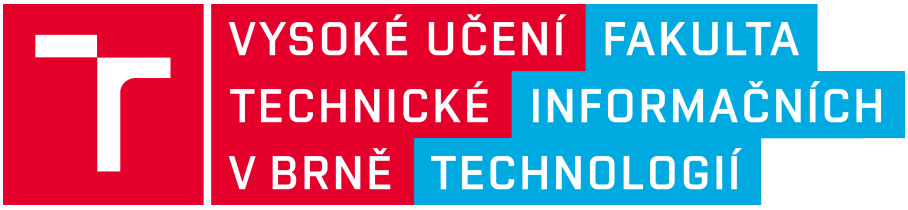


UPA 2022 – Projekt 2

Příprava dat a jejich popisná charakteristika

3. datová sada – Platy v IT

Vojtěch Fiala (**xfiala61**)
Vojtěch Giesl (xgiesl00)
Vojtěch Kronika (xkroni01)

1 Explorativní analýza

Tato dokumentace se zabývá explorativní analýzou a úpravami, které byly provedeny nad datovou sadou *Platy v IT*[1]. Jedná se o datovou sadu zaměřenou na platební ohodnocení zaměstnanců v oblasti IT, zejména v Německu, která krom samotných údajů o platu obsahuje i základní informace o respondentech. Datovou sadu tvoří celkem 1238 anonymních odpovědí.

1.1 Popis atributů

První část této dokumentace se bude zabývat popisem jednotlivých atributů, které datovou sadu tvoří. Pokud není uvedeno jinak, vždy se jedná o textový řetězec. Konkrétně se jedná o atributy následující –

Timestamp

Jedná se o časové razítko značící čas vzniku odpovědi.

Age

Číselná hodnota reprezentující věk respondenta. Průměr je *32.5* let.

Gender

Pohlaví respondenta. Většinu respondentů tvořili muži (84 %), zbylých 16 % tvořily ženy a ostatní.

City

Město, kde respondent pracuje. Většina respondentů (54 %) pracuje v Berlíně. Celkem respondenti uvedli 119 různých měst.

Position

Pracovní pozice, kterou respondent zastává. Nejvíce respondentů (387) uvedlo jako svoji pozici *Software Engineer*. Celkem bylo uvedeno 148 různých pozic.

[1] `https://www.kaggle.com/datasets/parulpandey/2020-it-salary-survey-for-eu-region`

Total years of experience

Textový řetězec reprezentující roky zkušeností, jak dlouho se respondent svému oboru věnuje. Nejvíce respondentů (138) uvedlo 10 let zkušeností. Atribut je numerického charakteru, nicméně kvůli několika odpovědím má i kategorické rysy.

Years of experience in Germany

Textový řetězec reprezentující číslo, kolik let daný respondent pracuje v Německu. Největší počet respondentů (195) uvádí 2 roky zkušeností v Německu. Atribut je numerického charakteru, nicméně kvůli několika odpovědím má i kategorické rysy.

Seniority level

Seniorita pozice, kterou daný člověk zastává. Největší počet respondentů (565) je na seniorních pozicích.

Your main technology / programming language

Hlavní technologie (či programovací jazyk), které se daná osoba věnuje. Nejvíce respondentů (184) uvádí *Javu*.

Other technologies/programming languages you use often

Vedlejší technologie, které respondenti často využívají. Nejčastěji byl uváden *Javascript / Typescript* (44 odpovědí).

Yearly brutto salary (without bonus and stocks) in EUR

Číslo vyjadřující roční hrubou mzdu (bez bonusů či případných odměn v podobě akcií) v eurech. Průměrná hodnota je *80 200 000*, což už na první pohled nevypadá jako reálná hodnota a skutečně není, je to dané odlehlými hodnotami. Více se odlehlým hodnotám věnuje kapitola 1.3. 75 % hodnot je nižších než *80 000* a medián je *70 000*.

Yearly bonus + stocks in EUR

Textový řetězec reprezentující číselnou hodnotu vyjadřující roční bonusy společně s akciemi společnosti, které respondent získal, uváděny v eurech. Nejvíce respondentů (227) nedostalo žádný bonus, tzn. dostali *0*. Atribut je numerického charakteru, nicméně kvůli několika odpovědím má i kategorické rysy.

Annual brutto salary (without bonus and stocks) one year ago. Only answer if staying in the same country

Číslo vyjadřující roční hrubou mzdu, kterou respondent vydělával před rokem. Respondenti měli odpovídat pouze v případě, že před rokem pracovali ve stejné zemi jako v době odpovědi. Průměrná hodnota je *630 000*, což obdobně jako u současné hrubé mzdy uvedené výše, nevypadá jako reálná hodnota. Medián je *65 000* a 75 % hodnot je nižších než *75 000*.

Annual bonus+stocks one year ago. Only answer if staying in same country

Textový řetězec reprezentující číselnou hodnotu bonusů, opět včetně případných akcií, jaké respondenti obdrželi před rokem. Nejvíce respondentů (200) uvedlo že, stejně jako o rok později, nedostali bonusy žádné (tzn. *0*). Atribut je numerického charakteru, nicméně kvůli několika odpovědím má i kategorické rysy.

Number of vacation days

Textový řetězec reprezentující číselnou hodnotu vyjadřující počet dní dovolené, na které respondenti měli nárok. Nejvíce respondentů (488) uvedlo, že měli nárok na *30* dní dovolené. Atribut je numerického charakteru, nicméně kvůli několika odpovědím má i kategorické rysy.

Employment status

Typ úvazku, na který respondenti pracují. Zdaleka nejčastěji (1190 odpovědí) respondenti uvedli, že jsou zaměstnancemi na plný úvazek (Full-time employee).

Contract duration

Doba, na kterou mají respondenti uzavřenou smlouvu se zaměstnavatelem. Zdaleka nejvíce (1159) lidí uvedlo, že mají smlouvu na dobu neurčitou (Unlimited

contract).

Main language at work

Jazyk, kterým respondenti v práci primárně komunikují. Celkem respondenti uvedli 14 jazyků, nejčastějším z nich byla angličtina (1020 odpovědí).

Company size

Velikost firmy, v jaké respondenti pracují. Na výběr bylo několik intervalů (0-10, 11-50, 51-100, 101-1000, 1000+). Nejvíce respondentů (448) uvedlo, že pracují ve firmě o 1000+ zaměstnancích. Hodnoty atributu jsou intervaly určující počet zaměstnanců ve firmě.

Company type

Označení, čím se společnost zabývá. Nejvíce respondentů (760) uvedlo, že se jejich společnost zabývá vývojem produktů.

Have you lost your job due to the coronavirus outbreak?

Odpověď na otázku, zda respondenti přišli o práci v důsledku koronavirové pandemie. Většina (1162) uvedla *no*, tedy ne.

Have you been forced to have a shorter working week (Kurzarbeit)? If yes, how many hours per week

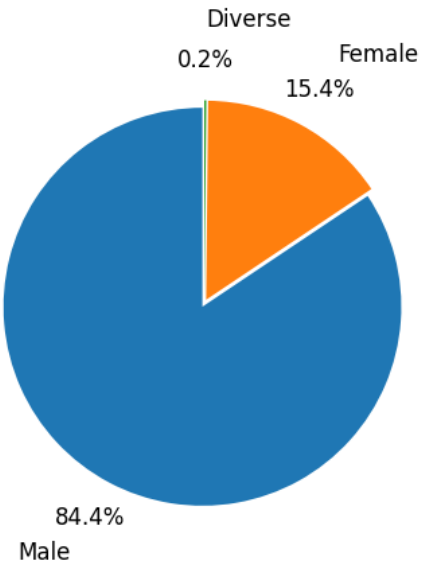
Číselná hodnota určující, zda byli respondenti nuceni pracovat na kratší uvázek (tzv. Kurzarbeit) a pokud ano, na kolik hodin týdně. Tento počet hodin odpovídá uvedené hodnotě. Průměrná hodnota je zkrácený uvázek na *12.9* hodiny, ale většina respondentů uvedla *0* hodin, takže není jasné, zda se na ně Kurzarbeit vztahuje, ovšem pouze na 0 hodin týdne, a nebo naopak Kurzarbeit neměli.

Have you received additional monetary support from your employer due to Work From Home? If yes, how much in 2020 in EUR

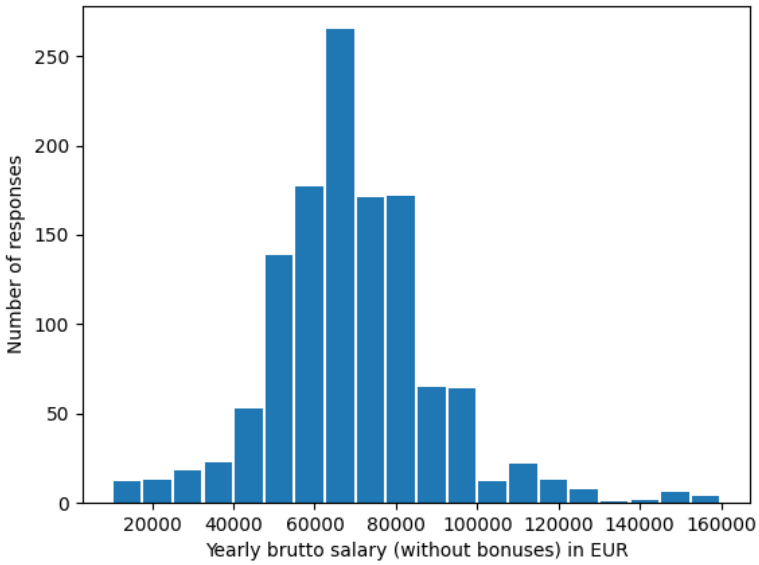
Otázka, zda respondenti obdrželi od zaměstnavatele finanční podporu kvůli práci z domu a pokud ano, tak kolik (v roce 2020, v eurech). Nejvíce respondentů (161) uvedlo, že neobdrželi nic (tzn. obdrželi *0*).

1.2 Rozložení hodnot

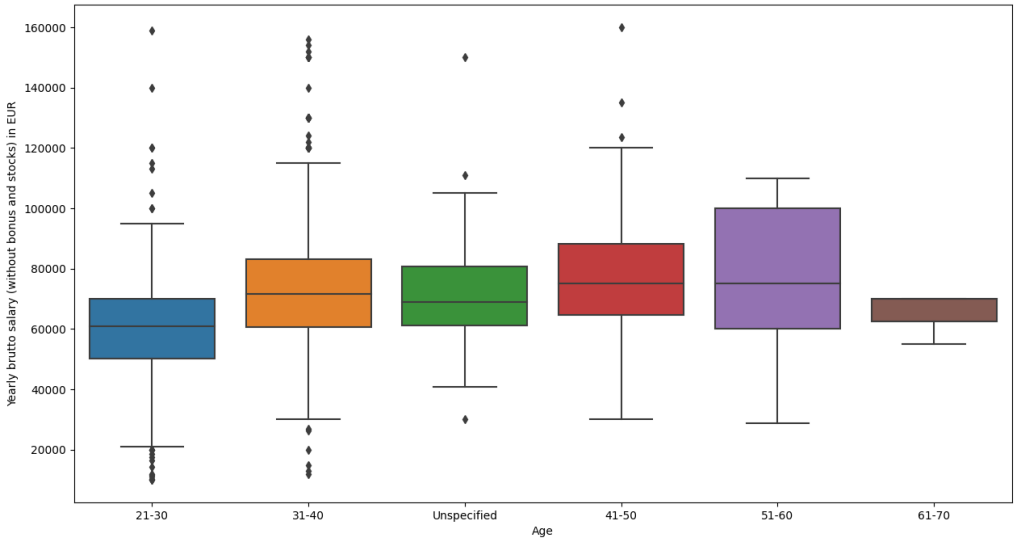
Následující řádky se budou zabývat rozložením hodnot vybraných atributů. Pro vizualizaci těchto rozložení budou použity různé typy grafů. Některé údaje byly za účely grafické vizualizace upraveny odstraněním krajních hodnot či úpravou atributů do vhodné podoby. Krajními hodnotami se více zabývá kapitola 1.3 a úpravou dat kapitoly 2.6 a 2.5.



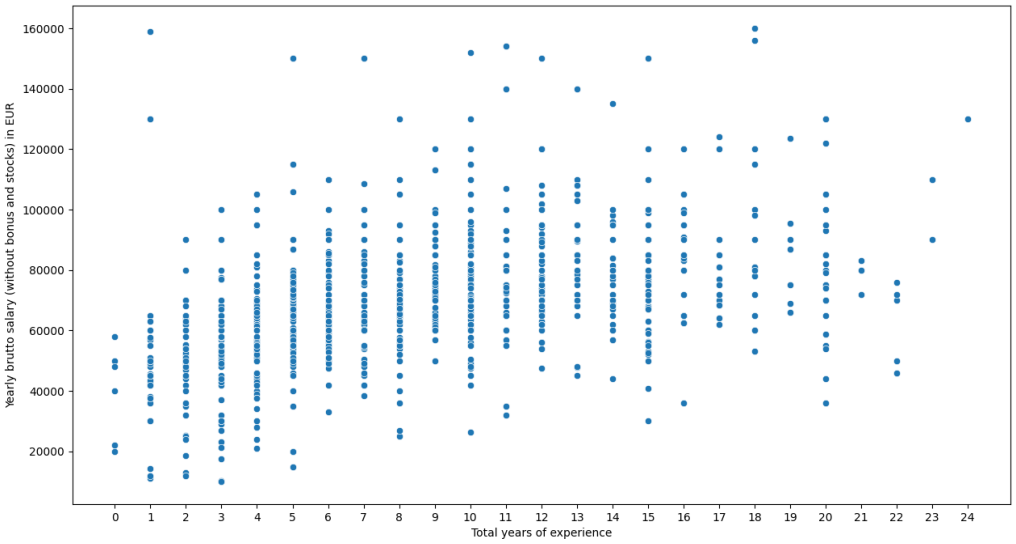
Obrázek 1: Koláčový graf zobrazující rozložení pohlaví respondentů



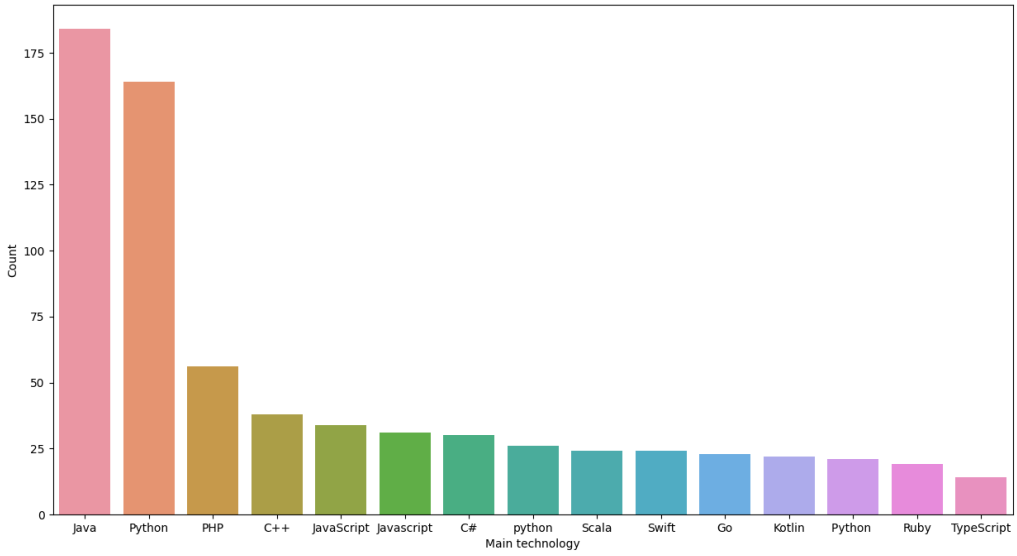
Obrázek 2: Histogram rozložení hrubé roční mzdy (v EUR)



Obrázek 3: Krabicový graf zobrazující hrubou roční mzdu (v EUR) v souvislosti s věkem



Obrázek 4: Bodový graf zobrazující hrubou roční mzdu (v EUR) v souvislosti se zkušenostmi



Obrázek 5: Graf zobrazující nejčastěji používané hlavní technologie

1.3 Odlehlé hodnoty

Jak bylo naznačeno již v předchozí kapitole 1.2, datová sada obsahuje hodnoty, které jsou nekonzistentní se zbytkem dat. To může být způsobeno buď chybou respondenta při vyplňování údajů, úmyslným uváděním nepravdivých informací a nebo zkrátka tím, že je respondent něčím odlišný.

Výpočet odlehlých hodnot je založen na mezikvartilním rozložení. Q1 a Q3 reprezentují 1. a 3. kvartil, IQR je mezikvartilním rozložení. Výpočet reprezentují následující rovnice:

$$IQR = Q3 - Q1$$
$$upper\_limit = Q3 + 1.5 \times IQR$$
$$lower\_limit = Q1 - 1.5 \times IQR$$

Hodnoty nižší než *lower_limit* či vyšší než *upper_limit* byly považovány za odlehlé.

Nejvýrazněji se odlehlé hodnoty projevují u údajů, kde měli respondenti zadávat číselnou hodnotu, tedy numerické atributy. Pouze takové atributy zde budou popsány. Jedná se o následující:

- Věk (Age) – standardní odchylka věku je *5.66*, nejvyšší věk je 69 let a nejnižší 20. Odlehlých hodnot je 40.
- Hrubý roční plat (Yearly brutto salary (without bonus and stocks) in EUR) – standardní odchylka dat je 2.82×10^9, což je hodnota více než *40 000* krát vyšší než medián (*70 000*). Maximální zadaná roční mzda je *100 000 000 000*, což je, jedná-li se o eura, téměř třetina hrubého domácího produktu České republiky a lze tedy říct, že se pravděpodobně nejedná o reálnou hodnotu. Celkově bylo odlehlých hodnot u toho atributu nalezeno 80.
- Hrubý roční plat před rokem (Annual brutto salary (without bonus and stocks) one year ago. Only answer if staying in the same country) – standardní odchylka je 1.68×10^7 a nejvyšší hodnota je *500 000 000*. Celkově bylo nalezeno 46 odlehlých hodnot tohoto atributu.
- Kurzarbeit (Have you been forced to have a shorter working week (Kurzarbeit)? If yes, how many hours per week) – standardní odchylka byla *15.27*, maximální hodnota byla *40* a odlehlé hodnoty nebyly nalezeny žádné.

Následující atributy původně nebyly numerické, nicméně po odstranění několika hodnot, které nebyly numerického charakteru, k těmto atributům lze jako k numerickým přistupovat. Jedná se o následující:

- Zkušenosti (Total years of experience) – standardní odchylka je *11.955* a průměrná hodnota *9.14* let. Nejvyšší hodnota je *383*, což bude nejspíše nereálná hodnota, neboť nejstarší člověk, jehož věk byl dokumentován, se dožil „pouhých" 122 let[2] a je proto nepravděpodobné, že by byl člověk s 383 roky zkušeností v IT. Nejnižší hodnota byla *0*. Medián byl 8 let. Bylo nalezeno celkem 22 odlehlých hodnot.

- Zkušenosti v Německu (Years of experience in Germany) – standardní odchylka je *3.716*, maximální hodnota *30*. Medián byly 3 roky. Odlehlých hodnot bylo celkově 47.

- Roční bonusy (Yearly bonus + stocks in EUR) – standardní odchylka je 1.745×10^8, maximální hodnota je 5×10^9. Medián byl *5 000*. Odlehlých hodnot bylo celkově 132.

- Roční bonusy před rokem (Annual bonus+stocks one year ago) – standardní odchylka je 2.020×10^6, maximální hodnota je 5×10^7. Medián byl *5 000*. Odlehlých hodnot bylo celkově 19.

- Počet dnů dovolené (Number of vacation days) – standardní odchylka je 10.761, maximální hodnota je 365. Medián byl *28*. Odlehlých hodnot bylo celkově 71.

1.4 Chybějící hodnoty

Tato část se bude věnovat chybějícím hodnotám. Bude se zjišťovat, zda-li řádek tabulky s údaji obsahuje pouze jednu chybějící hodnotu a nebo více. Vzhledem k povaze některých atributů je zřejmé, že chybějící hodnoty nemusí znamenat chybu. Jedná se o atributy, na které měli respondenti odpovědět pouze v případě, zda se na ně daná otázka vztahuje. Tyto atributy nebudou brány v případě chybějící hodnoty jako prázdné. Konkrétně se jedná o atributy následující –

- Have you received additional monetary support from your employer due to Work From Home? If yes, how much in 2020 in EUR

[2] `https://en.wikipedia.org/wiki/List_of_the_verified_oldest_people`

- Have you been forced to have a shorter working week (Kurzarbeit)? If yes, how many hours per week
- Annual bonus+stocks one year ago. Only answer if staying in same country
- Annual brutto salary (without bonus and stocks) one year ago. Only answer if staying in the same country

Celkově bylo zjištěno, že počet řádků, které obsahují alespoň jednu chybějící hodnotu, je 594. Z 1253 řádků je tedy kompletně vyplněných pouze 659, což je *52.5 %*. Počet řádků, které obsahují více než jednu chybějící hodnotu, je pak 199.

Byly zjištěny následující počty chybějících hodnot pro jednotlivé atributy:

- Timestamp – 0
- Age – 27
- Gender – 10
- City – 0
- Position – 6
- Total years of experience – 16
- Years of experience in Germany – 32
- Seniority level – 12
- Your main technology / programming language – 127
- Other technologies/programming languages you use often – 157
- Yearly brutto salary (without bonus and stocks) in EUR – 0
- Yearly bonus + stocks in EUR – 424
- Number of vacation days – 68
- Employment status – 17
- Contract duration – 29
- Main language at work – 16
- Company size – 18
- Company type – 25
- Have you lost your job due to the coronavirus outbreak – 20

1.5 Korelační analýza

Následující řádky se zabývají korelační analýzou. Byl počítán Pearsonův korelační koeficient[3]. Tento projekt se zabývá korelací pouze numerických atributů. Jak bylo zmíněno již v kapitole 1.1, ne všechny kategorické hodnoty jsou opravdu kategorické, některé mají numerický charakter a jsou kategorickými pouze z důvodu několika výjimek. Tento problém je řešen již v kapitole 1.3 a proto bude stejným způsobem řešen i zde, tzn. kategorické hodnoty budou odstraněny, aby se z atributu stal skutečně numerický a mohla proběhnout korelační analýza.

Pro přehlednost budou názvy atributů zkráceny následovně:

- Age = Age
- Salary = Yearly brutto salary
- Prev. salary = Annual brutto salary (without bonus and stocks) one year ago.
- Kurzarbeit = Have you been forced to have a shorter working week (Kurzarbeit)
- Exp. = Total years of experience
- German exp. = Years of experience in Germany
- Yearly bonus = Yearly bonus + stocks in EUR
- Previous bonus = Annual bonus+stocks one year ago
- Vacation = Number of vacation days

Výsledky korelace mezi jednotlivými, již originálně numerickými atributy, jsou následující:

	Age	Salary	Prev. salary	Kurzarbeit
Age		-0.017	-0.024	-0.003
Salary			0.999	-0.044
Prev. salary				-0.050

Všechny hodnoty vykazují velmi nízkou míru korelace, jedinou výjimkou je korelace mezi současným platem a platem před rokem. Tam je korelace velmi výrazná, což odpovídá očekávání.

Výsledky korelace mezi nově získanými atributy po odstranění kategorických jsou pak následující:

[3] `https://en.wikipedia.org/wiki/Pearson_correlation_coefficient`

	Exp	German exp.	Yearly bonus	Previous bonus	Vacation
Age	0.482	0.544	−0.022	−0.027	0.040
Salary	−0.019	−0.022	0.999	0.999	0.192
Prev. salary	−0.051	−0.031	0.999	0.999	0.524
Kurzarbeit	−0.003	0.003	−0.050	−0.054	0.007
Exp.		0.207	−0.052	−0.058	0.024
German exp.			−0.029	−0.037	0.050
Yearly bonus				0.999	0.195
Previous bonus					0.576

Ve výše uvedených numerických atributech, které byly získány úpravou původních ne-numerických, byly nalezeny nové korelace. Konkrétně to jsou tyto:

- Věk, Roky zkušeností, střední korelace
- Věk, Zkušenosti v Německu, střední korelace
- Roční hrubá mzda, Roční bonusy, velmi výrazná korelace
- Roční hrubá mzda, Roční bonusy před rokem, velmi výrazná korelace
- Roční hrubá mzda, Počet dnů dovolené, slabší korelace
- Roční hrubá mzda před rokem, Roční bonusy, velmi výrazná korelace
- Roční hrubá mzda před rokem, Roční bonusy před rokem, velmi výrazná korelace
- Roční hrubá mzda před rokem, Počet dní dovolené, střední korelace
- Roky zkušeností, Roku zkušeností v německu, slabší korelace
- Roční bonusy, Roční bonusy před rokem, velmi výrazná korelace
- Roční bonusy, Počet dnů dovolené, slabší korelace
- Roční bonusy před rokem, Počet dnů dovolené, silnější korelace

2 Dolovací úloha

Jako dolovací úloha byla zvolena již v zadání doporučená – *Predikce výše platu na základě ostatních atributů*.

2.1 Odstranění nerelevantních atributů

V první fázi přípravy dat bylo zapotřebí odstranit atributy, které nejsou relevantní pro zvolenou dolovací úlohu. Za účelem predikce platu jsme se rozhodli, že bude brána v potaz pouze hrubá roční výplata, bez dalších prémií či jiných bonusů. Konkrétně byly odstraněny následující atributy:

- Timestamp
- Yearly bonus + stocks in EUR
- Annual bonus+stocks one year ago. Only answer if staying in same country
- Have you lost your job due to the coronavirus outbreak?
- ”Have you been forced to have a shorter working week (Kurzarbeit)? If yes how many hours per week”
- ”Have you received additional monetary support from your employer due to Work From Home? If yes how much in 2020 in EUR”

2.2 Řešení chybějících hodnot

V datové sadě se nacházejí řádky, které neobsahují všechny atributy, viz 1.4. Aby bylo možné dosáhnout požadovaných výsledků, je zapotřebí hodnoty, které chybějí, doplnit. Toto doplňování se provádí dvěma způsoby.

V případě chybějících hodnot numerických atributů byly doplněny hodnoty pomocí mediánu tohoto atributu. Doplnění hodnoty mediánem by na výsledky nemělo mít vliv, neboť medián zbylých dat se nezmění. V případě chybějících kategorických atributů se doplnila hodnota *”undefined”*, což je pro účely dolovací úlohy nová třída, která se však nemá brát v potaz.

2.3 Řešení odlehlých hodnot

Odlehlé hodnoty jsou řešeny pomocí filtrování s využitím mezikvartilního rozpětí, tak, jak je popsáno v kapitole 1.3. Tím bude zajištěno, že transformace dat proběhnou bez problémů, které by odlehlé hodnoty mohly způsobit.

2.4 Úprava kategorických dat pro obě datové sady

Ještě před započetím transformace kategorických či numerických dat byla některá kategorická data upravena. Jelikož jsou v mnoha sloupcích různé hodnoty (napří-

klad ve sloupci *Your main technology / programming language* někteří respondenti uváděli dlouhý výčet všech technologií, kterým se hlavně věnují), rozhodli jsme se tyto hodnoty upravit.

Cílem této úpravy je, aby bylo možné data snáze klasifikovat. Například v rámci používaných technologií došlo k roztřízení jednotlivých technologií do obecnějších kategorií, konkrétně *front-end, back-end, cloud a mobile*.

Obdobně byly seskupeny i další kategorické atributy do předem nadefinovaných skupin. Rozdělení bylo založeno na určení nejlepší slovní shody[4] mezi konkrétní hodnotou a jednou z předdefinovaných skupin.

V případě, že některá z hodnot měla shodu příliš nízkou a tedy se ji nepovedlo spojit s některou z určených kategorií, byla přepsána na *"undefined"*.

Takto upravená data byla používána jak pro účely diskretizace numerických atributů, tak pro transformaci kategorických atributů.

2.5 Diskretizace numerických atributů

Cílem diskretizace numerických atributů bylo převést numerické atributy do kategorické podoby. Toho bylo dosaženo tím, že sloupce s numerickými hodnotami (které již byly vyčištěny od odlehlých hodnot) byly roztříděny do pěti rovnoměrně velkých intervalů na základě hodnot. Hodnoty v jednotlivých intervalech byly posléze přepsány na jednu z pěti kategorií – *low, moderate, medium, high, big* Pro sloupec *Company size* jsou používány hodnoty *tiny, small, medium, large, big*

2.6 Transformace kategorických atributů

V rámci transformace kategorických atributů na numerické došlo k odstranění, krom již zmíněných sloupců, také sloupce *City*, neboť pro pozdější transformaci by nebyl vhodný. Krom toho byly používány již upravené kategorické atributy, které byly popsány v rámci úpravy dat v kapitole 2.4.

Prvním krokem samotné transformace bylo, že atributy s numerickými rysy, které však numerické kvůli pár hodnotám nejsou (popsány v kapitole 1.1), byly upraveny a tyto hodnoty odstraněny.

Druhým krokem bylo převedení všech zbývajících kategorických atributů na numerické s využitím metody *One-Hot Encoding*[5], tedy z atributů obsahujících

[4] Pro testování podobnosti slov je použita knihovna `https://github.com/seatgeek/fuzzywuzzy`.

[5] `https://en.wikipedia.org/wiki/One-hot`

kategorické hodnoty byly vytvořeny atributy takové, že každá z původních kategorických hodnot je nyní atributem sama o sobě a obsahuje hodnoty buď 0 a nebo 1. Tato metoda byla zvolena z důvodu, že umí řešit problémy s atributy, které nelze jednoduše "srovnávat"a přesto je převede na numerický ekvivalent. Jedinou nevýhodou je, že počet nových sloupců je výrazně vyšší než původní, což ovšem bylo částečně vyřešeno již zmíněnou přípravou dat.

3 Návod ke spuštění a popis souborů

Součástí odevzdaného archivu jsou následující soubory:

- `main.py` – Python skript obsahující zdrojové kódy k explorativní analýze a úpravě dat, lze spustit (z kořenového adresáře) příkazem `python3 ./main.py`. Vyžaduje Python 3.10 a vyšší.
- `requirements.txt` – Textový soubor obsahující seznam knihoven nutných ke spuštění, lze použít společně s programem pip pro jejich stažení: `pip install -r requirements.txt`
- `IT_Salary_Survey_EU_2020.csv` – Stažená datová sada
- `categoric_transformed.csv.csv` – Upravená datová sada obsahující pouze numerické atributy
- `numeric_transformed.csv` – Upravená datová sada obsahující pouze kategorické atributy