ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

MASARYKŮV ÚSTAV VYŠŠÍCH STUDIÍ

BAKALÁŘSKÁ PRÁCE

Zpracování dat a nástroje Business Intelligence v týmu datové kvality

Data Processing and Business Intelligence Tools in the Data Quality Team

2024

Kilián Malý

Studijní program: Ekonomika a management

Vedoucí práce: doc. Ing. Tomáš Kubálek, CSc.

MALÝ, KILIÁN. *Zpracování dat a nástroje Business Intelligence v týmu datové kvality*. Praha: ČVUT 2024. Bakalářská práce. České vysoké učení technické v Praze, Masarykův ústav vyšších studií.

MASARYKŮV ÚSTAV
VYŠŠÍCH STUDIÍ
ČVUT V PRAZE

Prohlášení

Prohlašuji, že jsem svou bakalářskou práci vypracoval samostatně. Dále prohlašuji, že jsem všechny použité zdroje správně a úplně citoval a uvádím je v přiloženém seznamu použité literatury.

Nemám závažný důvod proti zpřístupňování této závěrečné práce v souladu se zákonem č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon) v platném znění.

V Praze dne: 25. 04. 2024

Poděkování

Rád bych touto cestou poděkoval doc. Ing. Tomáši Kubálkovi, CSc. za trpělivost, připomínky a pomoc, které mi pomohly k sepsání práce. Velký dík také patří mé rodině.

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

I. OSOBNÍ A STUDIJNÍ ÚDAJE

Příjmení: **Malý** Jméno: **Kilián** Osobní číslo: **509486**

Fakulta/ústav: **Masarykův ústav vyšších studií**

Zadávající katedra/ústav: **Institut manažerských studií**

Studijní program: **Ekonomika a management**

II. ÚDAJE K BAKALÁŘSKÉ PRÁCI

Název bakalářské práce:

Zpracování dat a nástroje Business Intelligence v týmu datové kvality

Název bakalářské práce anglicky:

Data Processing and Business Intelligence Tools in the Data Quality Team

Pokyny pro vypracování:

Práce se zaměřuje na řešení zpracování dat a využití metod business intelligence v týmu datové kvality v jedné z největších bank v České republice. Práce je rozdělena na teoretickou a praktickou část. Cílem teoretické části je popis bankovního sektoru, využití BI nástrojů, způsob ukládání dat a důležitost datové kvality. V praktické části se zodpovídají následující hypotézy: 1. Jaký vliv má tým datové kvality na chod společnosti? Jak by se absence týmu podepsala na chodu firmy?; 2. Jaké jsou klady a zápory významných BI nástrojů (Power BI a Tableau)?; 3. Na jakých pozicích má smysl vzdělávat zaměstnance v oblasti datové analýzy? A do jaké míry? Metody: sběr dat, transformace dat, analýza dat, vizualizace dat pomocí BI nástrojů.

Seznam doporučené literatury:

OLSON, Jack E. Data Quality: The Accuracy Dimension. Morgan Kaufmann, 2003. ISBN 978-1558608917.
KIMBALL, Ralph a Joe CASERTA. The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data. Paperback Press, 2004. ISBN 978-0764567575.
Lacko Ľuboslav, Business Intelligence v SQL Serveru 2008, Computer press, 2009.
JIRÁNEK, Karel. Datová kvalita, integrita a konsolidace dat v BI společnosti PREdistribuce, a.s.. Nárožní 2600/9a, 158 00 Praha 5, 2011. Diplomová práce. Vysoká škola ekonomie a managementu.
Data Quality vs Data Governance: Learn the Differences & Relationships!. Atlan [online]. 2023 [cit. 2023-10-30]. Dostupné z: https://atlan.com/data-quality-vs-data-governance. STEDMAN, Craig. Data quality. TechTarget [online]. 2022 [cit. 2023-10-30]. Dostupné z: https://www.techtarget.com/searchdatamanagement/definition/data-quality.

Jméno a pracoviště vedoucí(ho) bakalářské práce:

doc. Ing. Tomáš Kubálek, CSc. Masarykův ústav vyšších studií ČVUT v Praze

Jméno a pracoviště druhé(ho) vedoucí(ho) nebo konzultanta(ky) bakalářské práce:

Datum zadání bakalářské práce: **08.12.2023** Termín odevzdání bakalářské práce: **25.04.2024**

Platnost zadání bakalářské práce: _____

doc. Ing. Tomáš Kubálek, CSc. podpis vedoucí(ho) práce	Ing. Dagmar Skokanová, Ph.D. podpis vedoucí(ho) ústavu/katedry	prof. PhDr. Vladimíra Dvořáková, CSc. podpis děkana(ky)

III. PŘEVZETÍ ZADÁNÍ

Student bere na vědomí, že je povinen vypracovat bakalářskou práci samostatně, bez cizí pomoci, s výjimkou poskytnutých konzultací. Seznam použité literatury, jiných pramenů a jmen konzultantů je třeba uvést v bakalářské práci.

. _____ Datum převzetí zadání

_____ Podpis studenta

Abstrakt

Práce se zaměřuje na řešení zpracování dat a využití metod Business Intelligence v týmu Datové kvality v jedné z největších bank v České republice. Práce je rozdělena na teoretickou a praktickou část.

Cílem teoretické části je popis bankovního sektoru, využití BI nástrojů, způsob ukládání dat a důležitost datové kvality společně s regulatorním reportingem.

V praktické části se zodpovídají následující hypotézy: 1. Jaký vliv má tým datové kvality na chod společnosti? Jak by se absence týmu podepsala na chodu firmy?; 2. Jaké jsou klady a zápory významných BI nástrojů (Power BI a Tableau)?; 3. Na jakých pozicích má smysl vzdělávat zaměstnance v oblasti datové analýzy? A do jaké míry?

Metody: sběr dat, transformace dat, analýza dat, vizualizace dat pomocí BI nástrojů

Klíčová slova

Datová kvalita, regulatorní reporting, Business Intelligence, datová analýza, bankovní systém

Abstract

The thesis focuses on data processing solutions and the use of business intelligence methods in the Data Quality team in one of the largest banks in the Czech Republic. The work is divided into theoretical and practical parts.

The aim of the theoretical part is to describe the banking sector, the use of BI tools, the method of data storage and the importance of data quality.

In the practical part the following hypotheses are answered: 1. What impact does the data quality team have on the company's operations? How would the absence of the team affect the company?; 2. What are the pros and cons of major BI tools (Power BI and Tableau)?; 3. In what roles does it make sense to train employees in data analytics? And to what extent?

Methods: data collection, data transformation, data analysis, data visualization

Keywords

Data Quality, Regulatory Reporting, Business Intelligence, Data Analysis, Bank System

Obsah

Úvod

Neúprosné tempo vývoje technologií vyvolává reakci v bankovním sektoru, který se pomalu začíná transformovat z tradičního bankovnictví do světa, kde jedním z hlavních faktorů pro rozhodování hrají data, a s nimi spojená datová kvalita. Datům v bankovním sektoru, regulatornímu reportingu a nástrojům BI z pohledu datové kvality budou věnována teoretická část. Z důvodu anonymizace práce pokrývá teoretická část bankovnictví a jeho fungování v rámci celé České republiky, včetně regulatorního reportingu vůči ČNB, kde pro všechny banky platí stejná pravidla.

Praktická část je strukturovaná do tří části. V těchto částech práce také odkazuji na společnost X, která symbolizuje banku, v níž je práce psána.

Hlavní složku praktické části tvoří porovnání dvou vlajkových lodí na trhu vizualizačních BI nástrojů – Power BI a Tableau. Cílem porovnání jsou netechnické i technické rozdíly, jako připojení k datovým zdrojům a vytváření vztahů mezi jednotlivými tabulkami pomocí relací, joinů a union operací. Následně ukázky jednotlivých vizuálů v obou nástrojích, nebo možnosti vytváření nových sloupců v tabulkách. Datový zdroj pro vizualizace tvoří transformovaná a upravená datová sada.

Druhá část praktické části se zaměřuje na procesy měření kvality dat ve společnosti X. Jedním z procesů je samotné vytváření nástroje pro měření datové kvality a tato část práce také bude věnována tomuto tématu. Následuje diskuze o dopadech nízké datové kvality na fungování společnosti.

V třetí a poslední části se zaměřím na identifikaci pozic, kde by bylo vhodné zaměstnance vzdělávat v oblasti datové analýzy. Nejdříve analyzuji aktuální situaci pracovních nabídek v rámci společnosti X s podporou grafického zobrazení v Microsoft Excel, a poté se zkusím zodpovědět, v jaké části banky – Back-Office, Middle-Office, Front-Office by měl být kladen vyšší důraz na rozvoj dovedností v této oblasti.

TEORETICKÁ ČÁST

1 Data a bankovní sektor

1.1 Aktuální situace na bankovním trhu v ČR

Bankovní sektor v České republice, stejně jako v ostatních koutech světa, představuje naprosto klíčový prvek finančního systému země. Na území ČR funguje od ledna 1990 dvoustupňový bankovní systém, který zahrnuje jak centrální banku, tak komerční banky. Funkci centrální banky plní Česká národní banka, která vykonává měnovou politiku a dohled nad bankovním trhem. Na druhé úrovni komerční banky slouží jako primární zprostředkovatelé finančních transakcí mezi jednotlivci, podniky a dalšími institucemi. Komerční banky disponují širokou škálou produktů a služeb. Lze říci, že český bankovní sektor patří k těm nejstabilnějším v rámci Evropské unie, což také potvrzují zátěžové testy ČNB a prostředí na českém trhu je velmi konkurenční, což je dáno také různorodostí obchodních modelů.[1]

Ke konci roku 2020 bylo v České republice registrováno 49 subjektů s bankovní licencí, což odpovídá stejnému počtu, který byl i v roce 2018. Z celkového počtu jich je 37 pod kontrolou zahraničních vlastníků, přičemž 12 z nich je plnoprávnými bankami a zbylých 25 jsou pobočky zahraničních bank. Domácí vlastníci ovládají 12 bank, dvě z nich jsou se státní účastí.

Struktura bankovního sektoru:

- 4 velké banky
- 5 středně velkých bank
- 10 malých bank
- 25 poboček zahraničních bank
- 5 stavebních spořitelen[2]

1.1.1 Česká národní banka

Od 1. dubna 2006 vykonává dohled nad bankovním trhem v České republice Česká národní banka. ČNB sleduje dodržování bankovních předpisů, stanovuje požadavky na kapitálovou přiměřenost bank, provádí inspekce a poskytuje regulatorní rady a směrnice pro bankovní sektor v souladu s českými a mezinárodními zákony, avšak bankovní dohled nemá právo zasahovat do obchodního řízení bank, které je výlučně v kompetenci managementu. Cílem monitoringu je kontrolovat, zda banky jednají dle předpisů a vhodnými nápravami případně usměrňovat činnosti bank takovým způsobem, aby nedocházelo k aktivitám, které by poškodily zájmy klientů a především bankovní systém jako celek.[3; 4]

1.2 Data, informace a znalosti

Data sama o sobě jsou pouhým souborem faktů, čísel nebo symbolů, které lze zaznamenat a uložit. Představují surové informace, které samy o sobě nemají žádný význam. Zpracováním, interpretací a organizací dat se převedou data do smysluplné podoby, tím vzniká informace, která již smysl a hodnotu obsahuje. Znalosti vznikají tehdy, když dojde k hlubšímu pochopení informací, ty lze následně využít k rozhodování, vyvozování závěrů nebo řešení problémů. Jakmile se informace začnou prakticky využívat, jedná se o znalosti.[5]

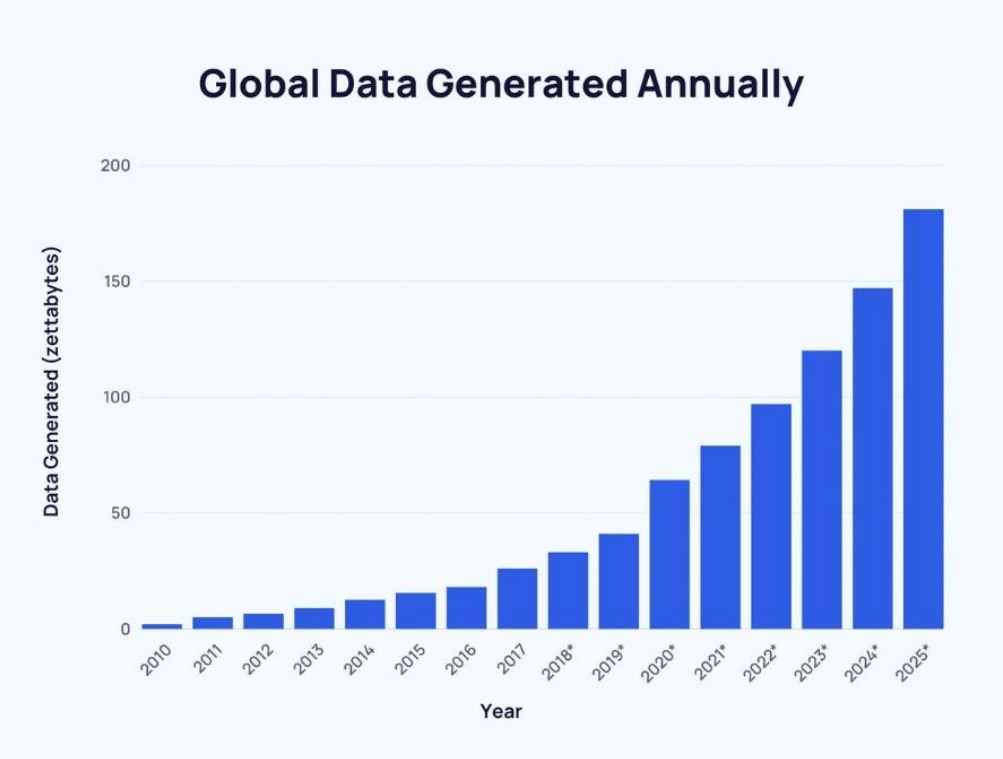
Propojení mezi daty, informacemi a znalostmi spočívá v tom, že data jsou základem pro informace, informace pak slouží jako základ pro znalosti, a znalosti nakonec vedou k akcím, inovacím. Tento proces transformace dat (od dat surových až po znalosti) je naprosto esenciální pro efektivní využití informačních zdrojů.

Ne nadarmo se říká, že v dnešní době digitalizace jsou data nové zlato. V posledních letech nezměrně roste důležitost a důraz na data. Též roste poptávka po pracovních pozicích, která se pojí

s prací s daty. Například ve formě datových analytiků, inženýrů nebo třeba datových vědců. Tento trend v posledních dekádách potvrzuje i bankovní sektor, kde nyní již data hrají neodmyslitelnou roli.

1.2.1 Množství dat

Odhaduje se, že každým dnem se ve světě vytvoří zhruba 2,5 triliónu bajtů dat (veškera digitální aktivita jako informace ze sociálních médií, e-mailů, internetových vyhledávačů, transakcí, sensorů z internetu atd...). Objem dat ve světě se nyní již neměří v gigabytech a terrabytech, ale v zettabytech, přičemž 1 zettabyt se rovná miliardě terrabytů. Na následující vizualizaci je vidět evidentní trend množství vygenerovaných dat ve světě v zettabytech.



OBR. 1 - GLOBAL DATA GENERATED ANNUALLY (ZDROJ: HTTPS://WWW.STATISTA.COM/STATISTICS/871513/WORLDWIDE-DATA-CREATED/)

Množství dat v období 2021 až 2025 bylo predikováno na základě růstu z předchozích let, nejedná se tedy o konkrétní hodnoty. Od roku 2010 do roku do roku 2016, tedy za období sedmi let bylo vytvořeno zhruba 69 zettabytů dat, což je pouze o 5 zettabytů více, než bylo vytvořeno za samotný rok 2020, kde se hodnota ustálila na 64 zettabytech. V posledních deseti letech největší meziroční růst v objemu dat nastal v roce 2020. Hlavním důvodem je celosvětová pandemie COVID-19. Během pandemie více lidí pracovalo vzdáleně a aktivněji nakupovali a používali domácí zařízení, která generují velká množství dat.[6]

Zajímavá statistika se opět váže k roku 2020, kdy bylo zjištěno, že pouze malá množství vytvořených dat jsou ukládána a zachovávána pro budoucí použití. Konkrétně se za tento rok jednalo o pouhá 2 %. Jak je již zmíněno výše, v roce 2020 bylo vytvořeno 64 zettabytů dat, celková kapacita uložišť dat byla však pouhých 6,7 zettabytů s tím, že celá tato kapacita nebyla určena jen k ukládání nově vytvořených dat, čímž se dostáváme k tomu, že generovat data stíháme exponenciálním tempem, ale zdaleka takovým tempem nerostou kapacity a množství datových uložišť, která by data uchovávala k budoucímu použití.[6]

Již v minulém desetiletí začaly banky přesouvat svá data do cloudových uložišť, především kvůli škálovatelnosti, snížení nákladů a možností pro inovace, avšak ještě pár let před tím bylo naprosto nemyslitelné, že by banky uchovávaly data mimo své vlastní servery. Velmi dlouhou dobu banky využívaly jen vlastní kapacity, ale s obrovským nárůstem dat se zdál pronájem externího datacentra jako vhodný nápad a ke všemu je pronájem externího datacentra rychlejší než budování vlastního uložiště dat, avšak stále si banky nechávají strategické části dat „u sebe“. V dnešní době tak podobnou formu outsourcingu využívá většina, ne-li všechny banky na našem území.[7]

Největším otazníkem nad cloudovými uložišti je samozřejmě riziko. Nespoléháte se jen sami na sebe, ale na externího dodavatele. Je tedy důležité, aby banky prováděly důkladné hodnocení rizik s cloudovými uložišti spojené, implementovaly kontrolní mechanismy a postupy pro správu a ochranu dat. Využívání externích datových uložišť samozřejmě podléhá přísné regulaci ze strany České národní banky.

1.2.2 Data v bankovnictví

To, jakým tempem roste množství dat ve světě nám již jasné je, ale jak tomu je v rámci bankovního sektoru České republiky? V této části se vraťme zpátky k tématu této práce a ukážeme si, jak je situace a jak se vyvíjí v rámci bankovnictví v ČR. A jaká data banka vůbec sbírají? Většina lidí si představí, že banka sbírá osobní údaje o klientech, jako jména, telefonní čísla, e-mailové adresy nebo místa bydliště. To je samozřejmě pravda, ale nejedná se jen o osobní údaje a rozhodně ne v tak malém měřítku.

Časy, kdy klienti banky stály fronty před pobočkami pro vložení peněz z výplatní pásky, jsou už dávno pryč. K osobnímu kontaktu mezi klientem a bankéřem nedochází zdaleka tak často a nahrazují jej o mobilní telefony, počítače či tablety, jejichž prostřednictvím mohou zákazníci sledovat stav účtu, posílat peníze, platit účty a není jediný důvod, aby klient vykročil ze dveří svého domu. Zde se však dostáváme k otázce „Jak klientům doporučovat bankovní produkty, když je osobně neznáme?“. Zde svou roli opět hrají data. Produkty jsou již v dnešní době nabízeny na základě sesbíraných dat o klientovi, jejich analýze a následně lze lépe doporučit a zacílit nové produkty.

Mezi kategorie dat, u kterých banky data sbírají, patří:

- osobní data,
- finanční údaje,
- transakční údaje,
- údaje o rizicích,
- záznamy o komunikaci,
- ostatní data (biometrické údaje…).[8]

Osobní data, neboli osobní údaje jsou takové údaje, které se týkají určité fyzické osoby a lze ji pomocí konkrétních údajů dohledat. V bankovních databázích se tedy dají dohledat jména, příjmení, rodná čísla, data narození, adresy, podpisy dle podpisového vzoru, telefonní čísla, e-mailové adresy, vzdělání, profese, informace o zaměstnání a další. Samozřejmě veškeré zpracování a nakládání s osobními údaji klientů je v souladu se zákonem č.101/2000 Sb., o ochraně osobních údajů. Banka tedy může osobní údaje získat pouze se souhlasem klientů.

Finanční údaje poskytují podrobné informace o aktivitách klientů, jako jsou příjmy, úspory, investice, úvěry, hypotéky a další finanční údaje. Tyto data poskytují především informace o chování klientů a toku peněz, což je také jeden z klíčových faktorů při posuzování rizikovosti klienta.

Transakční údaje obsahují informace o každé finanční transakci provedené bankovními klienty, včetně platebních operací, převodu prostředků, výběru z bankomatů a dalších transakčních aktivit.

Údaje o rizicích jsou kombinací několika typů rizik. Jedná se hlavně o úvěrová rizika, tržní rizika, operační rizika a Compliance rizika. Což v konkrétním případě může zahrnovat např. úvěrovou historii klientů, hodnoty zástavních prostředků nebo finanční rizika. Banky v tomto kontextu mohou také shromažďovat informace o tržních podmínkách, cenových indexech, úrokových sazbách či směnných kurzech. Tyto údaje jsou klíčové při ochraně klientů a minimalizaci ztrát společnosti.

Záznamy o komunikaci. Pokud jste v případě telefonátu s bankéřem nebo zaměstnancem v call centru upozorněni, že následující hovor může být monitorován, i tyto typy dat jsou bankovními společnostmi shromažďovány. Nejde však jen o telefonní hovory, ale také e-mailovou komunikaci mezi bankou a klientem, konverzace v online bankovnictví a další.

Ostatní data jsou taková data, které nelze zařadit do kategorií výše. Některé banky mohou sbírat například biometrické údaje jako součást ověřování svých klientů. Může se jednat o otisky prstů, rozpoznání obličeje, hlasové vzorky nebo oční rohovku.

1.2.3 Využití dat

A z jakého důvodu vůbec banky sbírají tolik dat? Poslat e-mail či dovolat se klientovi je jedna věc, na druhé straně jsou ovšem důvody, které odkrývají obrovský potenciál datových věd. Takže zde bych rád popsal pár příkladů, jakým způsobem se sbíraná data mohou využívat.

Segmentace klientů je proces, při kterém se klienti banky rozdělují do určitých skupin na základě všemožných charakteristik. Například na základě demografických informací (věk, pohlaví, příjem), transakční historie, vlastněných produktů nebo vzorců chování. Proces segmentace standardně probíhá za použití analytických nástrojů a algoritmů. Díky segmentaci klientů lze lépe pochopit a porozumět zákazníkům, na základě čehož lze vytvářet cílené marketingové kampaně, přizpůsobovat nabídku produktů a poskytovat personalizované služby.[9]

Prevence podvodů a bezpečnost. Podvody jsou závažným problémem pro všechny banky a existuje jich nezměrné množství. Přes bílé koně, podvodné emaily či SMS, vishing (podvodné telefonáty údajných bankéřů), přes praní špinavých peněz a dalších mnoho druhů. Co se týče praní špinavých peněz, v bankách již existují týmy či oddělení, které se přímo na tuto problematiku soustředí (Anti-Money Lanudering pod zkratkou AML). Ku příkladu Komerční banka v roce 2022 zachytila 606 kybernetických podvodů, což je vůči roku 2021 dvojnásobný nárůst a oproti roku 2018 pětinásobný nárůst. Komerční banka též uvedla, že neroste pouze počet pokusů o podvod, ale také velikost transakcí. Podrobné analýzy dat a vytváření prediktivních modelů i v tomto případě může pomáhat s identifikací podezřelých aktivit a předcházet jim na základě rozpoznání neobvyklých vzorců v transakcích.[9; 10]

Řízení rizik, ať už úvěrových, tržních nebo operačních může být také úkol pro datové vědy. Opět se může jednat o prediktivní modely, které analyzují historická data a na jejichž základě vyhodnocují rizika. V případě úvěrového rizika se jedná o hodnocení pravděpodobnosti, že dlužníci nebudou splácet své úvěry, což by v případě velkého objemu klientů může způsobit vážné dopady na celý bankovní sektor. Tržní riziko zahrnuje potenciální ztráty způsobené kolísáním úrokových sazeb, směnných kurzů či cen aktiv. Operační riziko se zaměřuje na rizika vyplívající z interních procesů, systémů či lidských chyb. Technologický pokrok také umožňuje provádět analýzy spojené s riziky v reálném čase.[9]

Virtuální asistenti a chatboti jsou již součástí většiny bank, zejména call center. Ty často bývají zahlcena opakujícími se dotazy, které by bylo jednoduché vyřešit prostřednictvím samoobsluhy. Čím více času věnují zaměstnanci těchto služeb odpovídáním na tyto požadavky, tím méně se mohou věnovat komplexnějším problémům, které jejich úsilí vyžaduje. V tomto případě na řadu přichází virtuální asistenti a chatboti. Ačkoliv z mého pohledu jsou schopnosti chatbotů a virtuálních asistentů stále velmi omezené a nedokážou nahradit lidskou činnost, konzistentními

kroky kupředu se tyto modely trénují a eliminují asymetrii mezi lidskou a strojovou činností. Samozřejmě tyto modely jsou trénovány na velkých objemech historických dat. Cílem trénování modelu je pokrýt všechny možné scénáře (což je v některých případech prakticky nemožné), aby model vždy věděl, jak si s určitou situací poradit.[9]

2 Regulatorní reporting

Termínem regulatorní reporting se označují procesy, kdy instituce je povinna pravidelně vykazovat informace o svých činnostech regulátorům. Těmi jsou zpravidla úřady, centrální banky či jiné instituce. Cílem regulátorů je, aby trhy, na kterých tyto instituce operují byly transparentní a odpovědné. Lze tedy říci, že regulatorní reporting redukuje mezery v informovanosti institucí a regulátorů. Regulatorní reporting se týká širokého spektra finančních institucí. Příkladem mohou být pojišťovny, banky nebo firmy obchodující s cennými papíry a každý druh společnosti má jiná pravidla na vykazování. V našem případě se však budeme bavit o regulatorním reportingu v bankovnictví.[11]

2.1 Finanční reporting, bankovní regulace a výkazy

V rámci práce je nutno odlišovat regulatorní reporting od bankovních regulací a finančního reportingu. Všechny tři termíny jsou vzájemně propojeny a spadají pod reportingový nebo regulatorní rámec, avšak je nutné odlišovat jeden termín od druhého.

2.1.1 Výkaz v rámci regulatorního reportingu

Pokud se bavíme o výkazu neboli reportu v rámci regulatorního reportingu, rozumí se jím soubor informací nebo formální dokument, který finanční instituce (v našem případě banky) musí pravidelně poskytovat regulátorům. Cílem tohoto typu reportingu je poskytovat regulátorům komplexní pohled na jejich činnost a rizika, na základě čehož regulátorům poskytují zdroje pro analýzy a posuzování systémových rizik. Výkazy jsou formulovány a vykazovány v souladu s regulačními standardy a předpisy platné v dané jurisdikci, také jsou regulátorům zasílány v předem stanovených lhůtách.

2.1.2 Finanční reporting

Existují dva základní rozdíly mezi regulatorním a finančním reportingem. Prvním rozdílem je typ adresáta, pro kterého jsou výkazy určeny. U finančního reportingu bývají hlavními konzumenty reportů účastníci trhu, jejichž kapitál podléhá riziku, tedy akcionáři, věřitelé, investoři. Na druhé straně u reportingu regulatorního jsou výkazy reportovány orgánům bankovního dohledu. Druhým z rozdílů je obsah informací, které jsou v reportech zasílány. V případě finančního reportingu tvoří obsah reportů především informace a data účetního typu, která jsou zpravidla odesílána v ne tak pravidelných intervalech v porovnání s reportingem regulatorním.

2.1.3 Bankovní regulace

Pojem bankovní regulace představuje soubor pravidel, zákonů, směrnic a standardů, které upravují chování bank. Bankovnictví je v současné době také nejvíce regulovanou součástí tržní ekonomiky, jelikož se systémově jedná o nejvýznamnější součást. Mezi typy bankovních regulací patří pravidla likvidity, praktiky řízení rizik, úvěrové postupy, opatření proti praní špinavých peněz a další. Lze tedy říci, že bankovními regulacemi jsou stanoveny pravidla a požadavky. Regulační reporting je mechanismus, který banky prosazují k dodržování těchto regulací prostřednictvím vykazování dat.

2.2 Aktuální situace v oblasti regulatorního reportingu

Jak jsem již zmiňoval, po ekonomické krizi v roce 2008 si svět začal uvědomovat vztah bankovnictví vůči světu a byla provedena nutná opatření, která by situacím a krizím pomáhala

zabránit. Jedním z nutných preventivních opatření je zvyšování frekvence a počtu výkazů, které jsou banky povinny adresátům zasílat.[12]

V oblasti celkového počtu zasílaných reportů lze jednoznačně vidět rostoucí trend i v posledních letech. Změna v počtu zasílaných reportů mezi rokem 2016 a 2018 odhalila nárůst o více než 351 % ročně. Nejvíce se změny dotkly výkazů týkajících se kapitálového trhu. Samotný nárůst výkazů kapitálového trhu mezi zmíněnými lety byl o 162,50 %. Na druhé straně ne všechny výkazy, co se počtu týče, zažívají vzestupný trend. Například výkazy bankovního dohledu zaznamenaly pokles o 9,80 %. V rámci průzkumu KPMG více než třetina dotazovaných uvedla, že mezi problémy sestavování reportů patří pozdní dodání dat potřebných k sestavení regulatorních reportů, což značně celý proces komplikuje.[12]

2.3 Typy regulatorních reportů

Typů regulatorních reportů existuje mnoho a v následující kapitole se zaměřím na tyto reporty:

- AnaCredit,
- CRS/FATCA,
- MKT,
- EBA,
- Likvidita.

2.3.1 AnaCredit

AnaCredit (Analytical Credit Datasets) vznikl jako iniciativa Evropské centrální banky v rámci reportingových požadavků za cílem vytvoření společné databáze s co nejvyšším detailem. Obsahem výkazů jsou úvěry a úvěrová rizika, což v rámci vysoké granularity znamená úvěr po úvěru. Vytvoření této databáze má vést k lepšímu dohledu nad finanční stabilitou a bude také sloužit jako podklad pro analýzy.[13; 14]

Povinnost vykazování se vztahuje na banky zemí v rámci Eurozóny, což však neznamená, že banky na území ČR AnaCredit nevykazují. ČNB na svých webových stránkách píše *„Vykazujícími osobami projektu AnaCredit jsou banky a pobočky zahraničních bank se sídlem ve zpravodajském členském státě.“* Jelikož v rámci České republiky má většina bank a poboček zahraničních bank sídlo ve zpravodajském státě, tedy ve státě, jež je členem Eurozóny, povinnost vykazovat AnaCredit se jich týká.[13]

Frekvence výkazů tohoto typu je na měsíční bázi, některé části jsou však vykazovány čtvrtletně. Předmětem vykazování jsou pohledávky za fyzickými osobami podnikateli a pohledávky za právnickými osobami. Dohromady se vykazuje 13 provázaných výkazů, z nichž je 12 AnaCredit a 1 report vykazuje referenční data systému RIAD.[13]

2.3.2 CRS/FATCA

CRS (Common Reporting Standard)/FATCA (Foreign Account Tax Compliance Act) představuje mezinárodní standard výměny informací. Standard byl zaveden Organizací pro hospodářskou spolupráci a rozvoj (OECD). Cílem vykazování CRS je zamezení daňových úniků a zajištění transparentnosti v oblasti finančních účtů. Banky jsou povinny systematicky shromažďovat informace o finančních účtech, které budou následně vykazovány. Na základě CRS dochází k automatické výměně informací s daňovými orgány po celém světě.

Jedním z mnoha daňových podvodů je ukrytí peněz na účtech v zahraničí, aniž by byly zdaněny ve státě daňové rezidence. Pomocí tohoto standardu by se mělo podobným situacím předcházet.

2.3.3 EBA

EBA (European Banking Authority) je agenturou Evropské Unie, která od roku 2011 usiluje o účinnou a jednotnou regulaci bankovního trhu. Také specifikuje EBA Reporting Framework, což je standard pro finanční instituce, v rámci něhož bankovní instituce v EU reportují se výkazy. Dělí se na několik dílčích částí, ale mezi hlavní dvě patří FINREP (Financial Reporting) a COREP (Common Reporting).

FINREP představuje soubor finančních výkazů, především rozvahy, výkazy zisků a ztrát a výsledků hospodaření. Některé reporty tohoto typu bývají dostupné i široké veřejnosti nebo například výhradně investorům, jejichž kapitál je zatížen rizikem.

COREP též představuje soubor reportů, avšak v porovnání s FINREP se COREP zaměřuje hlavně na kapitál. Výkazy tedy obsahují informace o solventnosti společnosti, kapitálové přiměřenosti, rizikových expozicích a další.

2.3.4 Likvidita

Mezi nejméně komplexní výkaz v rámci regulatorního reportingu v rámci mého výčtu patří likvidita. Řešit, zda je banka ochotna dostát svým krátkodobým závazkům, není otázka pouze bankovního trhu, ale jedná se o důležitý ukazatel pro všechny společnosti. Cílem vykazování je monitoring a dodržování integrity bankovního sektoru.

2.3.5 Nedodržování regulatorních povinností

V případě nedodržování povinností regulatorního reportingu mohou být banky pokutovány. Jeden z takových případů se udál v březnu roku 2023, kdy ČNB udělila Fio bance pokutu za chyby ve výkaznictví v hodnotě jeden milion korun. Prohřeškem byla oblast likvidity, dle vyjádření ČNB se jednalo o *„Zkreslení ukazatelů bylo natolik významné, že dávalo nesprávný obraz o likviditní pozici jak pro výkon dohledu ze strany ČNB, tak pro ostatní uživatele informací“.* V takovémto případě mohla ČNB pokutovat banku až do výše deseti procent čistého ročního obratu, což v tomto případě mohlo být až 316,8 milionu Kč.[15]

2.4 Vztah mezi regulatorním reportingem a datovou kvalitou

Poptávka po datech na granulární bázi vzrostla, a stále rychle roste, jako následek ekonomické krize, která se světem prohnala v roce 2008. Agregované výkazy zasílány regulátorům nestačily k získání podrobného přehledu ve vývoji, který se nakonec odehrál v globálním měřítku. Je nezbytné, aby byl vývoj ekonomické situace podrobně monitorován, aby bylo možné ekonomickým šokům a snahou krizím předcházet či je mírnit. A jak bylo již zmiňováno, neměnil se jen počet zasílaných reportů a granularita dat, ale také frekvence, což mělo obrovský vliv na datovou kvalitu. Před rokem 2008 banky měly spoustu času začít řešit kvalitu dat v sestavovaných reportech až při jejich sestavování, avšak kvůli vysoké frekvenci, velkému množství výkazů a vysokému detailu není nadále možné řešit kvalitu dat tímto způsobem. To, že jsou data kvalitní by mělo být vyřešené ještě před začátkem sestavování reportů. Kvalita dat tedy musí být řešena nepřetržitě a jak si s tímto nelehkým úkolem banky dokážou poradit opět odhalil průzkum od KPMG. [12]

Dle průzkumu totiž při sestavování reportů dochází hned k několika problémům a datová kvalita je jedním z nich. Může se jednat o nedodání vstupních dat, pozdní dodání dat potřebných pro sestavení reportu, nedostatečnou zastupitelnost zaměstnanců, kteří výkazy sestavují nebo IT problémy technického rázu. Avšak nejčastěji vyskytujícím se problémem, který uvedlo 72 % bankovních společností, jsou problémy kvality dat. Většina bankovních situací v rámci datové kvality vede deník korekcí. U velkých bankovních institucí je deník veden jak pro automatické, tak pro manuální korekce. Z celkového počtu respondentů 28 % aktivně usiluje o zlepšení datové kvality. Dalších 22 % respondentů využívá centrální tým specializovaný na řešení otázek týkající se datové

kvality. Na druhé straně 67 % oslovených bankovních institucí se zabývá otázkami datové kvality nezávisle v rámci svých vlastních týmů. Tato zjištění jen potvrzují, že problematika datové kvality je pro bankovní sektor esenciální a banky vynakládají značná úsilí na zajištění co nejvyšší úrovně kvality svých dat.[12]

Regulatorní reporting je tedy jedním z mnoha důvodů, proč by se datová kvalita neměla zanedbávat, ačkoliv je, často vnímán jako zlo, které negeneruje byznys.[12]

3 Business Intelligence

Business Intelligence (BI) nemá jednoznačný význam. Dle (Pour a kol., 2012) je BI: „sada procesů, know-how, aplikací a technologií, jejichž cílem je účinně a účelně podporovat řídicí aktivity ve firmě. Podporují analytické, plánovací a rozhodovací činnosti organizací na všech úrovních a ve všech oblastech podnikového řízení“.[16]

Obecně však lze říci, že se jedná o kombinaci business analytiky, vizualizace dat, data mining, datových nástrojů a infrastruktury. Díky použití těchto metodik BI umožňuje informované rozhodování na základě dat či informací, které z dat plynou (tzv. data-driven decission making). Dle (Luftmann & Kempaiah) bylo dokonce BI popsáno jako jedna z nejdůležitějších priorit pro CIO.[17]

3.1 Historie BI

První zmínka o Business Intelligence pochází z roku 1865, kdy jej Richard Millar Devens použil ve své knize "Cyclopædia of Commercial and Business Anecdotes". Pomocí termínu BI popisoval, jak sir Henry Furnese využíval shromážděných informací a na jejichž základě dělal rozhodnutí, čímž profitoval a porážel svou konkurenci. Dalším milníkem je rok 1958, kdy Hans Peter Luhn – vynálezce a konzultant společnosti International Business Machnines (IBM) – napsal článek, v němž popsal a vysvětlil potenciál shromažďování BI skrze využití technologií. Velký rozmach v oblasti nabídky BI byl pozorován v 80. letech 20. století. V tomto desetiletí si majitelé firem a podnikatelé začali uvědomovat důležitost informací, na kterou již dříve upozornil Hans Peter Luhn, a započal vývoj nástrojů, které by sběr a manipulaci s daty umožňovaly. Byly vyvinuty první systémy pro správu databází, mezi něž patří třeba Decision support systems (známé jako DSS). Spolu s tímto systémem byly taktéž vyvinuty datové sklady a systém OLAP, které umožňovaly práci právě s DSS.[18; 19]

Dnešní definice BI je však velmi široká a zmodernizovaná. Co znamená BI dnes, nebude s vysokou pravděpodobností znamenat za pár let. V dnešní době si pod BI můžeme představit procesy, nástroje a technologie, které slouží organizacím k shromažďování, analýze a transformaci dat na poznatky, které mají smysluplnou hodnotu.

3.2 BI jako vizualizační nástroje

V mé práci se zaměřuji na vizualizační nástroje pro analýzu dat. V následujících kapitolách budu používat termín „BI" pouze jako zkratku pro vizualizační nástroje. Takže veškeré zmínky o BI budou implicitně odkazovat na vizualizační nástroje.

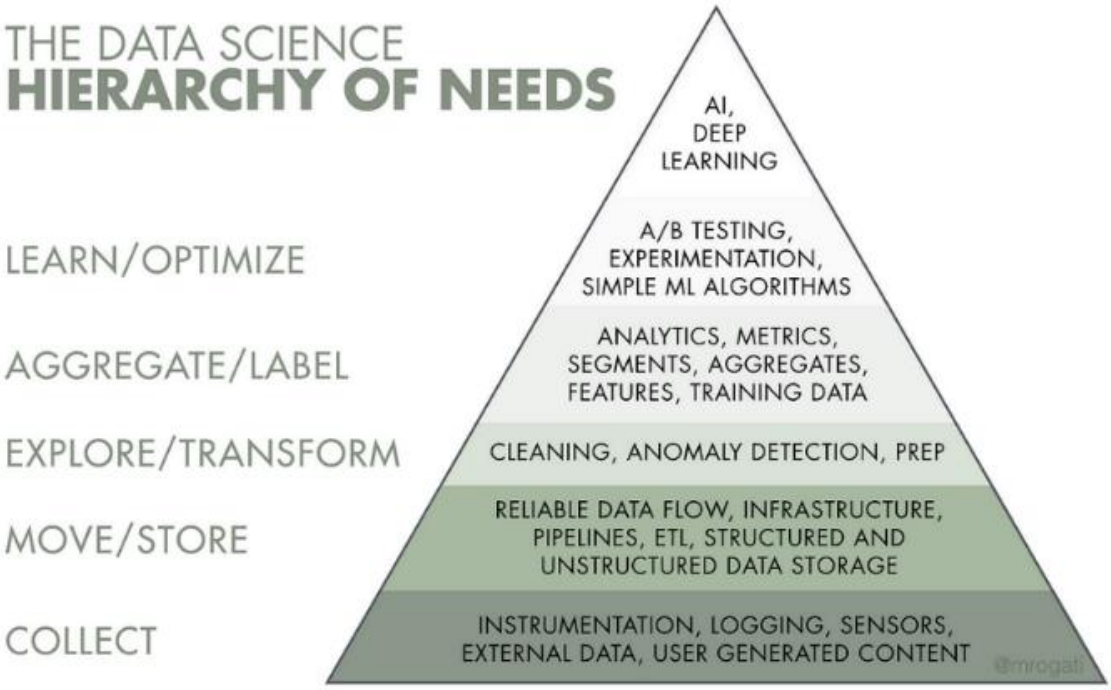
Vizualizační nástroje slouží, jak již název napovídá, k vizualizaci dat. Standardně se jedná o grafické prvky, jako mapy, grafy, tabulky, diagramy, prostřednictvím kterých se data interpretují a vizualizují. Hlavním účelem tedy je převádět obrovská kvanta dat do srozumitelné podoby. Často se setkávám s tím, že lidé namítají, proč vůbec BI nástroje potřebujeme, když máme spoustu skvělých tabulkových editorů, jako třeba Google Spreadsheets, případně Microsoft Excel, kde lze data mít velmi přehledně a vytvářet vizualizace je poměrně jednoduché, jaká je tedy přidaná hodnota BI?[20]

Největší užitek BI přináší při práci s big data, které obsahují řádově statisíce, miliony či i více dat. Problém např. u Microsoft Excelu je, že tento tabulkový editor zvládne zpracovávat maximálně kolem jednoho milionu záznamů a i řády deseti či statisíců se velmi podepisují na jeho výkonnosti. Dále možnosti vizualizačních nástrojů jsou oproti tabulkovým editorům velmi vyspělé a nabízejí velmi širokou škálu možností. BI nástroje jsou také přímo určeny pro spojení a zpracování dat z vícero možných zdrojů, ať už se jedná o CSV soubory, tabulkové editory, připojení k databázovým serverům, zpracování dat z NoSQL databází, nabídka je opravdu široká. Dalším rozdílem oproti konvenčním tabulkovým editorům jsou dynamické a interaktivní dashboardy, které sice dokážou i tabulkové editory, ale zdaleka ne v takové míře a hloubce jako BI. Co já ovšem vnímám jako jednu

ze stěžejních výhod – automatická aktualizace dat. BI nástroje umožňují pravidelné automatické aktualizace dat (např. každý den v 7 hodin ráno) a nemusí se ani jednat o aktualizace, nástroje zvládají pracovat i s live daty.

3.3 Zařazení BI v Pyramidě datové vědy

Termín „The Data Science Hierarchy of Needs“ poprvé ilustrovala Monica Rogati. V případě tohoto schématu jsem nedokázal dohledat přesný překlad do českého jazyka, avšak pro mě osobně je českým ekvivalentem „Hierarchie potřeb datové vědy“. Ovšem je nezbytné uvědomění, že se nejedná o hierarchii (ačkoliv tento termín je použit i v originálním názvu), ale o pyramidu. Pokud by se jednalo o hierarchii, znamenalo by to, že části na vrcholu schématu mají větší vliv na chod společnosti, společnosti o ně usilují a jsou nadřazené částmi pod nimi. To tak samozřejmě není, v tomto případě jde tedy o pyramidu, která vnímá spodní části jako esenciální předpoklady pro budování částí nad nimi. Takže například případě této pyramidy lze říci, že pokud společnost nemá vyřešené ani sbírání dat, nemůžou se věnovat ani vytváření modelů strojového učení, které jsou o pár úrovní výše.[21; 21; 22]



OBR. 2 – THE DATA SCIENCE HIERARCHY OF NEEDS (ZDROJ: HTTPS://MEDIUM.COM/@HUGH_DATA_SCIENCE/THE-PYRAMID-OF-DATA-NEEDS-AND-WHY-IT-MATTERS-FOR-YOUR-CAREER-B0F695C13F11)

Taktéž je důležité si uvědomit, že ne všechny společnosti musí nutně mít všechny úrovně tohoto schématu. Například, pokud se jedná o kamenný obchod s dekoracemi, který denně prodá 50 kusů produktů, ve většině situací bude stačit, pokud veškeré prodeje budou zaznamenávány a monitorovány v rámci tabulkového editoru, jehož hlavním cílem bude udržovat přehled o skladových zásobách, popřípadě jaké produkty se nejvíce prodávají, a naopak u kterých prodej stagnuje. Jde o to, že ve zmíněné situaci postačí jen první stupeň pyramidy, který ani nemusí být v komplexním měřítku. Každá společnost je jiná, ale s velkou jistotou lze říct, že většina bank po celém světě směřuje k tomu, aby všechny bloky měly své zastoupení.[22; 21]

3.3.1 Pracovní pozice datových věd

Také všechny pracovní pozice, které se pojí s datovými vědami spadají do tohoto spektra. Ve spodní části pyramidy (COLLECT a MOVE/STORE) se vyskytují softwarový inženýři či datoví inženýři, kteří zajišťují chod dat prostřednictvím organizace a ukládání dat, které budou použity pro další použití. Jestli organizace bude ukládat data ve formě CSV souborů či ve formě relačních databází (což je dle mého názorů nejčastěji využívaná forma uchování dat), je vedlejší. Tím to však nekončí, datoví a softwaroví inženýři se i nadále zabývají transformací a čištěním dat tak, aby byla použitelná v dalších fázích pyramidy.[22; 23]

Zhruba v polovině pyramidy (EXPLORE/TRANSFROM a AGGREGATE/LABEL) se vyskytují business analytici, datoví analytici, BI analytici či projektoví manažeři, jejichž práce je založena na tom, co připravili datoví a softwaroví inženýři. Jedná se tedy o hledání odpovědí opřená o data, většinou z business části společnosti. Například společnost či projektového manažera zajímá, zda je za posledních pět let vzestupný trend, co se týče prodaných úvěrů. Datoví analytici poté na základě dat tyto informace zjistí a v jednoduše pochopitelné formě předají odpovědným osobám, které na základě těchto dat mohou činit další rozhodnutí, čímž se opět vracíme k tzv. data-driven decission making, neboli rozhodování na základě dat. Důležitým předpokladem pro správné rozhodování na základě dat jsou správná a vysoce kvalitní data, v opačném případě může docházet tzv. k TITO (Trash-in, Trash-out), což znamená, že pokud jsou počáteční data znečištěná, budou znečištěné (nevalidní) i informace a znalosti, kterých pomocí těchto dat dosáhneme. Součástí prostřední části pyramidy je také vytváření vizualizací – užitečných dashboardů, která v efektivní a jednoduše pochopitelné formě předávají poznatky dále.[22; 23]

Vrchol pyramidy (LEARN/OPTIMIZE), tedy budování modelů strojového učení, AI modelů, lineárních regresí a deep learning nebo A/B testing. Tento druh práce zastávají datoví vědci nebo inženýři strojového učení. Cílem těchto modelů jsou vylepšení nových či již stávajících produktů, predikce chování klientů, hledáním vztahů mezi jednotlivými veličinami a jiné. Co bývá problémové u budování těchto modelů je množství dat. V ideálním případě společnosti potřebují řádově miliony, miliardy či i vícero dat specifického typu, aby na jejich základě bylo možné modely trénovat. Druhým velmi esenciálním požadavkem, stejně jako v prostřední fázi pyramidy, je kvalita a čistota dat.[22; 23]

Každá společnost definuje pracovní pozice jiným způsobem a co znamená datový analytik v jedné společnosti, často neznamená datový analytik ve společnosti druhé. Je třeba brát na vědomí, že zařazení pozic v rámci této pyramidy je orientační a mohou se prolínat mezi sebou.

3.4 Cíle BI

Z kapitoly výše je jasné, že BI zapadá zhruba do poloviny struktury pyramidy mezi vytváření datové infrastruktury, přípravy dat pro budoucí použití a vytváření statistických modelů či modelů strojového učení.

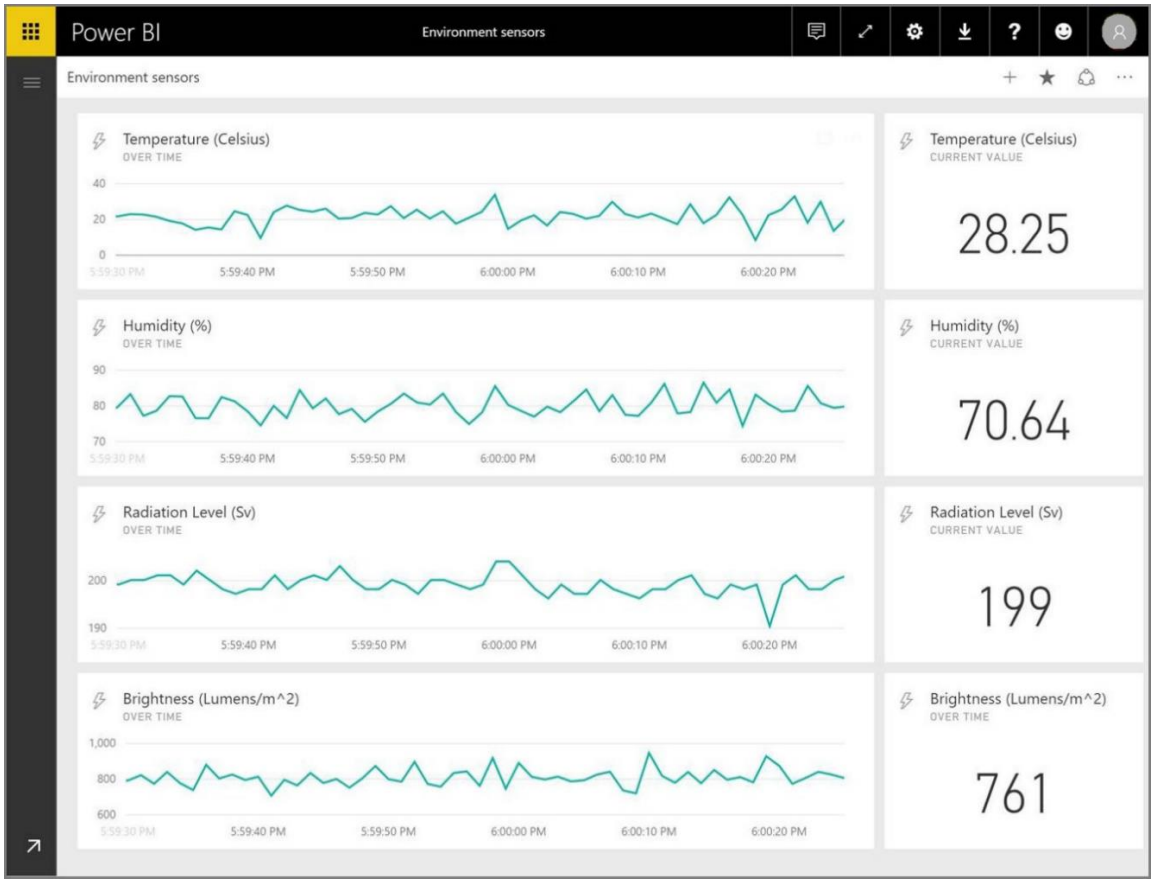
Cílem BI a také jednou z nejdůležitějších součástí vizualizací je však reporting, neboli vytváření dashboardů, které jsou dále reportovány ostatních členům týmu, nadřízenému či představenstvu. Dle (Pour a kol., 2018) reporting představuje *„komplexní systém informací a ukazatelů charakterizujících činnosti společnosti, poskytuje ve vhodné formě a včas podklady pro podporu rozhodování na všech stupních organizační struktury“.*[24]

Vytváření reportů a dashboardů v rámci organizace může být klíčové hned z několika důvodů, mezi něž patří:

- převedení Big Dat do srozumitelné formy,
- monitoring vývoje,
- vytváření dynamických reportů.

Převedení Big Data do srozumitelné podoby je prvním klíčovým prvkem**.** Zkoumat korelace a kauzality na obrovských objemech dat je velmi náročné, když jsou hodnoty uloženy ve formě řádků, sloupců. I pokud by člověk pouze četl data řádek po řádku, i po pročtení milionu záznamů by bylo velmi náročné vyvozovat závěry. Vizualizace dat umožňuje efektivně identifikovat vzorce a souvislosti, které by mohly být v tabulkové formě obtížně rozeznatelné

Monitoring vývoje, neboli streamování živých dat, je za mne osobně nejsilnější nástroj celého BI. Pokud dokážu sbírat data v reálném čase, mohu je také v reálném čase vizualizovat formou dashboardů. Může se jednat třeba o senzory, které sbírají údaje o počasí a data jsou poté živě (nebo klidně po intervalech, které si já sám určím) zobrazována. V následujícím dashboardu lze v pravé části vidět aktuální hodnoty teploty, vlhkosti, úrovně radiace a světelnosti, v levé části je vidět vývoj těchto hodnot. Pokud senzor naměří jiné hodnoty, na tomto dashboardu se změny okamžitě projeví.



OBR. 3 - ENVIRONMENT SENSORS REPORT V POWER BI (ZDROJ: HTTPS://LEARN.MICROSOFT.COM/CS-CZ/POWER-BI/CONNECT-DATA/SERVICE-REAL-TIME-STREAMING)

Vytváření dynamických reportů zajišťuje použití filtrů, parametrů a proměnných, což umožňuje uživatelům interaktivně data ovládat. Například pokud dashboard slouží pro vizualizaci prodaného zboží za poslední rok, uživatelé mohou pomocí filtrů vybrat specifické období a pokud tak učiní, všechny vizuály na daném dashboardu se přepočítají. A to je pouze jeden z mnoha případů dynamických reportů

3.5 Nabídka vizualizačních nástrojů

Na trhu BI nástrojů jich existuje nepřeberné množství a s rostoucí popularitou jich ještě přibývá. Na grafickém zobrazení Gartner, umístěném níže se vlajkové lodě v tomto oboru nachází v pravém horním kvadrantu. Mezi největší favority v této oblasti patří Power BI od společnosti

Microsoft nebo Tableau, které je produktem firmy Salesforce. Na předních příčkách se drží Qlik, Google Data Studio či Domo.

Figure 1: Magic Quadrant for Analytics and Business Intelligence Platforms

OBR. 4 - GARTNER MAGIC QUADRANT OF BI TOOLS (ZDROJ: HTTPS://WWW.ALIBABACLOUD.COM/BLOG/2023-GARTNER%C2%AE-MAGIC-QUADRANT%E2%84%A2-FOR-ANALYTICS-AND-BUSINESS-INTELLIGENCE-PLATFORMS_600141)

Rozdíly mezi jednotlivými nástroji nejsou obrovské, všechny fungují na podobném principu a hlavním rozdílem často bývá cenová politika nebo specifické požadavky, které si zákazníci na nástroj nárokují.

BI nástroje jsou v dnešní době neodmyslitelnou součástí všech organizací, které pracují s větším množstvím dat. Ačkoliv na trhu existují desítky BI nástrojů, které se jednotlivě mohou lišit v závislosti na ceně, licencování, uživatelské náročnosti či integraci, v kapitole 5 bych se rád podíval na dvě současné vlajkové lodě – Power BI a Tableau a následně porovnal jejich klady a zápory.

PRAKTICKÁ ČÁST

4 Porovnání Power BI a Tableau

Jak Power BI, tak Tableau vznikly až ve 21. století a cíle byly stejné – vytvořit jednoduché, intuitivní rozhraní, které by umožňovalo vytvářet vizuály i osobám, které nedisponují technickými dovednostmi. Jak je zmíněno již v podkapitole 3.5 „Nabídka vizualizačních nástrojů“, Tableau i Power BI patří mezi dva nejpopulárnější BI nástroje na celosvětovém trhu. Dle dat, kterými disponuje společnost Enlyft, 97 538 společností využívá Power BI a na druhé straně 120 182 společností využívá Tableau. Ze všech sektorů jsou oba dva nástroje nejvíce využívány ve službách informačních technologií s počtem zaměstnanců 50 – 200. Principy obou nástrojů jsou stejné, zakládají si na funkci Drag and Drop, která uživatelům umožňuje intuitivně skládat vizuály. Oba dva nástroje umožňují vytvářet desítky druhů vizuálů a dynamických dashboardů, které jsou navíc vysoce modifikovatelné.[25][26; 27]

Kvůli rozsáhlé funkčnosti obou nástrojů je nemožné pokrýt a porovnat všechny funkcionality. Proto jsem se rozhodl, že v rámci porovnání zaměřím pouze na vybrané aspekty, které rozdělím na technické a netechnické. Následně na transformaci a čištění datové sady, na jejichž základě dále vytvářím vizualizace v jednotlivých nástrojích pro porovnání.

V rámci netechnických aspektů se zaměřím na:

- certifikáty,
- cena/varianty,
- jazyk,
- operační systém.

V rámci technických aspektů jsem se rozhodl pro zaměření na:

- programovací jazyky,
- připojení k datovým zdrojům,
- relace mezi tabulkami,
- spojování tabulek pomocí joinů a union operace.

4.1 Netechnické rozdíly BI nástrojů

4.1.1 Certifikáty

Certifikáty ohledně znalosti BI nástrojů nabízí Salesforce pro Tableau i Microsoft pro Power BI. V případě Tableau je jich nabízeno dokonce 6 a patří mezi ně:

- Certified Tableau Desktop Specialist,
- Certified Tableau Data Analyst,
- Certified Tableau Server Associate,
- Certified Tableau Consultant,
- Certified Tableau Architect.[28]

Cena jednotlivých certifikátů se liší, ale pohybuje se od 100 USD to 250 USD. Na druhé straně Power BI poskytuje pouze jeden druh certifikátu – *„PL-300: Microsoft Power BI Data Analyst“* a cena činí 165 USD. Co se týče samotných zkoušek, zatímco u Power BI certifikátu lze zkoušku skládat v několika jazycích, u Tableau tomu tak není, např. Certified Tableau Desktop Specialist lze skládat pouze v angličtině a japonštině. Bohužel, ani v případě Power BI není k dispozici český jazyk.[28][29]

4.1.2 Cena

Cena BI nástrojů se liší v závislosti na variantě, kterou si chce uživatel pořídit. Power BI i Tableau nabízejí vícero možností a pro Power BI vypadají následovně:

- Power BI Desktop – zdarma: Individuální přístup k vytváření reportů na vlastním zařízení, což je vhodné pro domácí použití.
- Power BI Pro – 9,40 EUR/měsíc: Tato varianta poskytuje jednu licenci k vytváření reportů, které lze sdílet napříč organizací.
- Power BI Premium – 18,70 EUR/měsíc: Nejdražší varianta, která poskytuje vše, co Power BI nabízí.[30]

Tableau nabízí varianty 4, ovšem nedá se říci, že více je v tomto ohledu lépe:

- Tableau Public – zdarma: Obsahuje velmi limitované množství datových zdrojů, mezi nejvyužitelnější z nich patří Microsoft Excel, textové soubory, JSON a Microsoft Access. Vizualizace v rámci bezplatné Tableau verze lze pouze veřejně publikovat, tím pádem se nedoporučuje vizualizovat citlivá data, jelikož k nim může mít kdokoliv přístup. Tuto variantu vnímám jako vhodnou pouze pro prozkoumávání a zdokonalovaní se v rámci nástroje.
- Tableau Viewer – 15 EUR/měsíc: Tato varianta umožňuje pouze čtení Tableau reportů, takže pro individuální použití tato varianta nenajde využití.
- Tableau Explorer – 42 EUR/měsíc: Neumožňuje připojení k surovému datovému zdroji, lze pouze přistupovat a analyzovat k již publikovaným datovým zdrojům prostřednictvím Tableau Serveru.
- Tableau Creator – 75 EUR/měsíc: Verze, která umožňuje plný přístup, tedy vytváření vizualizací a dashboardů v rámci Tableau Desktop na základě vlastních datových zdrojů.[31]

4.1.3 Jazyk

Jazyk prostředí také odlišuje oba dva nástroje. Český jazyk je podporován pouze u Power BI. Při používání Tableau se tedy budeme muset spokojit s jazykem anglickým. Věřím, že tento rozdíl může být také pro spoustu lidí klíčový.

4.1.4 Operační systém

Operační systém, v tomto případě Windows a Mac OS je podporován pouze u Tableau. Pro Mac OS existují způsoby, jak pracovat s Power BI (například pomocí Dual Bootu), čímž lze doinstalovat Windows i na zařízeních běžících na Mac OS. Případně je dostupná také webová verze, ale i to může být nedostatkem. Pro uživatele, kteří pracují jen a pouze s Mac OS, může tato skutečnost velmi rychle ovlivnit rozhodovací proces.

4.2 Technické rozdíly BI nástrojů

4.2.1 Programovací jazyky

Programovací jazyky nejsou nutností, ale pokud chce uživatel z nástrojů dostat co nejvíce, cesta vede tímto směrem. A v tomto případě nutno vyzdvihnout Power BI, které kromě jazyku R, jež je hojně používán v datových vědách, nabízí použití jazyku DAX (Data Analysis Expression) a M. Oba dva jazyky – DAX i M lze najít pouze u některých programů od firmy Microsoft a každý plní svou jedinečnou funkci. Microsoft ohledně DAX na svých webových stránkách píše *„DAX je kolekce funkcí, operátorů a konstant, které lze použít ve vzorci nebo výrazu k výpočtu a vrácení jedné nebo více hodnot.“* Důležitá je zmínka, že se dají použít „ve vzorci“, jelikož vzorce jsou používány v rámci kalkulací, které napomáhají k transformaci dat a vytváření nových sloupců, které mohou vznikat jako deriváty ostatních sloupců.

Jazyk M má účel v jiné fázi vytváření reportu, konkrétně v Power Query, které se využívá při importování datového zdroje a je určen pro manipulaci dat, která pochází z různých datových zdrojů. Také pomocí M lze transformovat data před nahráním do datového modelu.

Tableau na druhé straně nabízí vyšší flexibilitu a vícero možností. Podpouje programovací jazyk R, stejně jako Power BI a kromě toho také Python, Javu, C a C++. Navíc podporuje jazyk MDX

(Multidimensional Expressions), jež byl vytvořen specificky pro Tableau. MDX je využíván pro práci s OLAP datovými zdroji a umožňuje složitější výpočty měr i kategorií. Avšak jak jsem již psal, znalost programovacích jazyků pro vytváření vizualizací není nutná a dle mého názoru ani obvyklá (vyjma DAX a M).

4.2.2 Připojení k datovým zdrojům

Jak Power BI, tak Tableau nabízejí nepřeberné množství datových zdrojů, odkud čerpat data pro vizualizace. Na jedné straně Power BI nabízí neuvěřitelných 192 druhů datových zdrojů, na straně druhé Tableau „pouze" 92. Ovšem tato čísla nejsou plně vypovídající. Abych uvedl na příkladu, společnost Tableau nabízí jako jednu ze svých 92 možností datových zdrojů také připojení k *„Oracle Eloqua,"* což je dceřiná společnost spadající pod Oracle Corporation, zabývající se automatizací marketingu. Power BI připojení přímo k tomu zdroji nenabízí, ale lze se připojit prostřednictvím jiného zdroje, a to přes „CData Connect Cloud". Porovnání počtu datových zdrojů v absolutních číslech tedy není relevantní a velmi těžko se odhaduje, jak moc má v tomto ohledu Power BI navrch, avšak oba dva nástroje podporují získání dat z těch nejzákladnějších zdrojů, které jsou uvedeny v tabulce níže.[32; 33; 34]

Typ zdroje	Zdroj
Azure	Azure Data Lake Storage Gen2, Azure SQL Database, Azure SQL Synapse Analytics, Databricks
Databáze	Actian Vectorwise, Amazon Athena, Amazon Redshift, Anaplan, Denodo, Dremio by Dremio, Exasol, Google BigQuery, IBM DB2, IBM PDA (Netezza), Impala, MariaDB, MarkLogic, Microsoft SQL Server, MongoDB, MySQL, Oracle, PostgreSQL
Soubor	Excel Workbook, Text/CSV, JSON File, PDF File
Online služby	Google Analytics, Intuit QuickBooks Online, LinkedIn Sales Navigator, Marketo
Ostatní	Hadoop, Google Sheets, Odata

TAB. 1 – UKÁZKA PODPOROVANÝCH ZDROJŮ POWER BI A TABLEAU (ZDROJ: HTTPS://LEARN.MICROSOFT.COM/EN-US/POWER-BI/CONNECT-DATA/DESKTOP-DATA-SOURCES, HTTPS://HELP.TABLEAU.COM/CURRENT/PRO/DESKTOP/EN-US/EXAMPLECONNECTIONS_OVERVIEW.HTM)

4.2.3 Vztahy mezi tabulkami

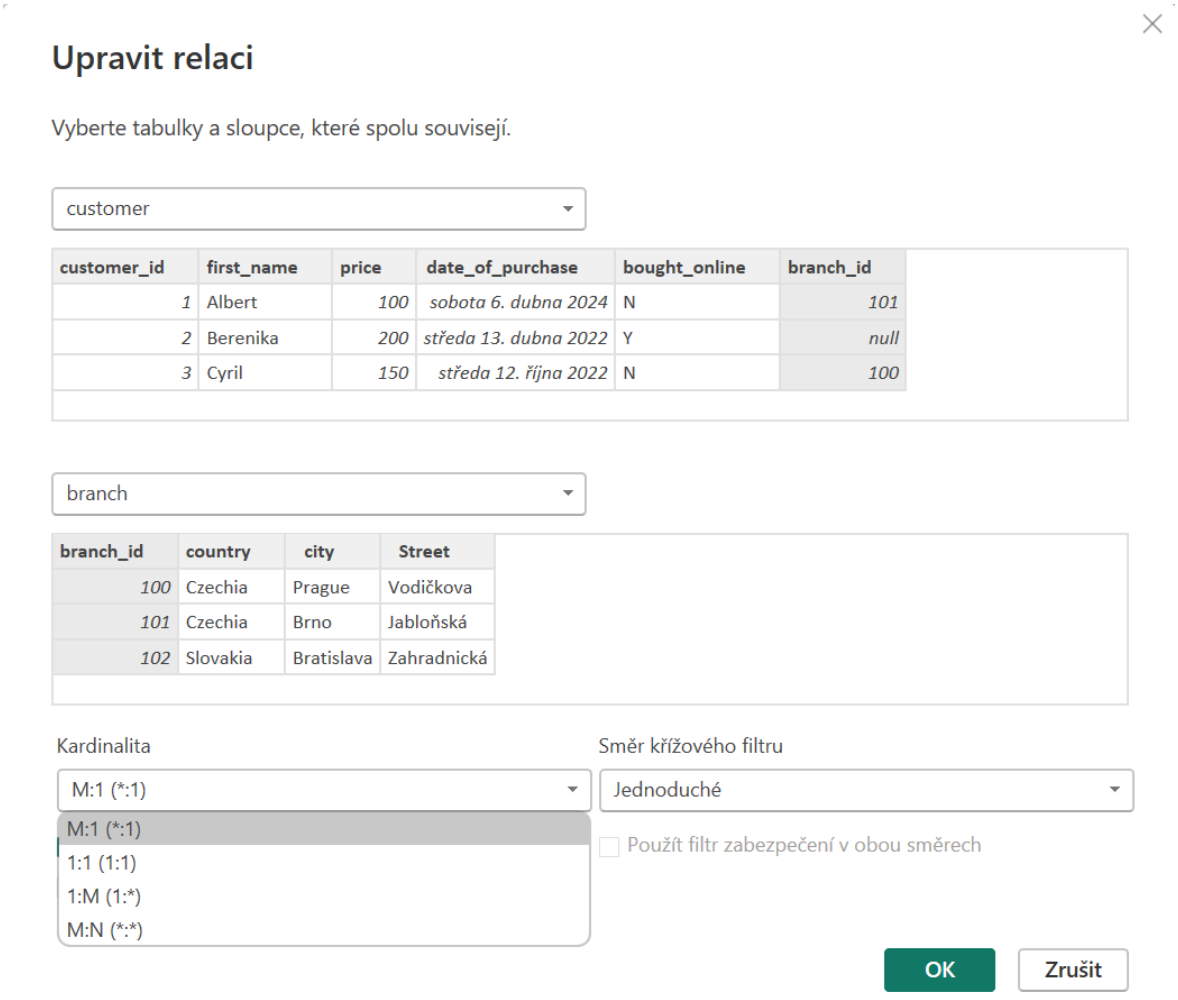
Vztahy, neboli relace, mezi tabulkami slouží k provázání dat, která spolu souvisí, mezi sebou. Relace jsou tvořeny mezi jednotlivými tabulkami (obvykle v rámci databáze) a každá vazba (relace) je definována kardinalitou na obou koncích. Při zobecnění relací dostaneme 3 základní typy a těmi jsou:

- Vazba 1:1 – nejjednodušší typ vazby. Definuje, že jednomu záznamu v tabulce jedné odpovídá právě jeden záznam v tabulce druhé. V takovém případě se primární klíč silnější entity používá jako cizí klíč v tabulce slabší. Příkladem je zaměstnanec a jeho služební auto. Právě jednomu zaměstnanci připadá jedno služební auto a na jedno služební auto připadá právě jeden zaměstnanec.
- Vazba 1:N – přiřazuje jednomu záznamu z první tabulky několik záznamů z tabulky druhé. Například vazba oddělení a zaměstnanec. Jednomu oddělení může být přiděleno několik zaměstnanců, ale jednomu zaměstnanci musí být přiděleno právě jedno oddělení.
- Vazba M:N – nejméně obvyklá vazba, pro kterou z logických důvodů bývá přidána tabulka třetí – spojovací. Vazba umožňuje vzájemně několika záznamům přiřadit několik záznamů. Typicky udávaným příkladem je student a předmět. Jeden student může být zapsán na několik předmětů a jeden předmět může být přidělen několika studentům. Jelikož tento typ vazby roznásobí

výsledky v obou tabulkách, využívá se tabulka spojovací, neboli vazební, která obsahuje atributy vztahu.[35]

Relace jsou v rámci Power BI přiřazovány automaticky v rámci importu dat. Avšak relace je možné vytvářet či upravovat manuálně v rámci záložky „Zobrazení modelu“. A nabízené relace jsou vidět na vizuálu níže. Power BI nabízí vybrat možnost:

- M:1 (*.1)
- 1:1 (1:1)
- 1:M (1:*)
- M:N (*.*)

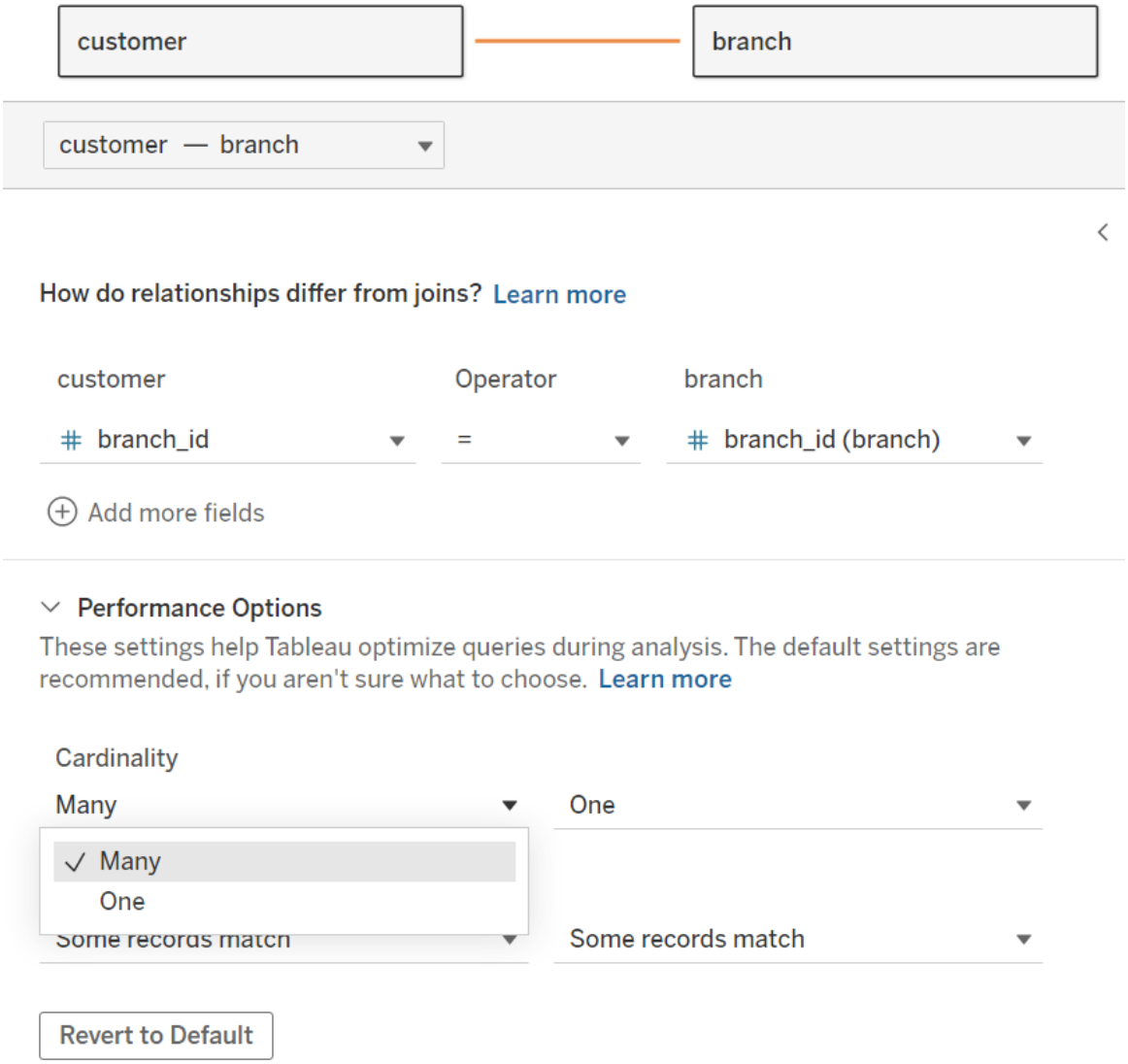


OBR. 5 - VYTVÁŘENÍ RELACÍ V POWER BI (ZDROJ: AUTOR)

V Power BI se v rámci relace nastavuje také směr křížového filtru, který lze nastavit „Jednoduché“ či „Obě“ a ovlivňuje, jakým způsobem se bude propisovat filtrování do obou tabulek. Pokud je směr křížového filtru nastaven na „Obě“, filtry se šíří z jedné tabulky do druhé a zpět, tedy pokud filtrujeme data v tabulce jedné, ovlivní to data i v tabulce druhé. V případě možnosti „Jednoduché“ jsou filtry aplikovány pouze jednosměrně.

Tableau šlo trochu jinou cestou, ale ve výsledku nabízí stejný princip. Kardinalita u jednotlivých relací mezi tabulkami je přiřazena, avšak z mé vlastní zkušenosti, ne vždy správně, proto mi přijde manuální nastavení relací jako vhodnější volba. V případě Tableau nenabízí číselné popisy kardinality, ale pouze „Many“ či „One“. Mezi dvěma tabulkami můžeme, stejně jako u Power BI, vytvořit dohromady čtyři typy relací:

- One to One
- One to Many
- Many to One
- Many to Many



OBR. 6 - VYTVÁŘENÍ RELACÍ V TABLEAU (ZDROJ: AUTOR)

Možnost směru křížového filtru v tomto případě Tableau nenabízí.

4.2.4 Spojování tabulek pomocí joinů a union operace

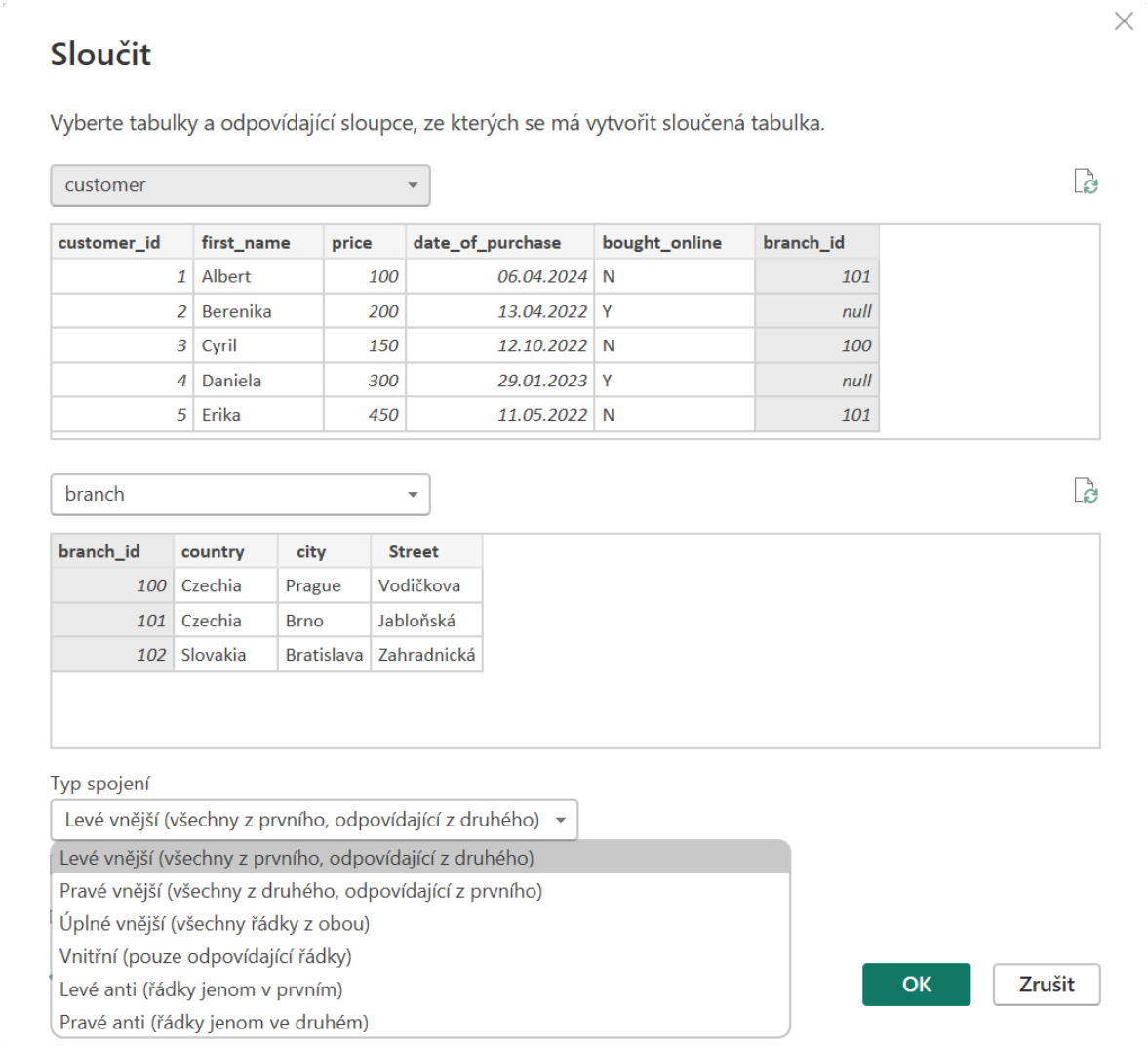
Tabulky nemusí být propojovány pouze pomocí relací mezi jednotlivými tabulkami. Druhou alternativou je vytváření joinů, které funguje podobně jako v SQL. Oba dva nástroje tyto možnosti nabízí, avšak trochu odlišným způsobem. Tabulky jsou pomocí joinů spojovány na základě společných sloupců a mezi základní typy joinů se řadí:

- Inner Join – obsahuje pouze společné hodnoty z obou tabulek.
- Left Join – obsahuje všechny hodnoty z levé tabulky a pouze jim odpovídající hodnoty z pravé tabulky.
- Right Join – obsahuje všechny hodnoty z pravé tabulky a pouze jim odpovídající hodnoty z levé tabulky.
- Full Outer Join – obsahuje všechny hodnoty z levé i pravé tabulky.

- Cross Join – každý záznam z levé tabulky spojí s každým záznamem z tabulky pravé, tedy vytvoří všechny možné kombinace jednotlivých záznamů. Tento typ se však u BI nástrojů nevyužívá.

V Power BI nelze spojovat tabulky pomocí joinů v rámci „Zobrazení modelu“, jako tomu bylo u relací, ale je nutné přejít do Power Query editoru, kde máme možnost „sloučit dotazy“. Joiny se vytváří v podobném okně jako relace, tedy po výběru společného pole lze vybrat typ spojení a na výběr je:

- Levé vnější spojení – Left Join
- Pravé vnější spojení (Right Join)
- Úplné vnější spojení (Full Outer Join)
- Vnitřní spojení (Inner Join)
- Levé anti spojení – zachová pouze řádky z levé tabulky, které nemají shodu s pravou tabulkou.
- Pravé anti spojení – zachová pouze řádky z pravé tabulky, které nemají shodu s levou tabulkou.



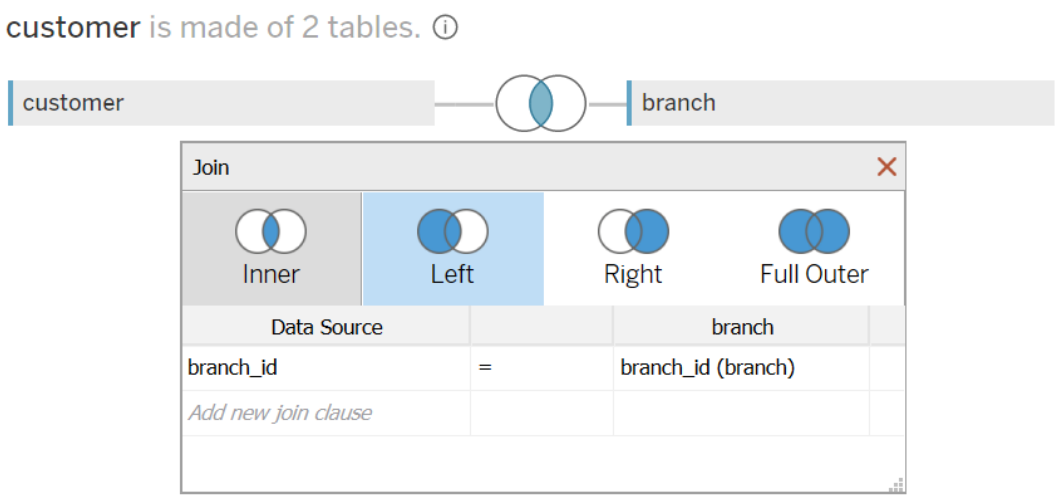
OBR. 7 - SLUČOVÁNÍ TABULEK V POWER BI (ZDROJ: AUTOR)

V Power BI tedy kromě tradičních joinů nalezneme také joiny, které nejsou zdaleka tak obvyklé (anti spojení). To může oproti Tableau nabízet určitou konkurenční výhodu.

V rámci nástroje Tableau máme k dispozici pouze tradiční joiny, ačkoliv pro většinu lidí bude i tato možnost naprosto dostačující. Jak je vidět na obrázku níže, k dispozici je:

- Inner (Join)
- Left (Join)

- Right (Join)
- Full Outer (Join)

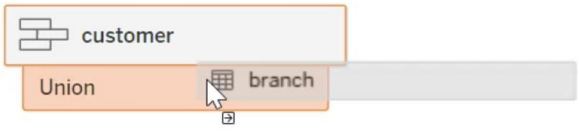


OBR. 8 – SLUČOVÁNÍ TABULEK V TABLEAU (ZDROJ: AUTOR)

Propojení v tomto případě lze udělat na základě jednoho či vícero sloupců. Po vybrání „Add new join clause“ bychom mohli vybrat další sloupce, na základě kterých chceme propojení učinit, případně vytvořit kalkulaci pro propojení tabulek.

Joiny spojují tabulky na základě společných polí. Další možností je tzv. union operace, která tabulky spojuje tím způsobem, že jednu tabulku „nalepí“ pod tu druhou. Tím pádem nám nevzniknou žádné nové sloupce, jako je tomu u joinů, ale pouze se zvýší počty řádků. Union funguje na podobném principu jako v SQL, s tím rozdílem, že v SQL je nutné mít stejný počet sloupců a v jednotlivých sloupcích stejné datové typy (například pokud první sloupec v jedné tabulce nese název „salary“ a je číselného typu, tak nepůjde napojit na druhou tabulku, kde první sloupec nese název „first_name“ a jde o textové pole). V Tableau tento předpoklad sice neplatí, ale i tak bych doporučil používat union podobně jako v SQL.

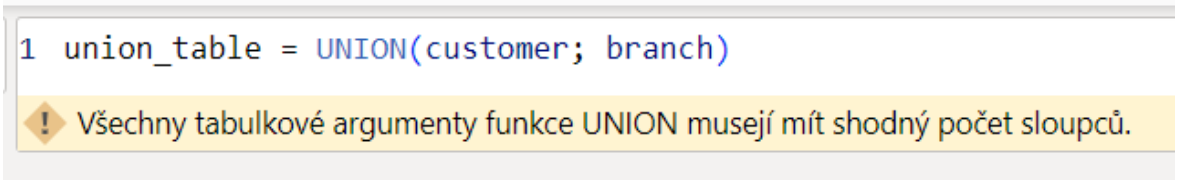
V případě Tableau lze provést union tabulek přetažením jedné tabulky na druhou v rámci vytváření datových zdrojů.



OBR. 9 – UNION OPERACE V TABLEAU (ZDROJ: AUTOR)

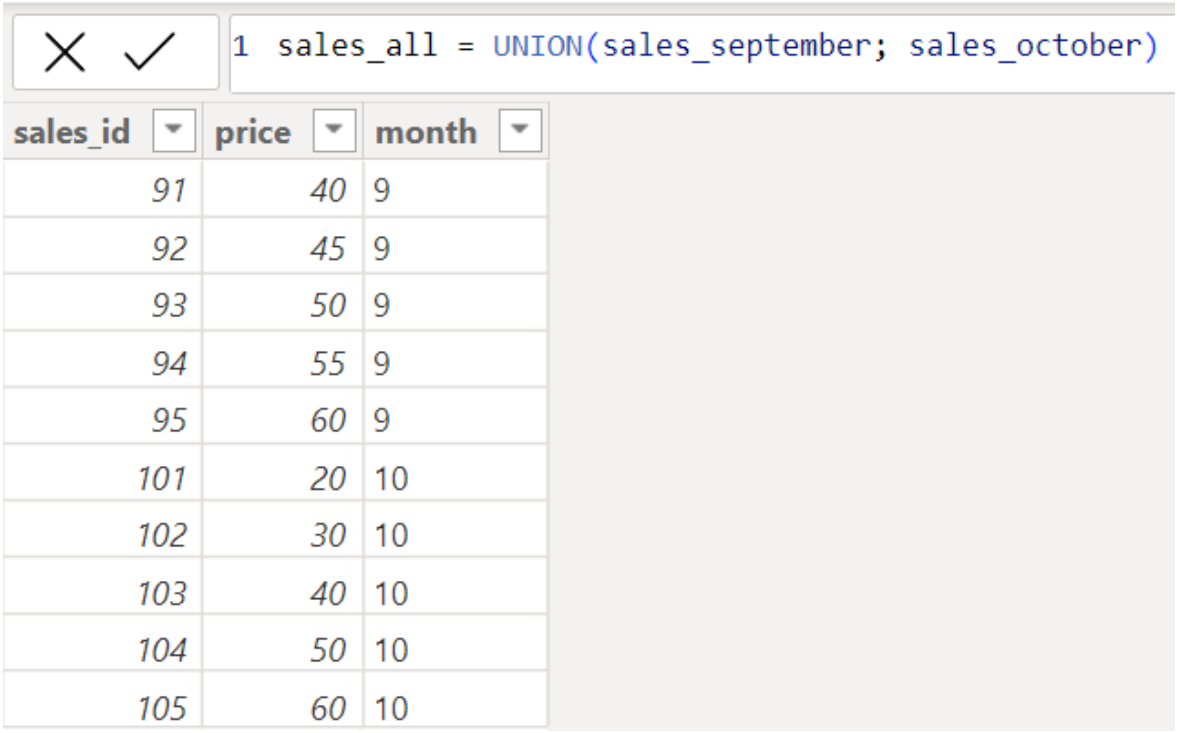
Na druhé straně Power BI nenabízí funkcionalitu vyloženě pro spojování tabulek tímto způsobem a je nutné zvolit alternativní cestu, kterou je jazyk DAX a použití funkce union. Pokud

chci vytvořit novou tabulku, musím tak učinit na kartě „Zobrazení tabulky“ a poté „Nová tabulka“. Poté lze napsat tento DAX příkaz a zde vidíme odlišnost mezi Tableau a Power BI.



OBR. 10 – CHYBA U UNION OPERACE V POWER BI (ZDROJ: AUTOR)

V rámci Power BI je nutné mít stejný počet sloupců pro obě tabulky (stejně jako je tomu u SQL), což je vhodné například pokud uchovávám prodeje z každého měsíce v jiné tabulce. Například jako v příkladu níže, kdy dvě tabulky, jedna prodeje ze září, druhá prodeje z října, spojím do jedné tabulky.



sales_id	price	month
91	40	9
92	45	9
93	50	9
94	55	9
95	60	9
101	20	10
102	30	10
103	40	10
104	50	10
105	60	10

OBR. 11 – UNION OPERACE V POWER BI (ZDROJ:AUTOR)

4.3 Zpracování dat a jejich vizualizace

Z důvodu anonymizace bude jako zdroj dat pro manipulaci s daty a následnou vizualizaci použit volně dostupný zdroj s názvem. „credit_risk_customers“. Jedná se o data, která pochází z 90. let z Německa. Data reprezentují údaje o pozorování kreditního rizika klientů, kteří požádali o úvěr. Data byla používána pro predikování, zda je klient rizikový pro podepsání úvěru či nikoliv. Jedná se o CSV soubor, který má 21 sloupců a 1 000 řádků.[36]

Cílem je datovou sadu transformovat takovým způsobem, aby byla připravena pro následnou vizualizaci pomocí Power BI a Tableau. Pro účely transformace, manipulace a čištění dat je primárně použita knihovna Pythonu Pandas, dále také Seaborn. Obsahem transformace je nalezení NaN polí, což jsou pole postrádající hodnotu, nalezení a odstranění duplicitních hodnot. Přejmenování, vytvoření a případně odstranění nerelevantních sloupců, případně jednotlivých hodnot, změna datových typů k posílení efektivity a další. Jako IDE (vývojové prostředí) používám Visual Studio Code.

4.3.1 Transformace datové sady

Prvním krokem, který je vidět v první buňce, je načtení knihoven, které budou pro transformaci použity. V tomto případě se jedná o Pandas a Seaborn, kterým jsou následně přiřazeny aliasy (pojmenování). V druhé buňce následuje načtení datové sady pomocí funkce *pd.read_csv()* a stanovení delimiteru, kterými jsou v CSV dokumentu odděleny hodnoty. Datová sada byla přiřazena do proměnné *df*. Pomocí funkcí *set_option()* lze specifikovat různé možnosti zobrazení dat. V mém případě jsem specifikoval zobrazení všech sloupců tabulky ve výpisu, u řádků zobrazení 50 řádků a u desetinných čísel zaokrouhlení na dvě desetinná místa.

```
# Importování knihoven
import pandas as pd
import seaborn as sns
```
✓ 5.3s

```
# Načtění dat a nastevení volitelných parametrů
df = pd.read_csv("credit_customers.csv", delimiter =";")
pd.set_option("display.max.columns", None)
pd.set_option("display.max.rows", 50)
pd.set_option("display.float_format", lambda x: "%.2f" % x)
```
✓ 0.0s

OBR. 12 – IMPORT PYTHON KNIHOVEN, NAČTENÍ DATOVÉ SADY A NASTAVENÍ FORMÁTU TABULKY (ZDROJ: AUTOR)

Druhým krokem je prvotní pohled na data pro pochopení struktury dat. Například hned v prvním sloupci s názvem „checking_status“ se vyskytují hodnoty, které jsou špatně čitelné (např. „0<=X<200“) a bude třeba přemapování těchto hodnot.

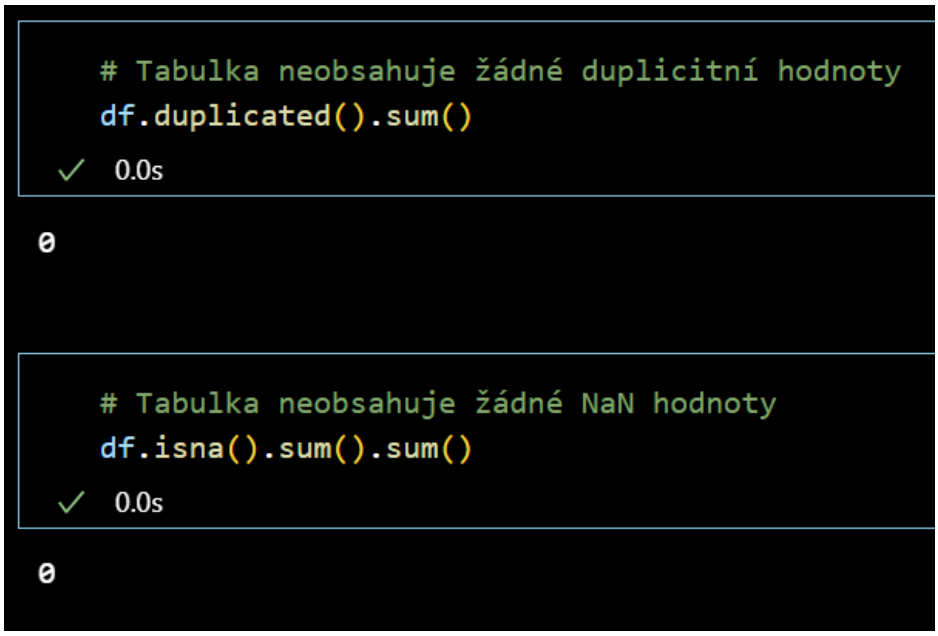
```
# První pohled na data
df
```
✓ 0.0s

	checking_status	duration	credit_history	purpose	credit_amount	savings_status	employment
0	<0	6.00	critical/other existing credit	radio/tv	1169.00	no known savings	>=7
1	0<=X<200	48.00	existing paid	radio/tv	5951.00	<100	1<=X<4
2	no checking	12.00	critical/other existing credit	education	2096.00	<100	4<=X<7
3	<0	42.00	existing paid	furniture/equipment	7882.00	<100	4<=X<7
4	<0	24.00	delayed previously	new car	4870.00	<100	1<=X<4
...	...	...	...	...	...	...	...
995	no checking	12.00	existing paid	furniture/equipment	1736.00	<100	4<=X<7
996	<0	30.00	existing paid	used car	3857.00	<100	1<=X<4
997	no checking	12.00	existing paid	radio/tv	804.00	<100	>=7
998	<0	45.00	existing paid	radio/tv	1845.00	<100	1<=X<4
999	0<=X<200	45.00	critical/other existing credit	used car	4576.00	100<=X<500	unemployed

1000 rows × 21 columns

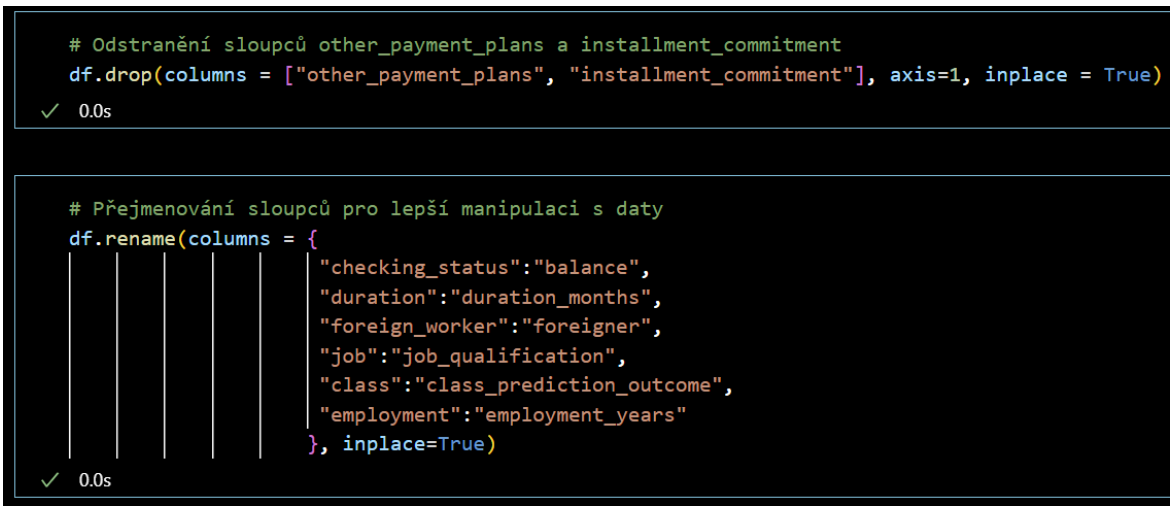
OBR. 13 – PRVOTNÍ POHLED NA DATOVOU SADU (ZDROJ: AUTOR)

Nejprve však následuje prozkoumání duplicitních a NaN hodnot. Tabulka však takovými hodnotami nedisponuje, což lze vidět na obrázku níže.



OBR. 14 – VYHLEDÁVÁNÍ DUPLICITNÍCH A NAN HODNOT POMOCÍ PANDAS (ZDROJ: AUTOR)

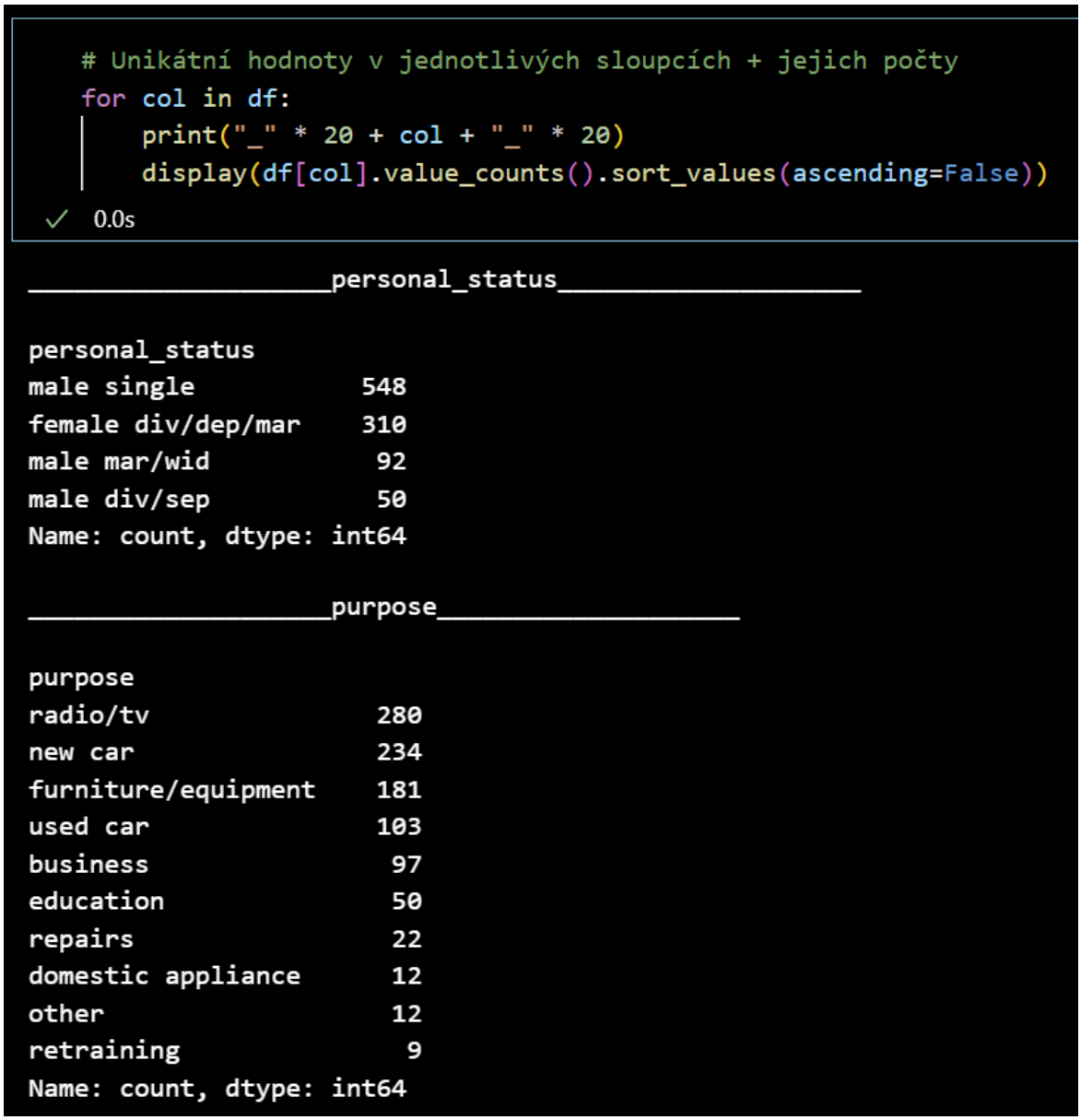
Pro lepší manipulaci s daty jsem se rozhodl pro odstranění sloupců „other_payment_plans“ a „installment_commitment“ a dále pro přejmenování šesti sloupců, aby byl lépe zachován význam uchovávaných dat:



OBR. 15 – ODSTRANĚNÍ A PŘEJMENOVÁNÍ SLOUPCŮ V PANDAS (ZDROJ: AUTOR)

Po prvotním naformátování tabulky je mým dalším cílem úprava hodnot v jednotlivých sloupcích. Cílem je opět lepší přehlednost a porozumění datům. Taktéž v této části upravuji datové typy jednotlivých sloupců.

Pro pohled na všechny typy hodnot a jejich četnosti jsem se rozhodl použít cyklus *for loop*, který iteruje přes pole *df.columns* a následně jednotlivé hodnoty přiřadí do df[], nad čímž se následně zavolá funkce *.value_counts().* Mezi jednotlivé iterace vkládám název proměnné *col*, která nese název sloupce obklopen středníky pro lepší orientaci ve výstupu. Část vrácených hodnot je vidět na obr. 16.



OBR. 16 – ZOBRAZENÍ ČETNOSTÍ JEDNOTLIVÝCH HODNOT VŠECH SLOUPCŮ POMOCÍ FOR LOOP V PANDAS (ZDROJ: AUTOR)

V případě obrázku výše je vidět, že sloupec „personal_status“ obsahuje nepřehledná data. Nejedná se pouze o osobní status, ale také o pohlaví a z tohoto důvodu jsem se rozhodl sloupec rozdělit a na základě delimiteru mezery a levá část bude použita pro nový sloupec s názvem „sex“. Jelikož zbývající pravá část bude po rozdělení obsahovat hodnoty, které nejsou jednoznačné (např. nedokážou určit rozdíl mezi hodnotami „div/sep/mar“, „mar/wid“ a „div/sep“), rozhodl jsem se pro smazání sloupce.

```
# Vytvoření sloupce sex a personal status ze sloupce personal status a následné ostranění sloupce personal status
df[["sex", "personal_status"]] = df["personal_status"].str.split(" ", expand = True)
df.drop(columns = "personal_status", axis = 1, inplace = True)

df.sex = df.sex.astype("category")
```
✓ 0.0s

OBR. 17 – VYTVOŘENÍ NOVÉHO SLOUPCE POMOCÍ .SPLIT() A NÁSLEDNÉ ODSTRANĚNÍ SLOUPCE V PANDAS (ZDROJ: AUTOR)

Úpravami projde valná většina sloupců, včetně sloupce „purpose“, jehož aktuální hodnoty viz obr. 15. Jedná se o kategorický sloupec a na místo stávajících 10 kategorií jsem se rozhodl vytvořit pouze 5 hodnot, tedy 5 kategorií. Například hodnoty „new car“ a „used car“ jsem zredukoval do jedné s názvem „car“, poté „education“ a „retraining“ do kategorie „self development“. Také typ sloupce byl změněn z „object“ na „category“, jak je vidět na obrázku níže.

```
# Vytvořeny nové kategorie ve sloupci purpose. Z původních 10 na 5.
df["purpose"] = df["purpose"].map(
    {
    "radio/tv":"consumer goods",
    "education":"self development",
    "furniture/equipment":"consumer goods",
    "new car":"car",
    "used car":"car",
    "domestic appliance":"consumer goods",
    "repairs":"other",
    "retraining":"self development",
    "other":"other",
    "business":"business"
    }
)
df.purpose = df.purpose.astype("category")
```
✓ 0.0s

OBR. 18 – VYTVOŘENÍ NOVÝCH HODNOT KATEGORICKÉHO SLOUPCE POMOCÍ .MAP() FUNKCE A ZMĚNA DATOVÉHO TYPU SLOUPCE V PANDAS (ZDROJ: AUTOR)

Podobné změny jsem taktéž učinil také u sloupců, kde to bylo potřeba. Výsledkem tabulky je nyní následující struktura a stačí už jen soubor vyexportovat jako CSV soubor, který následně bude používán jako datový zdroj pro BI nástroje.

Název sloupce	Popis sloupce	Hodnoty
balance	Množství prostředků na účtu (€)	unknown, negative, low, high
duration_months	Délka splatnosti žádaného úvěru v měsících	4 - 72
credit_history	Úvěrová historie klienta	existing paird, existing credit, previosly delayed, all paid, no credits
purpose	Účel žádaného úvěru	consumer goods, car, business, self development, other

credit_amount	Výše žádaného úvěru (€)	250-18424
savings_status	Úspory klienta na spořícím účetu nebo v dluhopisech	unknown, very low, low, medium, high
employment_years	Jak dlouho je již klient ekononmicky aktivní	less than 1, 1-3, 4-7, 7+
other_parties	Další strany zahrnuté v případě žádosti	none, guarantor, co applicant
residence_since	Kolik let klient žije v zemi	1, 2, 3, 4
property_magnitude	Majetek klienta pro krytí úvěru	car, real estate, life insurance, no known property
age	Věk klienta	19-75
housing	Bydlení klienta	own, rent, for free
existing_credits	Počet úvěrů klienta	1, 2, 3, 4
job_qualification	Druh zaměstnání	skilled, unskilled, highly qualified, unemployed
num_dependents	Počet rodinných příslušníků žijících s klientem	1, 2
own_telephone	Vlastní klient telefon?	Y, N
foreigner	Je klient cizinec?	Y, N
class_prediction_outcome	Predikované pole. Má klient nárok na úvěr?	good, bad
sex	Pohlaví klienta	male, female

TAB. 2 – ZÁKLADNÍ INFORMACE O VYEXPORTOVANÉ DATOVÉ SADĚ V PANDAS (ZDROJ: AUTOR)

Ještě před tím, než se přesuneme k samotným vizualizacím, Pandas nám nabízí i skvělé možnosti pro prvotní vizualizace a pochopení dat. Například pomocí funkce *.describe()* lze zobrazit základní statistické informace o numerických sloupcích. Jedná se o informace jako počet hodnot, průměr, směrodatná odchylka, minimum, dolní kvartil, medián, horní kvartil a maximum.

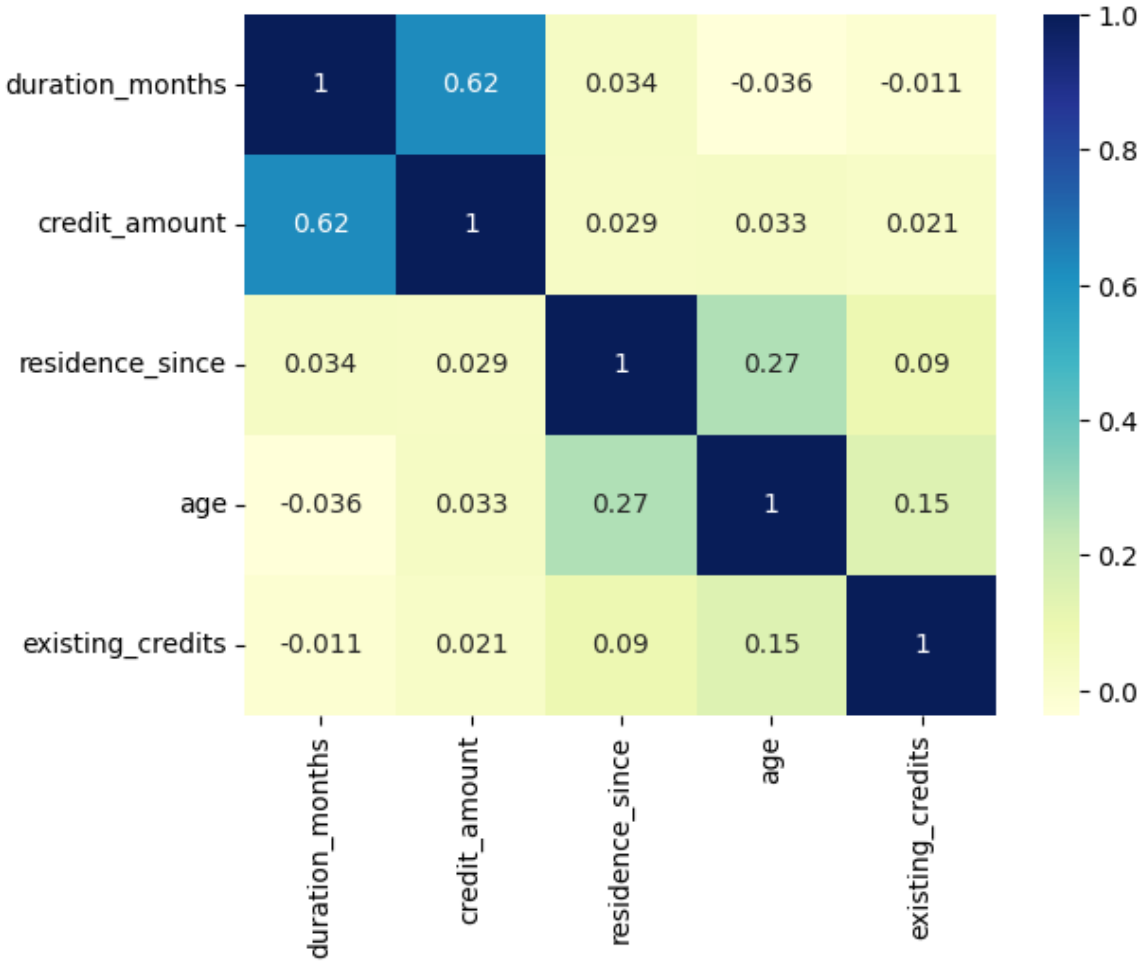
	duration_months	credit_amount	residence_since	age	existing_credits	num_dependents
count	1000.00	1000.00	1000.00	1000.00	1000.00	1000.00
mean	20.90	3271.26	2.85	35.55	1.41	1.16
std	12.06	2822.74	1.10	11.38	0.58	0.36
min	4.00	250.00	1.00	19.00	1.00	1.00
25%	12.00	1365.50	2.00	27.00	1.00	1.00
50%	18.00	2319.50	3.00	33.00	1.00	1.00
75%	24.00	3972.25	4.00	42.00	2.00	1.00
max	72.00	18424.00	4.00	75.00	4.00	2.00

OBR.19 – ZOBRAZENÍ STATISTICKÝCH HODNOT NUMERICKÝCH SLOUPCŮ POMOCÍ .DESCRIBE() FUNKCE V PANDAS (ZDROJ: AUTOR)

Další možností je získat prvotní pochopení dat pomocí vizualizací. Velmi mocným nástrojem je v tomto případě knihovna Seaborn a Matplotlib. V mém případě jsem se rozhodl pro vytvoření grafu korelace mezi numerickými sloupci pomocí následujícího kódu.

```
# Vytvoření korelační heatmap pomocí Seaborn
sns.heatmap(df.loc[:, df.dtypes == "int64"].corr(), cmap="YlGnBu", annot=True)
```
✓ 0.2s

OBR. 20 – VYTVOŘENÍ KORELAČNÍHO GRAFU NUMERICKÝCH SLOUPCŮ V SEABORN (ZDROJ: AUTOR)



OBR. 21 – KORELAČNÍ MAPA NUMERICKÝCH SLOUPCŮ V SEABORN (ZDROJ: AUTOR)

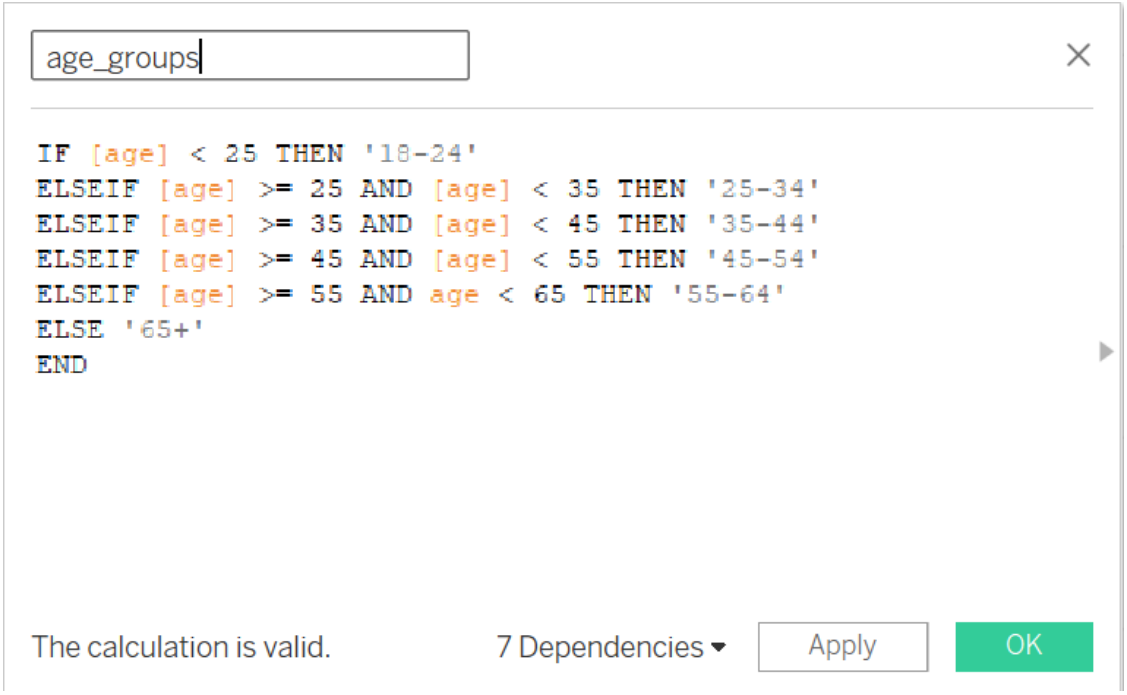
Nejvyšší korelaci lze pozorovat v rámci hodnot duration_months a credit_amount, což je také velmi logické, jelikož je pravděpodobné, že klienti s vyšším požadovaným množstvím půjčených prostředků budou úvěr déle splácet. Na druhé straně mezi určitými hodnotami korelace téměř neexistuje, například u mezi sloupci existing_credits a duration_months, ačkoliv bych očekával, že klient s vyšším počtem aktuálních úvěrů bude chtít nový, požadovaný úvěr splácet déle, očividně tomu tak není.

4.3.2 Vizualizace dat

V rámci práce a vizualizace dat pomocí BI nástrojů se zaměřím na vytvoření nového kategorického sloupce v obou nástrojích a následné vytvoření tří vizuálů, jimiž jsou sloupcový graf, prstencový graf a bodový graf.

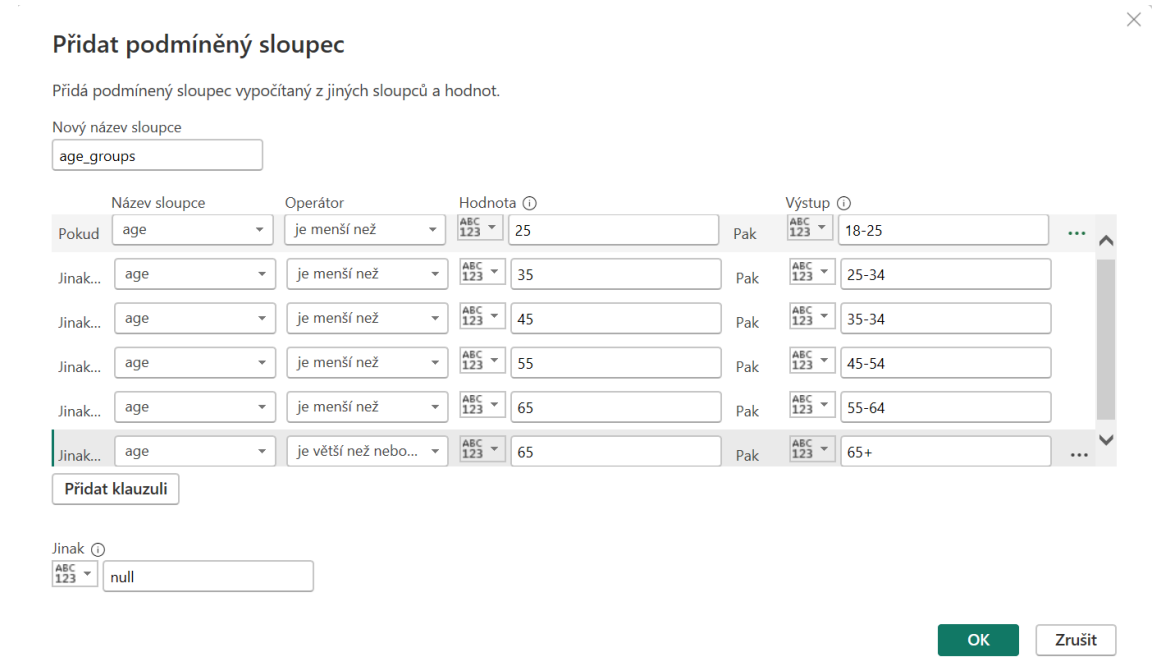
Jako první graf jsem zvolil graf sloupcový a mým cílem je vizualizovat počet klientů žádajících o úvěr za jednotlivé věkové kategorie, avšak pro tento případ ještě není kategorický sloupec věku vytvořen. V rámci Tableau je možné nový kategorický sloupec vytvořit pomocí „Create

Calculated Field“ v rámci jednotlivých sheetů. Pomocí podmínky if – else jsem vytvořil kategorické pole o šesti hodnotách s názvem „Age Groups“.



OBR. 22 – VYTVOŘENÍ KATEGORICKÉHO POLE V TABLEAU (ZDROJ: AUTOR)

Vytvoření Age Groups následuje také v Power BI, kde pro vytvoření existuje vícero možností. Jednou z nich je vytvoření sloupce pomocí jazyku DAX, avšak já jsem zvolil alternativní cestu, kterou je „Přidání podmíněného sloupce“, kde stejně jako v Tableau, lze stanovit pomocí podmínek a operátorů věkové kategorie.

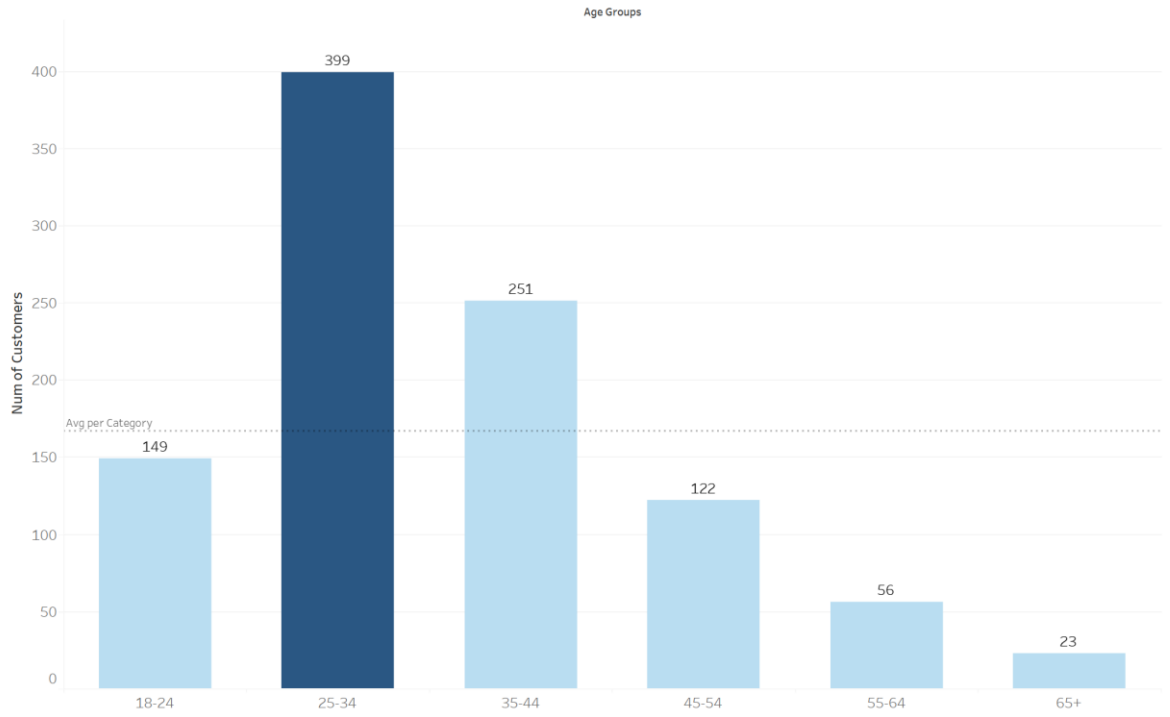


OBR. 23 – VYTVOŘENÍ KATEGORICKÉHO POLE V POWER BI (ZDROJ: AUTOR)

Cílem pro vytvoření sloupcového grafu je na ose x vizualizovat šest kategorií ze sloupce age_groups a na ose y počet jednotlivých klientů žádajících o úvěr, dále zobrazení popisků

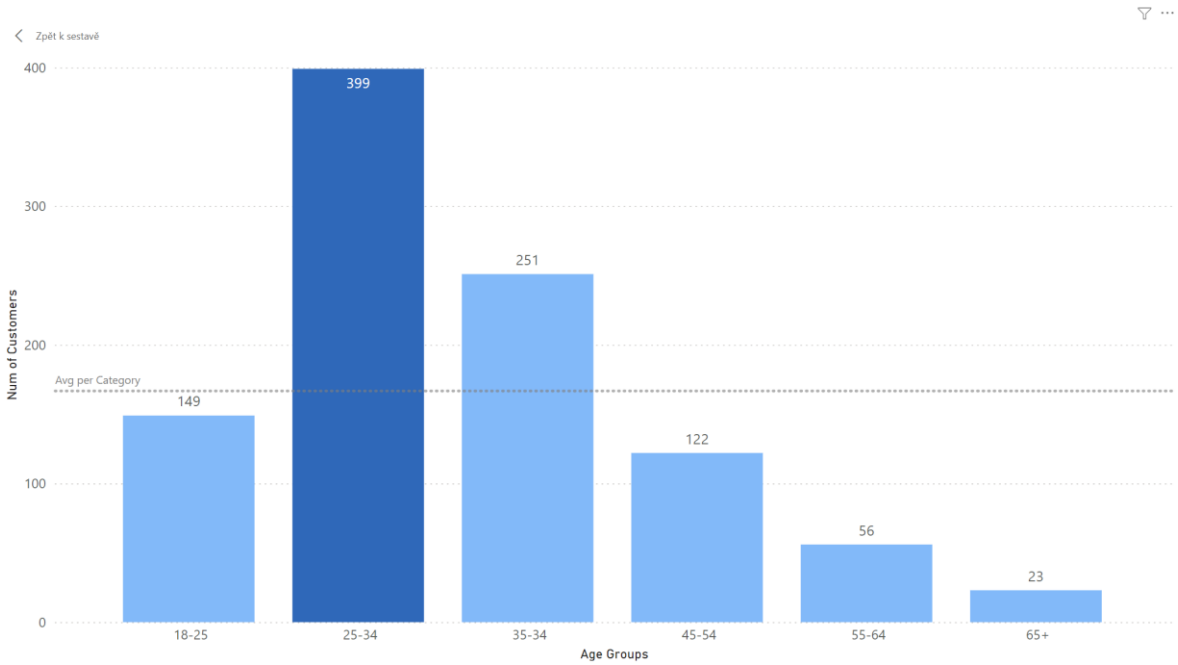
jednotlivých sloupců, průměrnou hodnotu počtu klientů za kategorii a zvýraznění sloupce/kategorie, pod kterou spadá nejvíce klientů.

Sloupcový graf v Tableau dle zadání:



OBR. 24 – VYTVOŘENÍ SLOUPCOVÉHO GRAFU V TABLEAU (ZDROJ: AUTOR)

Sloupcový graf v Power BI dle zadání:



OBR. 25 – VYTVOŘENÍ SLOUPCOVÉHO GRAFU V POWER BI (ZDROJ: AUTOR)

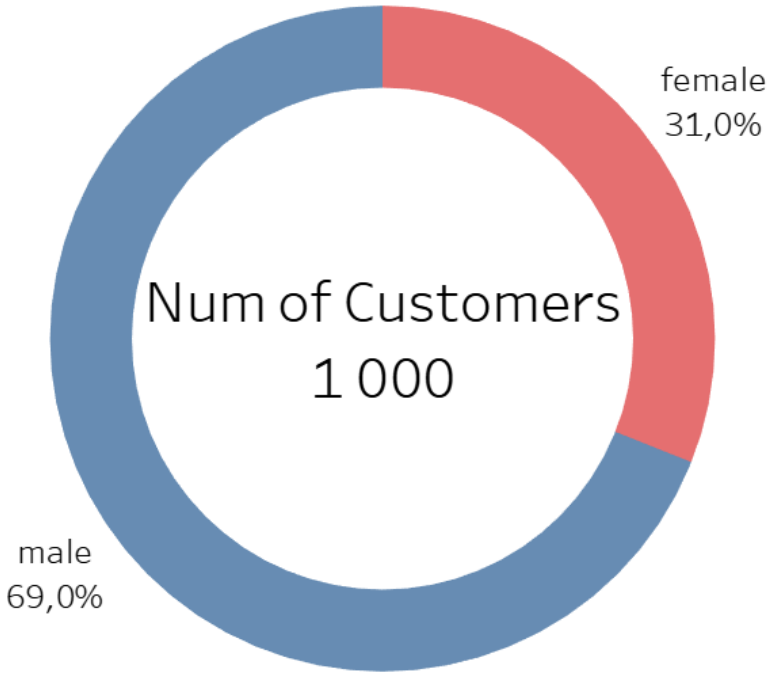
Oba dva vizuály vypadají velmi podobně, avšak samotné vytváření vizuálů neslo určité odlišnosti. Při vytváření čáry průměru v Power BI nešlo vytvořit celkového počtu za jednotlivé kategorie, která ve vizualizacích nese název „Avg per Category“. Avšak i tento nedostatek

funkcionality bylo jednoduché obejít. Pouze stačilo pomocí DAX vytvořit nový sloupec, který byl vypočten jako podíl celkové počtu zájemců a počtu unikátních hodnot ve sloupci „age_groups“. Poté již stačilo nově kalkulovaný sloupec s názvem „avg_count_per_category“ vložit jako pole pro základ při vytváření hodnoty a z pole vypočítat průměr.

OBR. 26 – VYTVOŘENÍ ČÁRY PRŮMĚRU V POWER BI (ZDROJ: AUTOR)

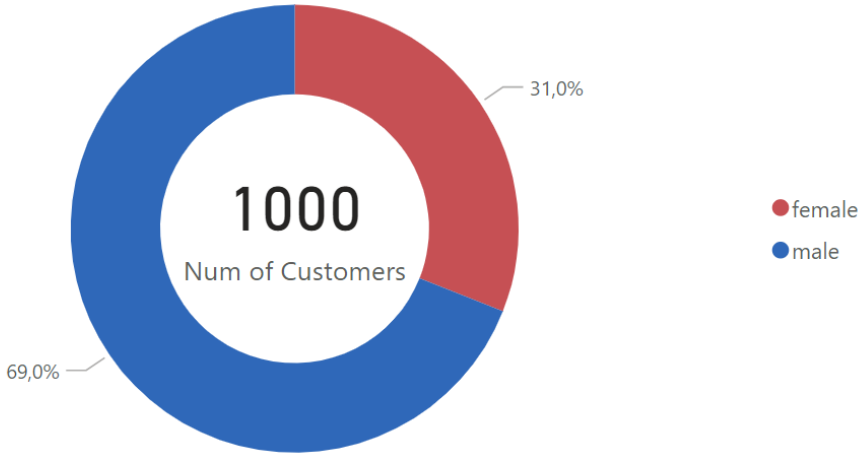
Jako cíl druhého vizuálu jsem vybral prstencový graf a jelikož není záhodné vytvářet koláčové či prstencové grafy pro kategorické sloupce s velkým množstvím unikátních hodnot, pro sloupce jako v případě pohlaví, které obsahuje dvě unikátní hodnoty, dává tento typ vizuálu smysl.

Prstencový graf dle zadání v Tableau:



OBR. 27 – VYTVOŘENÍ PRSTENCOVÉHO GRAFU V TABLEAU (ZDROJ: AUTOR)

Prstencový graf dle zadání v Power BI:



OBR. 28 – VYTVOŘENÍ PRSTENCOVÉHO GRAFU V POWER BI (ZDROJ: AUTOR)

V případě prstencových grafů byl postup vizualizace u obou nástrojů odlišný. Tableau dokonce vizualizaci prstencového grafu nenabízí, avšak existuje cesta, jak k grafu dojít. Jedná se o dva koláčové grafy vnořené do sebe. Vnitřní koláčový graf je pouze bílé pozadí, do kterého je vložen tooltip (text) s počtem klientů a textem „Num of Customers“. Vnitřní koláčový graf je též umístěn do popředí před vnější koláčový graf, tedy tu barevnou část grafu.

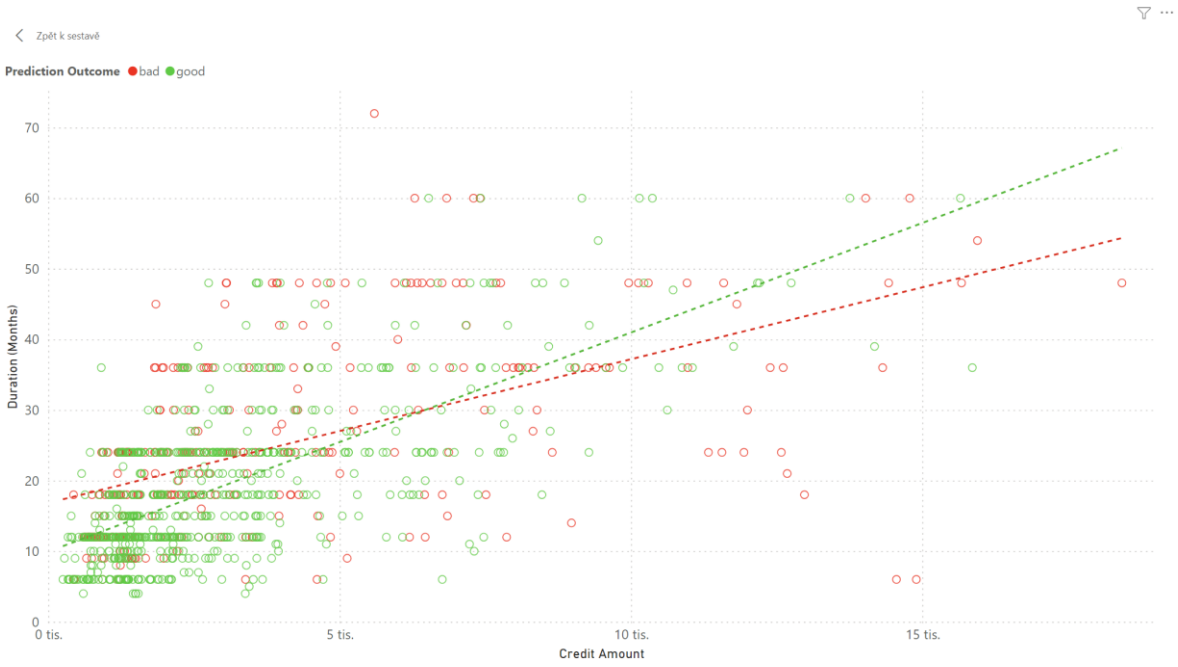
Na druhé straně Power BI nabízí vytvoření prstencového grafu, avšak dle mého názoru obsahuje značné mezery, jelikož i tak nenabízí vyplnění vnitřního prostoru prstenu. V tomto případě však nebylo nutné používat koláčový graf, stačilo pouze použít vizuál karty a do něj vložit počet záznamů společně s textem a následně ve formátování vizuálu zvýšit průhlednost na 100 %, tedy aby byl vidět pouze text a nikoliv pozadí karty.

Posledním vytvářeným vizuálem je bodový graf s vidinou vizualizace trendu. Na ose x bude zobrazena hodnota úvěru, o kterou klient žádá a na ose y splatnost úvěru v měsících. Mým cílem také je rozdělit body do dvou kategorií podle pole „class_prediction_outcome“, tedy zda byl klient v rámci predikce určen jako rizikový či nikoliv společně s vývojem trendu za obě dvě kategorie.

Bodový graf dle zadání v Tableau:

OBR. 29 – VYTVOŘENÍ BODOVÉHO GRAFU V TABLEAU (ZDROJ: AUTOR)

Bodový graf dle zadání v Power BI:



OBR. 30 – VYTVOŘENÍ BODOVÉHO GRAFU V POWER BI (ZDROJ: AUTOR)

V rámci vizualizací jsou zobrazeny dvě trendové linie, zelená pro klienty, kteří jsou predikováni jako vhodní adepti pro úvěr a červená linie pro ty, kteří byli vyhodnoceni jako rizikoví. Dle jednotlivých trendových linií lze dojít k závěru, že klienti kteří volili nižší částku úvěru s delší splatností byli spíše vyhodnoceni jako vhodní adepti.

Pokud bychom tyto tři vizuály chtěly zobrazit v rámci dashboardu, v Tableau je nutné dashboard vytvořit a vizuály do dashboardu zapracovat. U Power BI jsou všechny vizuály již v rámci dashboardu vytvářeny a není třeba dále cokoliv upravovat. Na dashboardech je také velmi obvyklé filtrování dat nejen na základě vybraných filtrů, ale také na základě již vytvořených vizuálů. Pokud

bychom v případě prstencového grafu kliknuly na část pohlaví, která reprezentuje ženy, všechny zbývající vizuály by taktéž byly filtrovány pouze pro hodnoty žen.

4.4 Závěr

Oba nástroje mají své výhody i nevýhody a nelze jeden postavit před druhý. V rámci technických aspektů probraných v této kapitole, žádné zásadní rozdíly mezi oběma nástroji nejsou, pokud pomineme absenci směru křížového filtru či anti spojení tabulek u Tableau. Hlavní rozdíly tkví v rámci netechnických aspektů. Zde má Power BI jednoznačně navrch v rámci ceny. Nejvybavenější verze Tableau je zhruba čtyřikrát dražší oproti Power BI a věřím, že toto může být rozhodující argument pro spoustu zájemců. Na straně druhé vnímám osobně jako velké negativum absenci Power BI Desktop ve výchozím nastavení na Mac OS. Nesmíme také opomenout prostředí v českém jazyce či DAX, který je opravdu hojně využívaný a dobře fungující pro transformaci dat. Tyto důvody mne přesvědčily, že právě Power BI má navrch oproti Tableau.

Co se týče jednotlivých vizuálů, ačkoliv nabídka a provedení jsou v určitých případech odlišná (jako u prstencového grafu či čáry průměru), častokrát je možné dojít k požadovanému výsledku jinou cestou, především díky vysoké flexibilitě a modifikovatelnosti u obou nástrojů. Co se týče vzhledů vizualizací, u Power BI působí decentněji a čistším dojmem.

5 Přínos týmu datové kvality pro společnost

Cílem společnosti X je držet datovou kvalitu co nejvýše, ačkoliv 100% kvalita dat je spíše utopická než realistická myšlenka. A jak se dají nekvalitní data rozpoznat od těch kvalitních? Jedná se o duplicitní záznamy, chybějící hodnoty, nevalidní hodnoty. Konkrétním příkladem tedy může být klient, který má v databázi dva totožné záznamy, nemá vyplněné jméno a narodil se v neexistujícím státě. Cílem datové kvality je chybám předcházet, případně chyby odhalovat a následně řešit.

K pochopení, jakým způsobem funguje datová kvalita v rámci organizace X, je třeba nastínit procesy, které za měřením datové kvality stojí a patří mezi ně:

- prevence,
- měření datové kvality,
- oprava nekvalitních dat.[37]

Prevencí se rozumí snaha o eliminaci problémů s datovou kvalitou. Například zavádění masek v systémech, úprava procesů pro eliminaci vzniku chyb nebo identifikace možného místa, kde má smysl data kontrolovat. Jinak řečeno implementovat kontroly datové kvality na místech, kde k chybám nemůže docházet postrádá smysl měření datové kvality.

Oprava nekvalitních dat je sama o sobě vypovídající. Jakmile pomocí měření datové kvality zjistíme, že data nesplňují požadovanou kvalitu, započnou opravy. Nejprve se identifikuje místo, kde se vyskytují nekvalitní data, následně se problém vzniku analyzuje a dohledá příčina vzniku a následně se příčina opraví, aby již chybná data nevznikala.

V rámci kapitoly 5 se však budu zaměřovat na prostřední bod – „Měření datové kvality“.

5.1 Měření datové kvality

Samotnými nástroji měření kvality dat jsou DQI kontroly. Ty se po fázi prevence vyvinou, následně otestují a pokud během tohoto procesu nevzniknou žádné problémy, tak se nasadí.

5.1.1 DQI

DQI (Data Quality Indicators), neboli indikátory datové kvality, jsou v organizaci X základními ukazateli, jež udávají přehled o datové kvalitě. DQI jsou tvořeny formou dotazovacích SQL scriptů, které se dotazují do databáze. Účelem DQI je pravidelné monitorování a hodnocení dat, které jsou v databázích uloženy.

5.1.2 Ukázka DQI

Každé DQI musí být správně popsané, včetně určení, nad jakými přesně záznamy se kontrola spouští, jaké z těchto kontrolovaných záznamů budou označeny jako chybné a do jaké kategorie toto DQI spadá.

V této podkapitole si ukážeme kontrolu datové kvality (DQI). Jedná se o kontrolu nad klientskými daty, konkrétně nad fyzickými osobami, jejímž cílem je měřit, jaké procento fyzických klientů má validně vyplněné rodné číslo (což by měli mít všichni klienti).

Také je nutné podotknout, že DQI je pouze ukázkové, tím pádem zadání DQI, SQL kód i výsledky jsou smyšlené a slouží pouze k nastínění, jakým způsobem DQI v rámci organizace fungují.

Tímto způsobem by tedy mohlo vypadat zadání DQI:

DQI 123	
Popis DQI	Fyzické osoby musí mít validní rodné číslo.
Kontrolované záznamy	Všechny fyzické osoby
Chybové záznamy	Fyzické osoby s nevalidním rodným číslem
Kategorie DQI	Klientská data

TAB. 3 – UKÁZKA ZADÁNÍ DQI (ZDROJ: AUTOR)

- **ID DQI** označuje univerzální identifikátor DQI (v tomto případě 123)
- **Popis DQI** udává, o čem kontrola pojednává.
- **Kontrolované záznamy** uvádí bázi této kontroly
- **Chybové záznamy** označují ty záznamy z báze, které jsou chybové
- **Kategorie DQI** definuje, pod jakou kategorii DQI spadá

A jak se pozná validní rodné číslo u mužů s českým občanstvím? Správně vyplněné rodné číslo musí:

- obsahovat 9 nebo 10 znaků
- obsahovat pouze numerické znaky
- mít pro klienty narozené před rokem 1954 přesně 9 znaků
- mít pro klienty narozené po roce 1953 přesně 10 znaků
- mít 5. a 6. číslici v rozsahu 1-31
- mít 3. a 4. číslici v rozsahu 1-12
- být dělitelné číslicí 11 beze zbytku
- být vyplněné

Jakmile rodné číslo nesplní minimálně jeden z těchto požadavků, jedná se o nevalidní záznam, jinými slovy se bude jednat o nekvalitní data, jelikož porušují logiku, kterou rodná čísla musí splňovat. Pro ženy a případně lidi, které se nenarodili na českém území platí trochu jiná pravidla a z toho důvodu jsem je z této ukázkové kontroly vyřadil, aby byl SQL script přehlednější.

Poté, co je DQI definováno, vytváří se implementace v SQL, která je náplní této kontroly a mohla by vypadat následovně:

```sql
/* ukázka DQI_123 */
SELECT
    customer_id id_klienta,
    first_name jmeno,
    personal_number rodne_cislo,
    chyba,
    CASE
        WHEN chyba=1 THEN 'RČ nemá 9 nebo 10 znaků'
        WHEN chyba=2 THEN 'RČ neobsahuje pouze numerické znaky'
        WHEN chyba=3 THEN 'RČ nemá 9 znaků pro klienty narozené před rokem
1954'
        WHEN chyba=4 THEN 'RČ nemá 10 znaků pro klienty narozené po roce 1953'
        WHEN chyba=5 THEN '5. a 6. pozice RČ není v rozsahu 1-31'
        WHEN chyba=6 THEN '3. a 4. pozice RČ není v rozsahu 1-12'
        WHEN chyba=7 THEN 'RČ není dělitelné číslem 11 beze zbytku'
        WHEN chyba=8 THEN 'RČ není vyplněno'
    END popis_chyby
FROM (
    SELECT
        customer_id,
        first_name,
        personal_number,
        natural_person,
        CASE
            WHEN LENGTH(personal_number) NOT IN (9,10) THEN 1
            WHEN
            LENGTH(personal_number)=9 AND TRANSLATE(personal_number,
'0123456789', '##########') != '#########'
                OR
            LENGTH(personal_number)=10 AND TRANSLATE(personal_number,
'0123456789', '##########') != '##########' THEN 2
            WHEN birth_year<=1953 AND LENGTH(personal_number)!=9 THEN 3
            WHEN birth_year>=1954 AND LENGTH(personal_number)!=10 THEN 4
            WHEN SUBSTR(personal_number, 5, 2) NOT BETWEEN 1 AND 31 THEN 5
            WHEN SUBSTR(personal_number, 3, 2) NOT BETWEEN 1 AND 12 THEN 6
            WHEN MOD(personal_number, 11) != 0 THEN 7
            WHEN personal_number IS NULL THEN 8
            ELSE NULL
        END chyba
    FROM customers
WHERE 1=1
    AND natural_person = 'yes'
    AND citizenship = 'cz'
    AND sex = 'male')
```

Výsledkem SQL by mohla být tabulka podobná této (jedná se o smyšlené hodnoty).

	ID_KLIENTA	JMENO	RODNE_CISLO	CHYBA	POPIS_CHYBY
1	1	Albert	980504123457	1	RČ nemá 9 nebo 10 znaků
2	2	Bert	001215ABCD	2	RČ neobsahuje pouze numerické znaky
3	3	Cyril	4805464888	3	RČ nemá 9 znaků pro klienty narozené před rokem 1954
4	4	Damián	010523548	4	RČ nemá 10 znaků pro klienty narozené po roce 1953
5	5	Erik	0113353258	5	5. a 6. pozice RČ není v rozsahu 1-31
6	6	Filip	0015207891	6	3. a 4. pozice RČ není v rozsahu 1-12
7	7	Gabriel	0011251234	7	RČ není dělitelné číslem 11 beze zbytku
8	8	Hynek	(null)	8	RČ není vyplněno
9	9	Chrudoš	390501452	(null)	(null)

OBR. 31 – VÝSLEDEK UKÁZKOVÉHO SQL KÓDU (ZDROJ: AUTOR)

Nyní, když víme celkový počet záznamů a počet chyb, můžeme jednoduše spočítat, jaké procento z báze je chybové. V tabulce výše je vidět, že pouze Chrudoš má validní rodné číslo (jelikož ve sloupce CHYBA se vyskytuje *null* hodnota). Také víme, že z 9 záznamů je 8 chybových, proto úroveň chybovosti činí cca 88,89 %.

Zároveň je v této SQL kontrole definováno 8 podmínek, kdy rodné číslo není validní a je možné, že nesplňuje vícero z daných podmínek, avšak označeno jako chyba bude pouze jednou a to při první podmínce, kterou poruší. Například RČ Alberta, které má hodnotu '980504123457' nesplňuje podmínek hned několik. Kromě chyby 1 – RČ nemá 9 nebo 10 znaků, také nesplňuje 4. podmínku, jelikož Albert byl narozen roku 1998 a jeho rodné číslo nemá 10 znaků a poté by také neprošlo 7. podmínkou, tedy dělitelností 11 beze zbytku.

Je nutné být si vědom toho, co daná DQI kontrolují. V tomto případě DQI nekontroluje, jestli má klient přiřazen správné rodné číslo, ale zda má klient přiřazen validní rodné číslo. Pokud by bylo rodné číslo zadáváno pracovníkem banky fyzicky a spletl by některé číslice, je velmi nepravděpodobné, že by byly zadány tím způsobem, aby splňovaly všechny podmínky této kontroly, avšak stát se to může. V takovém případě nelze z databáze zjistit, zda je skutečně rodné číslo přiřazeno správně či nikoliv. I v tomto případě mohou existovat další kontroly, které zjišťují, zda například první dvě číslice RČ odpovídají roku narození, další dvě čísla měsíci narození a 5. a 6. pozice dni narození, což může také odhalit chyby, které by nemusela odhalit tato kontrola. Všechny chyby jsou v tomto případě chybami, ale ne všechny správně označené záznamy odpovídají realitě.

Kontrolou tedy bylo zjištěno, že chybovost je v úrovni 88,89 %, tím pádem datová kvalita této kontroly činí cca 11,11 %. Každá kontrola vychází v určité datové kvalitě a kromě toho je pro každou kontrolu určen i práh datové kvality, pod který by neměla klesnout. Výsledky kontrol mohou tedy vypadat následovně. Pokud je kvalita nižší než prahová hodnota, označím ji v tomto případě červeně, pokud je však vyšší, označím ji zeleně. Opravy chybných dat jsou prioritní u kontrol, které vychází pod prahovou hodnotou.

ID DQI	Prahová hodnota	Kvalita	Počet záznamů	Chyby
DQI 123	99 %	11,11%	9	8
DQI 321	99 %	99,84%	54213	89

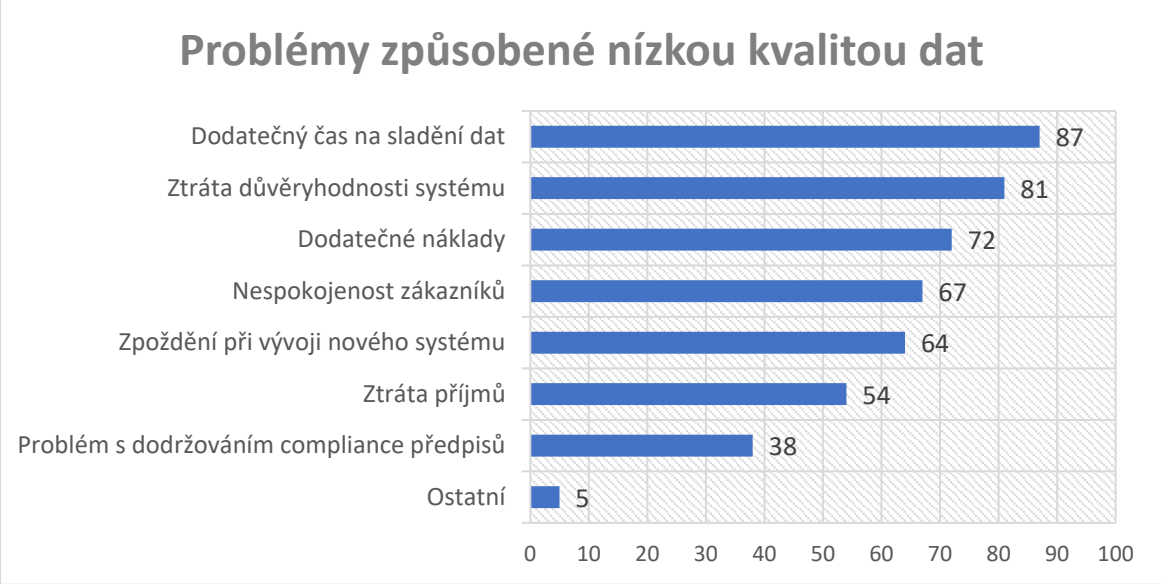
TAB. 4 – UKÁZKA VÝSLEDKŮ JEDNOTLIVÝCH DQI (ZDROJ: AUTOR)

U kontrol, které požadovanou kvalitu nesplňují se následně činí kroky, kterými se datová kvalita zlepší a zároveň se zajišťuje Root cause analysis (analýza kořenových příčin), aby se předešlo podobným problémům u budoucích dat.

5.2 Dopady nízké kvality dat

Neřešení, případně neoptimální řešení datové kvality ve společnosti s sebou nese určitá úskalí. Průzkumu TDWI ohledně datové kvality, který vedl Wayne W. Eckerson, se zúčastnilo 647 respondentů napříč všemi odvětvími.[38]

Následků, které nekvalitní data způsobují je mnoho. V následujícím grafu můžeme vidět základní problémy, které vznikají důsledkem nekvalitních dat. Výpis jednotlivých problémů je k vidění na ose y, procentuální rozložení na ose x. Nejčastějším problémem spojeným s nekvalitními daty je dodatečný čas strávený na sladění dat. S tímto problémem se setkalo 87 % respondentů.[38]



OBR. 32 – PROBLÉMY ZPŮSOBENÉ NÍZKOU KVALITOU DAT (ZDROJ: HTTP://DOWNLOAD.101COM.COM/PUB/TDWI/FILES/DQREPORT.PDF)

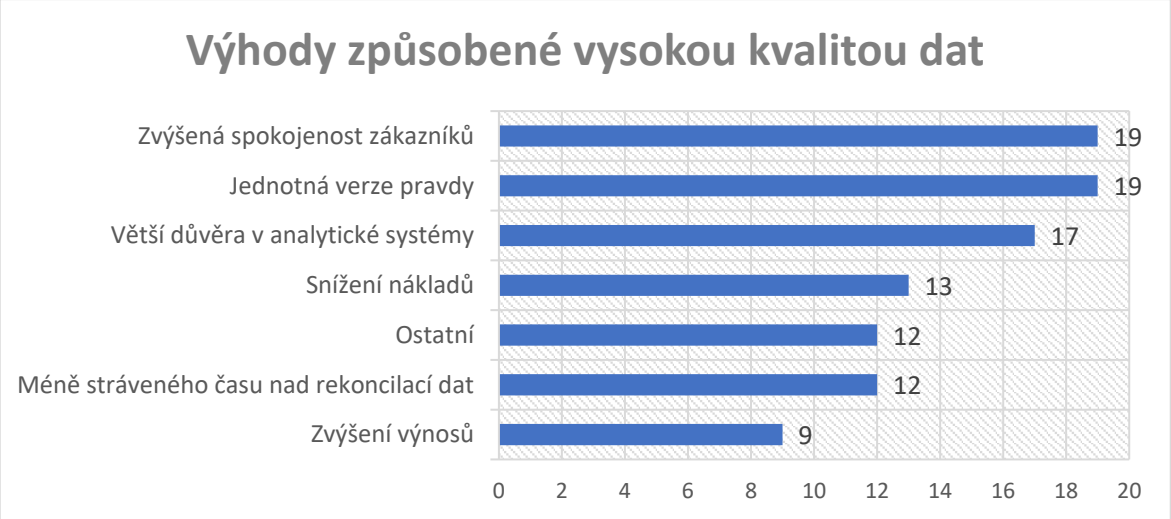
V rámci stejného průzkumu také 48 % společností konstatovalo, že datová kvalita dané společnosti je dobrá, popřípadě excelentní (pouze 7 % odpovědí reprezentovalo špatnou datovou kvalitu), což je samozřejmě dobré znamení. Následující otázka však navazovala na předchozí a 44 % respondentů odpovědělo, že datová kvalita je podle nich horší, než si myslí. Je zřejmé, že toto zjištění ukazuje na značný rozdíl mezi tím, jak společnosti datovou kvalitu vnímají a mezi skutečností.[38]

5.2.1 Důsledky nízké datové kvality

Mezi konkrétní případy, kdy nízká kvalita dat zapříčinila společnosti potíže, patří známá americká banka JPMorgan Chase, která v roce 2012 reportovala ztrátu 6 miliard USD kvůli rizikovým sázkám na finanční deriváty. Hlavním důvodem tohoto incidentu byla chyba v datech v oblasti řízení rizik společnosti, které měly před podobnými riziky chránit. Ve stejném roce trpěla incidentem také skotská banka The Royal Bank of Scotland. Kvůli rozsáhlému selhání IT, které bylo zapříčiněno především nízkou kvalitou dat, stálo banku 286 milionů USD, které vynaložili jako odškodnění pro klienty a dodatečné platby zaměstnancům. Důsledkem selhání totiž bylo zablokování účtů pro miliony zákazníků, což nejen způsobilo již zmíněné problémy, ale také vážně poškodilo pověst banky. V těchto dvou případech jsou banky trestány samy sebou, ale existují i případy, kdy za nízkou kvalitu dat udělují pokuty také regulátoři[39].

5.3 Dopady vysoké kvality dat

V rámci stejného průzkumu TDWI padly také dotazy na výhody způsobené kvalitními daty. A jak je vidět na hodnotách osy x, mnohem více jsou si lidi vědomi negativ než pozitiv, které jsou s kvalitou dat spojeny. Mezi nejčastěji zmiňované výhody v průzkumu patří zvýšená spokojenost zákazníků a jednotná verze pravdy. Především ta druhá výhoda je zvláště důležitá, pokud má společnost důvěru ve svá data, jedná se o mocný nástroj, který prezentuje pravdu a lze se spolehnout na rozhodování na základě dat.[38]



OBR. 33 – VÝHODY ZPŮSOBENÉ VYSOKOU KVALITOU DAT (ZDROJ: HTTP://DOWNLOAD.101COM.COM/PUB/TDWI/FILES/DQREPORT.PDF)

Čím dříve by společnosti začaly řešit datovou kvalitu, tím lépe. Důležité je také držet na paměti, že nejde pouze o prvotní sběr dat. Eckerson také upozorňuje, že data svou kvalitu časem ztrácí. Údajně 2 procenta všech čerstvých klientských dat ztratí svou kvalitu během prvního měsíce, jelikož klienti umírají, rozvádějí se, žení se, stěhují se a podobných případů je mnoho, a ne pouze u klientských dat. V rámci datové kvality je také esenciální data udržovat aktuální s postupem času.[38]

5.4 Závěr kapitoly

V dnešním světě je již nemyslitelné, že by rozsáhlé společnosti, jako jsou banky, zanedbávaly datovou kvalitu, případně na ni neměly přímo specializované týmy. Pokud by se pro to i tak rozhodly, nízká kvalita by se postupem času rozhodně snížila, situace by neměla dlouhého trvání, jelikož by zasáhli regulátoři a banka by neměla moc na výběr, než opět začít dbát na datovou kvalitu. Já osobně vnímám jako stěžejní výhodu úsporu nákladů a času, který je v dnešní době zvlášť cenný. Jak již bylo zmíněno několikrát v rámci práce, jedním z důvodů řešení datové kvality je regulatorní reporting. Avšak nejedná se pouze o činnosti uvnitř banky, datová kvalita se propisuje i navenek společnosti, jelikož zacházení s citlivými daty klientů vyžaduje také určitou úroveň vzájemné důvěry, která by neměla být poškozena. Ačkoliv důležitost datové kvality nelze častokrát najít v podvědomí široké veřejnosti, jedná se o mocný nástroj, který hraje velkou roli v rámci chodu celé společnosti.

6 Vzdělání zaměstnanců v oblasti datové analýzy

V rámci této části se zaměřuji na technické dovednosti, především dovednosti datové analýzy, kterými by měli zaměstnanci bank disponovat. Každá pozice v bance samozřejmě představuje individuální roli a nároky jsou jiné. I z tohoto důvodu jsem udělal analýzu aktuálně vypsaných pracovních pozic a požadavky na technické dovednosti v rámci společnosti X. Avšak nejdříve je třeba porozumět dovednostem, které by datoví analytici, dle mého názoru, měli splňovat.

6.1 Dovednosti datových analytiků

Jak bylo již zmíněno v kapitole „Zařazení BI v pyramidě datové vědy“, co znamená datový analytik pro jednu společnost, neznamená datový analytik pro společnost druhou. Definice jsou velmi odlišné napříč organizacemi. V rámci této kapitoly budu popisovat čtyři technické dovednosti, které jsou dle mého názoru nejdůležitější v oblasti datové analýzy, avšak to neznamená, že všichni datoví analytici těmito znalostmi disponují.

Datová analýza je specifická v tom, že se nejedná v porovnání s datovými inženýry nebo datovými vědci o tak technickou pozici, ačkoliv průnik do IT je jednoznačný. Datoví analytici musí zastávat jak technické dovednosti (Hard skills), tak měkké dovednosti (Soft skills).

6.1.1 Hard Skills

Já osobně dělím technické dovednosti datových analytiků do čtyř kategorií:

- tabulkový editor,
- SQL,
- vizualizační nástroje,
- programovací jazyk.

Tabulkový editor (Microsoft Excel či Google Sheets) vyžaduje asi nejméně úsilí, aby pracovník s nástrojem dokázal pracovat. Obrovská výhoda těchto editorů je intuitivní ovládání a přátelské uživatelské prostředí, což však neznamená, že se jedná o nástroj, který by nedosahoval vysokých kvalit v porovnání s ostatními. Pokud se bavíme o produktu společnosti Microsoft – Excelu, našlo by se spoustu pracovních pozic datových analytiků, kde práce obnáší jen a pouze Excel, což je v dnešní době docela výjimečné, ale desítky let nazpět to bylo naprosto běžné a normální.

SQL je standardizovaný dotazovací jazyk, který je používán v relačních databázích. Umožnuje procesovat informace mnohem rychleji a efektivněji než klasické tabulkové editory. Je nazýván dotazovacím jazykem, jelikož hlavní funkcí je psaní dotazů, pomocí kterých můžeme číst data z databáze. Kromě dotazování a čtení dat umožnuje SQL také vytváření, mazání, upravování záznamů, tabulek, schémat, databází a mnoho dalšího. Datoví analytici převážně využívají čtení dat z databáze, se kterými dále pracují, není však výjimkou, že se datoví analytici zabývají i správou databázových systémů.

SQL je používáno v relačních databázích, která ukládají data ve formě sloupců a řádků, kde sloupce definují datový typ a typ hodnoty, která je uložena a jednotlivé řádky reprezentují údaje. Existují však i další způsoby ukládání dat, kde hodnoty nejsou uloženy formou řádky/sloupce, ale například formou klíč: hodnota nebo formou uložení grafických dat. Jelikož se nejedná o standardní formu ukládání dat a ani nebývá pravidlem, že by se do těchto databází dotazovalo prostřednictvím SQL, neberu znalost a dotazování k NoSQL databázím jako klíčovou znalost datové analýzy.

Mezi nástroje SQL se řadí také procedurální jazyky, které rozšiřují jazyk SQL, avšak ty jsou nazývány ve všech systémech pro správu relačních databází jinak. Například u Oracle se jedná o PL/SQL (Procedural Language/Structured Query Language), pro Microsoft SQL Server se jedná o tzv. T-SQL (Transact-SQL). Procedurální jazyky umožňují vytváření procedur, triggerů nebo funkcí, které pomáhají v centralizaci logiky databáze, zabezpečení dat nebo zvýšení výkonu.

Vizualizační nástroje viz kapitola 4 *„Business Intelligence“* .

Programovací jazyky patří mezi ty technické dovednosti, jejichž osvojení zabere nejvíce času. A otázka toho, který programovací jazyk je ten pravý, není složitá. V rámci datové analýzy se používají primárně dva, a to Python nebo R. Využívají se i jiné, jako Java, Scala nebo Matlab, ale dle mého názoru pouze výjimečně. Co se týče Pythonu, tak v rámci datové analýzy stále získává na popularitě, především díky své jednoduchosti v porovnání s ostatními programovacími jazyky. K datové analýze se využívají především knihovny jako Numpy, Pandas, Matplotlib a Seaborn.

Knihovny Matplotlib a Seaborn nabízí alternativu k tradičním vizualizačním nástrojům jako Power BI a Tableau, avšak je mezi nimi zásadní rozdíl. Oproti velmi komplexním nástrojům, kterými BI nástroje jsou, slouží knihovny Pythonu především ke statickým vizualizacím a grafům, jejichž obrovskou výhodou však je vysoká flexibilita a možnost detailního přizpůsobení grafů dle uživatele.

Knihovna Numpy je základním stavebním blokem výpočtů v jazyce Python, která podporuje práci s vektory, maticemi, lineární algebrou, statistikou a matematickými operacemi. Nejčastěji se knihovna Numpy používá ve spojení s dalšími knihovnami, jako je třeba Pandas.

Pandas je knihovna používaná pro transformaci, čištění, filtrování dat, ale také načítání dat z několika souborů nebo spojování datasetů. Pandas se také často využívá k přípravě obrovských množství dat, které jsou dále použity pro strojové učení, což ale zpravidla nebývá mezi kompetencemi datových analytiků. Pomocí Pandas zde také vytvářet jednoduché vizuály, podobně jako v Matplotlib či Seaborn. V porovnání s databázovými systémy je Pandas jednoznačně rychlejší a výkonnější.

6.1.2 Soft Skills

Datoví analytici potřebují nejen technické dovednosti, ale neméně důležitou roli hrají také Soft Skills, neboli měkké dovednosti, mezi které paří široké spektrum dovedností, jimiž by datoví analytici měli disponovat. Já osobně bych potřebné Soft Skills rozdělil do následujících kategorií:

- komunikační dovednosti,
- spolupráce,
- analytické myšlení,
- flexibilita.

Komunikační dovednosti zaštiťují nejen komunikaci s cizími lidmi, ale také znalost cizích jazyků (především jazyka anglického), který je v bankovním světě naprosto nezbytný. Jelikož organizace X také působí v rámci nadnárodního trhu, je komunikační znalost anglického jazyka esenciální. A nejde jen o komunikaci s mezinárodními týmy, ze čtyř zmíněních technických dovedností v kapitole výše, minimálně tři z toho jsou programy nebo jazyky, které vychází z jazyka anglického. V tomto případě jsou i dokumentace často psány v tomto jazyce. Jelikož datoví analytici často prezentují nebo sdělují výsledky svých analýz zbylým členům týmu či nadřízenému, vnímám jako jednu z potřebných dovedností schopnost interpretace a prezentace výsledků.

Spolupráce je samozřejmě klíčová nejen v oblasti datových analytiků. V rámci korporátních společností málokdo pracuje pouze „sám na sebe“ a ve většině případů jsou datoví analytici rozmístěni do týmů v bance. Z tohoto hlediska je podstatná forma spolupráce ke správnému fungování týmu a dosahování společných cílů.

Analytické myšlení zahrnuje systematické analyzování a algoritmické uvažování, jejichž základy mohou vycházet ze znalosti matematiky a statistiky. Schopnost porozumění a rozkladu komplexních problémů na dílčí části. V tomto případě jde o dovednost, která se nejlépe rozvíjí až na samostatné pracovní pozici a může být tedy postupem času rozvíjena.

Flexibilita hraje v dnešním dynamickém světě roli důležitější nežli kdy dříve. A technologický pokrok je jen jedním z příkladů, dále může jít o legislativní požadavky ze strany státu či regulátorů se kterými je nutné držet krok, orientovat se a přizpůsobovat se.

6.2 Průzkum pracovních pozic

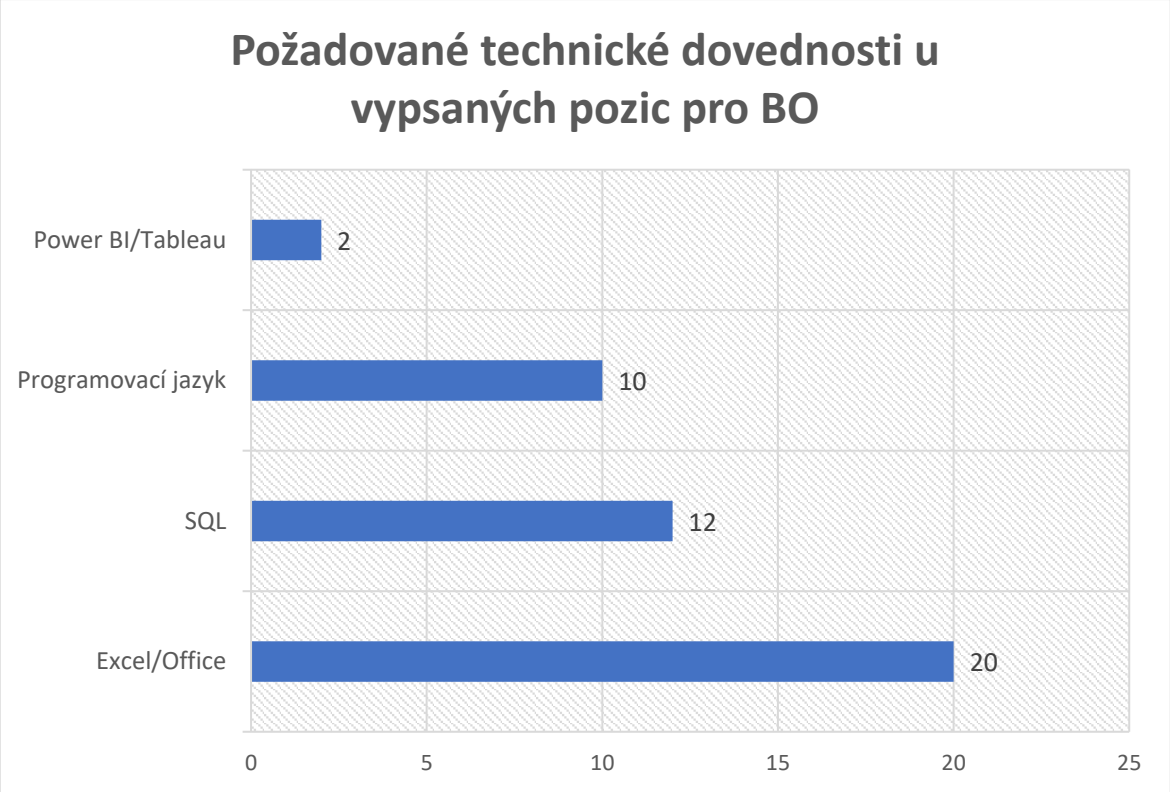
Jak je již zmiňováno v úvodu kapitoly, analyzoval jsem přes 130 pracovních nabídek, které byly ve společnosti X vypsány v rámci prvního čtvrtletí roku 2024. Jednalo se dohromady o 134 nabídek a předmětem zkoumání byly technické požadavky, které jsou vyžadovány.

Z celkem 134 pracovních pozic, které jsem analyzoval, pro mne bylo velmi překvapující zjištění, že zhruba 72 % vypsaných pozic neobsahuje žádné specifické technické dovednosti (nepočítám základní uživatelskou znalost PC nebo základní programy Office balíčku jako Word či PowerPoint). Tato situace je zapříčiněna především tím, že v prvním čtvrtletí roku 2024 společnost nenabírala ve velkém technicky orientované pracovní pozice. I z tohoto důvodu ze 134 se jednalo o 83 pracovních pozic Front-Office a pouze 51 pro Back/Middle-Office.

OBR. 34 - ROZDĚLENÍ NABÍZENÝCH PRACOVNÍCH POZIC V ORGANIZACI X DLE TECHNICKÝCH DOVEDNOSTÍ (ZDROJ: AUTOR)

Co se týče vypsaných pozic pro FO, neznamená to však, že by pracovní pozice neobsahovaly žádné požadavky na potenciální kandidáty, pouze se jedná, ve většině případů, o požadavky netechnického charakteru.

Porovnání vypsaných pracovních pozic pro MO a BO s technickými dovednostmi datový analytiků (tabulkový editor, SQL, programovací jazyky, vizualizační nástroje) vypadá následovně.



OBR. 35 – POŽADOVANÉ TECHNICKÉ DOVEDNOSTI U VYPSANÝCH PRACOVNÍCH POZIC V ORGANIZACI X PRO BACK-OFFICE (ZDROJ: AUTOR)

V celkově 51 inzerátech se dvakrát vyskytl požadavek na vizualizační nástroje, zejména Power BI a Tableau. 10 inzerátů také obsahovalo požadavek na znalost minimálně jednoho programovacího jazyku, nejčastěji se vyskytoval Python a hned za ním JavaScript, což je logické, jelikož oba jazyky patří mezi ty globálně nejpoužívanější. U 12 pracovních nabídek se vyskytl požadavek na znalost některé z verze SQL, v některých případech také na procedurální rozšíření. Nejčastěji požadované technické dovednosti se váží k Excelu či dalším nástrojům Office balíčku, požadavek na tento Hard Skill se vyskytl u 20 inzerátů. Samozřejmě i u zbytku pracovních pozic pro BO se vyskytovaly požadavky na určité technické dovednosti, avšak ty byly nerelevantní pro tento průzkum, a proto nejsou zahrnuty.

6.3 Front-Office, Middle-Office, Back-Office

Na otázku „Do jaké míry vzdělávat zaměstnance v oblasti datové analýzy?“ neexistuje jednoduchá a jednoznačná odpověď. Hlavní proměnnou jsou pracovní pozice či divize v rámci banky, ve kterých zaměstnanci pracují a pro zodpovězení dotazu je třeba banku rozdělit na Front-Office, Middle-Office a Back-Office, stejně jako v předchozí podkapitole.

6.3.1 Front-Office

FO, neboli Front-Office má v různých společnostech různé významy, avšak pojí je jedno. V případě bankovnictví jde však o tu část banky, která je zaměřena na generování příjmů, a to prostřednictvím klientů jak fyzických, tak i právnických osob. Obecně lze tedy říci, že se jedná o celé spektrum pracovních rolí zaměřené na klienty. To však neznamená, že Middle-Office či Back-Office nemají takový význam. Každá část banky má jiný účel a jedna by nemohla existovat bez druhé. Největší část pracovníků FO tvoří bankéři, kteří se dále mohou dělit, například na bankéře pro fyzické osoby a bankéře pro právnické osoby. Dále do této oblasti banky spadají úvěroví a hypoteční

specialisté, business konzultanti, zaměstnanci call centra a mnoho dalších zaměstnanců, kteří přichází do styku s klienty.

Ze všech částí banky (Front-Office, Middle-Office a Back-Office) se jedná o tu část, kde požadavky na technické pozice jsou nejmenší, což je vidět i z mé analýzy pracovních nabídek. Z 84 pracovních nabídek pouze pár z nich mělo požadavky na znalost Hard Skills a jednalo se nejčastěji o znalost práce s PC, případně programy Office balíčku. To samozřejmě neznamená, že tito pracovníci poté žádné technické dovednosti na svých pozicích nevyužívají. Využívají je, ale pouze v základním měřítku, nejčastěji se jedná o interní systémy a nástroje Office balíčku, jako Word, PowerPoint či Excel. Avšak je tento přístup správný? V dnešní pokrokové době technologií banka klade požadavky na digitalizaci a automatizaci procesů, což by jednoznačně mohlo umožnit také práci zaměstnanců FO, ovšem předpokladem pro to není hluboká znalost několika programovacích jazyků, ale znát alespoň základy a vědět, jaká je jejich síla a možnosti, poté může buďto sám bankéř přijít s vhodným řešením automatizace, popřípadě požádat o pomoc IT tým. Pokud však zaměstnanec neví, jaké procesy jdou automatizovat či zjednodušit, situace se v ničem nezmění. Takže co se týče programovacích jazyků, vnímám jako rezervu pro zlepšení zaměstnance FO edukovat v oblasti algoritmizace a optimalizace pomocí programovacích jazyků na základní úrovni a pochopení principu a možností programovacích jazyků.

Pokud přejdeme z programovacích jazyků na znalost tabulkových editorů, zde vnímám, že dobrá znalost zaměstnanců FO může být velmi obohacující. Pokud uvedeme na příkladu Microsoft Excel, který je nejrozšířenějším tabulkovým editorem, ačkoliv jde o intuitivní nástroj, možnosti nabízí téměř neomezené. A nemluvím pouze o základních znalostech, ale o vyhledávacích funkcích, textových funkcích, kontingenčních tabulkách a grafech, makrech, které mohou posloužit pro osobní analýzy, pochopení vzorů klientů, základní vizualizaci klientských portfolií a mnoho dalšího. Co se týče maker, ty opět mohou hrát skvělou roli v rámci optimalizace procesů. Nově také Microsoft začíná prosazovat možnost propojení programovacího jazyka Python a Excel. Funkce je v současné době přístupná pouze v beta přístupu.[40]

Ohledně vizualizačních nástrojů a SQL nesdílím stejný názor jako u Excelu a programovacích jazyků. Ačkoliv nabízí nepřeberné množství zdrojů, odkud lze získávat data, na jejichž základě lze vytvářet vizualizace. Dle mého názoru bývá nejčastější přístup k datům pomocí scriptů SQL, kterým se lze dotazovat do databáze. Tudíž předpokladem pro vytváření dynamických reportů je kromě znalosti vizualizačních nástrojů také znalost SQL a obojí popravdě nevnímám jako nutnost. Věřím tomu, že konzumenty určitých reportů mohou být i bankéři a vnímám jako potřebné, aby zaměstnanci FO rozuměli principu dynamických reportů a zvládali v reportech číst a orientovat se, což není nijak náročné. Avšak není žádoucí, aby zaměstnanci FO byli schopni samotné reporty vytvářet, případně psát SQL dotazy.

6.3.2 Middle-Office

Českým ekvivalentem pro Middle-Office (MO) může být „střední kancelář“ a ačkoliv se tento termín moc nepoužívá, z významu spojení slov Middle-Office je zřetelné, že jedná se o střední či prostřední část banky, která je umístěna mezi Back-Office a Front-Office. Hlavním cílem MO je podpora FO pomocí IT zdrojů, Compliance, řízení rizik a poskytováním právní podpory. Jedná se také o část banky, která má pod správou HR, účetnictví a administrativní pracovníky. Ekonomická krize v roce 2008 neměla dopad pouze na regulatorní reporting, ale také velký význam na MO, jelikož se značně zvýšil význam Compliance, Treasury nebo týmů řízení rizik. Back-Office a Middle-Office se ve spoustě částí svých agend překrývají, i z tohoto důvodu jsou technické dovednosti zaměstnanců obou částí podobné.

Do jaké míry by měli být vzděláváni zaměstnanci MO v oblasti datové analýzy není vůbec jednouché na zodpovězení, jelikož v mnoha případech se nejedná o vyloženě technické pozice. V případě právní podpory nevidím jediný důvod, proč by měla být snaha o prohlubování znalostí

těchto zaměstnanců, v případě Compliance, Treasury a řízení rizik. V případě, že by tyto týmy uchovávaly svá data či data, se kterými pracují v relačních databázích, je samozřejmě nutné ovládat SQL na úrovni SELECT operací a následně vytváření reportů pomocí BI, pomocí kterých by týmy jednoduše mohly data v databázích monitorovat. Určitě se v rámci zmíněných týmů najdou i pozice, u kterých je požadavkem znalost programovacích jazyků, avšak kvůli vysoké variaci pracovních pozic nelze přijít s jednoduchou a jednoznačnou odpovědí. Znalost programovacích jazyků pro datovou analýzu tedy nevnímám jako nedostatek, který je třeba zlepšit a na druhé straně Microsoft Excel vnímám jako nástroj, který v rámci téměř celého MO najde své využití.

6.3.3 Back-Office

Back-Office (BO), jakožto poslední zbývající část také tvoří podporu, a to pro MO a FO. Jedná se o část, která jak již název napovídá se nachází „vzadu“. Jde tedy o část, do které široká veřejnost nevidí a ani neuskutečňuje styk s klienty a nevytváří příjmy společnosti. Jedná se však o základní stavební kámen fungování a BO především zajišťuje hladký chod každodenních operací podniku a aby společnost fungovala jako celek. V rámci BO se vedou záznamy o obchodech a transakcích, vývoj a správa databází, automatizují procesy, řídí data a informace a mnoho dalšího. Jednoduše řečeno BO hraje důležitou roli v zajištění provozu a zároveň se jedná o pracovní pozice, které mají ze všech částí banky nejtechničtější charakter. Příkladem jsou BO vývojáři, softwaroví inženýři, techničtí pracovníci, datoví analytici či IT podpora.

Jelikož se většinou jedná o pracovní pozice technického charakteru, zaměstnanci mají velmi dobré povědomí o technických dovednostech datových analytiků a spoustu z nich alespoň část z nich ovládá. Požadované technické znalosti se na různých pozicích také velmi liší, takže nevnímám jako potřebu prohlubovat vědomosti v oblasti datové analýzy u těchto pracovníků a pokud náhodou tyto znalosti nemají a jsou po nich vyžadovány, již mají velmi dobré základy a nebude pro ně zdaleka tak náročné se novým dovednostem naučit, vzhledem k tomu, že spousta programovacích jazyků staví na stejných základech.

6.4 Závěr kapitoly

Z předchozích tří podkapitol je evidentní, že největší rezervu pro vzdělání v oblasti datové analýzy vnímám u zaměstnanců Front-Office. Po celou dobu našich životů budeme s velkou pravděpodobností vedeni k digitalizaci, zjednodušování a algoritmizaci všech procesů, abychom si svoji práci usnadnili co nejvíce. Právě v tomto směru může pomoci hlubší znalost tabulkových editorů či alespoň základní znalost programovacích jazyků, kterou z logických důvodu zaměstnanci Front-Office absentují. V rámci Middle-Office lze těžko soudit, jelikož tam jsou dle mého názoru zaměstnanci dost rozděleni. Spousta z nich má relativně dobré technické dovednosti, a také spousta z nich má minimální technické dovednosti. U té části zaměstnanců, kteří technickými dovednostmi nedisponují, bych doporučil to samé jako v rámci Front-Office. V rámci Back-Office nevnímám toto téma jako problém, jelikož většina zaměstnanců již alespoň základy v oblasti datové analýzy má, či minimálně má o nástrojích datové analýzy povědomí. A jakým způsobem se mohou zaměstnanci naučit těmto Hard Skills? Z mého pohledu by bylo optimální rozvíjet Hard Skills v rámci interních školení. Zaprvé by nevznikaly tak vysoké náklady jako ze strany externích školitelů a za druhé by se v rámci školení mohla probírat témata přímo spojená s náplní práce jednotlivých zaměstnanců.

Závěr

Na datovou kvalitu, jakožto nedílnou součást bankovní sféry, se klade vyšší důraz než kdy dříve. Důvodů je hned několik, mimo rozhodování na základě dat, snižování nákladů, dodržování obchodních cílů hraje nedílnou roli také vykazování vůči regulátorům ve formě regulatorního reportingu. Svět v oblasti dat se přitom řídí jednoduchým principem – GIGO, který je zkratkou „Garbage in, garbage out“, což hovoří o závislosti kvality výstupu na kvalitě vstupu. Pokud do rozhodovacího procesu vstupují data, je nutné, aby byla správná.

Jedním z mocných nástrojů, který se používá ve spojitosti s monitorováním dat, jsou BI vizualizační nástroje, jejichž hlavní úlohou je z velkých množství dat jednoduše interpretovat složitá data. Předpokladem pro použití BI nástrojů v rámci organizace je však již fungující a kvalitní datová infrastruktura. Z tohoto důvodu je také v rámci teoretické části práce vysvětleno zařazení BI v rámci schématu hierarchie datových potřeb.

Hlavní pozornost praktické části byla věnována kladům a záporům jednotlivých BI nástrojů. Kromě porovnání technických a netechnických aspektů jednotlivých nástrojů byla také transformována datová sada pomocí knihoven Pythonu, která byla následně použita jako zdroj pro jednotlivé vizuály, jež byly cílem porovnání. Vyjma vytváření grafů bylo také cílem vytvoření kategorického sloupce v obou nástrojích s ukázkou odlišných přístupů. V rámci samostatné vizualizace se jednalo o tři grafy – sloupcový graf s linií průměru, prstencový graf s vyplněnou vnitřní částí prstenu a bodový graf společně s ukázkou trendu za jednotlivé kategorie klientů. Ačkoliv Power BI i Tableau nabízí odlišné možnosti, vzhledem k vysoké flexibilitě a modifikovatelnosti nástrojů je možné dojít k podobným závěrům a nelze jednoznačně jeden nástroj postavit před druhý.

Následujícím cílem bylo zjištění vlivu týmu datové kvality na chod společnosti a jakým způsobem by se absence týmu projevila na chodu společnosti. V bakalářské práci jsem představil nástroj pro měření datové kvality ve společnosti X – indikátor datové kvality, včetně jeho popisu a ukázky ve formě SQL kódu. Absence týmu datové kvality by velmi rychle vedla ke ztrátě povědomí ohledně úrovně datové kvality v rámci společnosti a z dlouhodobého hlediska k nízké datové kvalitě, což společnost X nebylo udržitelné z důvodu regulatorních požadavků, které by společnost nebyla schopna plnit. Mimo to se nízká kvalita dat promítne do výše nákladů, které by byly vynakládány pro řešení a hledání alternativních způsobů. V neposlední řadě také v času, který by musel být stráven nad dodatečným sladěním dat. Mimo to nízká datová kvalita může poškodit jméno celé společnosti. Tým datové kvality je v rámci společnosti naprosto esenciální a není přípustná jeho absence.

Posledním cílem v rámci práce bylo nalezení vhodné odpovědi na otázku „*Na jakých pozicích má smysl vzdělávat zaměstnance v oblasti datové analýzy? A do jaké míry?*“. V rámci této kapitoly jsem pro jednoduchou kategorizaci rozdělil společnost na Back-Office, Middle-Office a Front-Office, následně jsem zanalyzoval a vizualizoval aktuální situaci nabízených pozic společností X v programu Microsoft Excel. Část banky, kde vnímám, že by vzdělání v oblasti datové analýzy přineslo nejvíce kladů je Front-Office, zejména se zaměřením na osobní využití ve formě základů automatizace.

Seznam použitých zdrojů

[1] ČESKÁ NÁRODNÍ BANKA. *Dvoustupňové uspořádání bankovního systému* [online]. [cit. 2024-04-25]. Dostupné z: https://www.historie.cnb.cz/cs/menova_politika/6_menova_politika_na_ceste_ke_standardu_vyspelych_zemi/1_dvoustupnove_usporadani_bankovniho_systemu/

[2] ČESKÁ BANKOVNÍ ASOCIACE. *Český bankovní sektor* [online]. [cit. 2024-04-17]. Dostupné z: https://cbaonline.cz/o-bankovnim-sektoru

[3] ČESKÁ NÁRODNÍ BANKA. *Regulace a dohled v oblasti finančního trhu* [online]. [cit. 2024-04-18]. Dostupné z: https://www.historie.cnb.cz/cs/regulace_a_dohled/regulace_a_dohled_v_oblasti_financniho_trhu_ii/

[4] *Zákon č. 6/1993 Sb.: Zákon České národní rady o České národní bance*. In: . 3/1993.

[5] TETTRA. *Data vs Information vs Knowledge: What Are The Differences?* [online]. SPILKER, Josh. 2023 [cit. 2024-04-25]. Dostupné z: https://tettra.com/article/data-information-knowledge/

[6] MARR, Bernard. How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read. *Bernard Marr & Co.* [online]. [cit. 2024-04-17]. Dostupné z: https://bernardmarr.com/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/

[7] NOVÁK, Daniel. Banky začínají „vyvádět" data na cizí servery. *E15* [online]. 2018 [cit. 2024-04-18]. Dostupné z: https://www.e15.cz/byznys/finance-a-bankovnictvi/banky-zacinaji-vyvadet-data-na-cizi-servery-1342223

[8] GLOBALBANKS. *How Are Dabases Used in Banking? [Banking Database 101]* [online]. ADDAMS, Oluwatosin. [cit. 2024-04-20]. Dostupné z: https://globalbanks.com/how-are-databases-used-in-banking/

[9] HITACHI SOLUTIONS. *Modern Data Analytics in Banking: Benefits, Outlook & More* [online]. [cit. 2024-04-18]. Dostupné z: https://global.hitachi-solutions.com/blog/big-data-banking/

[10] *Kybernetičtí podvodníci přitvrzují, proto se Komerční banka znovu zapojila do kampaně nePINdej* [online]. [cit. 2024-04-18]. Dostupné z: https://www.kb.cz/cs/o-bance/tiskove-zpravy/kyberneticti-podvodnici-pritvrzuji-proto-se-komercni-banka-znovu-zapojila-do-kampane-nepindej

[11] *Regulatory reporting in the financial world – the importance of compliance* [online]. [cit. 2024-04-17]. Dostupné z: https://br-ag.eu/2023/01/31/regulatory-reporting-in-the-financial-world-the-importance-of-compliance/

[12] SLÁNSKÝ, David a Marta KELNEROVÁ. *KPMG Česká republika, s.r.o.. Benchmarking regulatorního reportingu: v bankách v České republice.* [online]. 2019 [cit. 2024-04-18]. Dostupné z: https://assets.kpmg/content/dam/kpmg/cz/pdf/2019/Regulatorybenchmark-CZ.pdf

[13] ČESKÁ NÁRODNÍ BANKA. *Vykazování do AnaCredit* [online]. [cit. 2024-04-20]. Dostupné z: https://www.cnb.cz/cs/statistika/anacredit/vykazovani-do-anacredit/

[14] GRÜNBERG, Sebastian. *AnaCredit – Overview and implementation from an NCB`s point of view* [online]. [cit. 2024-04-17]. Dostupné z: https://www.bis.org/ifc/events/ifc_isi_2017/23_gruenberg_paper.pdf

[15] FORBES. *ČNB udělila pokutu Fio bance. Za chyby ve výkaznictví zaplatí milion korun* [online]. [cit. 2024-04-20]. Dostupné z: https://forbes.cz/cnb-udelila-pokutu-fio-bance-za-chyby-ve-vykaznictvi-zaplati-milion-korun/

[16] POUR, Jan a kol. *Business intelligence v podnikové praxi*. Praha: Professional Publishing, 2012. ISBN 978-80-7431-065-2.

[17] LUFTMAN, Jerry a Rajkumar KEMPAIAH. *An Update on Business-IT Alignment: "A Line" Has Been Drawn* [online]. 2008 [cit. 2024-04-18]. Dostupné z: https://www.researchgate.net/publication/220449609_IT_Governance_An_Alignment_Maturity_Perspective

[18] FOOTE, Keith D. *A Brief History of Business Intelligence* [online]. 2023 [cit. 2024-04-17]. Dostupné z: https://www.dataversity.net/brief-history-business-intelligence/

[19] THE NEW YORK TIMES. *HANS PETER LUHN, MENTOR, 68, DIES; Data-Processing Specialist Served F.B.M. 20 Years* [online]. [cit. 2024-04-17]. Dostupné z: https://www.nytimes.com/1964/08/20/archives/hans-peter-luhn-mentor-68-dies-dataprocessing-specialist-served-fbm.html

[20] TECHTARGET. *Self-service business intelligence (self-service BI)* [online]. ROBINSON, Scott. [cit. 2024-04-20]. Dostupné z: https://www.techtarget.com/searchbusinessanalytics/definition/self-service-business-intelligence-BI

[21] WILLIAMS, Hugh. The Pyramid of Data Needs (and why it matters for your career). *Medium* [online]. 2018 [cit. 2024-04-17]. Dostupné z: https://medium.com/@hugh_data_science/the-pyramid-of-data-needs-and-why-it-matters-for-your-career-b0f695c13f11

[22] ROGATI, Monica. *The AI Hierarchy of Needs* [online]. In: . 2017 [cit. 2024-04-17]. Dostupné z: https://hackernoon.com/the-ai-hierarchy-of-needs-18f111fcc007

[23] NOODLE CORPORATION. *What is the Hierarchy of Needs in Data Science?* [online]. DAVIES, Lucy. [cit. 2024-04-20]. Dostupné z: https://resources.noodle.com/articles/data-science-hierarchy-of-needs/

[24] POUR, jan a kol. *Self service business intelligence*. Praha: Grada, 2018. ISBN 978-80-271-0616-5.

[25] ENLYFT. *Companies using Tableau* [online]. [cit. 2024-04-18]. Dostupné z: https://enlyft.com/tech/products/tableau

[26] ENLYFT. *Companies using Microsoft power BI* [online]. [cit. 2024-04-18]. Dostupné z: https://enlyft.com/tech/products/microsoft-power-bi

[27] POP AUTOMATION. *Power BI and the History of Microsoft Business Intelligence* [online]. 2024 [cit. 2024-04-18]. Dostupné z: https://www.popautomation.com/post/power-bi-name-and-history

[28] SALESFORCE. *Get Tableau Certified* [online]. [cit. 2024-04-17]. Dostupné z: https://www.tableau.com/learn/certification

[29] MICROSOFT. *Microsoft Certified: Power BI Data Analyst Associate* [online]. 2024 [cit. 2024-04-18]. Dostupné z: https://learn.microsoft.com/en-us/credentials/certifications/power-bi-data-analyst-associate/?practice-assessment-type=certification

[30] MICROSOFT. *Ceny Power BI* [online]. [cit. 2024-04-18]. Dostupné z: https://powerbi.microsoft.com/cs-cz/pricing/

[31] SALESFORCE. *Decide the right mix of users for your team* [online]. [cit. 2024-04-18]. Dostupné z: https://www.tableau.com/pricing/teams-orgs

[32] ADATA. *Import Oracle Eloqua Data into the Power BI Service for Visualizations* [online]. [cit. 2024-04-17]. Dostupné z: https://www.cdata.com/kb/tech/eloqua-cloud-powerbi-service.rst

[33] MICROSOFT. *Data sources in Power BI Desktop* [online]. 2024 [cit. 2024-04-18]. Dostupné z: https://learn.microsoft.com/en-us/power-bi/connect-data/desktop-data-sources

[34] SALESFORCE. *CONTENTS Supported Connectors* [online]. [cit. 2024-04-18]. Dostupné z: https://help.tableau.com/current/pro/desktop/en-us/exampleconnections_overview.htm

[35] METABASE. *Database table relationships* [online]. [cit. 2024-04-18]. Dostupné z: https://www.metabase.com/learn/databases/table-relationships#many-to-many-relationship

[36] KAGGLE. *Credit_risk_customers* [online]. 2023 [cit. 2024-04-18]. Dostupné z: https://www.kaggle.com/datasets/ppb00x/credit-risk-customers/data

[37] RIMARČÍKOVÁ, Sabína. *Riešenie datovej kvality v bankovníctve*. Praha, 2022. Dostupné také z: https://vskp.vse.cz/86498_riesenie-datovej-kvality-v-bankovnictve??page=72. Bakalářská práce. Vysoká škola ekonomická v Praze.

[38] TDWI. *Data Quality and the Bottom Line: Achieving Business Success through a Commitment to High Quality Data* [online]. ECKERSON, Wayne W. 2002 [cit. 2024-04-18]. Dostupné z: http://download.101com.com/pub/tdwi/Files/DQReport.pdf

[39] DIGNA. *7 Most Dreadful Incidents Caused by Bad Data Quality in the Banking Sector* [online]. [cit. 2024-04-18]. Dostupné z: https://www.digna.ai/7-most-dreadful-incidents-caused-by-bad-data-quality-in-the-banking-sector

[40] SPREADSHEET POINT. *How To Use Python in Excel [Easy 2024 Guide]* [online]. 2023 [cit. 2024-04-18]. Dostupné z: https://spreadsheetpoint.com/excel/python-in-excel/

Seznam obrázků

Seznam tabulek