

Fakulta informačních technologií VUT v Brně

Příprava dat a jejich popisná charakteristika

Ukládání a příprava dat – projekt, 2. část

David Chocholatý, František Nečas, Ondřej Ondryáš
3. října 2023

Obsah

1 Cíle projektu a použitá datová sada	2
2 Explorativní analýza datové sady	2
2.1 Atributy datové sady a jejich počáteční čištění	2
2.2 Charakterizace numerických atributů	3
2.3 Charakterizace kategorických atributů	6
2.4 Identifikace odlehlych hodnot	8
2.5 Analýza chybějících hodnot	8
2.6 Korelační analýza	9
3 Příprava dat	9
3.1 Čištění dat	9
3.2 Úprava datové sady – redukce atributů	12
3.3 Transformace dat	13
4 Zobrazení výsledné očištěné sady	14
Přílohy	24

1 Cíle projektu a použitá datová sada

Cílem projektu je provést explorativní analýzu datové sady a upravit ji do podoby vhodné pro použití s dolovací úlohou. Úlohou je predikce výše platu na základě různých atributů.

Použita byla datová sada „IT Salary Survey for EU region (2018–2020)“¹. Pochází z výsledků anonymního internetového dotazníku, který byl publikován každoročně od roku 2015 (datová sada obsahuje záznamy pouze z let 2018 až 2020). V roce 2020 jej vyplnilo 1 253 respondentů. Dotazník byl zaměřen na IT specialisty v Německu, přesto ale cca 9 % respondentů pocházelo z jiné země.

Nahlédnutím na složení atributů v datových sadách odpovídajících jednotlivým rokům je možné zjistit, že se mezi roky otázky v průzkumu zásadně změnily. Pro použití všech sad by tedy bylo nutné provést integraci. Protože ta není požadavkem projektu, pro analýzu a přípravu byla zvolena pouze sada z roku 2020, která je nejrozsáhlejší (1 253 záznamů proti 991 a 765 z let 2019, resp. 2018), nejdetailejnější (má nejvyšší počet otázek – atributů) a také nejnovější (což je pro predikci jistě vhodnější).

2 Explorativní analýza datové sady

Nejprve bylo provedeno hrubé zhodnocení kvality dat, dostupných atributů a jejich hodnot. Ihned v počátku byly zjištěny zjevné nekonzistence dat týkající se zejména nenumerických hodnot numerických atributů a různými podobami významově shodných hodnot kategorických atributů. Ty byly pro další explorativní analýzu odstraněny.

2.1 Atributy datové sady a jejich počáteční čištění

Atributy jsou v datové sadě pojmenovány stejně jako otázky v původním dotazníku, pro snadnější práci s nimi bylo tedy vhodné nejprve provést jejich přejmenování. Původní i nové názvy atributů jsou shrnuty v tabulce 2. Sada obsahuje 23 atributů, z toho 11 numerických a 12 kategorických.

Po načtení datového souboru byly jako typ object uloženy významově numerické atributy *TotalExperience*, *GermanyExperience*, *YearlyBonus*, *PreviousYearlyBonus*, *VacationDays* a *WFHSupport*. To naznačovalo, že v souboru nachází objekty s neplatnými (nečíselnými) hodnotami těchto atributů. Náhledem na tyto objekty byly odhaleny dva významné zdroje nekonzistenčí:

1. použití desetinné čárky místo majoritně zastoupené desetinné tečky,
2. výskyt upřesňujících textových komentářů k numerickým hodnotám.

V obou případech šlo o jednotky výskytů. Oba druhy nekonzistenčí byly odstraněny manuální transformací hodnot. Problematickým se může jevit atribut *VacationDays*, ve kterém se kromě číselných hodnot devětkrát vyskytla také odpověď „unlimited“ (neomezeno). Vzhledem k tomu, že tento atribut nebyl v dalších fázích použit, jeho úprava není významná, nabízí se ale buď nahrazena za vhodnou hodnotu vyjadřující „dostatečně velký“ počet dní dovolené (např. 365 ve smyslu celého roku), nebo nahrazení zpočátku prázdnou hodnotou a poté např. vhodným výpočtem ze sousedních hodnot.

Šum byl identifikován také u kategorických atributů. V rámci počátečního čištění zde byly ručně odšuměny atributy *EmploymentStatus*, *LostJobInCovid* a částečně *MainTechnology* a

¹<https://www.kaggle.com/datasets/parulpandey/2020-it-salary-survey-for-eu-region>

Seniority – ve všech případech šlo opět o jednotky či nižší desítky „příliš přesně“ uvedených textových odpovědí v dotazníku.

2.2 Charakterizace numerických atributů

Atribut	Výskytů	Avg	Med	Mod	StdDev	IQR
Age	1 226	32,51	32	30	5,67	6
TotalExperience	1 237	9,06	8	10	11,88	7
GermanyExperience	1 220	3,71	3	2	3,64	4
YearlySalary	1 253	$8,02 \cdot 10^7$	70 000	60 000	$2,83 \cdot 10^9$	21 200
YearlyBonus	824	$6,09 \cdot 10^6$	5 000	0	$1,74 \cdot 10^7$	20 000
PreviousYearlySalary	885	632 245	65 000	65 000	$1,68 \cdot 10^7$	20 000
PreviousYearlyBonus	613	103 030	5 000	0	$2,02 \cdot 10^6$	36 400
VacationDays	1 184	30,62	28	30	29,60	3
KurzarbeitHours	373	12,97	0	0	15,28	30
WFHSupport	458	443,95	200	0	855,20	600

Počet výskytů, aritmetický průměr, medián, modus, směrodatná odchylka, mezikvartilové rozpětí

Atribut	Min	$q_{0,25}$	Med	$q_{0,75}$	Max	IQR
Age	20	29	32	35	69	6
TotalExperience	0	5	8	12	383	7
GermanyExperience	0	1	3	5	30	4
YearlySalary	10 001	58 800	70 000	80 000	$1 \cdot 10^{11}$	21 200
YearlyBonus	0	0	5 000	20 000	$5 \cdot 10^9$	20 000
PreviousYearlySalary	11 000	55 000	65 000	75 000	$5 \cdot 10^8$	20 000
PreviousYearlyBonus	0	0	5 000	36 400	$5 \cdot 10^7$	36 400
VacationDays	0	27	28	30	365	3
KurzarbeitHours	0	0	0	30	40	30
WFHSupport	0	0	200	600	10 000	600

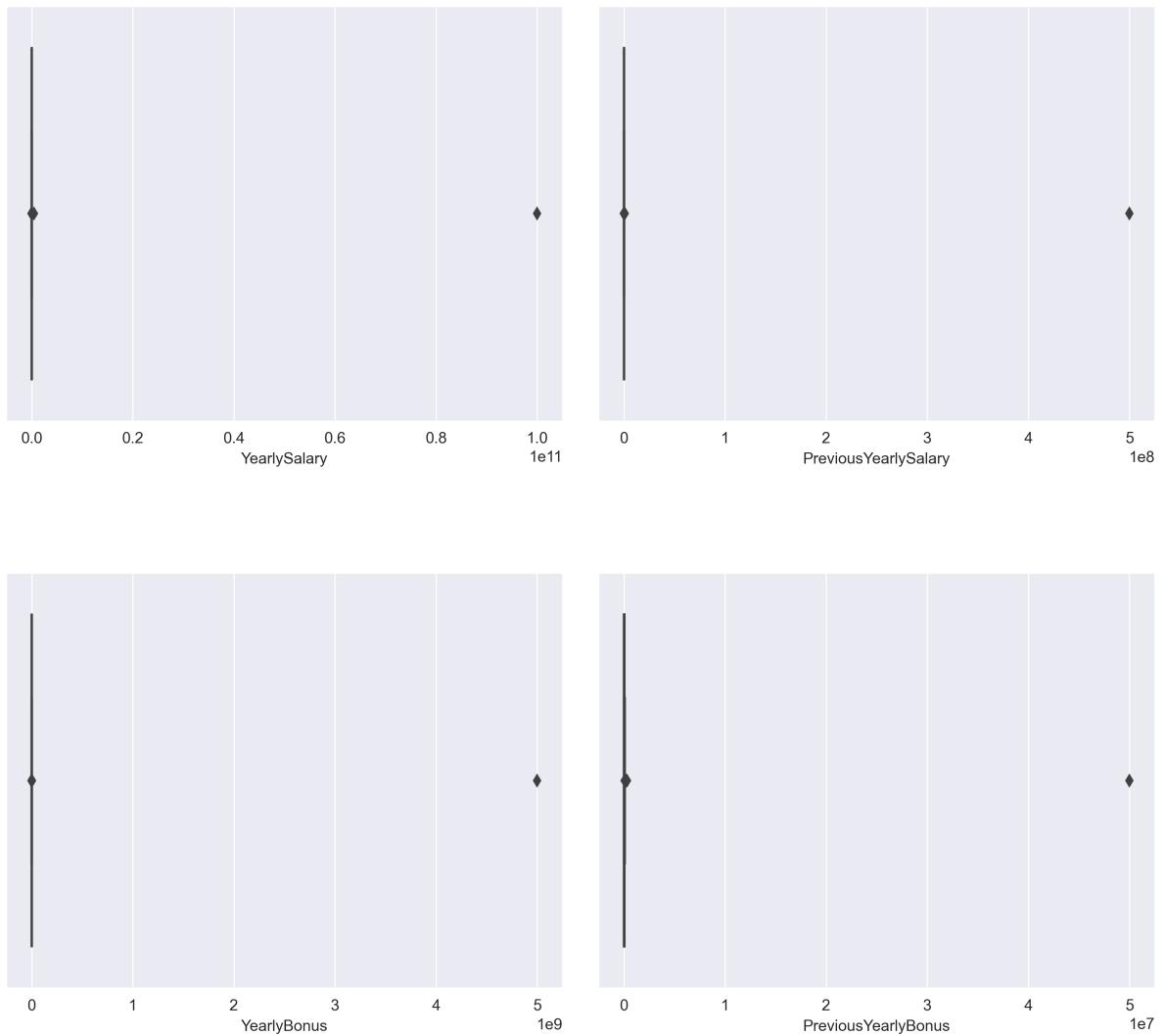
Minimum, 1. kvartil, medián, 3. kvartil, maximum, mezikvartilové rozpětí

Tabulka 1: Popisné charakteristiky numerických atributů v datové sadě.

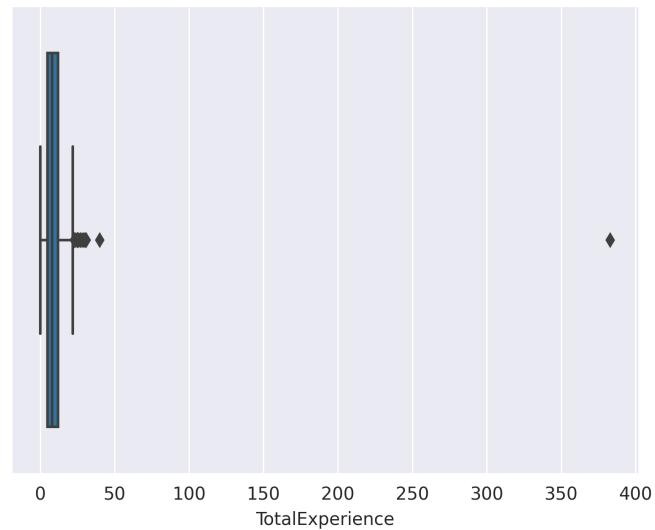
Tabulka 1 shrnuje základní popisné charakteristiky hrubě očištěných numerických atributů. Na první pohled si lze povšimnout podezřelých hodnot atributů *YearlySalary*, *YearlyBonus*, *PreviousYearlySalary* a *PreviousYearlyBonus*, které mají řádově vyšší maximální hodnoty a extrémně odchýlené aritmetické průměry od mediánu, zjevně se v nich budou nacházet značně odlehlé hodnoty, které data vychylují. Obdobná situace zřejmě nastává i u atributu *TotalExperience*, i když zde je vychýlení nižší. Z grafů 1 a 2 je možné odhadnout, že ve všech případech půjde o jednotky takto odlehlých hodnot.

O něco pravidelnější jsou hodnoty atributů *Age* a *GermanyExperience*. V případě věku rozdělení připomíná normální rozdělení se středem v hodnotě 30, ačkoliv by se pak v datech nacházelo relativně mnoho odlehlých hodnot, viz obr. 3. Některé ze známých rozdělení by jistě bylo možné vysledovat i pro druhý jmenovaný atribut.

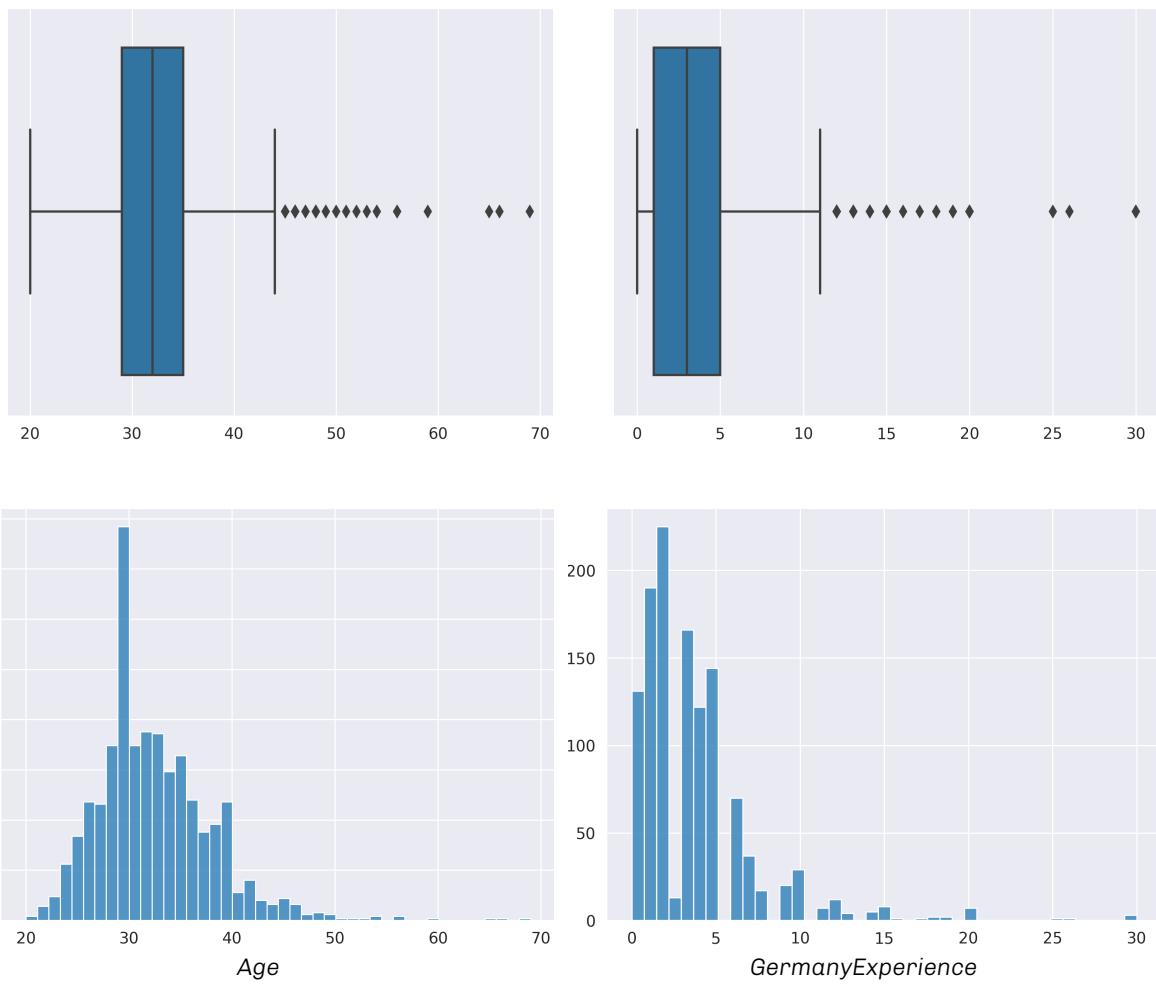
Atributy *VacationDays*, *KurzarbeitHours* a *WFHSupport* považujeme z hlediska úlohy za nevýznamné (viz 3.2), proto zde nejsou dále rozvedeny.



Obrázek 1: Krabicové grafy podezřelých atributů týkajících se mzdý. Ve všech čtyřech případech zřejmě způsobují vychýlení jednotky odlehčit hodnot. Z těchto grafů samotných však není zřejmé, zda jde o tytéž datové objekty.



Obrázek 2: Krabicový graf podezřelého atributu *TotalExperience*.



Obrázek 3: Krabicové grafy a histogramy atributů *Age* a *GermanyExperience*. Vertikální osa u histogramů značí počet výskytů.

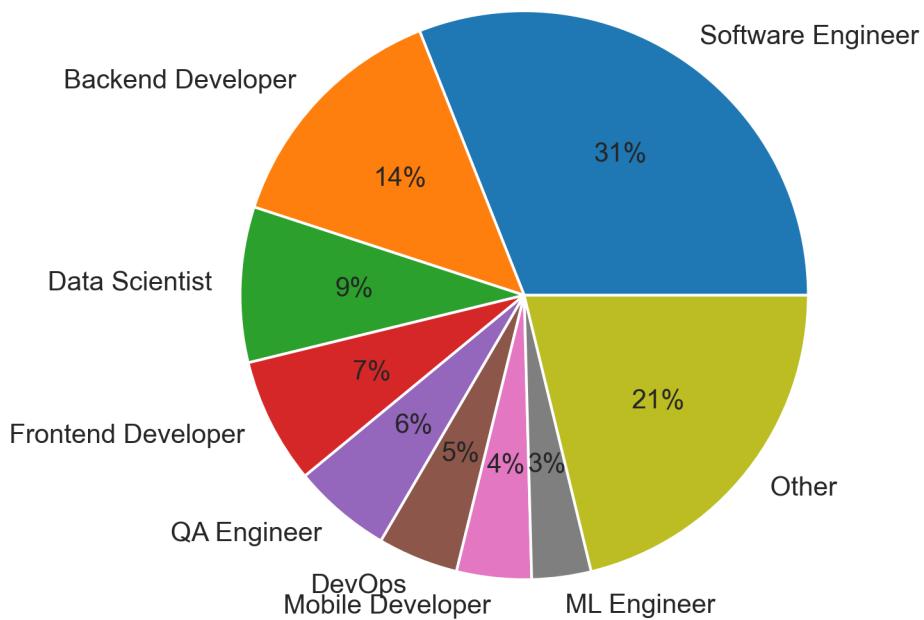
2.3 Charakterizace kategorických atributů

Atribut	Výskytů	Hodnot	Nejčastější hodnota	Výskytů	%
Gender	1 243	3	Male	1 049	84,53
City	1 252	118	Berlin	681	54,48
Position	1 247	148	Software Engineer	387	31,00
Seniority	1 238	9	Senior	565	45,71
MainTechnology	1 125	241	Java	193	17,19
SecondaryTechnology	1 096	562	Javascript / Typescript	44	4,02
EmploymentStatus	1 236	5	Full-time employee	1 195	96,68
ContractDuration	1 224	3	Unlimited contract	1 159	94,68
MainLanguage	1 237	14	English	1 020	82,51
CompanySize	1 235	5	1000+	448	36,17
CompanyType	1 228	63	Product	760	61,83
LostJobInCovid	1 233	2	No	1 167	94,64

Atribut *Gender* obsahuje tři hodnoty – *Male*, *Female* a *Diverse*, poslední jmenovaný však obsahují pouze dva objekty z 1 243, pro další zpracování byly tyto odstraněny.

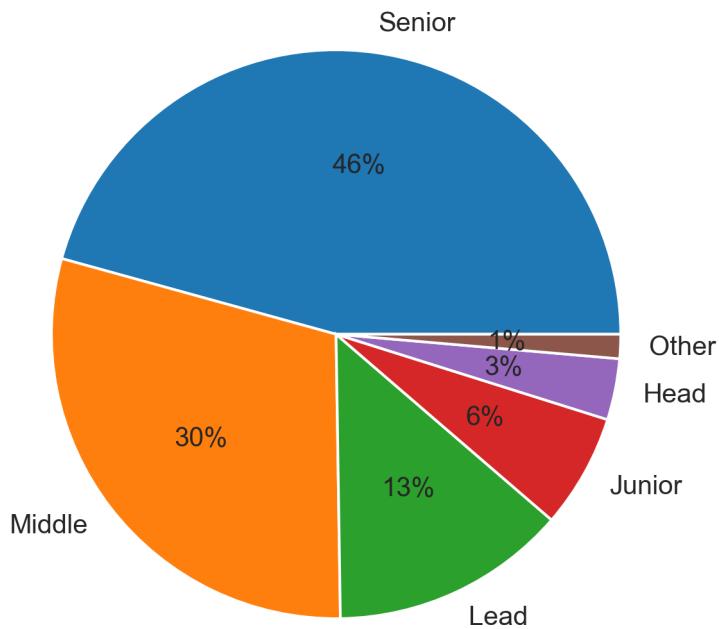
Většinu hodnot atributu *City* nabývají pouze jednotky objektů. Více než 100 výskytů mají pouze hodnoty Berlin (681) a Munich (236). Více než 10 výskytů mají hodnoty Frankfurt (44), Hamburg (39), Stuttgart (26) a Cologne (19). Nutno podotknout, že u tohoto atributu nedošlo k počátečnímu hrubému čištění – některé hodnoty se zde jistě opakují (zejména vlivem různých pravopisných podob stejného názvu města nebo překlepů).

Nad atributem *Position* bylo vhodné provést ruční spojení některých hodnot: více než 10 výskytů má pouze 11 hodnot ze 14, více než 40 výskytů pak 8 z nich, jak naznačuje graf 4.



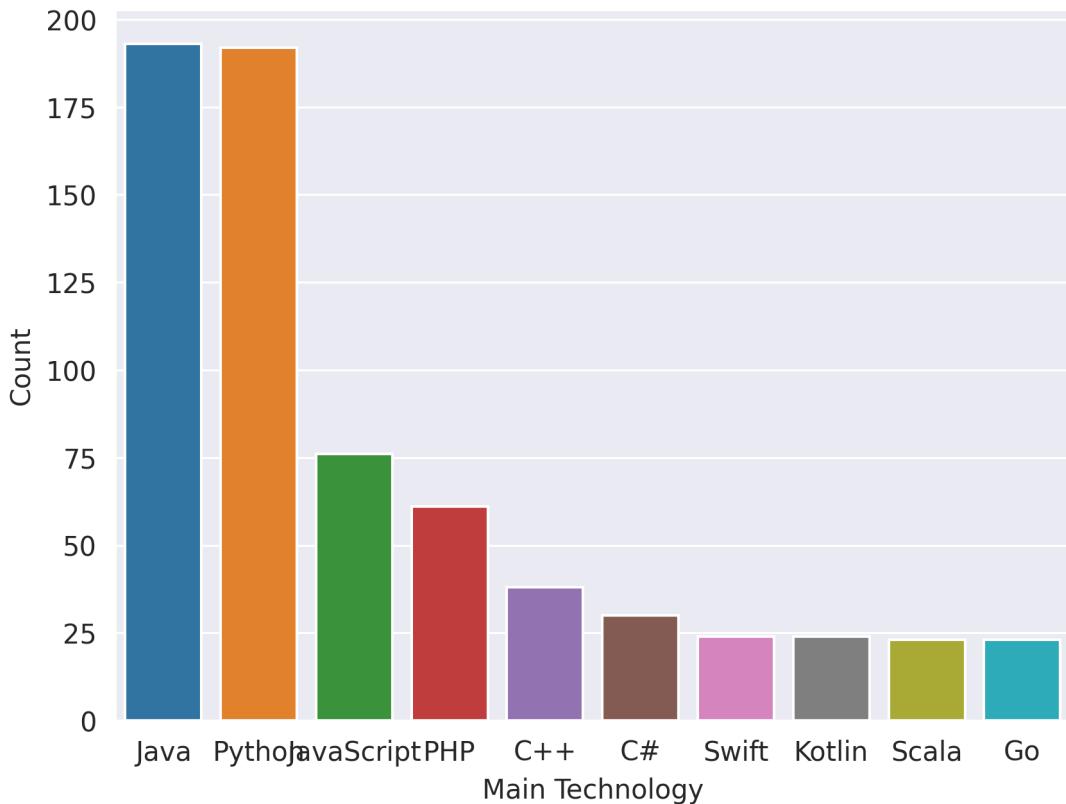
Obrázek 4: Rozložení hodnot atributu *Position* s více než 40 výskyty. Méně zastoupené hodnoty jsou shrnuty v kategorii *Other*.

Atribut *Seniority* nabývá především hodnot Senior (565), Middle (365), Lead (166), Junior (80) a Head (43). Méně zastoupeny jsou hodnoty Manager (11), Student (3), Intern (2) a Self employed (1). Nutno podotknout, že pro tento atribut byly při počátečním čištění nahrazeny některé odpovědi (např. Director, VP) nahrazeny hodnotou Manager.



Obrázek 5: Rozložení hodnot atributu *Seniority* s více než 11 výskyty.

Podstatným atributem pro predikci bude *MainTechnology* označující hlavní programovací jazyk respondenta. Při počátečním čištění byly odstraněny zejména některé překlepy. I tak ale atribut zůstává s 241 různými hodnotami značně různorodý, bylo by vhodné provést další ruční čištění. V některých případech respondenti také uvedli více různých hodnot – takové objekty je možné z datové sady vyřadit.



Obrázek 6: Graf zobrazující 10 nejběžnějších hlavních technologií mezi respondenty a jejich počet výskytů.

2.4 Identifikace odlehlých hodnot

Již na předchozích grafech jsme si mohli všimnout, že se u několika atributů vyskytují odlehlé hodnoty, buď zcela nesmyslné hodnoty (400 let celková zkušenost), nebo možné, ale silně odlehlé hodnoty. Další příklady takových hodnot můžeme vidět v sekci 4. Rozumným ořezáním rozsahu hodnot takových atributů se můžeme obou typů odlehlých hodnot. Můžeme použít například pravidlo tří sigma.

Zároveň si také můžeme povšimnout, že některé kategorické atributy, například *Gender*, obsahují pár vzorků s minimálně zastoupenými třídami, například třída *Diverse* u atributu *Gender*. Tato hodnota se v celé datové sadě vyskytuje pouze 2krát, tedy její jakékoliv porovnání s více zastoupenými majoritními třídami ztrácí přesnost, neboť se nejedná o reprezentativní vzorek dané minoritní třídy.

2.5 Analýza chybějících hodnot

Jak můžeme vidět na obrázku 7, u mnohých atributů chybí velké množství hodnot. Jedná se převážně o atributy, které mohly být pro respondenty náročné na vyplnění, a tedy je raději nechali prázdné. Například si nemuseli být jistí, jak mají správně odpovědět, proto nechali hodnotu raději nevyplněnou. Jedná se hlavně o atributy *KurzarbeitHours*, *WFHSupport*, *PreviousYearlyBonus* a *YearlyBonus*. U těchto atributů chybělo tolik hodnot, že snaha o jejich jakoukoliv opravu a doplnění by zaváděla více klamných hodnot, než by jich řešila. Zřejmě tedy má smysl tyto atributy úplně z datové sady vyřadit, neboť z nich nedokážeme získat smysluplné informace.

U atributu *PreviousYearlySalary* sice chybí stále ještě značné množství hodnot, ale protože

naopak u atributu *YearlySalary* nechybí téměř žádné hodnoty, nabízí se možnost na základě vysoké korelace mezi platy v předchozím roce a aktuálním roce doplnit platy v předchozím roce pomocí platů v aktuálním roce.

U chybějících hodnot u atributů *MainTechnology* a *SecondaryTechnology* existuje taková různorodost hodnot, že není možné rozumně odhadnout chybějící hodnoty ani je nikterak approximovat.

2.6 Korelační analýza

Na obrázku 8 můžeme vidět korelací různých atributů. Datová sada je sice silně zašuměná, ovlivněná chybějícími hodnotami a odlehlými hodnotami.

Přesto můžeme již nyní vidět silnou korelaci mezi atributy *Age* a *TotalExperience* nebo *GermanExperience*. Vyneseme-li si atributy *Age* a *TotalExperience* do samostatného grafu, získáme graf z obrázku 9. Vidíme, že data jsou zašuměná odlehlou hodnotou. Již takto ale můžeme usuzovat, že mezi oběma atributy lze vidět pravděpodobně silnou lineární závislost. Uvědomíme-li si, jaké informace atributy obsahují, můžeme zkonstatovat, že tyto výsledky jsou očekávatelné, nicméně nemůžeme říci, že by atributy byly duplicitní.

3 Příprava dat

Po zanalyzování kontextu a datové sady bude třeba datovou sadu připravit pro dolovací úlohu. Zvolili jsme si dolovací úlohu ze zadání (Predikce výše platu na základě ostatních atributů), ovšem vybrali jsme pouze některé atributy vhodné pro tuto úlohu.

3.1 Čištění dat

Datová sada je značně zašuměná, obsahuje mnohé atributy s chybějícími hodnotami, atributy s odlehlými nebo nesmyslnými hodnotami i atributy s nejednoznačnými či nezpracovatelnými hodnotami.

Pro získání datové sady pro dolovací úlohu musíme tedy datovou sadu očistit v následujících krocích.

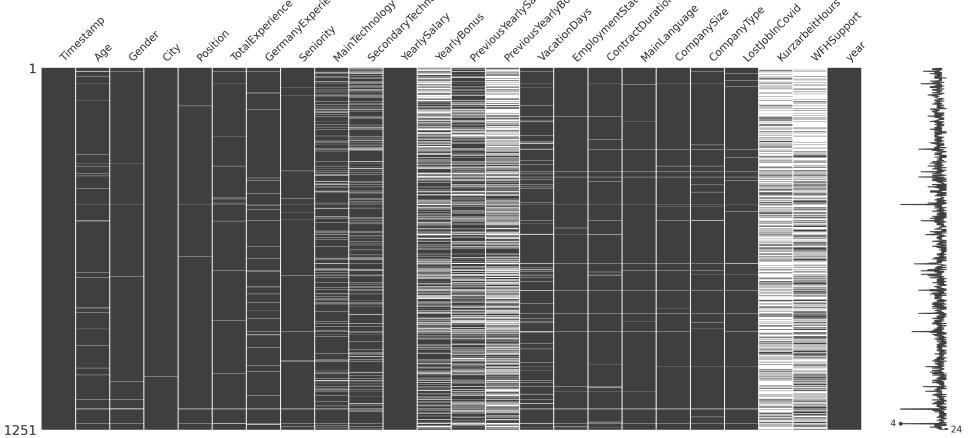
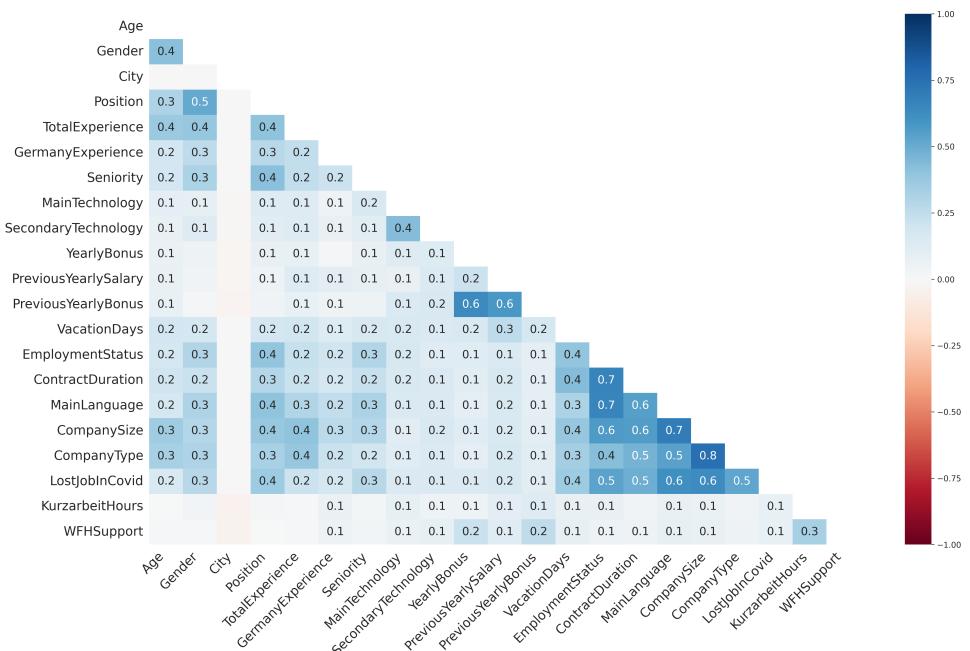
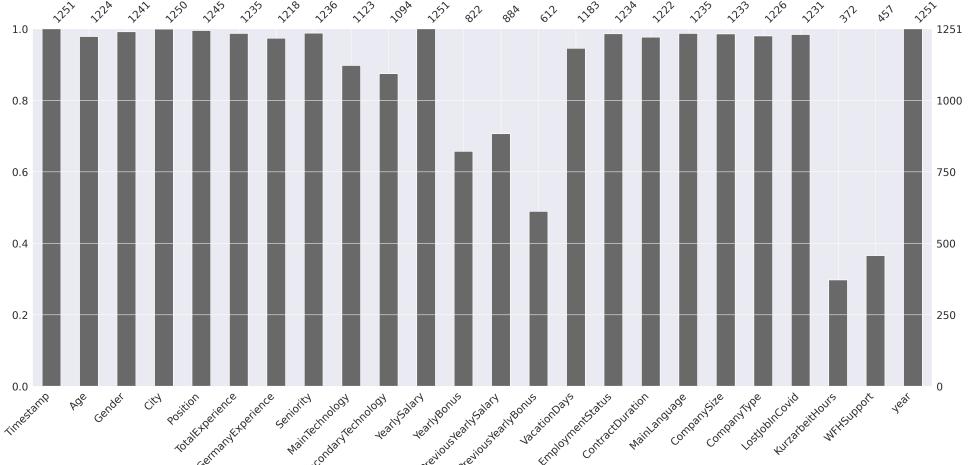
Ošetření chybějících hodnot

Pro atribut *Age* jsme místo chybějících hodnot použili aritmetický průměr, protože z histogramu rozložení můžeme vidět, že se jedná o normální rozložení na poměrně úzkém intervalu.

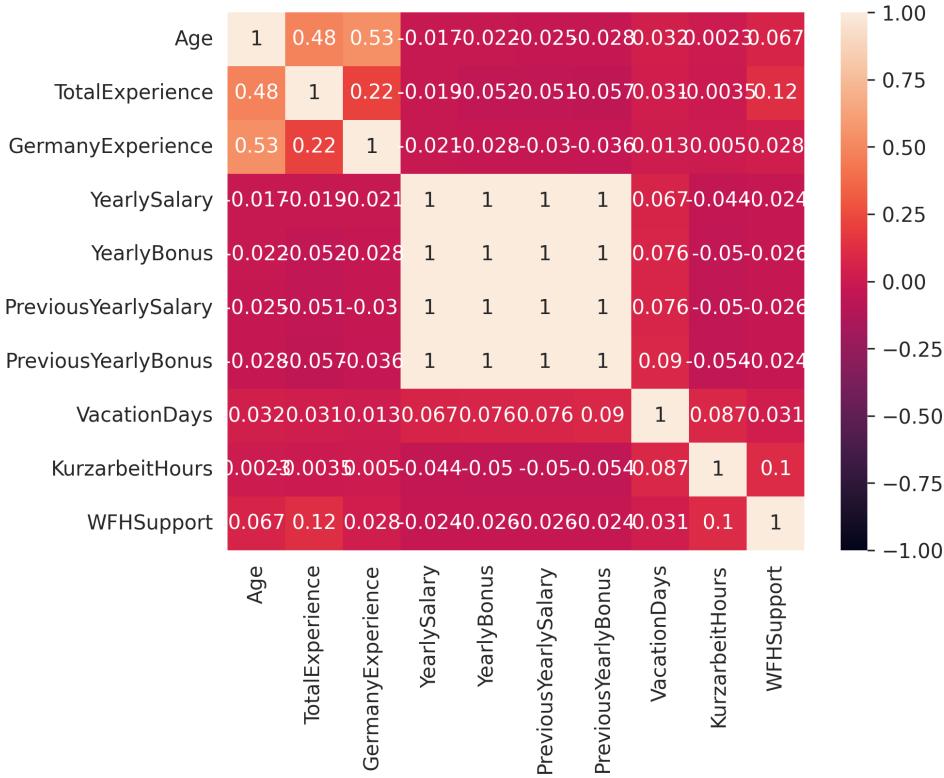
Připady chybějící hodnoty atributu *Gender* jsme odstranili z datové sady, protože z předchozí analýzy bylo zřejmé, že se jedná o malý počet objektů.

Kvůli vysoké korelace mezi platy předchozího roku (atribut *PreviousYearlySalary*) a aktuálními platy (atribut *YearlySalary*) jsme se rozhodli nahradit chybějící hodnoty u atributu *PreviousYearlySalary* odpovídající hodnotou platu z aktuálního roku. Tato úprava je také podpořena faktom, že typicky se plat mezi roky na stejně pozici (bez povýšení) zásadně nemění, maximálně v jednotkách procent.

Podobně, kvůli jisté korelace mezi atributy *GermanExperience* a *TotalExperience* jsme nahradili chybějící hodnoty u atributu *GermanExperience* hodnotou atributu *TotalExperience*. Tuto úpravu je možné udělat zejména protože většina respondentů je z Německa, jak bylo zmíněno výše.



Obrázek 7: Grafy chybějících hodnot u jednotlivých atributů. U některých atributů chybí hodnoty ve velkém množství (KurzarbeitHours), jinde naopak téměř žádné (City).



Obrázek 8: Korelační analýza nad atributy neočištěné sady.

U chybějících hodnot atributu *EmploymentStatus* jsme předpokládali, že daná osoba je zaměstnancem na plný úvazek, tedy jsme za chybějící hodnoty dosadili hodnotu *Full-time employee*, neboť se jedná o nejčastější hodnotu.

Protože používané technologie zřejmě ovlivňují platy významně, rozhodli jsme se odstranit případy s chybějícími technologiemi. Při jakémkoliv doplnění by nám v daných případech chyběl důležitý údaj, který bude pravděpodobně značně ovlivňovat vlastní plat osoby, tedy nemá smysl tuto hodnotu jakkoliv approximovat či doplňovat.

V takto očištěné sadě nám zůstalo již jen minimum stále chybějících hodnot. Bez újmy na obecnosti můžeme tedy tyto vzorky odstranit, neboť jejich ztrátou nepřijdem o téměř žádnou informaci.

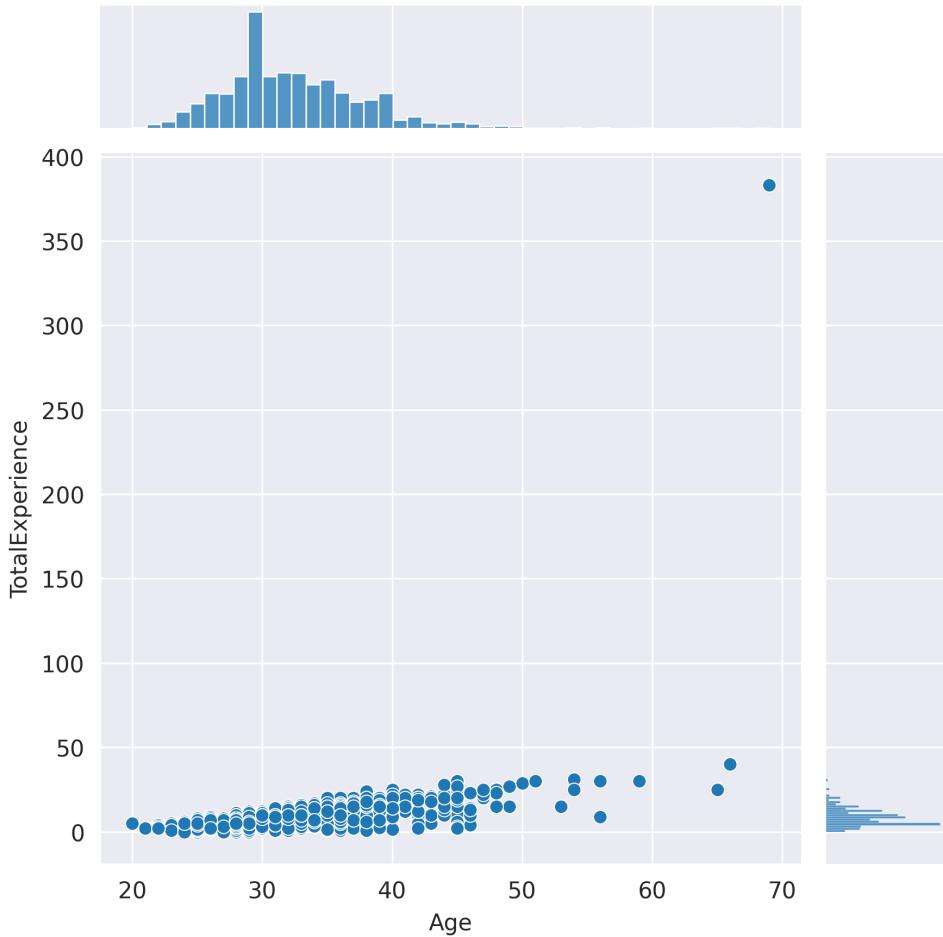
Odstranění odlehčích hodnot

Odlehle hodnoty, které nebyly vyřešeny v předchozích krocích, jsme odhalili u atributů *YearlySalary*, *PreviousYearlySalary* a *TotalExperience*.

Pro atributy platů jsme nejdříve rozhodli, že platy nad 1 000 000 € jsou buď nemožné nebo tak silně odlehle, že pro naši úlohu nemají význam, tedy jsme případy s ročními platy vyššími než 1 000 000 € odřízli.

Následně jsme pro všechny atributy s odlehlymi hodnotami (oba atributy s platy i atribut celkové zkušenosti) ořezali metodou tří sigma za předpokladu normálního rozdělení hodnot atributů. Díky tomu byl například z datové sady odstraněn objekt s platem 800 000 €, což je nejspíše plat možný, ovšem značně odlehly².

²Bez předchozího ořezání pomocí konstanty by tento plat nebyl vyhodnocen jako odlehly na základě pravidla tří sigma, neboť datová sada obsahovala vskutku extrémní hodnotu, která vychylila střední hodnotu i odchylku.



Obrázek 9: Korelační analýza nad atributy *Age* a *TotalExperience* vykazujícími silnou korelací.

3.2 Úprava datové sady – redukce atributů

Za významné jsme rozhodli následující atributy, na které jsme se nadále v datové sadě omezili: (1) *Age*, (2) *Gender*, (3) *City*, (4) *Position*, (5) *TotalExperience*, (6) *GermanyExperience*, (7) *Seniority*, (8) *MainTechnology*, (9) *YearlySalary*, (10) *PreviousYearlySalary*, (11) *EmploymentStatus*, (12) *CompanySize* a (13) *CompanyType*.

Tyto atributy se nám jeví jako nejvíce ovlivňující platy jednotlivých osob a pro dolovací úlohu mohou tedy mít největší význam. Ostatní atributy buď dle předchozí analýzy nebo našeho odhadu nepředstavují zajímavá data, která by mohla mít vliv na plat osob.

Některé ze zahozených atributů by se sice mohly hodit do datové sady pro dolovací úlohu, ale jejich hodnoty jsou natolik různorodé, že je není možné smysluplně vyhodnotit danou dolovací úlohou a nepřináší tedy v závěru žádný užitek pro celkovou datovou sadu. Jedná se například o atribut *SecondaryTechnology*. Zde lidé vypisovali v prostém textu množství různých technologií, které považovali za sekundární ke své hlavní náplni práce, ale každý respondent uváděl technologie z trochu jiného pohledu, například technologie z různých úrovní tak, že jakákoli analýza nad těmi dala z pohledu dolování byla v podstatě minimální nebo nemožná. V případě snahy o použití tohoto atributu by nejspíše bylo nutné použít některou z pokročilých metod pro přípravu textových dat.

3.3 Transformace dat

Nad daty byly provedeny následující transformace připravující datovou sadu pro dolovací úlohy, jednou pro dolovací úlohu vyžadující kategorické hodnoty, podruhé numerické hodnoty.

Diskretizace numerických atributů

Diskretizaci numerických atributů jsme provedli metodou plnění.

Hodnoty následujících atributů jsme převedli do následujícího počtu košů o stejné šířce:

- *Age*: 10 košů,
- *TotalExperience*: 10 košů,
- *GermanyExperience*: 10 košů,
- *YearlySalary*: 50 košů a
- *PreviousYearlySalary*: 50 košů.

Transformace kategorických atributů a normalizace

Pro transformaci bylo třeba převést následující kategorické atributy:

- *CompanySize*: Jelikož velikost společnosti má jasně dané uspořádání, toto uspořádání zachováme i pro numerická data následovně:
 - kategorické hodnotě *up to 10* přiřadíme numerickou hodnotu 0,
 - kategorické hodnotě *11-50* přiřadíme numerickou hodnotu 1,
 - kategorické hodnotě *51-100* přiřadíme numerickou hodnotu 2,
 - kategorické hodnotě *101-1000* přiřadíme numerickou hodnotu 3,
 - kategorické hodnotě *1001+* přiřadíme numerickou hodnotu 4,
- pro následující atributy bez daného uspořádání použijeme *one hot encoding* přiřazující všem novým hodnotám čísla od 0 do N , kde N je počet unikátních hodnot daného atributu: (1) *Seniority*, (2) *EmploymentStatus*, (3) *CompanyType*, (4) *City*, (5) *Position*, (6) *MainTechnology* a (7) *Gender*.

Takto jsme všechny kategorické atributy převedli na numerické a u těch atributů, kde má smysl zachovat uspořádání hodnot (*CompanySize*), jsme upořádání zachovali.

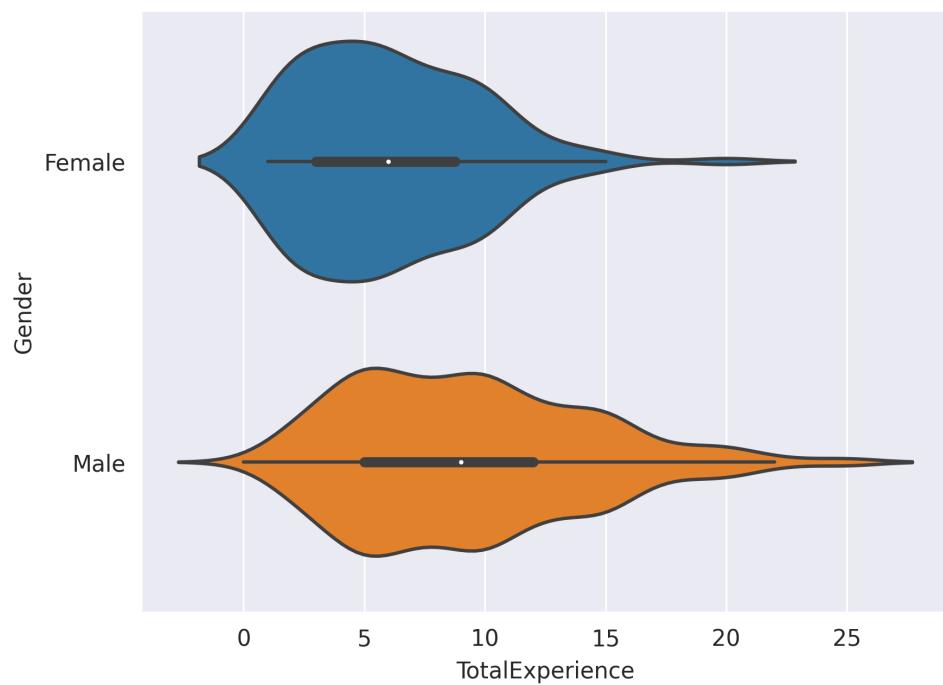
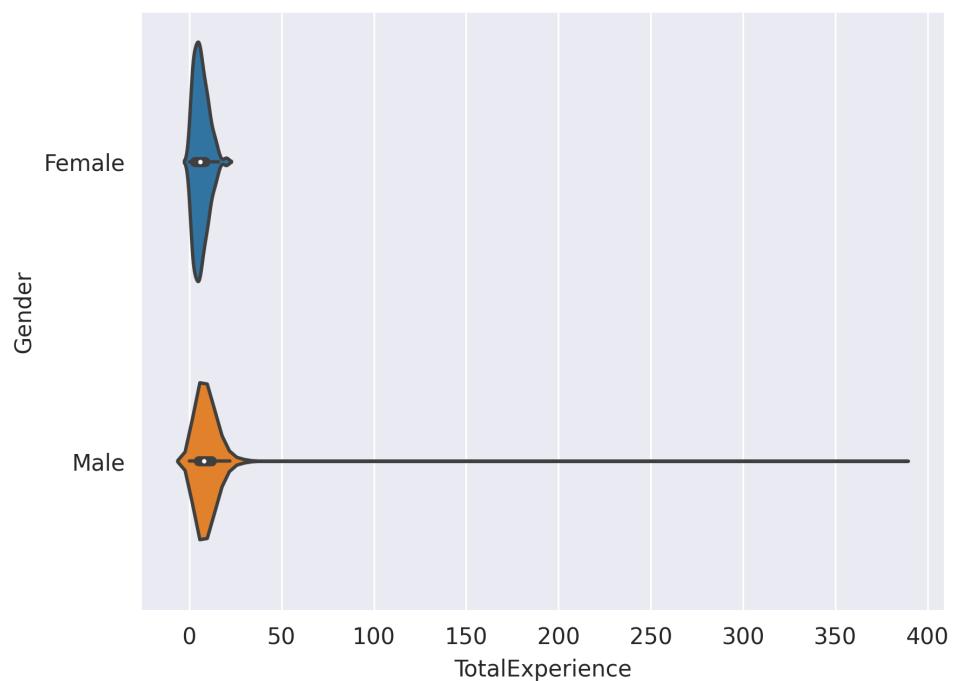
Pro normalizaci jsme si zvolili úlohu ze zadání (*Predikce výše platu na základě ostatních atributů*), ale pouze pro atributy, které mají definované uspořádání hodnot, tedy původní numerické atributy a kategorický atribut *CompanySize* převedený na numerický atribut. Ostatní atributy nemají definované uspořádání, proto z pohledu normalizace vzdálenost dvou hodnot nemá vliv na celkové ohodnocení. Na námi zvolené atributy jsme použili *z-score* normalizaci.

Pokud by se při dolovací úloze ukázalo, že bude užitečné porovnávat námi zvolené atributy i s ostatními atributy, které nebyly normalizovány, můžeme tyto zbývající atributy převést pomocí *z-score* normalizace taktéž, ale musíme mít na paměti, že tyto atributy nemají definované uspořádání, a tedy dolovací úlohy předpokládající rozdílnou váhu různě vzdálených hodnot jednoho atributu budou s takto upravenými atributy pracovat jako s uspořádanými

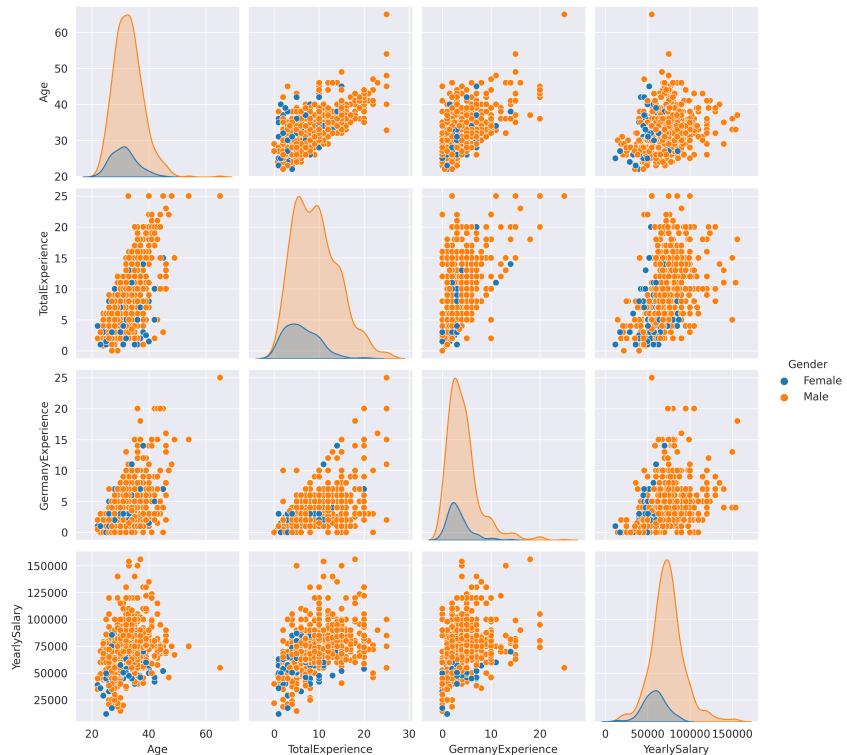
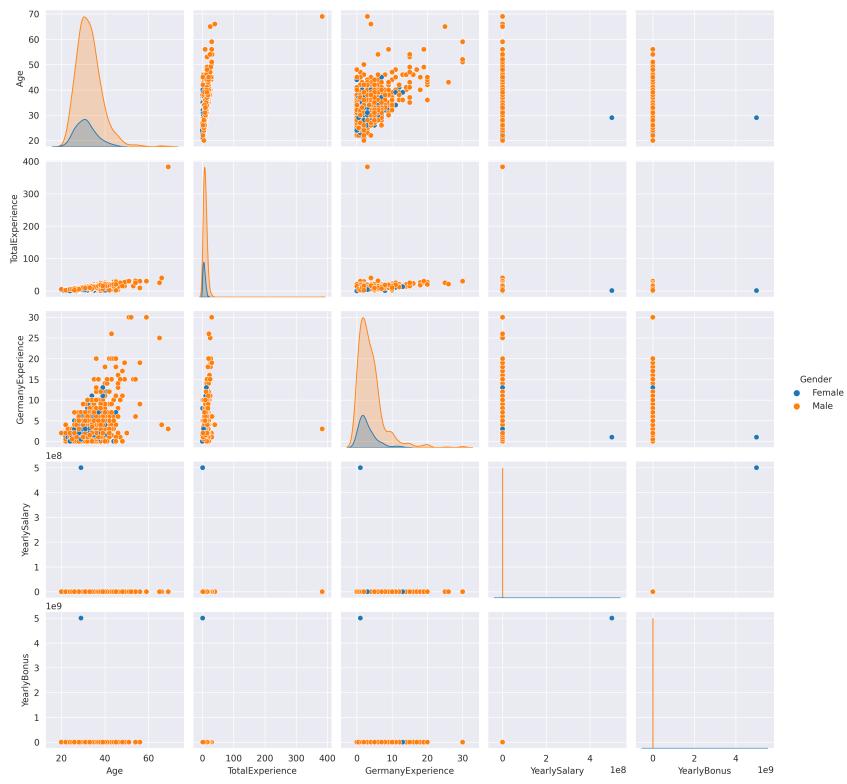
hodnotami. Bylo by třeba experimentálně vyhodnotit, který z přístupů je lepší pro konkrétní zadání úlohy a vybranou dolovací metodu.

4 Zobrazení výsledné očištěné sady

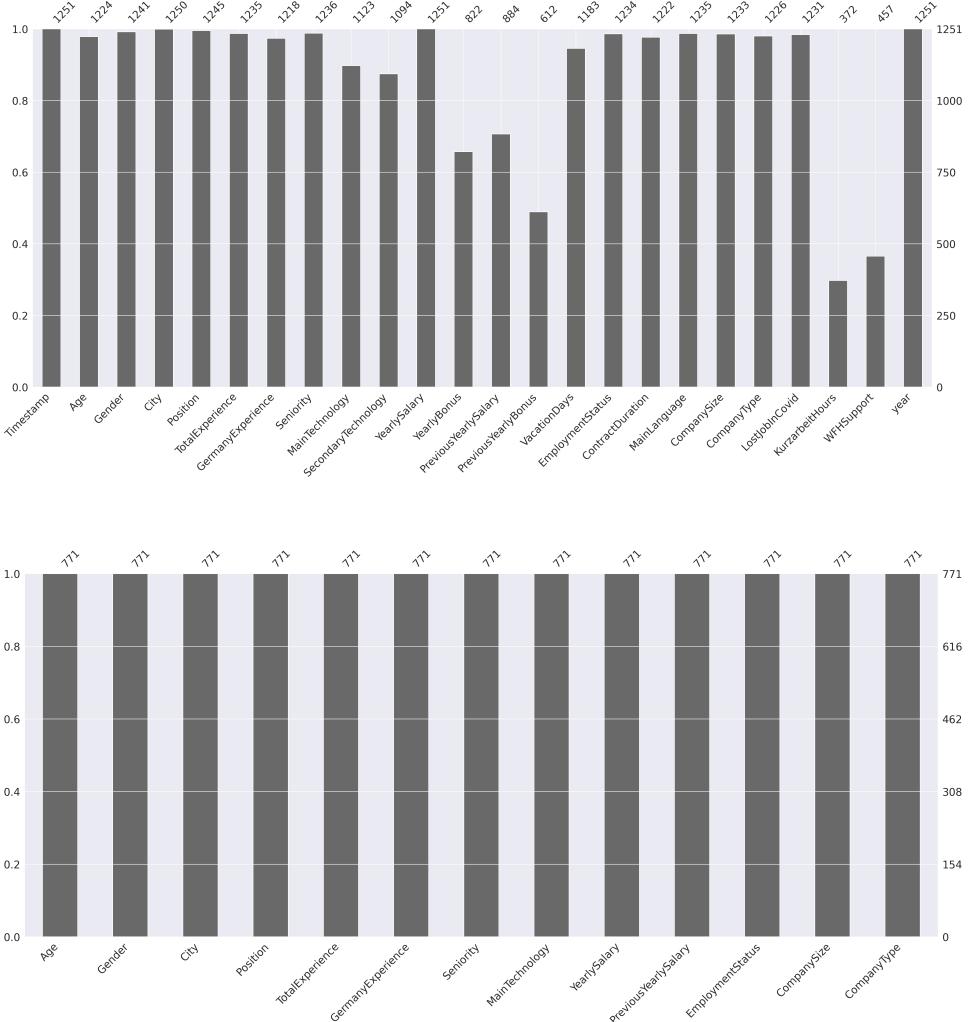
Vzhledem k nízké kvalitě datové sady, a tedy i výstupních grafů z explorativní analýzy, byla analýza znova spuštěna po očištění dat popsaném v sekci 3.1 pro získání lepší představy o reálném rozložení hodnot.



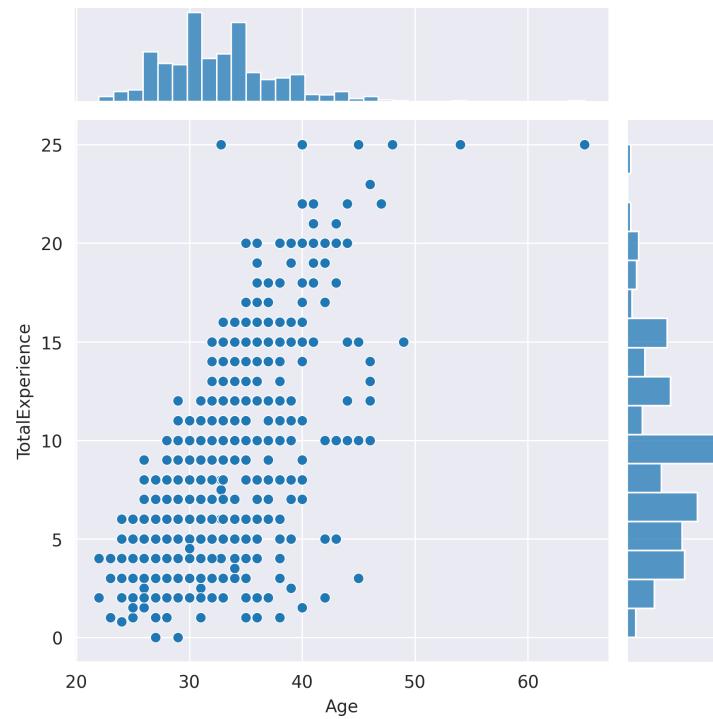
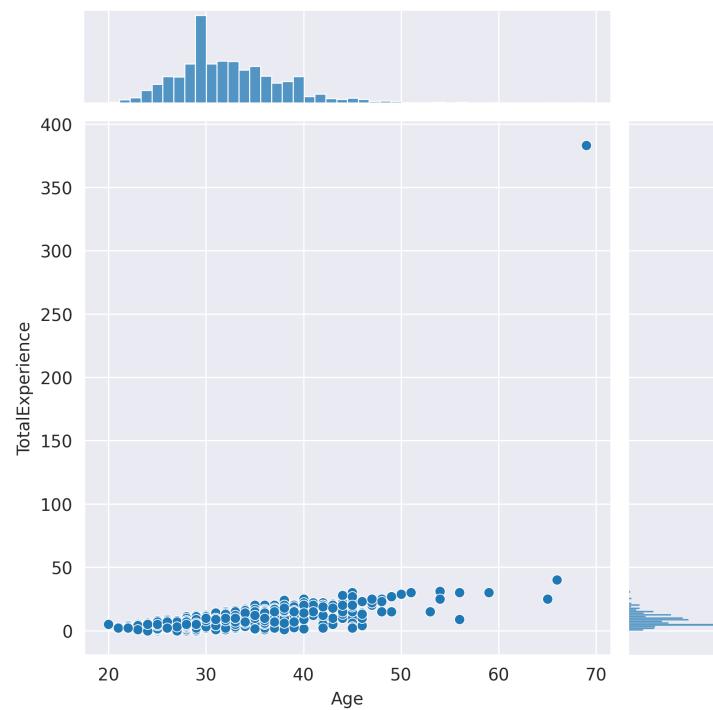
Obrázek 10: Porovnání houslových grafů původní a očištěné sady. Můžeme si povšimnout, že bez odlehlých hodnot se nám x-ová osa zmenšila, tedy nyní můžeme nad grafem vyvozovat závěry.



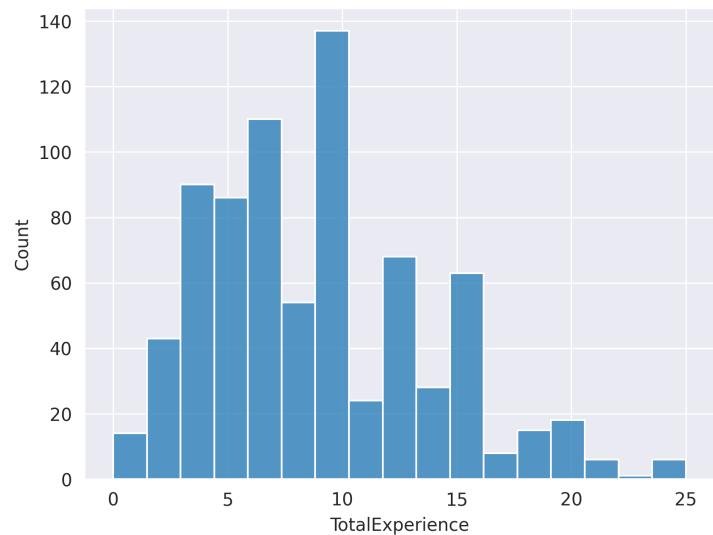
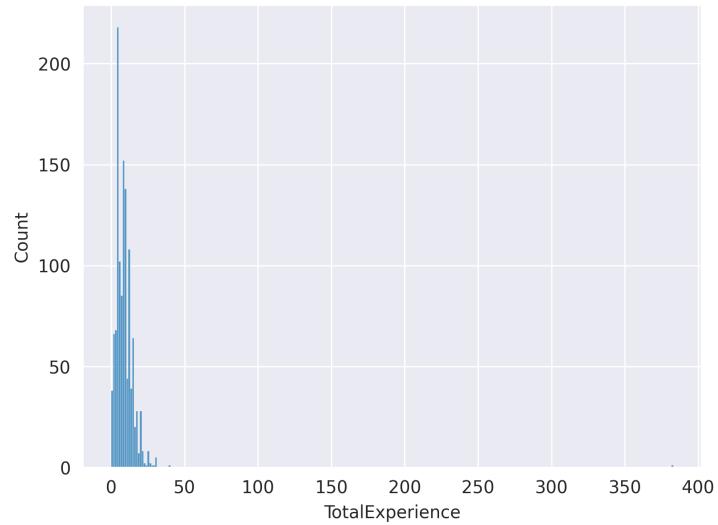
Obrázek 11: Kvůli zašumění dat bylo v podstatě nemožné vyvozovat závěry o korelacích mezi atributy. Díky vyčištění datové sady jsme nyní schopni určit, u kterých atributů pozorujeme silnou a naopak slabou závislost.



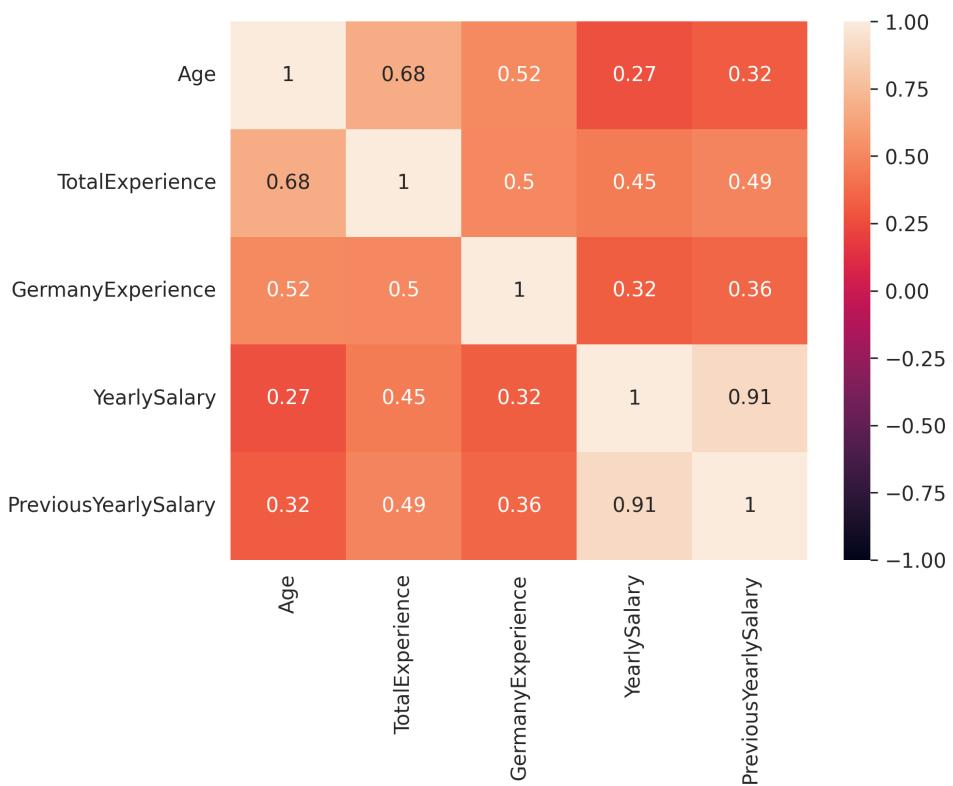
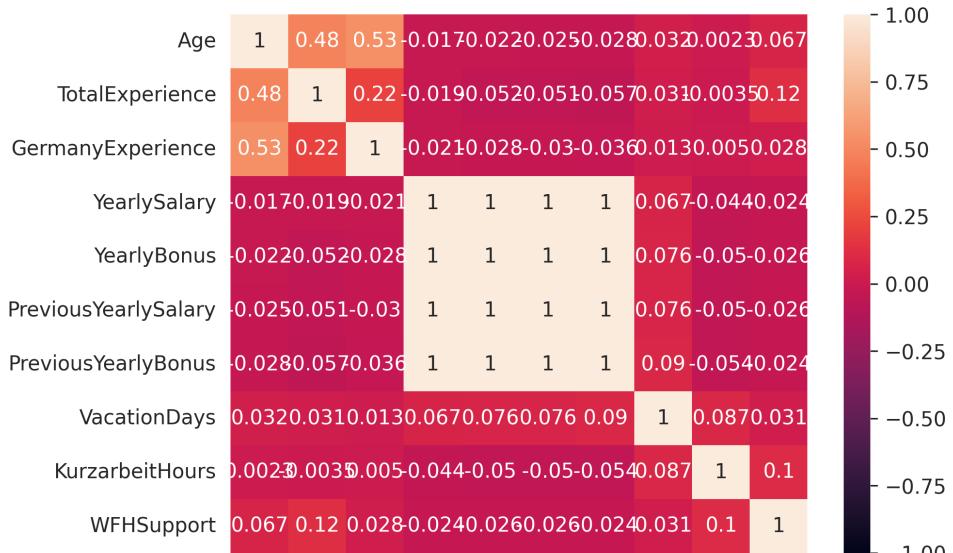
Obrázek 12: Porovnání grafů chybějících hodnot před očištěním datové sady a před vyřešením chybějících hodnot a po jejím očištění. Můžeme pozorovat, že nyní v datové sadě již není jediná chybějící hodnota a při celkovém počtu vzorků 771 (tj. přibližně dvě třetiny oproti vstupní datové sadě). Také byly odstraněny některé atributy, u kterých chyběla většina hodnot nebo které byly pro naši úlohu nezajímavé.



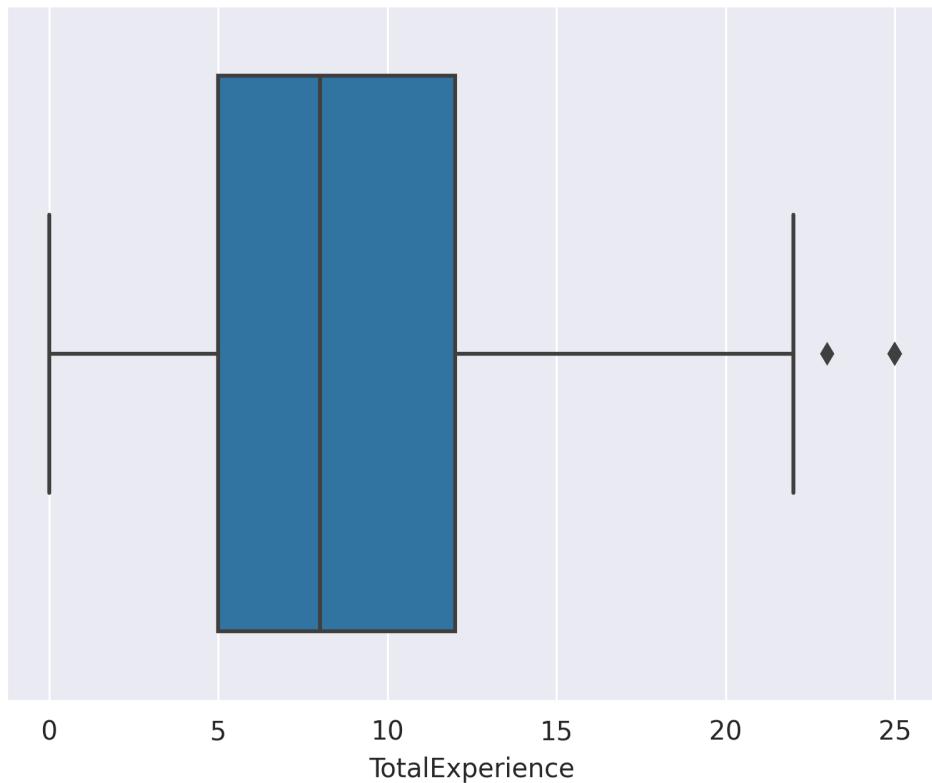
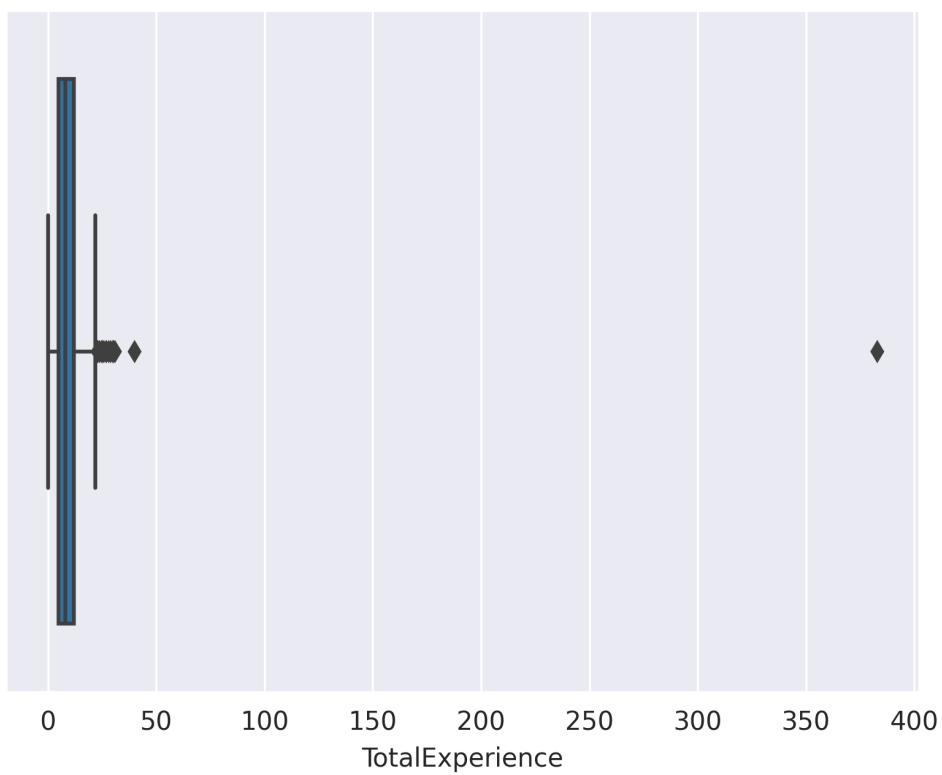
Obrázek 13: Porovnání grafů závislosti věku na celkové zkušenosti osob před očištěním datové sady a po něm. Podle očekávaní můžeme po očištění datové sady vidět, že věk koreluje s celkovou zkušeností, aniž by nás málky odlehle hodnoty.



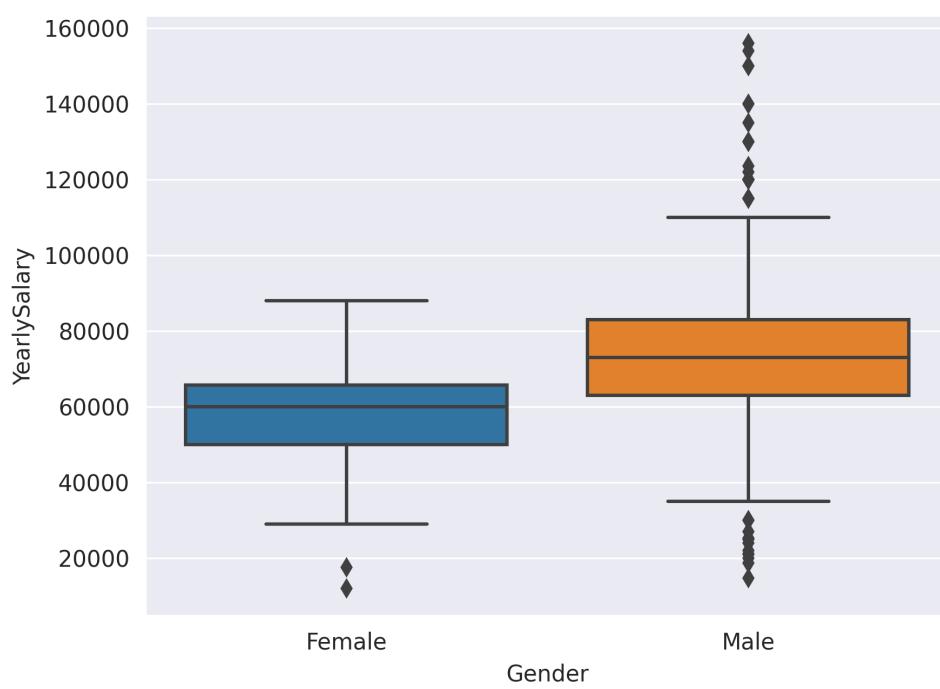
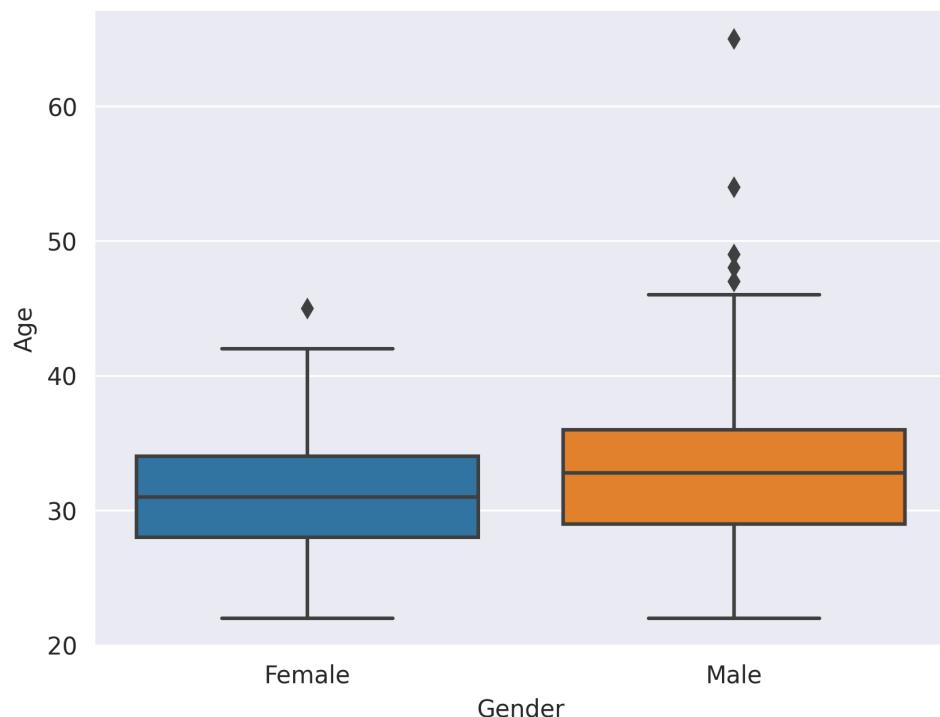
Obrázek 14: Porovnání histogramů celkové zkušenosti osob před očištěním datové sady a po něm. Jak můžeme vidět, po očištění datové sady již v histogramu můžeme pozorovat normální rozložení, kterého bychom si před očištěním pravděpodobně nevšimli.



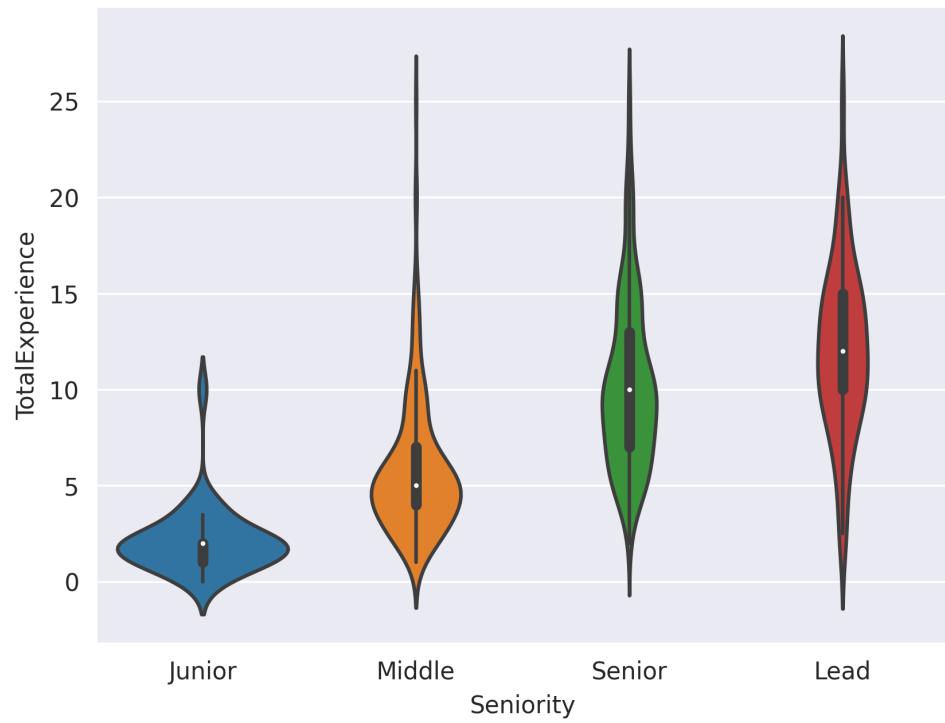
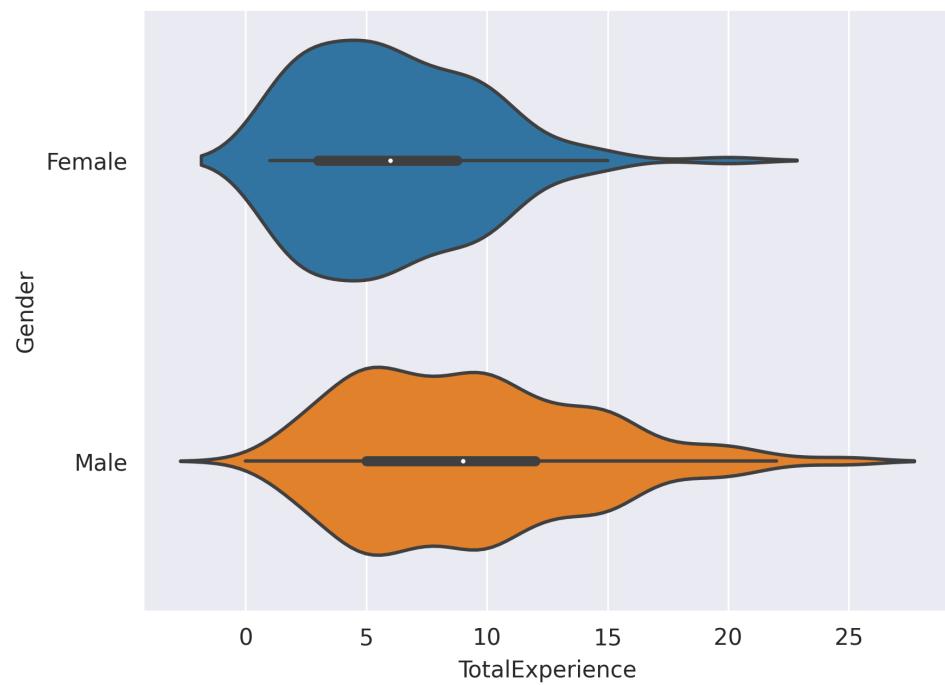
Obrázek 15: Porovnání matic korelace jednotlivých atributů před očištěním datové sady a po něm. Informace zjištěné z očištěné sady nám dávají mnohem lepší představu o reálné situaci.



Obrázek 16: Porovnání grafů celkové zkušenosti osob před očištěním datové sady a po něm. Krabicové grafy jsou před očištěním v podstatě nečitelné. Nyní se již z grafu můžeme dozvědět, co potřebujeme.



Obrázek 17: Krabicové grafy nad očištěnými daty zobrazující věk a roční mzdu na základě pohlaví.



Obrázek 18: Houslové grafy nad očištěnými daty zobrazující množství zkušeností na základě seniority a pohlaví.

Přílohy

Původní název atributu	Výskytů	Nový název
Timestamp	1 253	Timestamp
Age	1 226	Age
Gender	1 243	Gender
City	1 253	City
Position	1 247	Position
Total years of experience	1 237	TotalExperience
Years of experience in Germany	1 221	GermanyExperience
Seniority level	1 241	Seniority
Your main technology / programming language	1 126	MainTechnology
Other technologies/programming languages you use often	1 096	SecondaryTechnology
Yearly brutto salary (without bonus and stocks) in EUR	1 253	YearlySalary
Yearly bonus + stocks in EUR	829	YearlyBonus
Annual brutto salary (without bonus and stocks) one year ago. Only answer if staying in the same country	885	PreviousYearlySalary
Annual bonus+stocks one year ago. Only answer if staying in same country	614	PreviousYearlyBonus
Number of vacation days	1 185	VacationDays
Employment status	1 236	EmploymentStatus
Contract duration	1 224	ContractDuration
Main language at work	1 237	MainLanguage
Company size	1 235	CompanySize
Company type	1 228	CompanyType
Have you lost your job due to the coronavirus outbreak?	1 233	LostJobInCovid
Have you been forced to have a shorter working week (Kurzarbeit)? If yes, how many hours per week	373	KurzarbeitHours
Have you received additional monetary support from your employer due to Work From Home? If yes, how much in 2020 in EUR	462	WFHSupport

Tabulka 2: Atributy v datové sadě. Sloupec Výskytů označuje počet záznamů, ve kterých byl atribut neprázdný. Podbarvené řádky značí numerické atributy, ostatní jsou kategorické.