UPA - projekt 2

Timotej Ponek, xponek00

Bod 1 - stiahnutie datasetu, príprava prostredia

Stiahol som si datovú sadu ‘Platy v IT’ dostupnú z
https://www.kaggle.com/datasets/parulpandey/2020-it-salary-survey-for-eu-region

Pre vývoj som Python Jupyter Notebook

Bod 2 - Exploratívna analýza

Podukol 1 & 2

prozkoumejte jednotlivé atributy datové sady, jejich typ a hodnoty popis jednotlivých atribútov dátovej sady, typ a hodnota (počet hodnôt, najčastejšia hodnota, rozsah hodnôt) + rozloženie hodnôt jednotlivých atribútov vykreslené pomocou vhodných grafov

Popis atribútov

Dátová sada obsahuje nasledujúce atribúty, zapísané ako dvojica (atribút - dátový typ):

- Timestamp - string
- Age - float
- Gender - string
- City - string
- Position - string
- Total years of experience - string
- Years of experience in Germany - string
- Seniority level - string
- Your main technology / programming language - string
- Other technologies/programming languages you use often - string
- Yearly brutto salary (without bonus and stocks) in EUR - float
- Yearly bonus + stocks in EUR - string
- Annual brutto salary (without bonus and stocks) one year ago. Only answer if staying in the same country - string
- Annual bonus+stocks one year ago. Only answer if staying in same country - string
- Number of vacation days - float
- Employment status - string
- Contract duration - string
- Main language at work - string
- Company size - string (ordinálny)
- Company type - string
- Have you lost your job due to the coronavirus outbreak? - string
- Have you been forced to have a shorter working week (Kurzarbeit)? If yes, how many hours per week - string
- Have you received additional monetary support from your employer due to Work From Home? If yes, how much in 2020 in EUR - string

Niektoré atribútu napriek tomu, že by mali nadobúdať iba číselných hodnôt nadobúdajú hodnoty textovej. Dátovú sadu bude treba očistiť a zbaviť irelevantných hodnôt.

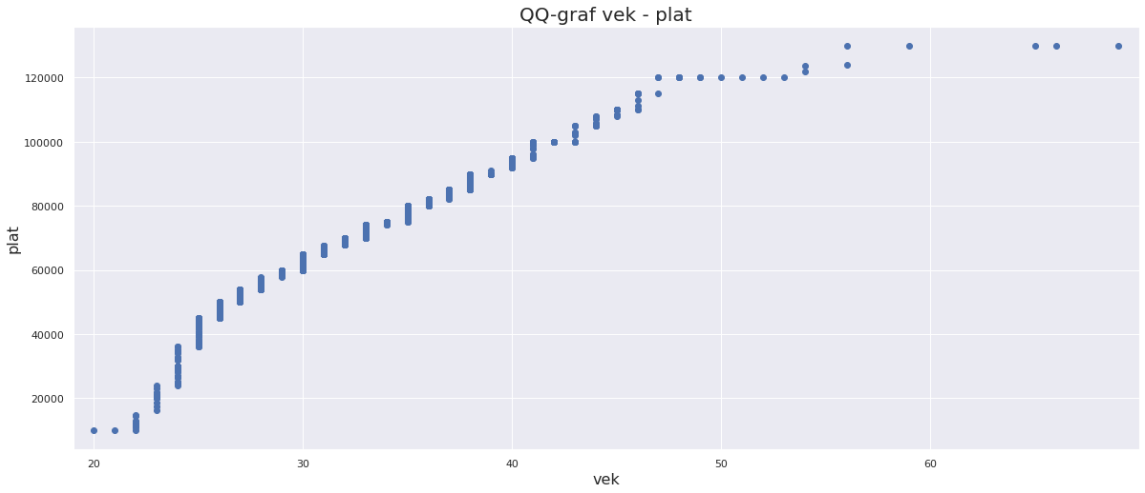
1. Časová značka (Timestamp)

- počet hodnôt 1253
- unikátne hodnoty 1248
- najčastejšia hodnota 24/11/2020 13:55:19
- frekvencia najčastejšej hodnoty 2

Graf pre tento atribút je viac menej podstatný, ide o záznam ktorý určuje čas v ktorom bol dotazník vyplnený

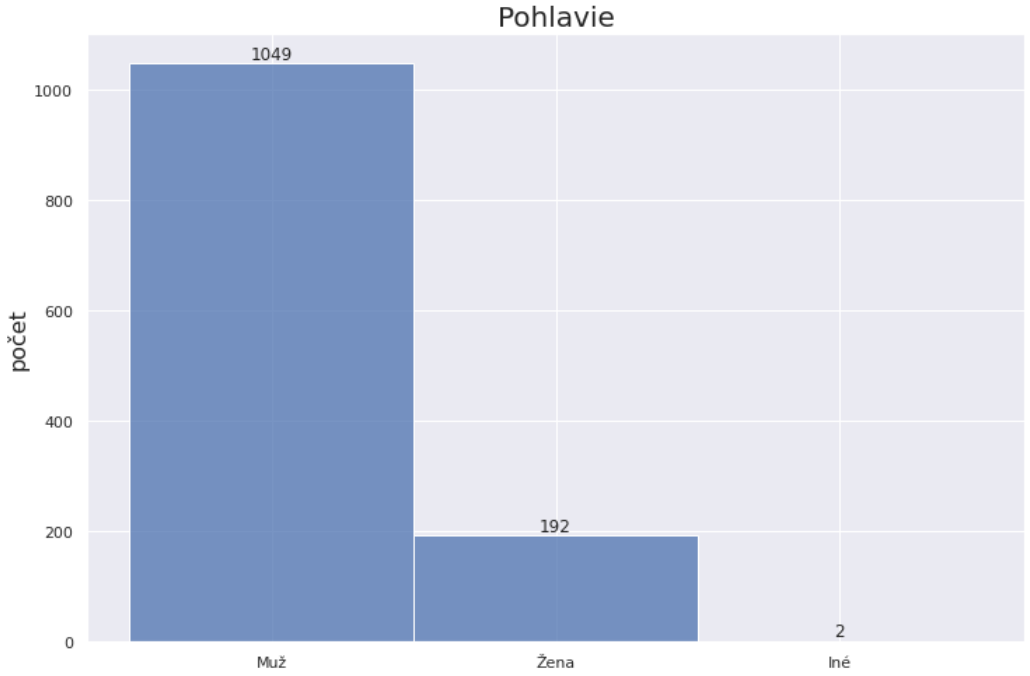
2. Vek (Age)

- počet 1226
- priemer 32.509788
- smerodajná odchýlka 5.663804
- minimum 20
- 25% 29
- 50% 32
- 75% 35
- maximum 69



3. Pohlavie (Gender)

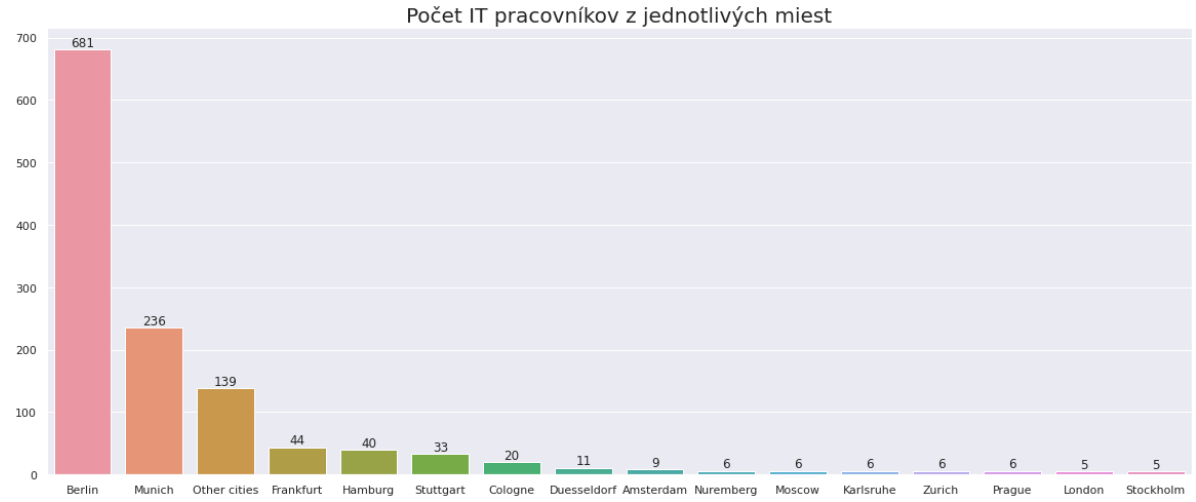
- počet hodnôt 1243
- unikátne hodnoty 3
- najčastejšia hodnota Male (Muž)
- frekvencia najčastejšej hodnoty 1049



4. Mesto (City)

- počet hodnôt 1253
- unikátne hodnoty 119
- najčastejšia hodnota Berlin
- frekvencia najčastejšej hodnoty 681

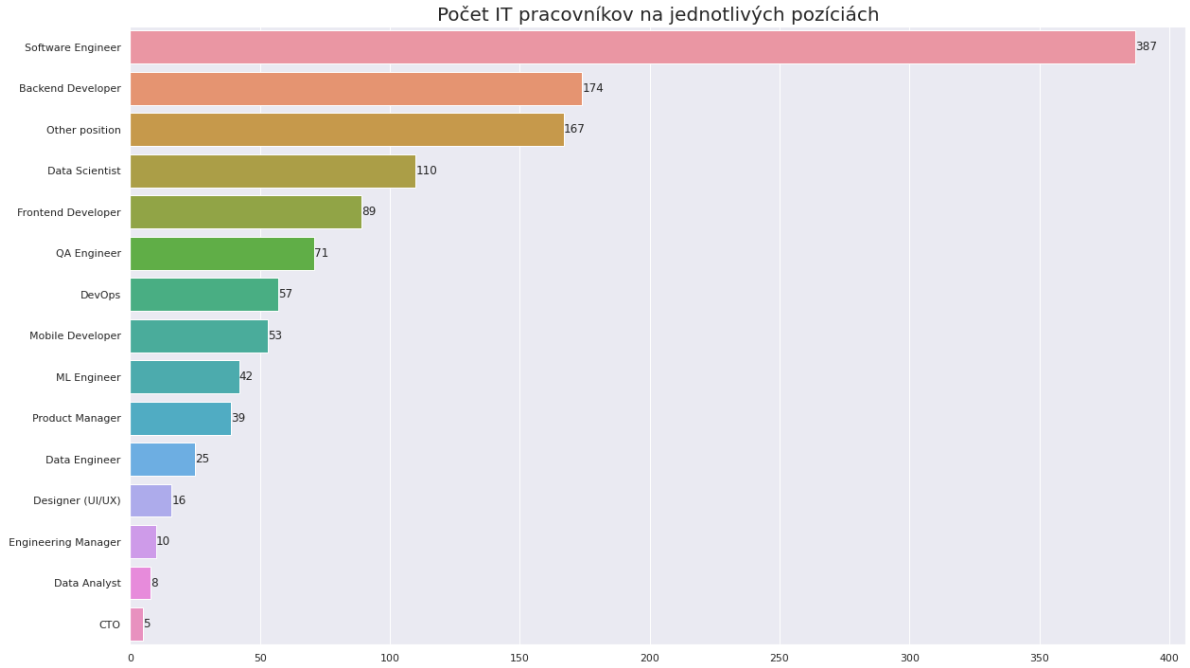
Pre sprehľadnenie vykreslenia boli na nasledujúcom grafe mestá z ktorých bolo menej ako 5 účastníkov prieskumu zredukované do stĺpca “Other cities”



5. Pozícia (Position)

- počet hodnôt 1247
- unikátne hodnoty 148
- najčastejšia hodnota Software Engineer
- frekvencia najčastejšej hodnoty 387

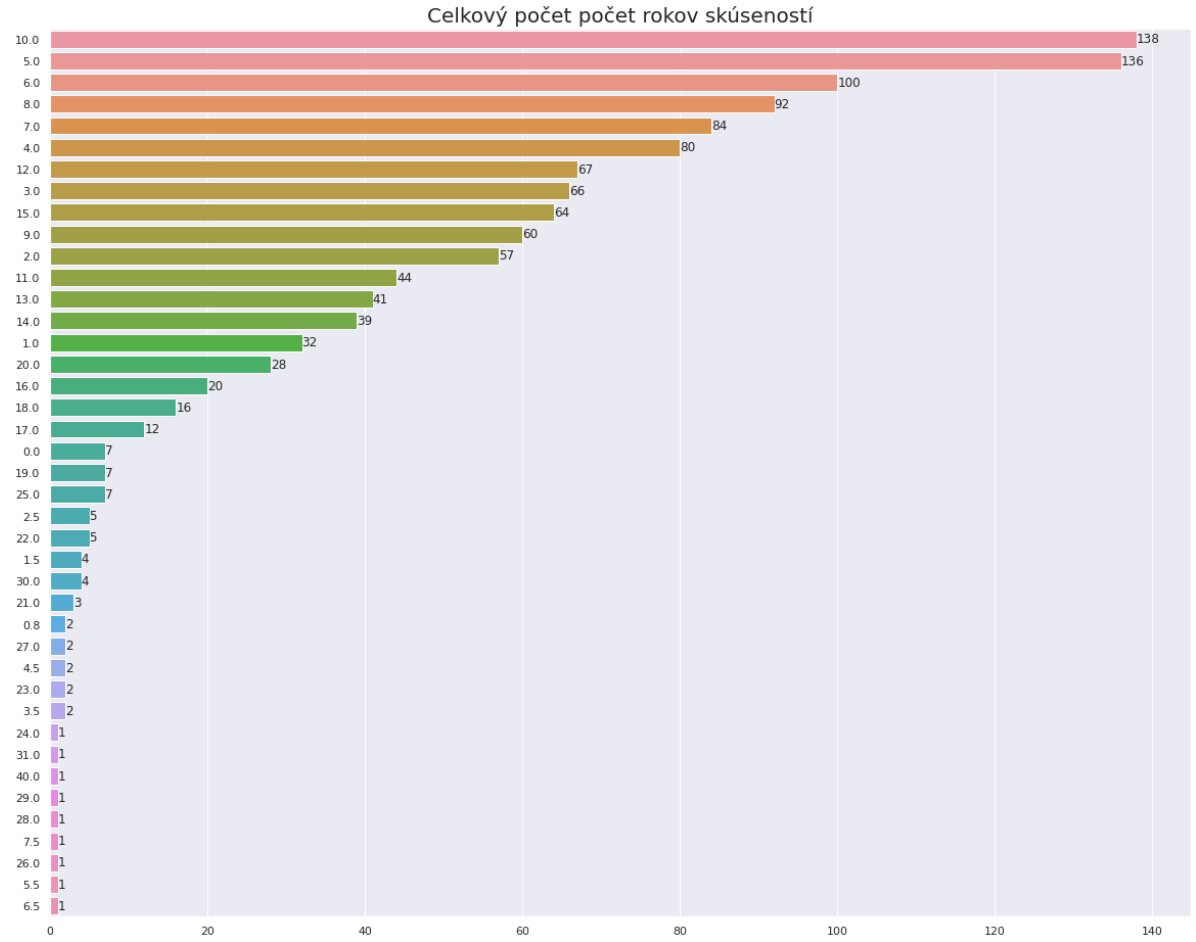
Pre sprehľadnenie vykreslenia boli na nasledujúcom grafe pozície na ktorých bolo menej ako 5 účastníkov prieskumu zredukované do stĺpca “Other position”



6. Celkový počet rokov skúseností / odpracovaných rokov (Total years of experience)

Stĺpec bol najskôr očistený a následne textové hodnoty prevedené na float

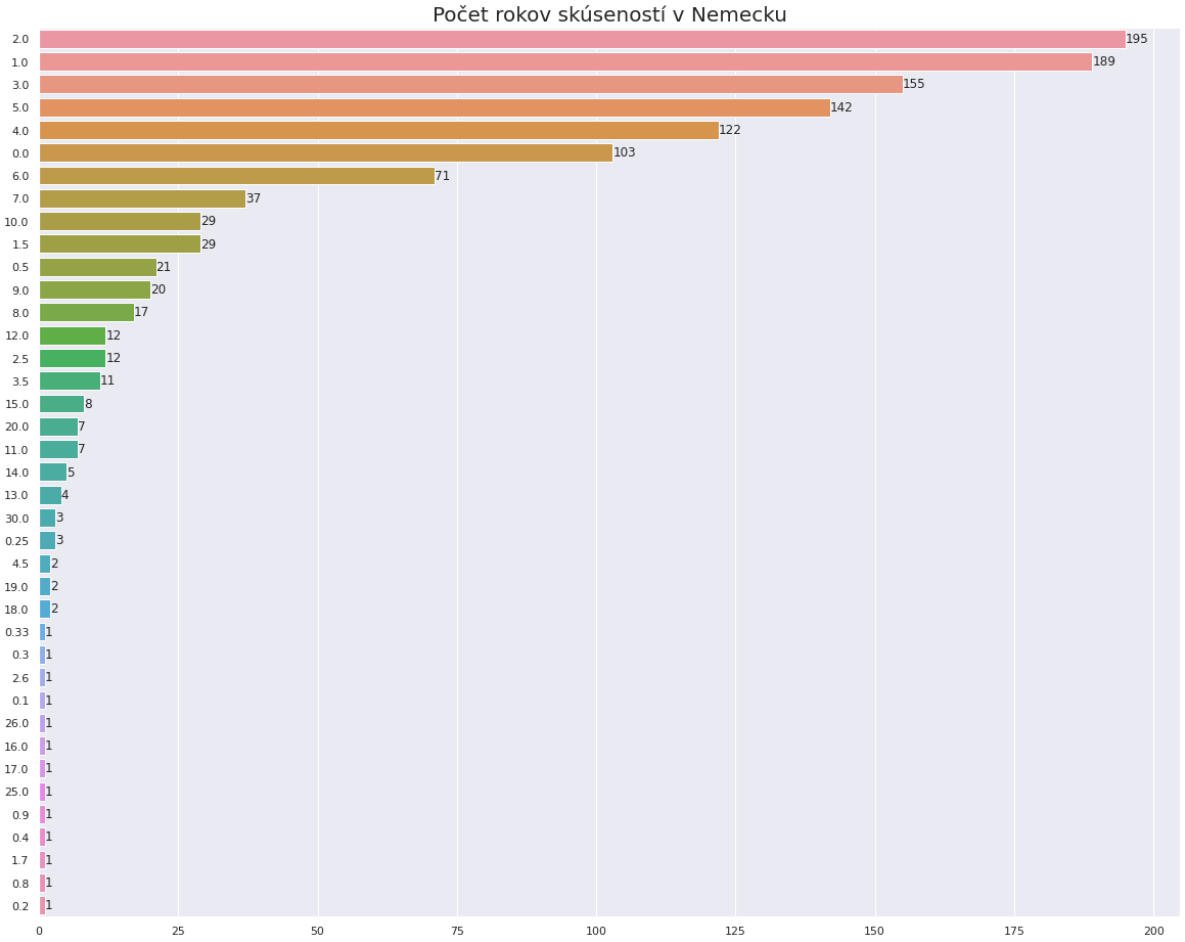
- počet 1237
- priemer 8.748262
- smerodajná odchýlka 5.663804
- minimum 0
- 25% 5
- 50% 8
- 75% 12
- maximum 40



7. Počet rokov skúseností / odpracovaných rokov v Nemecku (Total years of experience in Germany)

Stĺpec bol najskôr očistený a následne textové hodnoty prevedené na float

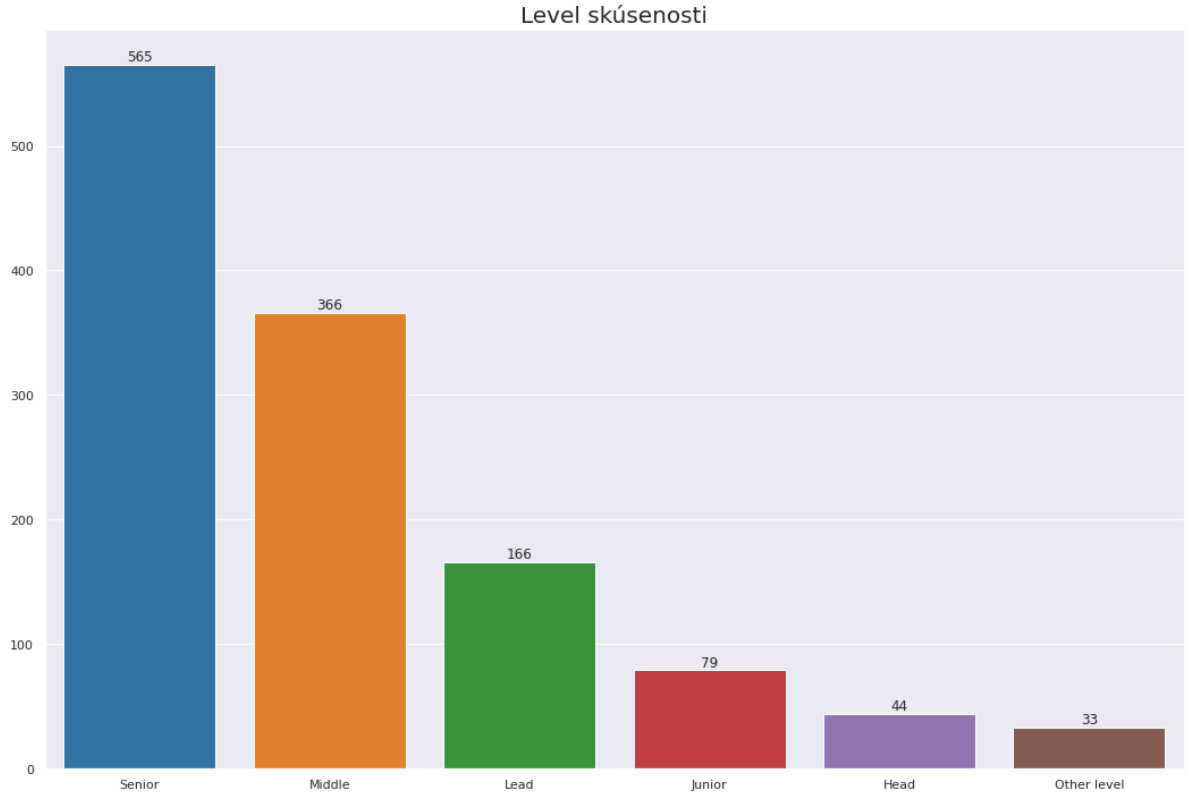
- počet 1221
- priemer 3.704816
- smerodajná odchýlka 3.639877
- minimum 0
- 25% 1
- 50% 3
- 75% 5
- maximum 30



8. Level skúsenosti (Seniority level)

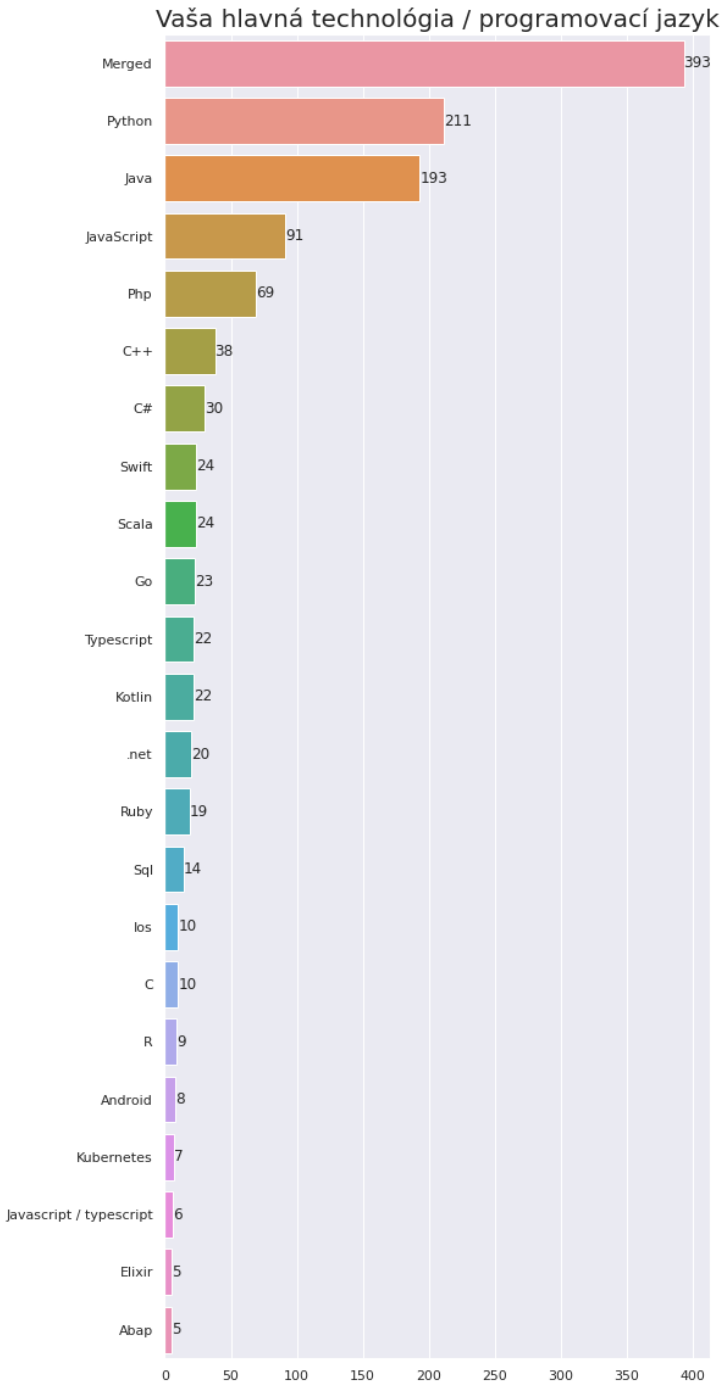
- počet hodnôt 1241
- unikátne hodnoty 24
- najčastejšia hodnota Senior
- frekvencia najčastejšej hodnoty 565

Pre sprehľadnenie vykreslenia boli na nasledujúcom grafe levely skúsenosti u ktorých bolo menej ako 5 účastníkov prieskumu zredukované do stĺpca “Other level”



9. Vaša hlavná technológia / programovací jazyk (Your main technology / programming language)

- počet hodnôt 1126
- unikátne hodnoty 256
- najčastejšia hodnota Java
- frekvencia najčastejšej hodnoty 184



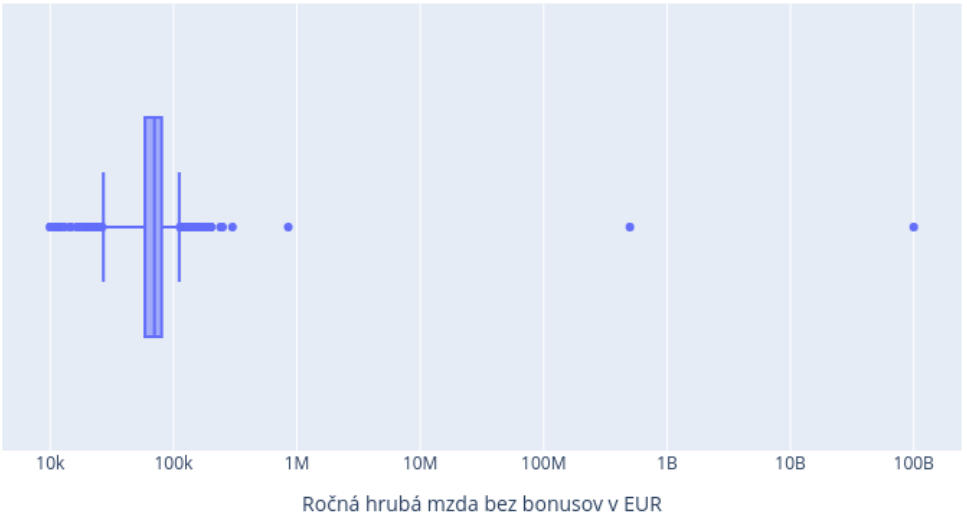
10. Ďalšie technológie / programovacie jazyky ktoré často používate (Other technologies/programming languages you use often)

- počet hodnôt 1096
- unikátne hodnoty 562
- najčastejšia hodnota Javascript / Typescript
- frekvencia najčastejšej hodnoty 44

Tento stĺpec obsahuje tak veľa unikátnych hodnôt že pre neho nebude vykreslený graf

11. Ročná hrubá mzda (bez bonusov a akcií) v eurách - (yearly brutto salary (without bonus and stocks) in EUR)

- počet 1253
- priemer 80 279 040
- smerodajná odchýlka 2 825 061 000
- minimum 10001
- 25% 58800
- 50% 70000
- 75% 80000
- maximum 100000000000



12. Ročný bonus + akcie v EUR (yearly bonus + stocks in EUR)

- počet hodnôt 829
- unikátne hodnoty 168
- najčastejšia hodnota 0
- frekvencia najčastejšej hodnoty 227

13. Ročná hrubá mzda (bez bonusu a akcií) pred rokom. Odpovedajte iba ak zostávate v rovnakej krajine (Annual brutto salary (without bonus and stocks) one year ago. Only answer if staying in the same country)

- počet 885
- priemer 632 245.90
- smerodajná odchýlka 16 805 080
- minimum 11 000
- 25% 55 000
- 50% 65 000
- 75% 80 000
- maximum 500 000 000

14. Ročný bonus + akcie pred rokom. Odpovedajte iba ak zostávate v rovnakej krajine (Annual bonus+stocks one year ago. Only answer if staying in same country)

- počet hodnôt 614
- unikátne hodnoty 131

- najčastejšia hodnota 0
- frekvencia najčastejšej hodnoty 200

15. Počet dní pracovného voľna (Number of vacation days)

- počet hodnôt 1185
- unikátne hodnoty 45
- najčastejšia hodnota 30
- frekvencia najčastejšej hodnoty 488

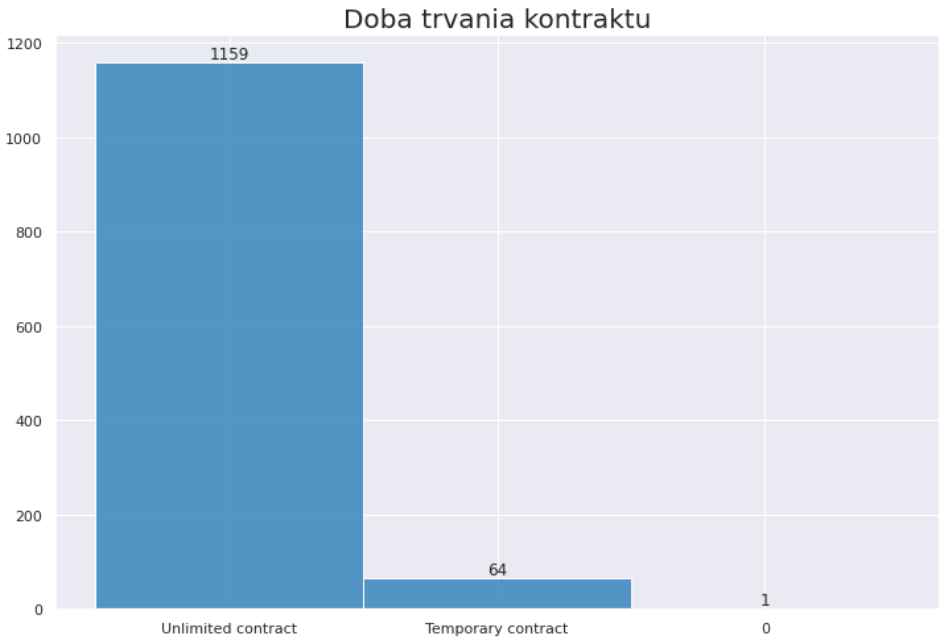
16. Ako zamestnaný (Employment status)

- počet hodnôt 1236
- unikátne hodnoty 11
- najčastejšia hodnota Full-time employee
- frekvencia najčastejšej hodnoty 1190

Pre sprehľadnenie vykreslenia boli na nasledujúcom grafe typy pracovného úväzku u ktorých bolo menej ako 5 účastníkov prieskumu zredukované do stĺpca “Other”

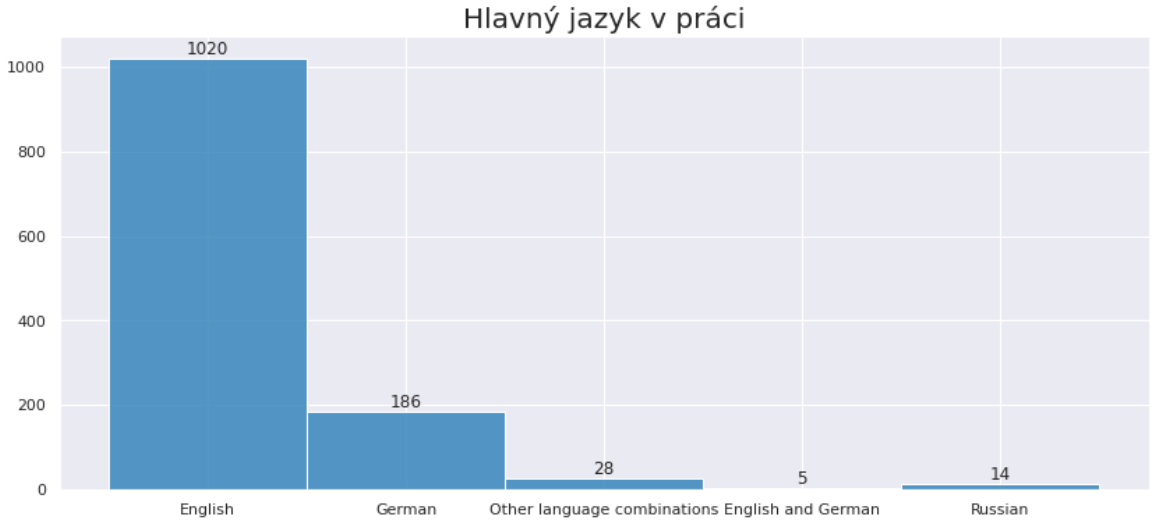
17. Doba trvania kontraktu (Contract duration)

- počet hodnôt 1224
- unikátne hodnoty 3
- najčastejšia hodnota Unlimited contract
- frekvencia najčastejšej hodnoty 1159



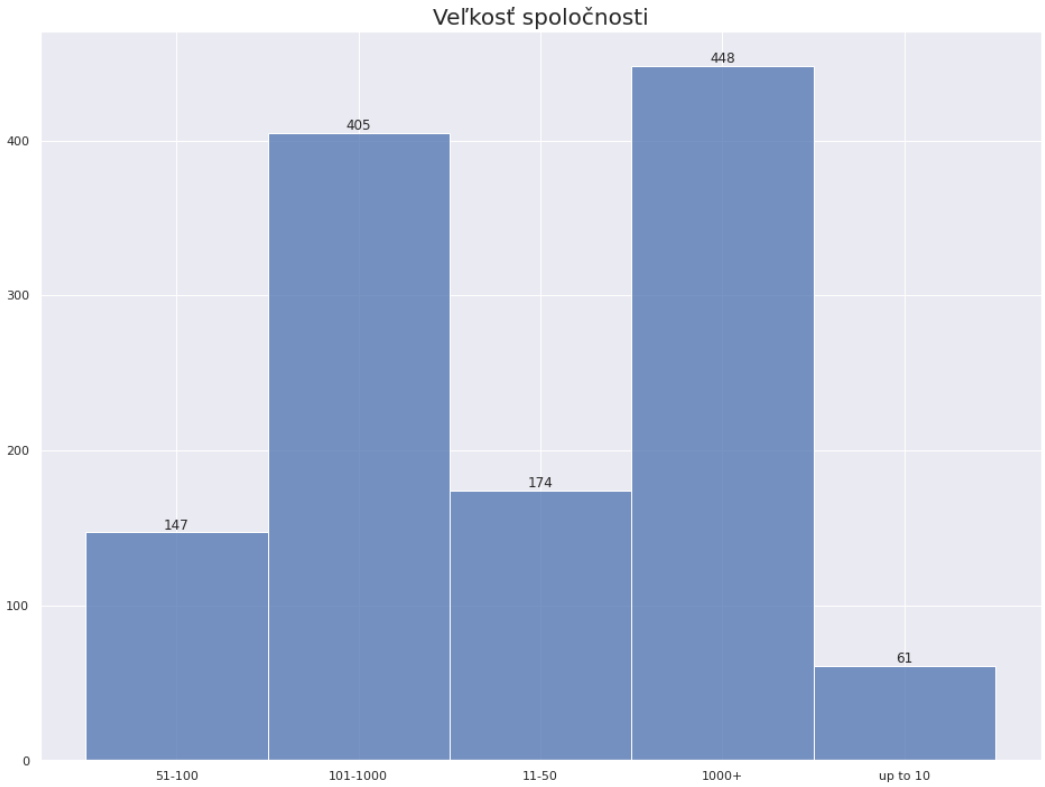
18. Hlavný jazyk v práci (Main language at work)

- počet hodnôt 1237
- unikátne hodnoty 14
- najčastejšia hodnota English
- frekvencia najčastejšej hodnoty 1020



19. Veľkosť spoločnosti (Company size)

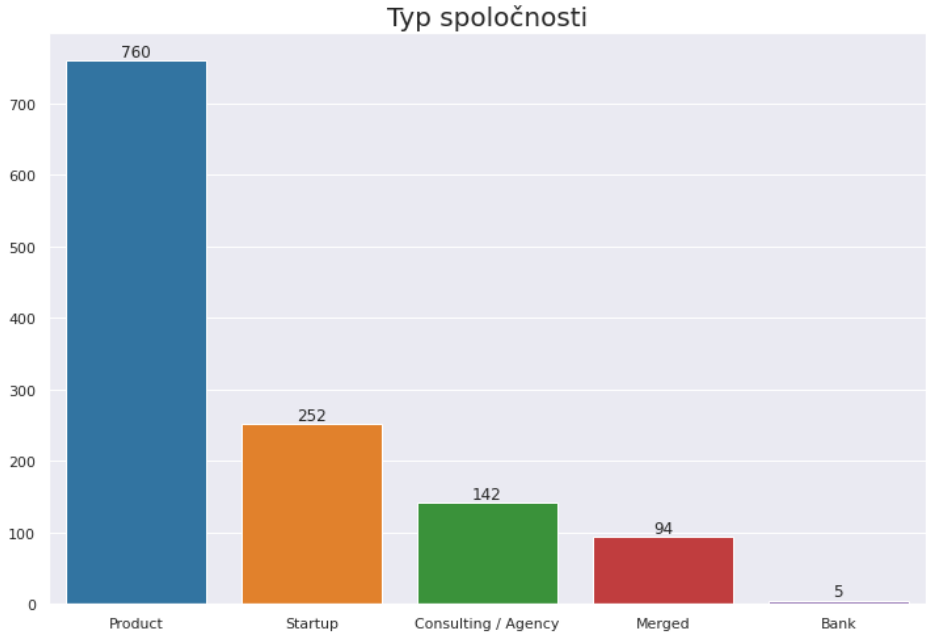
- počet hodnôt 1235
- unikátne hodnoty 5
- najčastejšia hodnota 1000+
- frekvencia najčastejšej hodnoty 448



20. Typ spoločnosti (Company type)

- počet hodnôt 1228
- unikátne hodnoty 63
- najčastejšia hodnota Product
- frekvencia najčastejšej hodnoty 760

Pre sprehľadnenie vykreslenia boli na nasledujúcom grafe typy spoločnosti u ktorých bolo menej ako 5 účastníkov prieskumu zredukované do stĺpca “Merged”

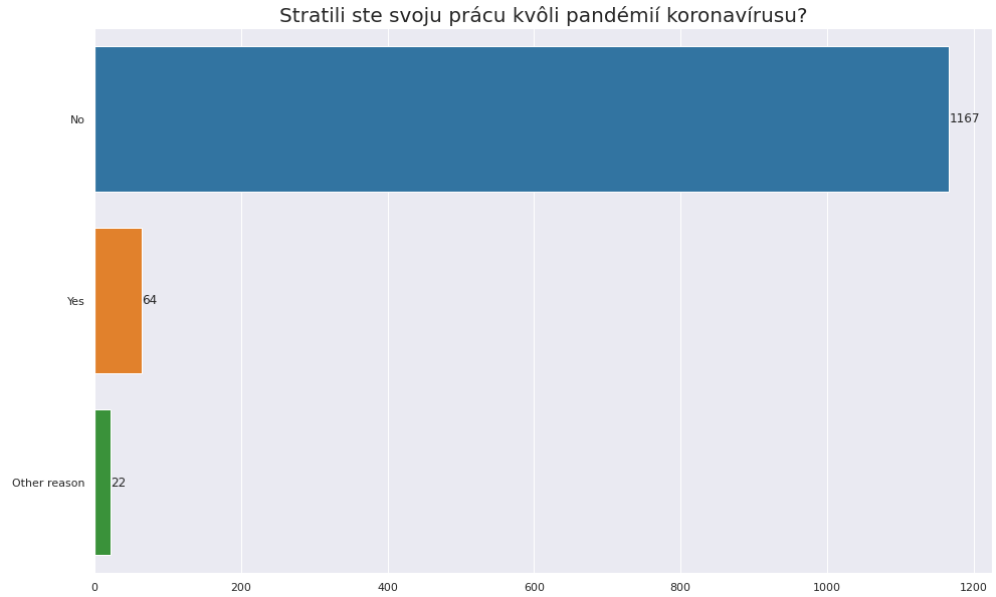


21. Stratili ste svoju prácu kvôli pandémií koronavírusu (Have you lost your job due to the coronavirus outbreak?)

- počet hodnôt 1233
- unikátne hodnoty 10

- najčastejšia hodnota No
- frekvencia najčastejšej hodnoty 1162

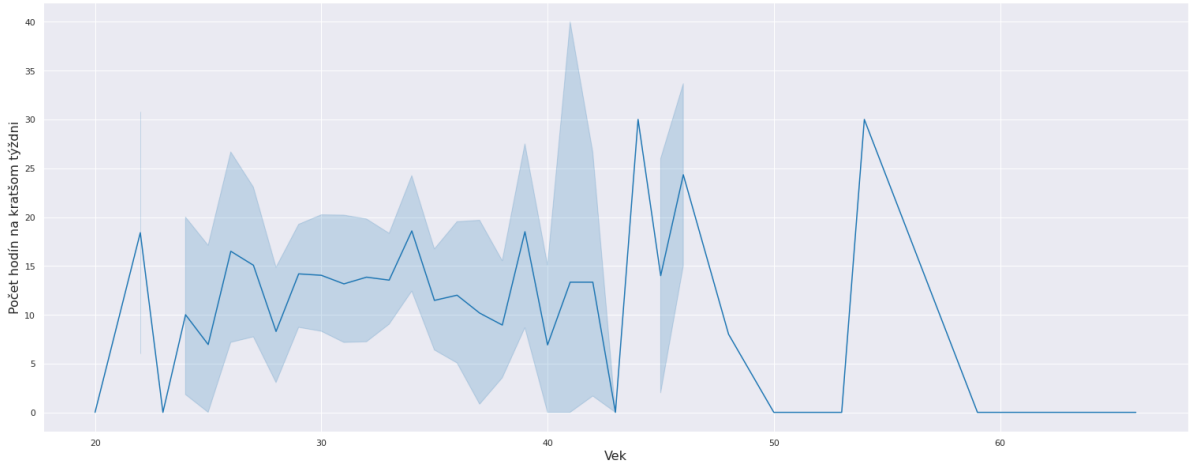
Pre sprehľadnenie vykreslenia boli na nasledujúcom grafe hodnoty u ktorých bolo menej ako 5 účastníkov prieskumu zredukované do stĺpca “Other reasons”



22. Boli ste nútený mať kratšie pracovné týždne (Kurzarbeit)? Ak áno, koľko to bolo hodín za týždeň? (Have you been forced to have a shorter working week (Kurzarbeit)? If yes, how many hours per week)

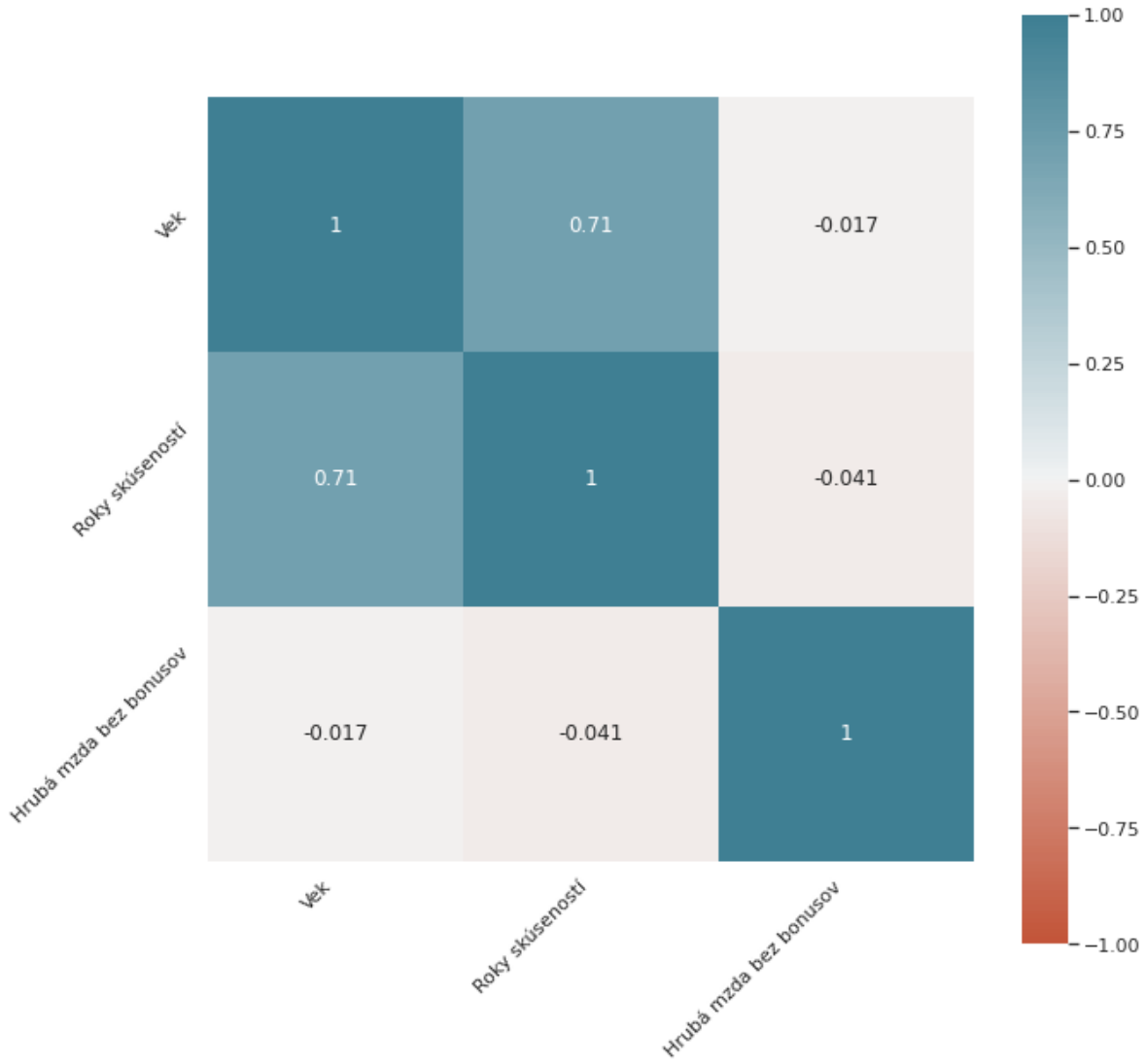
- počet 373
- priemer 12.967828
- smerodajná odchýlka 15.275174
- minimum 11 000
- 25% 0
- 50% 0
- 75% 30
- maximum 40

Lineplot:



Podukol 5 - preveďte korelačnú analýzu

Korelačná analýza atribútov `"Age", "Total years of experience", "Yearly brutto salary (without bonus and stocks) in EUR"`
`Heatmapa:`



Môžeme z nej vyčítať že korelácia medzi vekom a rokmi skúseností je vysoká, zatiaľčo korelácia medzi vekom a mzdou a mzdou a skúsenosťami je nízka, ide až do záporu.

Bod 3

Pre dolovacie algoritmy boli vytvorené dva csv súbory:
dis.csv a cat.csv

dis.csv obsahuje kategorické atribúty transformované na numerické

cat.csv obsahuje numerické atribúty transformované na kategorické