R - PROGRAMMING

Analysis of Data jobs 2020-2022

Többváltozós adatelemzési modellek
S.J Helena

Picture was created by Helena Simon with Canva Artificial intelligence

Tartalomjegyzék

Az Adattábla

A datasetemet a kaggle-ről töltöttem le: https://www.kaggle.com/datasets/ruchi798/data-science-job-salaries/data , 2023. 10.26-án.

Összesen 11 változóval rendelkezik az adattáblám, és 607 megfigyelési egységet tartalmaz.

A megfigyelési egységeim különböző nyitott data scienchez köthető állásokat mutatnak be, amiket az https://ai-jobs.net/ - en lehet találni. Látni fogjuk, az adattáblában hogy minden sor tartalmaz évet, tapasztalati szintet, fizetést, a cég elhelyezkedését, és a cég méretét is – állásonként azaz soronként is nézhetjük.

A változóim a következőek:

- **work_year:** The year the salary was paid.
- **experience_level**: - ordinális The experience level in the job during the year with the following possible values:
    - EN Entry-level
    - Junior
    - MI Mid-level
    - Intermediate SE Senior-level
    - Expert EX Executive-level
    - Director
- **employment_type**: - nominálisThe type of employement for the role:
    - PT Part-time
    - FT Full-time
    - CT Contract
    - FL Freelance
- **job_title**: The role worked in during the year.
- **salary**: The total gross salary amount paid.
- **salary_currency:** currency of the salary paid as an *ISO 4217 currency code.
- **salary_in_usd:** The salary in USD (FX rate divided by avg. USD rate for the respective year via fxdata.foorilla.com).
- **employee_residence**: Employee's primary country of residence in during the work year as an **ISO 3166 country code.
- **remote_ratio**: intervallum skála The overall amount of work done remotely, possible values are as follows:
    - 0 No remote work (less than 20%)
    - 50 Partially remote
    - 100 Fully remote (more than 80%)
- **company_location**: - nominális The country of the employer's main office or contracting branch as an **ISO 3166 country code.
- **company_size**: - ordinális The average number of people that worked for the company during the year:
    - S less than 50 employees (small)
    - M 50 to 250 employees (medium)
    - L more than 250 employees (large)

Az oszlopaim mérési skálái:

- **work_year** intervallum skála
- **experience_level**: ordinális skála
- **employment_type**: nominális skála
- **job_title**: nominális skála
- **salary**: arányskála skála
- **salary_currency:** nominális skála
- **salary_in_usd:** arányskála skála
- **employee_residence**: nominális skála
- **remote_ratio**: intervallum skála
- **company_location**: nominális skála
- **company_size**: ordinális skála

Az adatok típusai R studio környezetben:

```
'data.frame':   607 obs. of  12 variables:
 $ X                 : int  0 1 2 3 4 5 6 7 8 9 ...
 $ work_year         : int  2020 2020 2020 2020 2020 2020 2020 2020 2020 2020 ...
 $ experience_level  : chr  "MI" "SE" "SE" "MI" ...
 $ employment_type   : chr  "FT" "FT" "FT" "FT" ...
 $ job_title         : chr  "Data Scientist" "Machine Learning Scientist" "Big Data Engineer" "Prod
uct Data Analyst" ...
 $ salary            : int  70000 260000 85000 20000 150000 72000 190000 11000000 135000 125000 ...
 $ salary_currency   : chr  "EUR" "USD" "GBP" "USD" ...
 $ salary_in_usd     : int  79833 260000 109024 20000 150000 72000 190000 35735 135000 125000 ...
 $ employee_residence: chr  "DE" "JP" "GB" "HN" ...
 $ remote_ratio      : int  0 0 50 0 50 100 100 50 100 50 ...
 $ company_location  : chr  "DE" "JP" "GB" "HN" ...
 $ company_size      : chr  "L" "S" "M" "S" ...
```

Leíró statisztikai elemzés

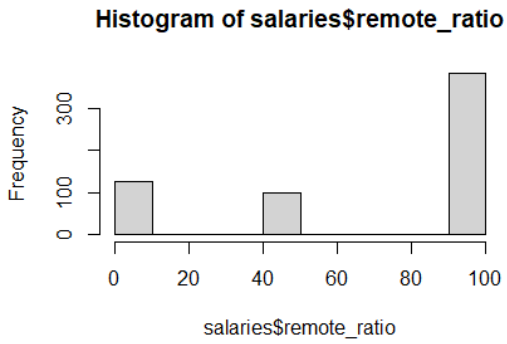
Gyakorisági táblázat – numerikus változó Kilistázza a változó összes lehetésges értékét és megadja, hogy melyik érték hányszor fordul elő a vizsgált adattáblában → megadja a változó értékeinek a gyakoriságát.

Remote cégek gyakoriságának megoszlása:
table(salaries$remote_ratio)

```
  0  50 100
127  99 381
```

Több mint az állást hirdető cégek fele remote munkavégzési állást hirdet

Histogram: – (csak numerikus lehet) függvényben megvizsgáljuk, hogy milyen a gyakorisági eloszlása a remote állásoknak:
hist(salaries$remote_ratio)



Az eredményben láthatjuk, hogy a 607 állásból több mint a fele, 100% os remote állást ajánl

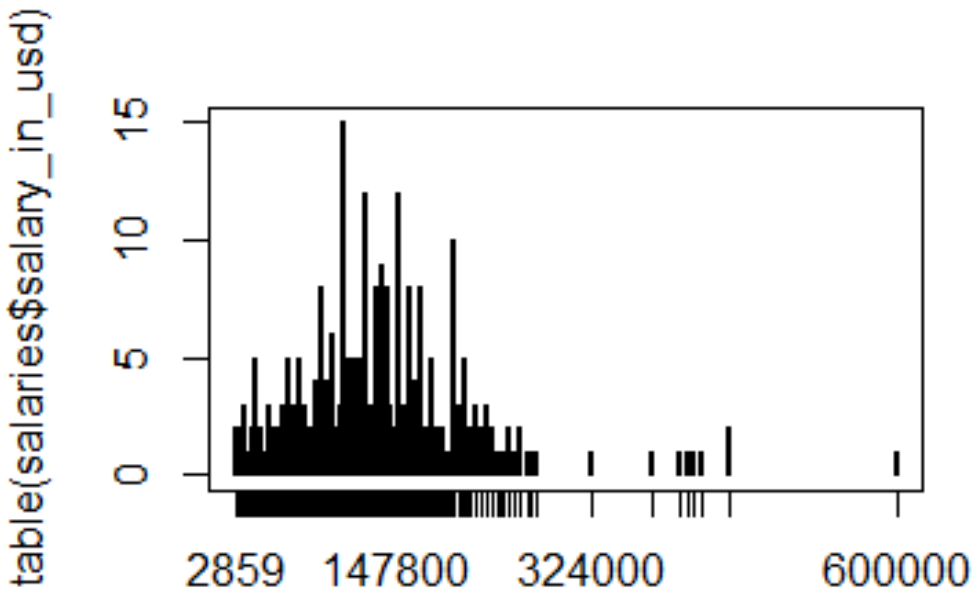
Helyzetmutatók (Módusz, Medián, Átlag) **salary_in_usd**

Módusz: egy változónak a legtöbbször előforduló értéke - ezt a gyakorisági táblázatból a legegyszerűbb leolvasni. Itt átváltunk egy másik változóra, és ezzel folytatjuk a továbbiakban

table(salaries$salary_in_usd) – láthatjuk , hogy ebben az esetben nem annyira alkalmas nekünk ez e módszer, mivel sok különböző adattal rendelkezünk, ezért nem átlátható az eredményünk

2859	4000	5409	5679	5707	5882	6072	8000	9272	9466	10000	10354	12000	12103	12901	13400	15966	16228	16904	18000	18053
1	2	1	1	1	1	2	1	1	1	2	1	3	1	1	1	1	1	1	1	1
18442	18907	19609	20000	20171	21637	21669	21844	21983	22611	24000	24342	24823	25000	25532	26005	28016	28369	28399	28476	28609
2	1	1	5	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1
29751	30428	31615	31875	32974	33511	33808	35590	35735	36259	36643	37236	37300	37825	38400	38776	39263	39916	40000	40038	40189
1	1	1	1	3	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1
40481	40570	41689	42000	42197	43331	43966	45391	45618	45760	45807	45896	46597	46759	46809	47282	47899	48000	49268	49461	49646
1	1	1	1	1	1	2	1	1	1	3	1	1	1	1	1	1	1	1	2	1
50000	50180	51064	51321	51519	52000	52351	52396	53192	54000	54094	54238	54742	54957	55000	56000	56256	56738	58000	58035	58255
5	1	1	1	1	1	3	1	1	1	1	1	1	3	2	1	1	1	3	1	1
58894	59102	59303	60000	60757	61300	61467	61896	62000	62649	62651	62726	63711	63810	63831	63900	64849	65000	65013	65438	65949
1	2	1	5	1	2	1	1	1	1	1	2	1	1	2	1	1	2	1	3	2
66022	66265	67000	68147	68428	69000	69336	69741	69999	70000	70139	70500	70912	71444	71786	71982	72000	72212	72500	73000	74000
1	1	1	1	1	1	1	2	1	2	1	1	1	1	1	1	1	1	1	1	1
74130	75000	75774	76833	76940	76958	77364	77684	78000	78526	78791	79039	79197	79833	80000	81000	81666	82500	82528	82744	82900
1	4	1	3	2	1	1	1	1	4	2	1	1	2	8	1	2	1	2	1	1
84900	85000	86703	87000	87425	87738	87932	88654	89294	90000	90320	90700	90734	91000	91237	91614	93000	93150	93427	93700	94564
1	4	1	1	1	1	4	3	1	6	5	1	2	1	1	2	1	1	1	2	1
94665	95550	95746	96113	96282	98000	98158	99000	99050	99100	99360	99703	100000	100800	101570	102100	102839	103000	103160	103691	104702
1	1	1	1	1	1	3	2	1	1	1	1	15	1	1	2	1	1	1	1	2
104890	105000	105400	106000	106260	108800	109000	109024	109280	110000	110037	110500	110925	111775	112000	112300	112872	112900	113000	113476	114047
1	5	1	1	2	1	1	1	2	5	1	1	1	1	1	1	1	4	1	1	1
115000	115500	115934	116000	116150	116914	117104	117789	118000	118187	119059	120000	120160	120600	122346	123000	124190	124333	125000	126000	126500
5	1	2	1	1	1	1	2	1	1	1	12	1	1	1	3	1	1	3	1	2
127221	128875	129000	130000	130026	130800	132000	132320	135000	136000	136600	136620	136994	137141	138000	138350	138600	140000	140250	140400	141300

Tapasztalatok gyakoriságának ábrázolása plottal
plot(table(salaries$salary_in_usd))



Itt már láthatjuk, hogy ami táblázatban leolvashatatlan, az plotban- már látható és értelmezhető eredményt hoz ki.

Medián: egy változó mediánja az az érték, aminél a változó értékeinek 50%-a kisebb, másik fele pedig nagyobb.
median(salaries$salary_in_usd)

101570

Átlag: változó elemeinek összege osztva az adatok elemszámával.
mean(salaries$salary_in_usd)

 # 112297.9

Megfigyelhetjük, hogy feljebb a medián értéke 101 ezer dollár míg az átlagé 112 ezer dollár, 10 ezer dollár különbségük van.

Kvantilisek (öröklik a medián tulajdonságait)

Nézzük meg, hogy a fizetéseknek USD-ben hová esik a Q1-es értéke, Q2,Q3-as értéke.

quantile(salaries$salary_in_usd, probs = 0.25) # *25% 62726* , tehát a fizetések első negyede eléri a 60 ezer dollárt

quantile(salaries$salary_in_usd, probs = 0.50) # *50% 101570* megegyezik az általunk régebben kiszámolt medián értékkel

quantile(salaries$salary_in_usd, probs = 0.75) # *75% 150000* ezt csak úgy érdekeségből lett kiszámolva

Summary

A summary függvényben láthatunk minden numerikus változóra levezetve a Mediánt, Átlagot, Maximumot, Minimumot, és a Kavrtilisből pedig a 1Q,3Q-t. Abban az esetben, ha nem numerikus változóról van szó, akkor csak feltünteti az adattipusokat.

summary(salaries)

```
       X           work_year    experience_level   employment_type     job_title             salary         salary_currency
 Min.   :  0.0   Min.   :2020   Length:607         Length:607         Length:607         Min.   :    4000   Length:607
 1st Qu.:151.5   1st Qu.:2021   Class :character   Class :character   Class :character   1st Qu.:   70000   Class :character
 Median :303.0   Median :2022   Mode  :character   Mode  :character   Mode  :character   Median :  115000   Mode  :character
 Mean   :303.0   Mean   :2021                                                          Mean   :  324000
 3rd Qu.:454.5   3rd Qu.:2022                                                          3rd Qu.:  165000
 Max.   :606.0   Max.   :2022                                                          Max.   :30400000
 salary_in_usd    employee_residence  remote_ratio     company_location   company_size
 Min.   :  2859   Length:607         Min.   :  0.00   Length:607         Length:607
 1st Qu.: 62726   Class :character   1st Qu.: 50.00   Class :character   Class :character
 Median :101570   Mode  :character   Median :100.00   Mode  :character   Mode  :character
 Mean   :112298                      Mean   : 70.92
 3rd Qu.:150000                      3rd Qu.:100.00
 Max.   :600000                      Max.   :100.00
```

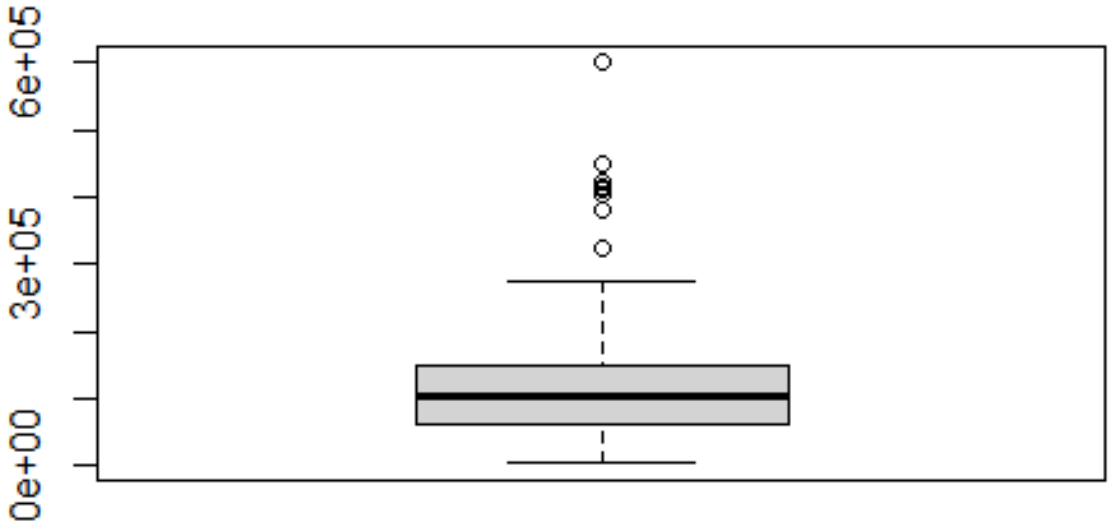
Describe: a describe alapvetően a **numerikus változók** alaposabb leíró statisztikai elemzésére alkalmas

```
> describe(salaries)
                   vars   n      mean        sd median   trimmed      mad  min      max    range  skew kurtosis       se
X                     1 607    303.00    175.37    303    303.00   225.36    0      606      606  0.00    -1.21     7.12
work_year             2 607   2021.41      0.69   2022   2021.51     0.00 2020     2022        2 -0.73    -0.66     0.03
experience_level*     3 607      3.13      1.03      3      3.28     1.48    1        4        3 -1.04    -0.10     0.04
employment_type*      4 607      2.99      0.24      3      3.00     0.00    1        4        3 -4.14    45.81     0.01
job_title*            5 607     21.96     10.49     18     21.00     7.41    1       50       49  0.88     0.40     0.43
salary                6 607 324000.06 1544357.49 115000 118919.11 68706.65 4000 30400000 30396000 13.98   244.57 62683.54
salary_currency*      7 607     14.03      4.38     17     14.67     0.00    1       17       16 -1.03    -0.38     0.18
salary_in_usd         8 607 112297.87  70957.26 101570 106157.63 62906.72 2859   600000   597141  1.66     6.26  2880.07
employee_residence*   9 607     41.41     18.27     56     43.66     0.00    1       57       56 -0.67    -1.22     0.74
remote_ratio         10 607     70.92     40.71    100     76.08     0.00    0      100      100 -0.90    -0.90     1.65
company_location*    11 607     36.89     16.03     49     39.07     0.00    1       50       49 -0.77    -1.09     0.65
company_size*        12 607      1.81      0.65      2      1.76     0.00    1        3        2  0.21    -0.73     0.03
>
```

 # n: Az elemszáma a változónak. Mint látjuk, ez itt nem egységes, hiszen nem azonos számú érettségiző volt a háromtárgyból.A hiányzó értékeket tehát nem számolta bele a változó elemszámába.
#mean: átlag értéke
#sd: A szórás értéke (itt is az N − 1-gyel osztós verziót számolja)
#median: A medián értéke
#trimmed: A nyesett átlag értéke úgy, hogy a változó értékeinek alsó és felső 10%-át kihagyja az átlagszámolás során.
#mad: Mean Absolut Deviation –> a szórás abszolútértékkel (és nem négyzet + gyökvonással) számolt értéke.
#min, max és range: A változó értékeinek minimuma és maximuma, valamint ezek különbsége, mint a változó teljesterjedelme:
#range = max - min
#skew: Az α3 mutató értéke
#kurtosis: Az α4 mutató értéke
#se: Az átlag standard hibája (standard error)

Doboz ábra, kiugró értékek, alakmutatók

Doboz Ábra:
boxplot(salaries$salary_in_usd)



Alakmutató:
A doboz ábra egy jobbra elnyúló eloszláshoz tartozik: Az adatok középső 50%-a (a doboz) az adatok minimumához közelebb van, mint a maximumához, és felfelé találhatók pontok, azaz kilógó értékek.

Kerítések:
summary(salaries$salary_in_usd)

```
> summary(salaries$salary_in_usd)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   2859   62726  101570  112298  150000  600000
>
```

Tukey féle kerítés: Tukey-féle kerítések segítségével jelzi az adatsor kilógó értékeit. A Tukey-féle kerítések alapján egy változó azon értékei minősülnek kilógónak, amelyek a változó középső 50%-hoz, tehát az IKT-hez képest túl messze helyezkednek el (akár felfelé, akár lefelé).

Alsó kerítés: Q1- 1,5* (Q3-Q1) = 62726-130911 = -68 185 alatt lefelele kilóg

Felső kerítése: Q3 + 1,5*(Q3-Q1) = 150000 + 130911 = 280 911 felett már kilógó értéknek számít

Külső kerítés:

Alsó kerítés: Q1- 3* (Q3-Q1) = 62726 - 261 822 = -199 096

Felső kerítése: Q3 + 3*(Q3-Q1) = 150000 + 261 822 = 411 822

Gyakorisági táblázat – nem numerikus (experience_level)

- **table(salaries$company_size)** - # cég méretek gyakorisága

```
  L   M   S
198 326  83
```

  - A nyitott álláshirdetéseket hirdető cégekből a fele 50 - 250 alkalmazottal rendelkezik, ellenben láthatjuk, hogy a 607 cégből csak 83 olyan cég van, ahol kevesebb mint 50 ember dolgozik, ebből következtethetünk arra, hogy a nagyobb cégeknél van több esélyünk elhelyezkedni.

- **table(salaries$experience_level)** - # tapasztalatok gyakorisága

```
 EN  EX  MI  SE
 88  26 213 280
```

  - legtöbb nyitott állásnál Senior tapasztalattal rendelkező embert keresnek ennek gyakorisága: 280, de második leggyakoribb értéknél is Mid senior alkalmazottat keresnek. Ezek alapján kevesebb esély van egy Entry level pozit találni.

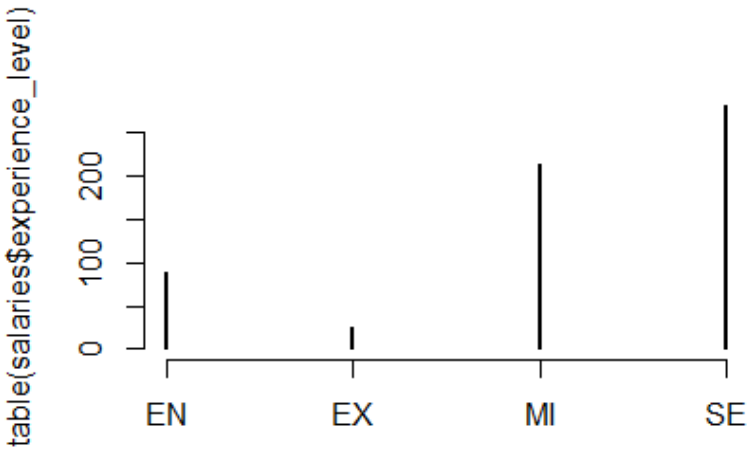
Helyzetmutatók – nem numerikus

Módusz:
table(salaries$experience_level) - # tapasztalatok gyakorisága

```
 EN  EX  MI  SE
 88  26 213 280
```

Tapasztalatok módusza 280 – Senior pozíciók

Tapasztalatok gyakoriságának ábrázolása plottal:
plot(table(salaries$experience_level))



Láthatjuk, hogy a senior pozik a legtöbbször elforduló értékek módusza – az állások legtöbbször előfrorduló esetben a senioroknak szólnak

Medián: – egy nem numerikus értéknek is van mediánja:
median(salaries$experience_level) # MI láthatjuk, hogy a tapasztalattal rendelkező állások mediánja Mid Senior

Summary:

Summary függvény esetében a nem numerikus változókat át kell konvertálnunk:

salaries$experience_level<- as.factor(salaries$experience_level)

summary(salaries)

```
       X           work_year    experience_level employment_type     job_title             salary        salary_currency
 Min.   :  0.0   Min.   :2020   EN: 88           Length:607         Length:607         Min.   :   4000   Length:607
 1st Qu.:151.5   1st Qu.:2021   EX: 26           Class :character   Class :character   1st Qu.:  70000   Class :character
 Median :303.0   Median :2022   MI:213           Mode  :character   Mode  :character   Median :  115000   Mode  :character
 Mean   :303.0   Mean   :2021   SE:280                                                 Mean   :  324000
 3rd Qu.:454.5   3rd Qu.:2022                                                          3rd Qu.:  165000
 Max.   :606.0   Max.   :2022                                                          Max.   :30400000
 salary_in_usd    employee_residence  remote_ratio    company_location   company_size
 Min.   :  2859   Length:607         Min.   :  0.00   Length:607         Length:607
 1st Qu.: 62726   Class :character   1st Qu.: 50.00   Class :character   Class :character
 Median :101570   Mode  :character   Median :100.00   Mode  :character   Mode  :character
 Mean   :112298                      Mean   : 70.92
 3rd Qu.:150000                      3rd Qu.:100.00
 Max.   :600000                      Max.   :100.00
```

Eredményül a lehívott summary függvényben az expreience level-ben mostmár láthatjuk, hogy a gyakorisági előfordulásokat írja le.

Intervallumbecslés

Az intervallumbecslések feltételei:

- A mintakiválasztási módjának véletlennek kell lennie
- Statisztikai mutató legyen torzítatlanul becsülhető – az átlag, medián, arány eleve torzítatlanok viszont a szórás tud torzítani és ha torzít akkor lefele.
    - Torzítatlan szórás: = gyök(szum((Érték-Átlag)^2)/N-1)

Az átlagos konfidencia intervallum = mintaátlag +- Standard hiba

Az általunk megkapott intervallum azt fogja megadni, hogy 90 %-os pontossággal **mekkora fizetés várható** a data science területen dolgozó embereknek.

Ez egy intervallum értéket fog adni, felső és alsó határral fog rendelkezni.

Numerikus változóra:

```
groupwiseMean(salary_in_usd ~ 1,
              data = salaries,
              conf = 0.90,
              na.rm=TRUE)
```

- teljes intervallum 107560.6 USD - 117035.2 USD - 90% os pontossággal ennyi fizetés várható egy data sciencel foglalkozó munkakaeresőnek ezen az oldalon (ezt az adatot manuális számolás után kaptam meg, viszont függvény segítségével egyszerűbben is lel ehet futtani az általam kért számításokat, így már egy kerekített eredményt kapok válaszul, ez pedig: 108 000 USD – 117 000 USD közötti összeget ad meg.

```
groupwiseMedian(salary_in_usd ~ 1,
                data = salaries, conf =  0.90)
```

- ebben a számolásban a medián intervallumot számoltam, értékül

```
  n Median Conf.level Bca.lower Bca.upper
607 102000        0.9     99000    109000
```

- az eredményemen is jól látszik, hogy nincs nagy eltérés az átlag és a medián között, ami azt mutatja meg, hogy nincsenek nagy kiugró értékek, mivel ha lenne akkor az átlag más eredményt mutatna, és a medián mivel nem hat rá annyira a kiugró érték mutatná, hogy nagy különbség van az eredmények között.

Minőségi változóra:

```
groupwiseMean(salary_in_usd ~ experience_level,
              data = salaries,
              conf = 0.90,
              na.rm=TRUE)
```

- ebben a számolásban kíváncsi voltam arra, hogyha a fizetéseket ha minőségi változó alapján akarjuk kiszámolni, azoknak a fizetési átlagát nézni, akkor milyen eredmények jönnek ki.

```
  experience_level   n   Mean Conf.level Trad.lower Trad.upper
1               EN  88  61600        0.9      53800      69500
2               EX  26 199000        0.9     160000     239000
3               MI 213  88000        0.9      80800      95200
4               SE 280 139000        0.9     133000     144000
```

- itt láthatjuk igazán, hogy mekkora fizetési különbség van tapasztalat szinten. Egy átlag Entry lvl data val foglalkolkozó szakember átlagosan 61 ezer USD-t keres, aminek az alsó határa 53 ezer is lehet míg a felső 69 ezer, fizetésben ezt elég soknak mondanám.
- Azt is láthatjuk a kiszámolt intervallumok között hogy egy data expereted fizetése 160 ezer USD és 239 ezer USD között van. Továbbra is nagyon nagy a különbség.
- Személy szerint ez az intervalumbecslés nagy segítséget tud nyujtani egy kezdő dataval foglalkozó szakembernek, mivel interjúk alkalmával betudja bizonyítani, hogy a kért fizetés reális.

```
groupwiseMedian(salary_in_usd ~ experience_level,
                data = salaries, conf =  0.90)
```

experience_level	n	Median	Conf.level	Bca.lower	Bca.upper
EN	88	56500	0.9	49500	64400
EX	26	171000	0.9	136000	216000
MI	213	76900	0.9	69700	80000
SE	280	136000	0.9	128000	140000

- ugyanezt az intervallumbecslést mediánnal végeztem el, és pontosan ezestben láthtjuk a medián és az átlag különbségét, ami abban mutatközik meg, hogy itt már a medián intervallum felső határa nem 70 ezernél van hanem csak 64-ezer dollárnál, ugyanígy az experted átlag felső határ 240 ezer míg ha medián intervallum felső határt nézek az csak 216 ezer lett. Azt jelenti, hogy a mi esetünkben bölcsebb a medián értéket figyleembe venni, mert az átlag fizetések tartalmaznak pár kiugró értéket. Ezesetben egy interjún az átlag alapján meghatározni a fizetést akár jelentheti azt is, hogy elutasítják, mert többet kért az illető mint amennyi a reális lenne. Személy szerint én a mediánt intervallum eredményeket nézném.

Arány intervallumbecslés:

```
groupwiseMean(Amerika ~ "1",
              data = salaries,
              conf = 0.90,
              na.rm=TRUE)
```

Mean	Conf.level	Trad.lower	Trad.upper
0.585	0.9	0.552	0.618

Az első számolásunkból az derül ki, hogy átlagosan a álláslehetőségek ezen az oldalon 58 % Amerikába van. Ezen kívül láthatjuk az elsó határt is ami azt jelneti, hogy az állások 55% százaléka mindenképpen Amerikába lesz, ami a való életben azt jelenti, hogy ha ezen az oldalon keresünk állást akkor minden második állás Amerikába lesz.

```
groupwiseMean(Amerika ~ job_title,
              data = salaries,
              conf = 0.90,
              na.rm=TRUE)
```

	job_title	n	Mean	Conf.level	Trad.lower	Trad.upper
1	3D Computer Vision Researcher	1	0.000	0.9	NaN	NaN
2	AI Scientist	7	0.571	0.9	0.1790	0.964
3	Analytics Engineer	4	1.000	0.9	1.0000	1.000
4	Applied Data Scientist	5	0.600	0.9	0.0778	1.120
5	Applied Machine Learning Scientist	4	0.750	0.9	0.1620	1.340
6	BI Data Analyst	6	0.833	0.9	0.4970	1.170
7	Big Data Architect	1	0.000	0.9	NaN	NaN
8	Big Data Engineer	8	0.125	0.9	-0.1120	0.362
9	Business Data Analyst	5	0.400	0.9	-0.1220	0.922
10	Cloud Data Engineer	2	0.500	0.9	-2.6600	3.660
11	Computer Vision Engineer	6	0.333	0.9	-0.0915	0.758
12	Computer Vision Software Engineer	3	0.667	0.9	-0.3070	1.640
13	Data Analyst	97	0.732	0.9	0.6570	0.807
14	Data Analytics Engineer	4	0.250	0.9	-0.3380	0.838
15	Data Analytics Lead	1	1.000	0.9	NaN	NaN
16	Data Analytics Manager	7	1.000	0.9	1.0000	1.000
17	Data Architect	11	0.818	0.9	0.5970	1.040
18	Data Engineer	132	0.644	0.9	0.5750	0.713
19	Data Engineering Manager	5	0.600	0.9	0.0778	1.120
20	Data Science Consultant	7	0.286	0.9	-0.0727	0.644
21	Data Science Engineer	3	0.000	0.9	0.0000	0.000
22	Data Science Manager	12	0.833	0.9	0.6320	1.040
23	Data Scientist	143	0.587	0.9	0.5190	0.656
24	Data Specialist	1	1.000	0.9	NaN	NaN
25	Director of Data Engineering	2	0.500	0.9	-2.6600	3.660
26	Director of Data Science	7	0.286	0.9	-0.0727	0.644
27	ETL Developer	2	0.000	0.9	0.0000	0.000
28	Finance Data Analyst	1	0.000	0.9	NaN	NaN
29	Financial Data Analyst	2	1.000	0.9	1.0000	1.000
30	Head of Data	5	0.400	0.9	-0.1220	0.922
31	Head of Data Science	4	0.750	0.9	0.1620	1.340
32	Head of Machine Learning	1	0.000	0.9	NaN	NaN
33	Lead Data Analyst	3	0.667	0.9	-0.3070	1.640
34	Lead Data Engineer	6	0.500	0.9	0.0494	0.951
35	Lead Data Scientist	3	0.333	0.9	-0.6400	1.310
36	Lead Machine Learning Engineer	1	0.000	0.9	NaN	NaN
37	Machine Learning Developer	3	0.000	0.9	0.0000	0.000
38	Machine Learning Engineer	41	0.390	0.9	0.2600	0.520
39	Machine Learning Infrastructure Engineer	3	0.333	0.9	-0.6400	1.310
40	Machine Learning Manager	1	0.000	0.9	NaN	NaN
41	Machine Learning Scientist	8	0.625	0.9	0.2780	0.972
42	Marketing Data Analyst	1	0.000	0.9	NaN	NaN
43	ML Engineer	6	0.333	0.9	-0.0915	0.758
44	NLP Engineer	1	1.000	0.9	NaN	NaN
45	Principal Data Analyst	2	0.500	0.9	-2.6600	3.660
46	Principal Data Engineer	3	1.000	0.9	1.0000	1.000
47	Principal Data Scientist	7	0.571	0.9	0.1790	0.964
48	Product Data Analyst	2	0.000	0.9	0.0000	0.000
49	Research Scientist	16	0.250	0.9	0.0540	0.446
50	Staff Data Scientist	1	1.000	0.9	NaN	NaN

Itt szerettem volna megtalálni, hogy Amerikán belül mely pozik milyen arányban oszlanak meg, azt láthatjuk hogy a Data science és Data Engineer felkapott állás.

Hipotézis vizsgálat:

Míg a konfidencia intervallumok - van egy mintavételem a teljes sokaságból, és ebből kiszámolok valamilyen statisztikai mutatót (átlag, medián ...) és megnézem, hogy mit mond el nekem a statisztikai mutató, statisztikai mutató sokasági értékéről.

A hipotézis vizsgálatnál másfajta logikai rendszer lesz - van egy előfeltevésem egy stat. mutató soksasági értékéről pl a data analyst pozíciók több int 60 % a Amerikában található miután megvan a feltételem, azután veszek egy mintát, azért, hogy az előfeltételemet ez igazolja vagy cáfolja, ami eldönti hogy igaz vagy hamis-e az előfeltétel

előfeltétel = hipotézis

Az én feltételem: **Átlag(salary_in_usd) >90000**
#minta az átlagban
mean(salaries$salary_in_usd) # ez igaz mert 112 ezer dollár az átlag

- egy hipotézis vizsgálat eldönthető konfidencia intervallummal is pl: ha kijönn hogy 20 ezerrel eltérő lehet az érték akkor is benne vagyunk az általunk felállított feltételezésben

A hipotézis vizsgálat lépése: (lényege, hogy a mintavételi hiba nagyságába betudható-e az értéknek amit kiszámolunk)

R Stúdió környezetben ez 3 lépésre egyszerüsíthető:

1. Állítás --> nullhipotézist csinálunk, illetve alternatív hipotézist csinálunk

2. p-érték nevű mutatószám számítása (számítunk egy p értéket eldönti, hogy melyik hipotézis jó)

3. p-érték alapján el lehet dönteni, hogy a null vagy az alternatív hipotézis igaz, ebből meglehet mondani, hogy az alap állítás igaz vagy hamis (valós állításba is tudjuk helyezni)

Vizsgálat:

Állítás: átlag fizetés > mint 90 ezer

M (sokasági átlag MÖ)
H0 - null hipotézis - ha törik ha szakad, azt mondja, hogy az átlag egyenlő azzala számmal ami a hipotézisben szerepel
H1 - alternatív hipotézis

	M>90000	M=90000	M!=90000	M<=90000	M<90000	M>=90000
H0	M=90000	M=90000	M=90000	M=90000	M=90000	M=90000
H1	M>90000	M!=90000	M!=90000	M>90000	M<90000	M<90000

Az első lépés mindig arról fog szólni, hogy az előfeltevésemet átfogalmazom null és alternatív hipotézis párrá.
1. H0: Átlag (salary_in_usd) = 90000 || H1: Átlag (salary_in_usd) > 90000 --> H1 szeretem
2. p-érték számolás lelesz, ahol figyelembe veszem a megfigyelt adatokat, mintákat
   - számolása attól függ milyen statisztikai mutatót számolok
   - p-érték (R függvénye attól függ, mi a stat. mutató MOST: átlag)

```
t.test(salaries$salary_in_usd, alternative = "greater", mu = 90000)
```

p érték p-value = 2.059e-14 --> 0.0000000002059
p érték megmutatja, hogy mi a valószínűsége hogy a H0-t elutasítani hibás döntés →
- ha azt mondom a null hipotézisre, hogy hamis akkor mennyi a valószínűsége hogy hibás

Mi a valószínűsége, hogy a H0-t elutasítani hibás döntés?

Ha a p érték kisebb, mint a kritikus szint (általában 0,05), akkor a kutatók elutasítják a nullhipotézist, és elfogadják az alternatív hipotézist.

Döntés: **p-érték kicsi --> H0 diszlájk --> H1 lájk --> eredeti állításunk a H1 ben volt --> állításunk igaznak vehetjük**

A fizetések átlaga szignifikánsan nagyobb mint 90000

Kétváltozós kapcsolatvizsgálat

Kapcsolatvizgsálat: 2 változó . x , y →Mennyire befolyásolja Y-t az X?

- például mennyire befolyásolja az árakat a company location?
- mennyire befolyásolja a fizetéseket az experience?

Egyik oszlop mennyire befolyásolja a másikat

- amiért ez eltud bonyolódnia, a mérési skálák esete - a kapcsolatvizsgálat esete mennyire befolyásolja az egyik változó a másikat

3 féle statisztikai mutatóval tudjuk megadni, attól függően hogy x,y nak milyen a mérési skálája

1. eset **Vegyes kapcsolat:** egyik változó Nominális (ordinálist is ide értem) + intervallum vagy arány mérési skálájú: arány + nominális → vegyes kapcsolat
2. eset **Asszociációs kapcsolat** - mindkét nominális vagy ordinális → nomin(ord) + nomin (ord)
3. este **Korrelációs kapcsolat**: interv/arány +interv/arány

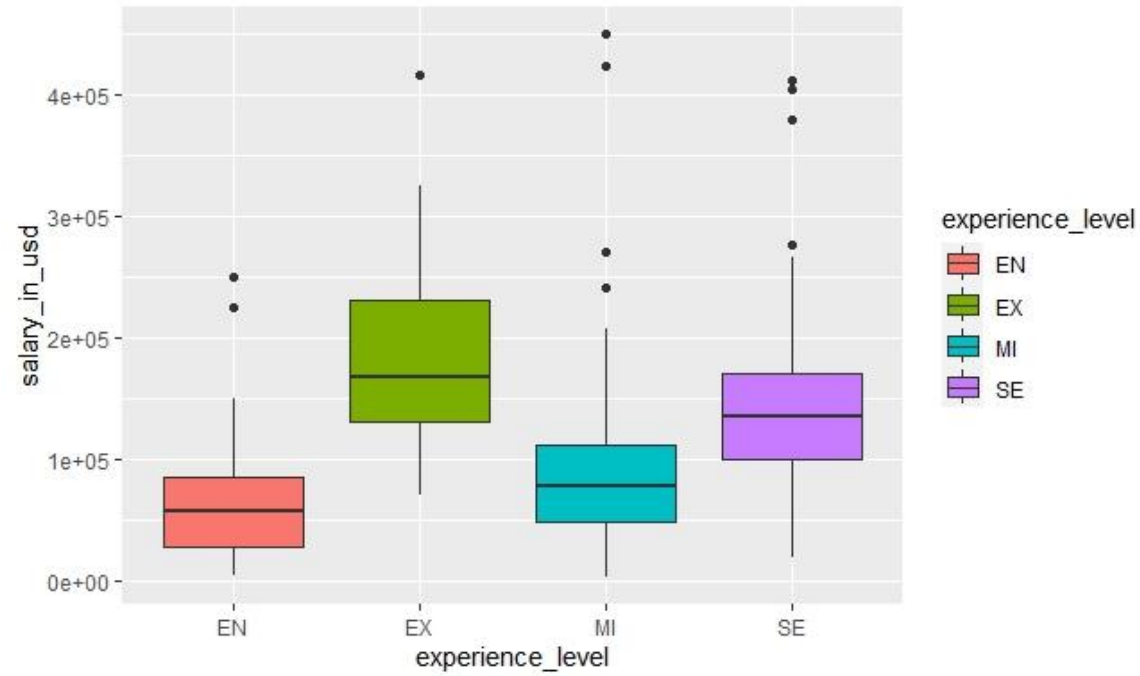
Ezek az esetek döntik el melyik statisztikai mutatókat használhatom majd a kapcsolat vizsgálatakor

Folyamata 3 fő lépés:
1. mi a kapcsolat jellege → mely országokba nagyobb a fizetés , hol jellemzőek entry level pozik → Ábrákkal válaszoljuk meg - mindegyikhez kapcsolódik egy ábra típus
2. magyarázó erő két változó között - statisztikai mutatóval mérjük le ami leírja hogy a két változó között x,y között milyen erős a kapcsolat (a megfigyelt adatok körében értelmezhető)
3. megnézzük, hogy a kapcsolat létezik-e szebb szóval Szignifikáns-e a nem megfigyelt adatok körében - azaz a sokaság → hipotézis vizsgálat

Vegyes kapcsolat: Fizetés és Tapasztalatok kapcsolata

Kapcsolat jellege Az ábra kapcsolat jellegéről (átlag, medián,) - a teljes ár eloszlás hogyan különbözik tapasztalatonként → doboz ábra külön mindegyikre

```
ggplot(salaries, aes(y=salary_in_usd, x=experience_level, fill=experience_level)) +
  geom_boxplot()
```



Eredmény:

- láthatjuk, hogy a dobozok nagyjából hasonló magasságúak kivéve az expert tapasztalattal rendelkezőket
- a mid seniornál van olyan kilógó érték is, ami magasabb egy expert-nél is.
- ezen kívül, ami még érdekes lehet hogy az entry lvl poziknál van olyan kilógó érték ami eléri az expert szint fizetését is.

Az eloszlások jobbra elnyúlóak mindegyik felfele nyúlik, legmagasabb értékek az expert résznél vannak

Szépen növekednek a fizetések ahogyan tapasztalattal gyarapszik az ember, nagy növekedés a senior és az expertnél található

Kapcsolatok erőssége a mutatószám, amivel megtaláljuk a kapcsolat szorosságát (variancia - hányados = h^2)

```
aov(salary_in_usd~experience_level , data = salaries)
```

- SSB= SS Between = 6.714309e+11
- SSR = SSResiduals = 2.141492e+12
- SST = SSTotals = 6.714309e+11 + 2.141492e+12 = ssb +ssr = 2.8129229e+12

variancia (Ár) = SST/(N-1)

```
var(salaries$salary_in_usd)*(606-1)
```

- Az eredménye pont az mikor az ssb és az ssr -t összeadnám → 2.812923e+12 → szórás négyzet teteje - minden elem eltérése az átlagtól a négyzeten összeadva

SST = SZUM((Érték-Átlag)^2) - a fizetések az összesített szórása - ennyivel térnek el az átlagtól egy négyzetes skálán

Residuál 2.141492e+12 az a rész ami a maradványérték, az össz. információ a residuálja,az a rész ami nem magyarázható meg a csoportok közti eltéréssel, másnéven ami a csoportokon belüli eltéréssel jön

H^2 = SSB/SST → 0-1 közt arányszám érték % ban

```
6.714309e+1/(6.714309e+1+2.141492e+12)
```

3.135342e-11 → 0.000000003135342 % összes információ

Az eredmény azt jelzi, hogy az SSB közel van egymáshoz nagyon és így SSB kicsi így az SSR-nek nagynak kell lennie, azaz egy megfigyelés a csoport átlagához képest elég messze van, a csoport átlagok közelednek

- **H^2 <10% gyenge - ez a mi esetünkbe gyengének számít**
- 10% <= H^2 <= 50% --> KÖZEPES
- 50% <H^2 --> ERŐS/SZOROS

Szignifikancia – Hipotézisvizsgálat
Hipotézis vizsgálatot elvégezve kiderül az eremdény:

- H0: H^2 = 0 - A kapcsolat Nem szignifikáns a sokaságban
- H1: H^2 > 0 - A kapcsolat Szignifikáns marad
- p-érték → welch-korrekciózott F-próba

```
oneway.test(salary_in_usd~experience_level , data = salaries)
```

- p-value <2.2e-16 →H0 elutasítható minden szokásos alfa szignifikancia szinten → a **kapcsolat szignifikáns a sokaságban**

Megfelelő elemszám: minden csoportban legyen legalább 100 megfigyelés

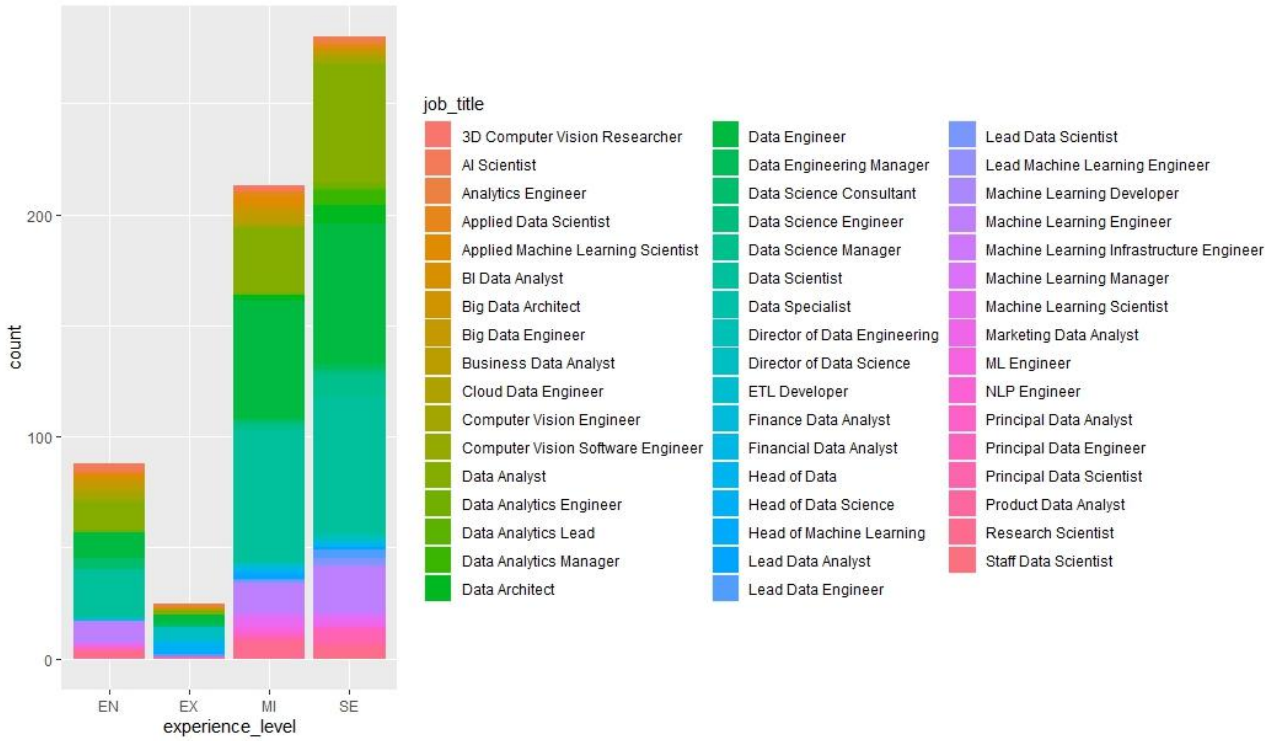
summary(salaries$experience_level)

- sajnos a mi esetünkben a feltétel nem teljesült, mert az expert lvl-ben 25 db van és az entry lvl is csak 88
- más esetben a ritka kategóriákat össze lehetne vonni de a beadandó esetében megengedett, hogy ne teljesüljön

Asszociáció kapcsolat: experience lvl + job title megnézzük hogyan oszlanak el a meghirdetett állások tapasztalatok alapján, milyen kapcsolatba vannak egymással

Kapcsolat jellege - halmozott oszlop diagram - egyik nominális változón belül megnézem a másik változó arányait

```
ggplot(salaries, aes(x=experience_level, fill=job_title)) +
  geom_bar()
```



- látjuk, hogy az állások hogyan oszlanak meg tapasztalati szintenként

- láthatjuk, hogy legkevesebb állás az expertnél található, és experten belül a többihez képest kevesebb principal dataval foglalkozó állás található, míg arányosítva a hed of data itt a legnagyobb
- arányokat tekintve mind az entry, mid senior, és senior között hasonló az eloszlás az állások kapcsán, többségük, data analyst, science, engineer állások 70 % ba körülbelül. A maradék pedig machine learningel foglakozó állások és Ai területen dolgozók nyitott pozicióját mondanám.
- ami a legjobban szembetűnik tapasztalati szinten - az állások darabszáma:
    - ha nem vesszük figyelembe a közel 25 db expert állást, láthatjuk, hogy a harmadik helyen a négyből az entry level áll kb 80 db-al, míg mid seniorként körülbelül 210 állás elérhető. Akiknek a legjobban kedvez a piac a seniorok akiknek kb 280 db nyitott állás áll rendelkezésükre.

Kapcsolatok erőssége

```
(table(salaries[,c("experience_level","job_title")]))
```

együttes gyakorisági táblázat

Cramér együttható:

```
cramer.v(table(salaries[,c("experience_level","job_title")]))
```

- c = 0.4655748 közepes a kapcsolat erőssége
- Krámér együttható 0-1 közötti mutatószám -> de ez sem százalékos! c
    - <0.3 → gyenge
    - 0.3 <= c<= 0.7 → közepes
    - 0.7 <c → erős vagy szoros

Közepes erősségű kapcsolatban vannak a tapasztalati szintű állások

Szignifikancia – Hipotézisvizsgálat

H0: C = 0 - A Kapcsolat nem szignifikáns

H1: C > 0 – A Kapcsolat szignifikáns

- p-érték→ khi-négyzet próbából jön
- előfeltétel: együttes gyakorisági táblában minden elem >= 10 - a mi esetünkben ez nem fog teljesülni

```
chisq.test(table(salaries[,c("experience_level","job_title")]))
```

- p-value < 2.2e-16 → minden szokásos szignifikancia szinten H0 elvethető

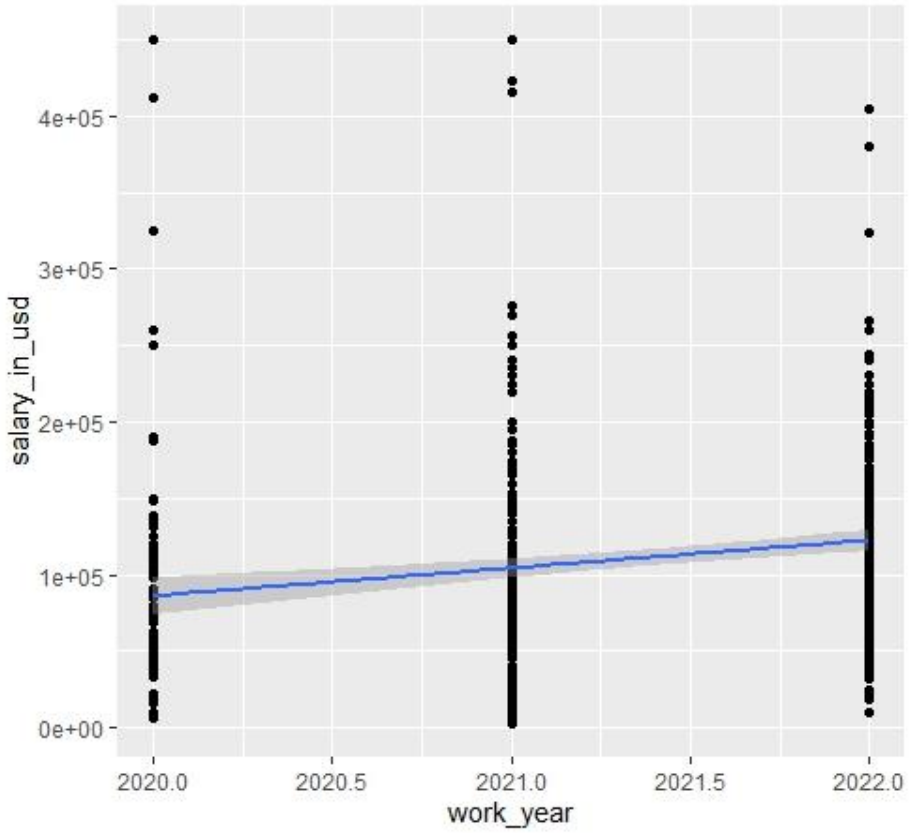
A kapcsolat szignifikáns

Korrelációs kapcsolat: arányskálás, intervallumskálás változók : évek és a fizetések

Kapcsolat jellege mivel két numerikus van ezért az ábrázolás pont diagrammon fog történni

- x: work_year y: salary_in_usd

```
ggplot(salaries, aes(x=work_year, y = salary_in_usd)) +
  geom_point() + geom_smooth(method =  lm)
```



- egyirányú pozitív emelkedésű tengely
- látható egy átlagos fizetés emelkedés 2020-2022 között viszont azt is láthatjuk, hogy 2021-ben a kiugró értékekből volt akár 450 ezer dolláros fizetés is viszont ez 2022 ben már csak a max 400 ezres volt.
- elmondható, hogy ezek alapján a legjobb évet 2021 zárta mert a kiugró értékeknek 2020-as hoz hasonló volt, míg a stabil fizetési szakasz az a 2022-re hasonlít.

Összességébe véve felfelé stagnáló fizetést vehetünk észre az évek haladtával.

Szürke sáv: Az egyenesnek a 95 %-os konfidencia intervalluma - a nem megfigyelt fizetések értelemében

- ha kevés megfigyelés van akkor tág, ha sok akkor szűk - ezek alapján látszik, hogy a 2020-as állásokból kevesebb volt mint a 2022-ből

Egyirányú pozitív emelkedésű tengely, ami azt jelenti, hogy az évek múlásával a fizetések növekedő tendenciát mutatnak.

Kapcsolatok erőssége ezekre a trendvonalakra, mennyire jól illeszkednek rajta a pontok

- korrreláció mutató - egyenesre való illeszkedés mérése: (Pearson-féle korreláció) = r
  - -1 és +1 közötti mutató szám -1 <= r <= +1

```
cor(salaries$work_year, salaries$salary_in_usd)
```

  - r=0.1845471 - az előjellel az irányt jelöli - azaz pozitív korreláció egy irányú kapcsolat - pozitív meredekségű egyenes -> pozitív korreláció

- korreláció abszolút értékét a kapcsolat szorosságának a leírására használjuk
  - |r| --> 0-1 de nem százalékos
  - határok: 0.3 > közepes
  - 0.7 > erős

- korreláció négyzetre emelése →az évek múlása hány százalékba magyarázza a fizetés emelkedését?
  - R-négyzet = r^2 → %-ban → határok 10%, 50%
    - 0.1845471^2
  - Eredmény 0.03405763 → az évek múlása 3 %-ban magyarázza a fizetés emelkedést

Szignifikancia – Hipotézisvizsgálat
Pontokra legjobban illeszkedő egyenesek = regresszió egyenesek - lineáris regresszió

```
lm(salary_in_usd ~ work_year, data = salaries)
```

- work year: ez a meredekség
- tengely metszet: -366
- becsült salaries = 18171* work year + (-36619860)
  - -36619860: modell becslés, ha nem telne az idő akkor -36619860 lenne a fizetés?
- 18717: Ha 1% telnek az évek akkor 18717 egységgel növeli a mortalitásra adott egységet.

H0: B1 = 0 | kapcsolat nem szignifikáns
H1: B1 <> 0 | kapcsolat szignifikáns

- p-érték # regressziós egyenes értékét nézzük summary függvényben

```
summary(lm(salary_in_usd ~ work_year, data = salaries))
```

- Reidual standard error: 67070-on 604 az átlagos becslési hiba
- p-value: 4.809e-06 - H0 elvethető → H1: kapcsolat szignifikáns

A kapcsolat szignifikáns

*This standard establishes internationally recognized codes for the representation of currencies that enable clarity and reduce errors. Currencies are represented both numerically and alphabetically, using either three digits or three letters. Some of the alpha betic codes for major currencies are familiar, such as “EUR” for Euros. Fortunately, ISO 4217 covers everything from Afghanis to Zambian Kwacha as well.

*The International Organization for Standardization (ISO) created and maintains the ISO 3166 standard – *Codes for the representation of names of countries and their subdivisions*.[1] The ISO 3166 standard contains three parts:

ISO 3166-1 – *Codes for the representation of names of countries and their subdivisions – Part 1: Country codes*[2] defines codes for the names of countries, dependent territories, and special areas of geographical interest. It defines three sets of country codes:

ISO 3166-1 alpha-2 – two-letter country codes which are also used to create the ISO 3166-2 country subdivision codes and the Internet country code top-level domains.

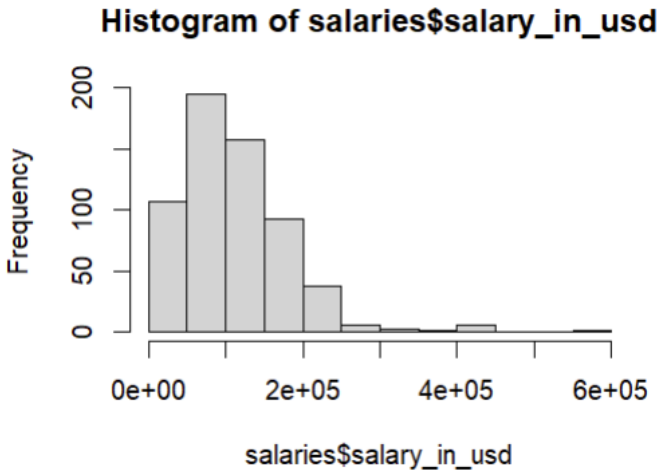
- több helyen a statisztikai mutatószámokat nem értelmezed az adatok kontextusában, csak a jegyzetből vett általános értelmezéseket másolod be. De ezek a bemásolások nem kellenek, én is ismerem pl. a korreláció értelmezési határait. Hanem csak az érdekes egy statisztikai elemzésben, hogy az adott adatbázis kontextusában tudod-e értelmezni az eredményeket, és
- **ezeket az értelmezéseket közérthető formába tudod-e önteni**.
- **Továbbá, sokszor nincsenek a gondolatok teljes mondatban megfogalmazva**, és néhány helyen olyan formai dolgok nehezítik a megértést, mint hogy nyitsz a szövegben egy zárójelt, és nem zárod be.

A **'remote ratio' változó inkább ordinális,** mert a leírás alapján nem minden "50" értéknél pontosan 50% a távmunka aránya. **Ezek inkább nagyságrendi kategóriáknak tekinthetők**. Amúgy intervallum biztosan nem lehet a változó, mert a 0% távmunka állapota egy egyértelműen értelmezhető dolog, nem pedig önkényesen adott szint.

- A leíró statisztika fejezetnél a **numerikus változó elemzésével az a gond, hogy nem egységesen egy változó kerül elemzésre a tanult mutatókkal, eszközökkel, hanem össze-vissza csapong az elemzés a változók között, így nincs bemutatva pl. az, hogy egy konkrét változó hisztogramjáról leolvasható következtetések hogyan jelennek meg ugyan azon változó alakmutatóiban**.
- **Konkrét értelmezések kellenek, hogy használható elemzése legyen a vége!**
- A doboz ábránál hasonló a gond: **nem elemzed, hogy akkor most mit számítasz tényelegesen kilógó értéknek**: kiszámoltad a belső és a külső kerítéseket is, de nem **elemzed**
- **a doboz ábra alapján, hogy melyik kerítések választják le reálisabban a kilógó értékeket.**

"a tapasztalattal rendelkező állások mediánja Mid Senior" --> értelmezés: a vizsgált állások fele legalább mid senior tapasztalatot követel meg + megjegyzendő, hogy a kis értékkészlet miatt a medián erre a változóra nem informatív

Módusz – gyakoriság hisztogrammal ábrázolva



Megfigyelhető, hogy az **alacsonyabb** fizetési kategóriákban **gyakoriak** az értékek, míg a **magasabb** fizetési tartományokban **ritkábban** találunk értékeket. Ezen kívül jól látszik, hogy a medián értékünk is a 100 ezres kategóriai körül lesz. Ez pontosabban azt jelenti, **hogy az állások többsége 200 ezer** dollárt fizet míg a kiugró értékekkel kapcsolatban lehet találni olyan állást, ami 600 ezer dollárt is fizet. Elég nagy a különbség a két érték között.

```
salary_in_usd
Min.   :  2859
1st Qu.: 62726
Median :101570
Mean   :112298
3rd Qu.:150000
Max.   :600000
```

median(salaries$salary_in_usd)
101570

mean(salaries$salary_in_usd)
 # 112297.9

Az átlag és a medián közötti különbség azt sugallja, hogy a fizetések **eloszlása nem egyenletes**, és lehetnek olyan kiugró értékek, amelyek az átlagot emelik.:

A medián értéke 101 570 dollár, míg az átlag 112 297,9 dollár. A két érték közötti különbség 10 727,9 dollár. Ez azt jelzi, hogy az adathalmazban található fizetések nem egyenletesen oszlanak el, és a kiugró értékek befolyásolják az átlagot

Szórás: 70957.26 -

A szórás értéke 70,957.26 ez alapján látható, hogy a fizetések változatossága jelentős. **Az adatok szétszórtak az átlagtól, ami azt mutatja, hogy a keresetek közötti különbségek jelentős mértékűek lehetnek.** Az adathalmazban széles skálán mozognak a fizetések.

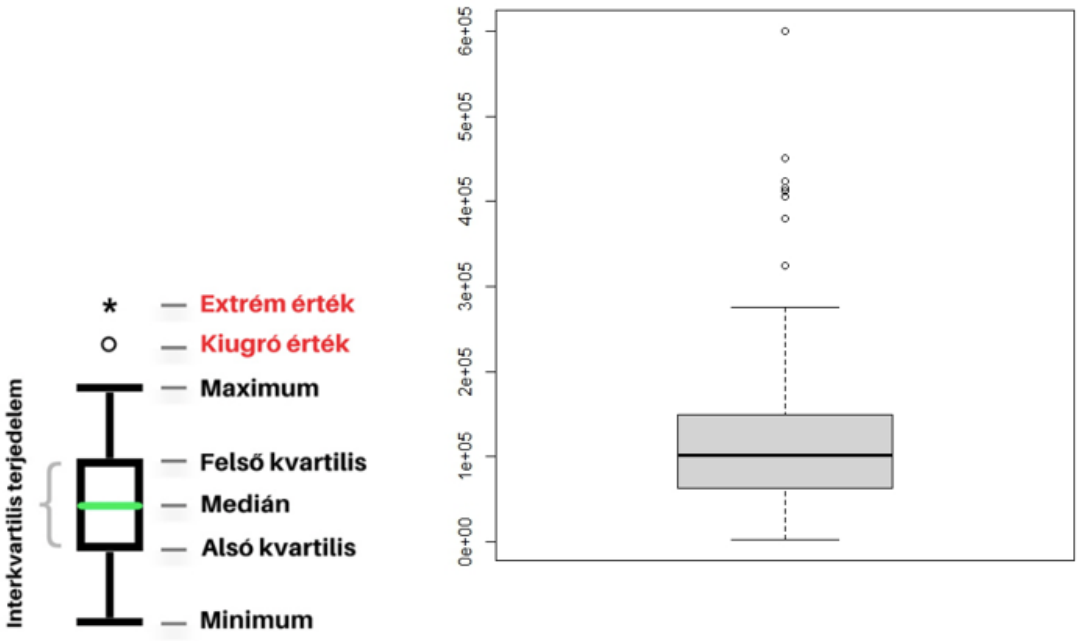
Kvantilisek
Q1:25% 62726 ,
Medián: 50% 101570
Q3: 75% 150000

A kvantilisek értékei az alsó 25%-ban 62,726 dollár, az átlag alacsonyabb, mint a medián, ami azt jelzi, hogy az alacsonyabb fizetési kategóriákban a keresetek kisebbek a medián 101,570 dollár, míg a felső 25%-ban az átlag 150,000 dollár felső 25% -ban az átlag magasabb, ami a magasabb fizetési kategóriákban mutatkozó magasabb kereseteket tükrözi. **Ez azt sugallja, hogy az alacsonyabb fizetési kategóriákban a keresetek kisebbek, míg a magasabb kategóriákban magasabbak.**

Kerítés:
Alsó kerítés: Q1- 1,5* (Q3-Q1) = 62726-130911 = -68 185
Felső kerítése: Q3 + 1,5*(Q3-Q1) = 150000 + 130911 = 280 911

Az alsó kerítés -68,185 dollár, míg a felső kerítés 280,911 dollár. Azok az értékek, amelyek a **felső kerítés felett vannak, kiemelkedően magas fizetéssel rendelkeznek.**

Alakmutató – doboz ábra:



Doboz Ábra:

Az ábra alapján megállapítható, hogy az eddigi számolásaim alátámasztják, hogy az adatok eloszlása meglehetősen széles skálán mozog.

Az alsó és felső kerítések közötti tartomány nagyon tág, ami arra utal, hogy az adatokban **jelentős szórás van.**

Az adatok között található minimum és maximum értékek is jelzik, hogy vannak **kiugróan alacsony és magas értékek is.**

A medián értéke 101570 dollár, ami azt jelenti, hogy az adatok középső értéke ebben a tartományban helyezkedik el.

Az átlag értéke pedig 112297.9 dollár, ami azt mutatja, hogy az adatok átlagosan ebben a tartományban helyezkednek el**, de a kiugró értékek miatt ez az átlag meglehetősen magas.**

A szórás értéke 70957.26 dollár, ami azt jelenti, hogy az adatok **nagyfokú szóródást mutatnak**. Ez azt jelzi, hogy az adatok között jelentős eltérések vannak, és nincs egyértelműen meghatározható trend vagy mintázat.

A kvantilisek értékei is azt mutatják, hogy az adatok **eloszlása nem egyenletes**. Az első kvantilis (Q1) értéke 62726 dollár, ami azt jelenti, hogy az adatok alsó negyedének értéke ebben a tartományban helyezkedik el. A harmadik kvantilis (Q3) értéke pedig 150000 dollár, ami azt jelenti, hogy az adatok felső negyedének értéke ebben a tartományban helyezkedik el.

Összességében elmondható, hogy az adatokban jelentős szórás és kiugró értékek vannak.

Az átlag konfidencia-intervallum értelmezésénél sehol nem derül ki az elemzésből, hogy a vizsgált **mutatónk az átlagfizetés...**

pl.: "90%-os pontossággal ennyi fizetés várható egy data sciencel foglalkozó munkakaeresőnek ezen az oldalon" **--> sehol nem derül ki, hogy itt az átlagfizetésre adunk becslést**!

Az arány intervallumbecslésnél nagyon ki kell emelni, hogy a NaN-ok, és a negatív alsó határok a nem értelmezhető **elemszámok (1-3 db) miatt vannak!**

Az intervallumbecsléseknél nagyon hiányzik az intervallumok metszéspontjainak összevetése: pl. 90% valószínűséggel nézve az EX és SE szintek medián fizetése nem különbözik egymástól a nem megfigyelt állások sokaságában**. – átlag értéke 107560.6 USD és a mediáné 102000 Usd**

"láthatjuk, hogy a dobozok nagyjából hasonló magasságúak kivéve az expert tapasztalattal rendelkezőket" --> azért a senior medián fizetése is magasabb, mint az alatta lévő két szint felső kvartilise! Erre indirekt módon később magad is utalsz! **- rosszul fogalmaztam, mert arra gondoltam hogy a dobozok szélessége hasonló egymáshoz**

H^2 el van számolva: = 6.714309e+11/2.8129229e+12 = 0.239 = 23.9% --> közepesen szoros kapcsolat!

Asszociációnál a rengetegféle állást nagyon jó lett volna összevonni. Sokkal szebben értelmezhető lett volna a halmozott oszlopdiagram is! **– több számolást is csináltam, de nem tettem bele hanem kiválasztatottam egyet (országok – fizetés, fizetés – tapasztalat, országok - pozik, tapasztalat – pozik, tapasztalat – országok)**

Az egész kapcsolatvizsgálatra igaz ez a megjegyzésem, de a korrelációs részre nagyon is. Annyit elég leírni, hogy "mivel a 0.18 korreláció abszolút értékben 0.3 alatti, így a kapcsolat gyengének minősíthető". Kész. Ehhez képest bevágtad a jegyzetből az értelmezési határokat, és annyit sem írsz le, hogy a Te korrelációd hova esik.

Tengelymetszet értelmezése rossz. Hülye az értelmezés, így én nem is tenném meg, de ha ragaszkodunk hozzá, akkor azt jelenti, hogy Krisztus urunk születésekor a modell becslése szerint egy data scientist fizetése kb. -36.6 millió dollár (y_kalap, ha az x=0).

Hibás a meredekség értelmezése: **Évente átlagosan 18171 dollárral nő a data scientistek fizetése.**