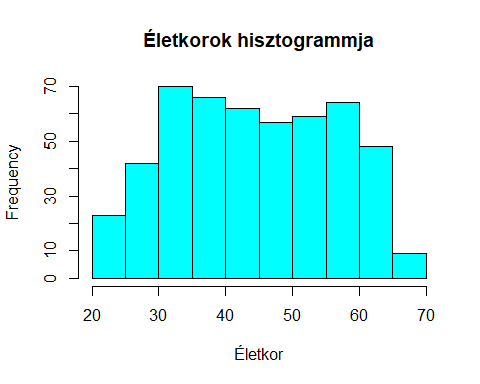
A testmagasságok minimuma: 147cm, maximuma pedig 199cm, amik szintén teljesen normális értékek.

Végül pedig a jövedelmek minimuma: 105ezer forint, maximuma pedig 3.273 millió forint, amik ugyancsak teljesen normális értékek.

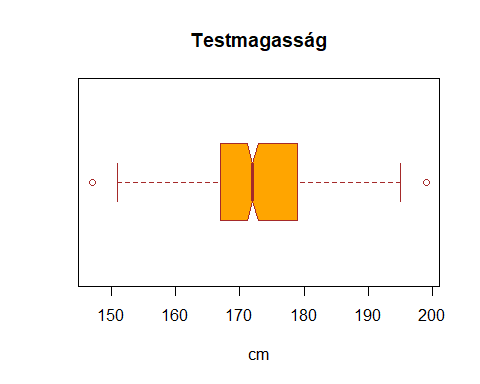
### Az adathalmazunk most már hibamentes, kezdhetjük az elemzést.

# Leíró statisztikai elemzés

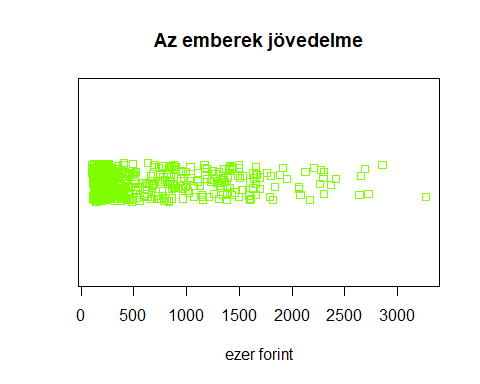
## Egyszerű ábrák



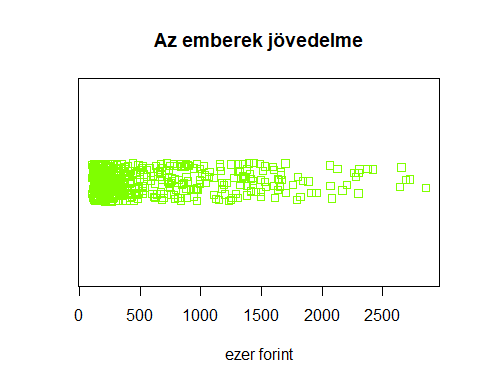
Az emebrek átlagos kora 44.8 év az átlagtól egyes emberek életkora 12.1 évvel tér el. Az emberek negyede legfeljebb 35 éves és leggalább 55 éves.



Az emberek átlagos testmagassága 172.6 cm, az átlagtól az egyes emberek magassága 5%-kal ami 8.8 cmrel tér el. Az emberek fele 167 cm és 179 cm közt van. Az ábra két végén kiugró értéknek jelenhet meg az adat, de szeintem a 147cm és a 199 cm teljesen életszerű normális magasság, ezért benne hagyom az adatok közt.



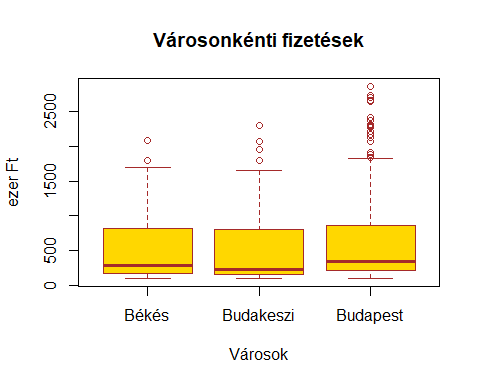
Az emberek egyik felének legfeljebb, másik felének legalább 314 ezer forint a jövedelme. Az átlag jövedelem = 590.44 ezer forint.Ábrán is jól látható, hogy az átlag az alsó kvartilisben található és a maximum jövedelem pedig eléggé kigró adat, ezért most korrigálok és kicserélem a maximum jövedelmet a jövedelmek átlagára. Úgy gondolom, hogy ekkora jövedelem nem elképzelhetetlen, de az adathalmaz szempontjából mégis ez a jó döntés.



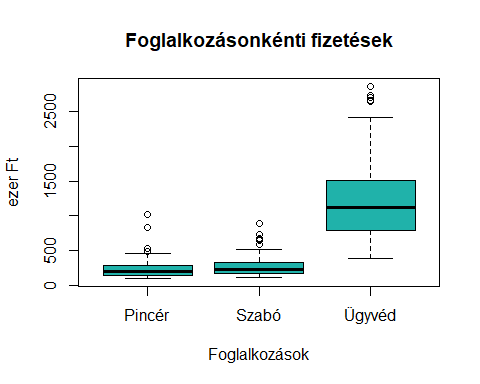
Két végletet figyelembe véve a legtöbb jövedelemmel rendelkező ember 27.22-szer keres annyit mint a legekevesebb jövedelemmel rendelkező ember, a különbség 2.75millió forint.

## Specifikus ábrák

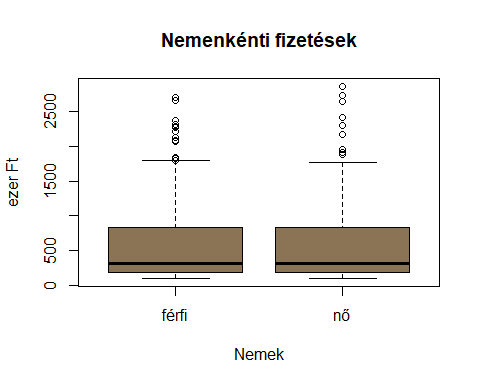
Nézzük meg a fizetéseket különböző foglalkozások, települések és nemek esetén.



Jól látszik az ábrán, hogy bár az emberek háromnegyed részének Budapesten 863.5 Ft, Békésen 801.75 Ft és Budakeszin 797 Ft a maximum jövedelme, de Budapesten sokkal több a felfelé kiugró adat.

` 

A foglalkozásokat megfigyelve tisztán látszik, hogy az ügyvédek átlagjövedelme meghaladja a pincérek vagy szabók átlagjövedelmét. A legkevesebb jövedelemmel rendelkező ügyvéd jövedelme(388) is nagyobb, minta pincérek felső kvartilise (283)



A nemenkénti boxplot válasz ad arra a kérdésre, hogy a nők vagy a férfiak keresnek átlagosan többet. A férfiak átlagfizetése 615 ezer forint, míg a nőké 563 ezer forint.

##   
## férfi nő   
## 208 292

A megkérdezett emberek nagyobb része nő volt.

##   
## Békés Budakeszi Budapest  
## Pincér 36 34 99  
## Szabó 30 32 94  
## Ügyvéd 44 44 87

A gyakorisági táblázat megmutatja, hogy az adott helyen az adott foglalkozásból hány ember szerepel az adathalmazban. Jól látható, hogy a megkérdezett emberek nagyjából fele, egész pontosan 280 ember fővárosban él, a legtöbb megkérdezett embernek pedig ügyvéd a foglalkozása.

### konfidenciaintervallumok

Magasság alsó konfidenciaintervalluma 171.7830058, felső intervalluma 173.3369942

Jövedelem alsó konfidenciaintervalluma 534.9952741, felső intervalluma 635.1607259

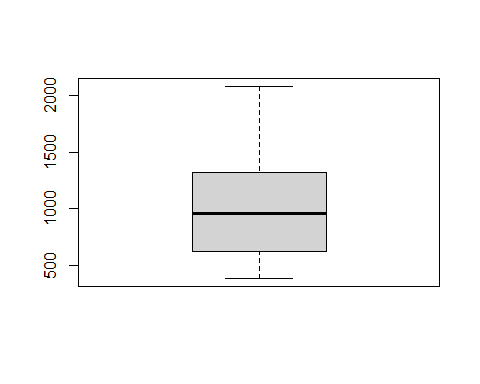
# Hipotézisvizsgálat normális eloszlást feltételezve

## Békési ügyvédek jövedelem

Azt szeretnénk vizsgálni, hogy a békési ügyvédek jövedelme kevesebb-e mint 1 millió forint.

H0: m = m0  
h1: m < m0

Segítségképp kirajzolunk egy ábrát



Egymintás, ismeretlen szórású, normális eloszlású eset -> t-próbát fogunk alkalmazni.

##   
## One Sample t-test  
##   
## data: jövbeuv  
## t = -0.078309, df = 43, p-value = 0.469  
## alternative hypothesis: true mean is less than 1000  
## 95 percent confidence interval:  
## -Inf 1100.94  
## sample estimates:  
## mean of x   
## 995.0682

A próba alapján 95%-os megbízhatósággal H0-t elfogadjuk, azaz nem tudjuk elvetni

95%-os megbízhatóság esetén azt mondhatjuk, hogy a békési ügyvédek fizetése egyenlő 1 millió forinttal

## Budakeszi és Békés jövedelme

Ismerjük a Békésen élő emberek jövedelmét illetve a Budakeszin elő emberek jövedelmét. Állíthatjuk-e azt, hogy Békésen jobban élnek-e az emebrek? H0 m1=m2  
H1 m1>m2

Kétmintás, ismeretlen szórású, normális eloszlású eset -> F-próbát fogunk alkalmazni.

##   
## F test to compare two variances  
##   
## data: dfbk$jövedelem and dfbe$jövedelem  
## F = 1.2432, num df = 109, denom df = 109, p-value = 0.2574  
## alternative hypothesis: true ratio of variances is not equal to 1  
## 95 percent confidence interval:  
## 0.8523482 1.8132187  
## sample estimates:  
## ratio of variances   
## 1.243179

Az F-próba p-értéke nagyobb minden értelmes szignifikanciaszintnél, ezért nem tudjuk elvetni a szórások egyezőségét. Így kétmintás t próbát fogunk alkalmazni.

##   
## Two Sample t-test  
##   
## data: dfbk$jövedelem and dfbe$jövedelem  
## t = 0.21524, df = 218, p-value = 0.5851  
## alternative hypothesis: true difference in means is less than 0  
## 95 percent confidence interval:  
## -Inf 125.5423  
## sample estimates:  
## mean of x mean of y   
## 538.9636 524.4909

P-érték 58% ami nagyobb minden értelmes szignifikanciaszintnél, gyenge döntés 95%-os megbízhatóság eseten nem állíthatjuk, hogy jobban élnek Békésen mint Budakeszin.

## Pincérek és Szabók jövedelme

Ismerjük az adathalmazban a pincérek és a szabók jövedelmét. Állíthatjuk-e azt, hogy a jövedelmükben nincs eltérés?

H0: m0 = m1  
H1: m0 != m1

Mivel a szórások ismeretlenek ezért F próbára lesz szükség.

##   
## F test to compare two variances  
##   
## data: jövpi and jövsz  
## F = 1.3887, num df = 279, denom df = 109, p-value = 0.04794  
## alternative hypothesis: true ratio of variances is not equal to 1  
## 95 percent confidence interval:  
## 1.003028 1.880987  
## sample estimates:  
## ratio of variances   
## 1.388692

Mivel a p-érték kisebb mint 5%-os szignifikanciaszintnél, ezért el tudjuk vetni a szórások egyezőségét. Welch próbát kell végrehajtanunk

##   
## Welch Two Sample t-test  
##   
## data: jövpi and jövsz  
## t = 1.4146, df = 233.25, p-value = 0.1585  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -34.57083 210.63641  
## sample estimates:  
## mean of x mean of y   
## 626.9964 538.9636

A p-érték 15,8%, így például 5%-os elsőfajú hiba valószínűsége esetén nem tudjuk elvetni H0-t, Gyenge döntés.  
95%-os megbízhatósági szintnél azt állíthatjuk, hogy a szabók és a pincérek jövedelme egyezik.

## Budapesti ügyvédek és békési ügyvédek összehasonlítása

Megvizsgáljuk a budapesti és a békési ügyvédek jövedelmeit. Először eldöntjük, melyik állítást érdemes vizsgálni.

## [1] 1438 1510 2068 1305 1327 415 1415 1178 2858 838 520 1430 2418 1335 2373  
## [16] 723 1763 1415 887 835 865 818 1822 2277 1655 630 790 900 1292 1887  
## [31] 1658 1360 712 970 2300 2693 782 1170 2725 1198 1598 1380 947 2208 878  
## [46] 1348 1160 1172 2127 1232 863 882 747 2640 523 1368 1432 1840 665 1912  
## [61] 780 1655 1635 1352 1613 747 2167 1495 1192 590 892 1068 843 1213 975  
## [76] 438 2267 1603 1642 1042 2653 1538 880 870 1698 1670 2307

## [1] 540 994 624 499 815 675 762 1346 1011 1324 836 992 842 628 1310  
## [16] 388 603 1527 1274 1491 1261 486 945 1412 843 414 1156 1700 1408 1247  
## [31] 1254 525 1530 1793 752 1108 555 465 676 1380 494 850 2080 968

##   
## Shapiro-Wilk normality test  
##   
## data: jövbkuv  
## W = 0.91605, p-value = 0.003521

## [1] 1382.782

## [1] 995.0682

Mivel a dudapesti ügyvédek átlagjövedelme nagyobb ezért azt éri meg vizsgálni. H0: m0=m1  
H1: m1>m0

A minták függetlenek és a szórások nem ismertek, ezért előzetes F-próbát kell először végrehajtani a szórások egyezőségére vonatkozóan.

##   
## F test to compare two variances  
##   
## data: jövbpuv and jövbeuv  
## F = 2.0463, num df = 86, denom df = 43, p-value = 0.0109  
## alternative hypothesis: true ratio of variances is not equal to 1  
## 95 percent confidence interval:  
## 1.184949 3.374680  
## sample estimates:  
## ratio of variances   
## 2.04629

Mivel a p-érték 1% ezért el tudjuk vetni a H0-t azaz a szórások egyezőségét.

##   
## Welch Two Sample t-test  
##   
## data: jövbpuv and jövbeuv  
## t = 4.3156, df = 115.96, p-value = 1.683e-05  
## alternative hypothesis: true difference in means is greater than 0  
## 95 percent confidence interval:  
## 238.7509 Inf  
## sample estimates:  
## mean of x mean of y   
## 1382.7816 995.0682

Mivel a p-érték 1% ezért el tudjuk vetni a H0-t, erős döntés született Azt állíthatjuk, hogy a budapesti ügyvédek jövedelme nagyobb mint a békési ügyvédeké 95%-os megbízhatósági sznten

# Hipotézisvizsgálat nem normális eloszlást feltételezve

## Békési szabók és budakeszin élő szabók jövedelme

Arra vagyunk kíváncsiak, hogy a békési szabók és a budakeszin élő szabók jövedelme megegyezik-e?  
H0: m0 = m1  
H1: m0 != m1

##   
## Shapiro-Wilk normality test  
##   
## data: dfbesz$jövedelem  
## W = 0.83315, p-value = 0.0002806

##   
## Shapiro-Wilk normality test  
##   
## data: dfbksz$jövedelem  
## W = 0.92545, p-value = 0.0293

Normalitásvizsgálattal megvizsgáltuk, hogy egyik minta sem származik normális eloszlásból.

## Warning in wilcox.test.default(dfbesz$jövedelem, dfbksz$jövedelem): cannot  
## compute exact p-value with ties

##   
## Wilcoxon rank sum test with continuity correction  
##   
## data: dfbesz$jövedelem and dfbksz$jövedelem  
## W = 484.5, p-value = 0.9551  
## alternative hypothesis: true location shift is not equal to 0

Mivel a p érték nagyobb mint 5% ezért nem tudjuk elvetni H0-t 95%-os megbízhatósággal azt állíthatjuk, hogy a Budakeszin és a Békésen élő szabók jövedelme megegyezik.

# lineáris model illesztése

## Stringet tartalmazó mezők lecserélése

## Kovarianciamátrix létrehozása

## kor testmagasság jövedelem ferfi no pincer szabo ugyved  
## kor 1.000 -0.015 0.376 -0.012 0.012 -0.193 -0.052 0.242  
## testmagasság -0.015 1.000 0.006 0.523 -0.523 0.050 -0.001 -0.049  
## jövedelem 0.376 0.006 1.000 0.045 -0.045 -0.439 -0.379 0.804  
## ferfi -0.012 0.523 0.045 1.000 -1.000 -0.003 -0.069 0.070  
## no 0.012 -0.523 -0.045 -1.000 1.000 0.003 0.069 -0.070  
## pincer -0.193 0.050 -0.439 -0.003 0.003 1.000 -0.481 -0.524  
## szabo -0.052 -0.001 -0.379 -0.069 0.069 -0.481 1.000 -0.494  
## ugyved 0.242 -0.049 0.804 0.070 -0.070 -0.524 -0.494 1.000  
## budapest 0.008 -0.041 0.083 -0.028 0.028 0.037 0.058 -0.093  
## budakeszi -0.024 0.071 -0.043 0.051 -0.051 -0.032 -0.024 0.056  
## bekes 0.015 -0.022 -0.057 -0.017 0.017 -0.012 -0.045 0.056  
## budapest budakeszi bekes  
## kor 0.008 -0.024 0.015  
## testmagasság -0.041 0.071 -0.022  
## jövedelem 0.083 -0.043 -0.057  
## ferfi -0.028 0.051 -0.017  
## no 0.028 -0.051 0.017  
## pincer 0.037 -0.032 -0.012  
## szabo 0.058 -0.024 -0.045  
## ugyved -0.093 0.056 0.056  
## budapest 1.000 -0.599 -0.599  
## budakeszi -0.599 1.000 -0.282  
## bekes -0.599 -0.282 1.000

## Lineáris model illesztése, minden paraméterrel

##   
## Call:  
## lm(formula = jövedelem ~ kor + testmagasság + ferfi + no + pincer +   
## szabo + ugyved + budapest + budakeszi + bekes, data = rdf)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -964.86 -141.95 -15.85 113.59 1472.10   
##   
## Coefficients: (3 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -130.070 316.878 -0.410 0.68164   
## kor 8.840 1.191 7.423 5.06e-13 \*\*\*  
## testmagasság 4.841 1.853 2.612 0.00927 \*\*   
## ferfi -48.767 33.227 -1.468 0.14283   
## no NA NA NA NA   
## pincer -934.009 34.812 -26.830 < 2e-16 \*\*\*  
## szabo -927.066 35.003 -26.486 < 2e-16 \*\*\*  
## ugyved NA NA NA NA   
## budapest 187.569 34.977 5.363 1.26e-07 \*\*\*  
## budakeszi 17.719 41.880 0.423 0.67241   
## bekes NA NA NA NA   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 309.9 on 492 degrees of freedom  
## Multiple R-squared: 0.7086, Adjusted R-squared: 0.7044   
## F-statistic: 170.9 on 7 and 492 DF, p-value: < 2.2e-16

## [1] 8.000 5744.112

Szignifikáns mezőket megtartjuk, a nem szignifikánsakat kidobjuk.

##   
## Call:  
## lm(formula = jövedelem ~ kor + testmagasság + pincer + szabo +   
## budapest, data = rdf)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -994.27 -135.14 -14.88 114.35 1481.73   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 94.175 278.591 0.338 0.7355   
## kor 8.880 1.190 7.460 3.92e-13 \*\*\*  
## testmagasság 3.447 1.573 2.191 0.0289 \*   
## pincer -929.762 34.683 -26.807 < 2e-16 \*\*\*  
## szabo -921.184 34.785 -26.482 < 2e-16 \*\*\*  
## budapest 178.595 28.093 6.357 4.68e-10 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 310 on 494 degrees of freedom  
## Multiple R-squared: 0.7072, Adjusted R-squared: 0.7043   
## F-statistic: 238.7 on 5 and 494 DF, p-value: < 2.2e-16

## [1] 6.000 5742.461

Ez tűnik a legjobb választásnak, de megpróbáljuk még kihagyni a testmagasságot.

##   
## Call:  
## lm(formula = jövedelem ~ kor + pincer + szabo + budapest, data = rdf)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -999.7 -144.2 -13.2 119.8 1456.2   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 688.302 63.989 10.757 < 2e-16 \*\*\*  
## kor 8.880 1.195 7.432 4.74e-13 \*\*\*  
## pincer -925.356 34.757 -26.623 < 2e-16 \*\*\*  
## szabo -918.890 34.903 -26.327 < 2e-16 \*\*\*  
## budapest 175.803 28.172 6.240 9.38e-10 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 311.2 on 495 degrees of freedom  
## Multiple R-squared: 0.7044, Adjusted R-squared: 0.702   
## F-statistic: 294.9 on 4 and 495 DF, p-value: < 2.2e-16

## [1] 5.000 5745.295

Rosszabb lett a model, maradunk az előzőnél.

### Az ember jövedeleme függ az ember korától, testmagasságától, attől hogy szabó-e vagy pincér és hogy budapesti-e

## Logikusabb megközelítés

Azokat az adatokat hagyom bent amik szerintem fontosak a jövedelemhez

##   
## Call:  
## lm(formula = jövedelem ~ kor + testmagasság + ferfi + ugyved +   
## budapest, data = rdf)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -964.94 -143.27 -17.59 119.36 1472.25   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1057.386 317.686 -3.328 0.000939 \*\*\*  
## kor 8.847 1.185 7.468 3.73e-13 \*\*\*  
## testmagasság 4.871 1.848 2.636 0.008658 \*\*   
## ferfi -48.758 33.152 -1.471 0.142000   
## ugyved 930.673 30.277 30.738 < 2e-16 \*\*\*  
## budapest 178.806 28.032 6.379 4.11e-10 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 309.3 on 494 degrees of freedom  
## Multiple R-squared: 0.7085, Adjusted R-squared: 0.7055   
## F-statistic: 240.1 on 5 and 494 DF, p-value: < 2.2e-16

## [1] 6.000 5740.338

Látható, hogy a modellem kicsit jobb mint az előző esetben, tényleg ezektől az adatoktól függ a jövedelema legjobban.

### Az adatok pedig a: kor, testmagasság, férfi nem, budapesti lakhely

## Becslés

Becsléshez létrehozok egy modelt minden paraméterrel.

Egyesével megnézzük egy 178 cm magas 43 éves férfi milyen településen, melyik szakmával mennyi jövedelmet ér el.

Egy 178 cm magas 43 éves budapesti pincér jövedelme: 307.78

Egy 178 cm magas 43 éves budapesti szabó jövedelme: 314.72

Egy 178 cm magas 43 éves budapesti ügyvéd jövedelme: 1241.79

Egy 178 cm magas 43 éves Budakeszin élő pincér jövedelme:137.93

Egy 178 cm magas 43 éves Budakeszin élő szabó jövedelme:144.87

Egy 178 cm magas 43 éves Budakeszin élő ügyvéd jövedelme:1071.94

Egy 178 cm magas 43 éves békési pincér jövedelme:120.21

Egy 178 cm magas 43 éves békési szabó jövedelme:127.15

Egy 178 cm magas 43 éves békési ügyvéd jövedelme:1054.22