Univerzita Hradec Králové

Fakulta informatiky a managementu

DIPLOMOVÁ PRÁCE

2020 Bc. Natália Rakovská

Univerzita Hradec Králové
Fakulta informatiky a managementu
Katedra informatiky a kvantitativních metod

Porovnání kombinovaných nástrojů pro vizualizaci a analýzu dat

Diplomová práce

Autor: Bc. Natália Rakovská

Studijní obor: Informační Management

Vedoucí práce: prof. RNDr. Hana Skalská, CSc.

Hradec Králové listopad 2020

Prehlásenie:

Prehlasujem, že som diplomovú prácu spracovala samostatne a s použitím uvedenej literatúry.

vlastnoručný podpis

V Hradci Králové dne 16.11.2020 Bc. Natália Rakovská

Poďakovanie:

Ďakujem vedúcej mojej diplomovej práce prof. RNDr. Hane Skalskej, CSc. za cenné rady, metodické vedenie práce, trpezlivosť a čas, ktorý mi venovala.

Rada by som sa tiež touto cestou poďakovala mojej rodine za podporu a pomoc počas celého univerzitného štúdia.

Anotácia

Hlavným cieľom práce je porovnať dva BI self-service nástroje, ktoré boli vybrané na základe stanovených kritérií z každoročného hodnotenia zostaveného spoločnosťou Gartner. Teoretická časť je zameraná na analýzu dostupných zdrojov ku kľúčovým pojmom, ktorými sú okrem iných dátová veda, vizualizácia a analýza dát, business intelligence a BI self-service nástroje. Praktická časť je spracovaná na základe metodiky CRISP DM a teda rozdelená podľa jednotlivých fáz tejto metodiky. V tejto časti sú stanovené kritériá pre porovnanie vybraných nástrojov a kritéria pre výber vhodného data setu, vykonané čistenie a príprava data setu. Dva vybrané nástroje sú porovnané z hľadiska dostupných analytických funkcií, vizualizácií a používateľskej prívetivosti. Zistené rozdiely a odporúčania sú zhrnuté v závere práce.

Annotation

Title: Evaluation and comparison of Self-Service BI Tools

The main goal of this Diploma Thesis is the comparison of two self-service BI tools selected according to specific criteria from the annual evaluation of self-service BI tools issued by Gartner. The theoretical part is focusing on the analysis of available resources for the key concepts, which are data science, data analysis and visualization, business intelligence and self-service BI tools. The practical part follows the division into phases of CRISP-DM methodology. This section defines criteria for comparing selected tools as well as the criteria for selection of a suitable data set. Selected data set is cleaned and prepared for analysis. The two selected tools are compared, focusing mainly on the availability of analytical functions, various visualizations, and user-friendliness. Discovered differences and recommendations are summarized in the conclusion of this thesis.

Obsah

Zoznam obrázkov

Zoznam skratiek

IT – informačné technológie

CRISP DM – Cross-industry standard process for data mining

BI – Business Intelligence

GDPR – General Data Protection Regulation

KPI – Key Performance Indicators

MS – Microsoft

SQL – Structured Query Language

1. Úvod

Témou diplomovej práce je porovnanie vybraných, tzv. self-service nástrojov pre vizualizáciu a analýzu dát. S rapídnym rastom objemu dát, ktoré je potrebné spracovať je nutné presunúť analýzu z človeka na technológie. Keďže objem dát stále narastá rozširuje sa aj potreba analyzovať a vizualizovať dáta inými používateľmi ako dátovými expertmi, čo je jedným z dôvodov vzniku self-service BI nástrojov.

Teoretická časť práce je zameraná na analýzu dostupných zdrojov ohľadom tém súvisiacich s vizualizáciou a analýzou dát. Pre pochopenie existencie a využitia self-service BI nástrojov je potrebné poznať pojmy ako dátová vedy, typy dát, data mining, data warehouse a je tiež dôležité poznať úlohu dát v podniku. Ďalšia časť práce sa zaoberá samotným pojmom vizualizácie dát, jej významom a súčasným stavom využitia vizualizácie v praxi. Posledná teoretická časť približuje pojem Business Intelligence, definuje self-service BI nástroje ako aj ich vymedzenie voči tradičným nástrojom business inteligencie.

Praktická časť práce sa opiera o metodiku CRISP DM a jej hlavným cieľom je porovnanie vybraných self-service BI nástrojov. Sú tu stanovené atribúty, na ktoré je zamerané porovnanie vybraných nástrojov, ťažiskom porovnania sú dostupné možnosti vizualizácie a ponúkané analytické funkcie v jednotlivých nástrojoch ako aj celková používateľská prívetivosť.

Tretia podkapitola praktickej časti je usporiadaná podľa jednotlivých fáz metodiky CRISP DM. Vo fáze „Pochopenia dát“ je vysvetlený proces výberu vhodného data setu, na ktorom je dostatočne možné demonštrovať všetky základné funkcie vybraných nástrojov. V nasledujúcej fáze „Prípravy dát“ je vybraný data set očistený a upravený do podoby, ktorá umožní efektívne využitie na analýzu obsiahnutých dát. V tretej fáze Modelovania je zahrnutá samotná komparácia dostupných analytických funkcií a dostupnosť ich využitia pre používateľa s minimálnymi znalosťami oblasti dátovej vedy. V poslednej využitej fáze metodiky CRISP DM nazývanej „Vyhodnotenie“ sú zhodnotené objavené rozdiely medzi vybranými nástrojmi ako aj intuitívnosť ich používania pre začiatočníkov.

V závere práce sú zhrnuté výsledky dosiahnuté počas vypracovávania jednotlivých kapitol a uvedené odporúčania ohľadom vhodnosti využitia vybraných self-service nástrojov.

2. Ciele práce

Práca s dátami a ich vizualizácia sa v súčasnosti stáva dominantou v manažérskej praxi. Často nie je jednoduché vybrať vhodný nástroj na vizualizáciu dát tak, aby správne podporil rozhodovanie nad dátami. Z toho dôvodu je práca orientovaná hlavne prakticky a jej hlavným cieľom je porovnať vybrané tzv. „self-service“ BI nástroje na vybranom konkrétnom data sete a porovnať ich výhody, zhodnotiť ich rozdiely, prípadne uviesť ideálne scenáre využitia každého z nich. Práca bude taktiež obsahovať zhrnutie základnej teórie vzťahujúcej sa k téme vizualizácie dát. Taktiež sa venuje významu vizualizácie dát v podnikovej praxi (ale aj významu pre jednotlivcov) v súčasnej dobe.

Aby bol splnený celkový cieľ práce je potrebné splniť nasledujúce podciele:

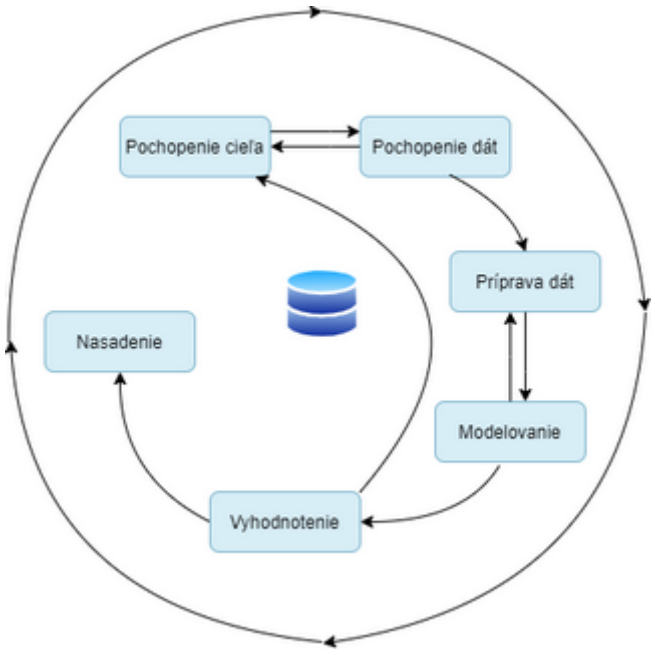
- Charakterizovať BI a potrebu vizualizácie dát
- Charakterizovať rôzne typy dát od klasického ponímania po Big Data
- Objasniť význam rôznych dát v podnikovej praxi
- Charakterizovať vizualizáciu dát
- Špecifikovať vhodné formy vizualizácie pre dané typy dát
- Analyzovať súčasný stav využívania vizualizácie dát v praxi
- Vysvetliť vhodné a nevhodné vizualizovanie dát
- Urobiť prehľad nástrojov BI
- Vybrať podľa kritérií dva self-service BI nástroje
- V praktickej časti vybrať vhodný data set pre prácu vo vybraných BI nástrojoch
- Vykonať prípravu a čistenie dát zvoleného data setu
- Urobiť komparáciu vybraných nástrojov z rôznych hľadísk (analytické funkcie, ponúkané typy grafov pre vizualizáciu)
- Sumarizovať výsledky komparácie

3. Metodika spracovania

Teoretická časť práce používa niektoré metódy vedy a výskumu. Opiera sa najmä o analýzu dostupných zdrojov a rešerš týchto zdrojov. Týmito spôsobmi sú vypracované vysvetlenia a zhrnutia k základným pojmom týkajúcim sa dátovej vedy, vizualizácie a analýzy dát, ktoré sú jej súčasťou.

Praktická časť práce je spracovaná podľa metodiky CRISP, kde boli použité tie fázy metodiky, ktoré sú aplikovateľné vzhľadom k využívaniu self-service nástrojov BI.

Metodika CRISP-DM sa využíva ako univerzálny postup pri riešení rôznych úloh získavania znalostí z dát. Popisuje štandardný proces získavania znalostí s dôrazom na kľúčovú časť procesu, ktorou je data mining. V súčasnosti patrí k najpoužívanejším metodikám. „Metodika CRISP-DM (Cross Industry Standard Process for Data Mining) vznikla v rámci Európskeho výskumného projektu, ktorého cieľom bolo navrhnúť univerzálny postup použiteľný v najrôznejších komerčných aplikáciách. Cyklus podľa metodiky CRISP-DM pozostáva zo šiestich fáz, medzi ktorými existujú vzťahy. Výsledok dosiahnutý v jednej fáze ovplyvní voľbu nasledujúceho kroku. Často sa treba k niektorým krokom a fázam vracať (napríklad príprava dát, modelovanie)“ [1] Obrázok 1 schematicky zobrazuje usporiadanie jednotlivých fáz CRISP metodiky.



Obrázok 1 CRISP DM metodika, spracované podľa [1]

Týmito fázami sú [2] :

- Pochopenie cieľa (Business Understanding) – táto počiatočná fáza je zameraná na pochopenie cieľov projektu a požiadaviek z obchodného hľadiska a následne na využitie týchto znalostí na zostavenie definície problému pre data mining a náčrtu predbežného plánu ako problém riešiť,
- Pochopenie dát (Data Understanding) – fáza začína zberom dát, pokračuje oboznámením sa s dátami, identifikovaním problémov kvality dát a hľadaním prvých vzorov v dátach alebo bodov a hypotéz, pre ďalšie skúmanie,
- Príprava dát (Data Preparation) – táto fáza zahŕňa všetky aktivity spojené s čistením a prípravou dát potrebných pre vytvorenie finálneho data setu,
- Modelovanie (Modeling) – v tejto fáze sú aplikované vybrané techniky modelovania a ich parametre sú kalibrované na optimálne hodnoty. Pre jeden problém data miningu môžu existovať rôzne techniky [3], čo si môže vyžadovať návrat k príprave dát,
- Vyhodnotenie (Evaluation) – táto fáza hodnotí vytvorený model (resp. viac modelov) v predošlých fázach, ktoré sú na vysokej úrovni z hľadiska dátovej analýzy. Cieľom tejto fázy je následne zhodnotiť, či niektorý z podnikových problémov nebol opomenutý alebo nedostatočne vyriešený. Model je potrebné skontrolovať aj z hľadiska uskutočnených krokov, ale aj z hľadiska dosiahnutia požadovaných cieľov. Na konci tejto fázy by malo byť uskutočnené rozhodnutie, ktorý z modelov data miningu bude nasadený (resp. ak ide o jeden model, či je vhodný na použitie).
- Nasadenie (Deployment) – vo všeobecnosti projekt nekončí vytvorením modelu, je potrebné získané znalosti upraviť do podoby pochopiteľnej pre používateľa, ktorý daný model nasadí do podniku. Táto fáza má v každom projekte rozličnú náročnosť: od jednoduchých riešení v podobe niekoľkých reportov až po implementovanie opakovateľného procesu data miningu pre celý podnik.

Nie je však nutné dodržiavať toto najčastejšie používané poradie fáz. Celý proces so všetkými fázami sa cyklicky opakuje, čo je na obrázku znázornené vonkajším kruhom. Proces teda nezačína prvou fázou a nekončí nasadením, ale je nutné sa vracať k jednotlivým fázam, priebežne ich dopĺňať a prehodnocovať. Metodika CRISP DM

umožňuje cyklický vývoj a zdokonaľovanie modelu na získavanie znalostí v dátach, čo je jej nespornou výhodou.

V praktickej časti práce je použitá selekcia a komparácia. Selekcia je použitá na výber vhodného data setu. Komparácia ako „metóda zisťovania zhodných alebo rozdielnych vlastností pozorovaných entít“ [4] je použitá na porovnanie dvoch vybraných BI self-service nástrojov. Tieto nástroje sú porovnávané najmä z hľadiska funkcionalít, ktoré ponúkajú na prípravu a čistenie dát, na analýzu a vizualizáciu dát.

4. Klasické rozdelenie dát a možnosť ich vizualizácie

Táto kapitola sa venuje dôležitým pojmom, ktoré súvisia s vizualizáciou a analýzou dát. Ide predovšetkým o pojmy týkajúce sa dát, ich klasifikácie a vedy o dátach, ktoré sú potrebné na pochopenie významu vizualizácie dát a aj na dosiahnutie hlavného cieľa práce.

4.1 Dátová veda

Dátová veda je jedným zo základných pojmov, ktoré môžu mať rôzne definície v závislosti na tom, s akým ďalším vedným odborom práve spolupracuje. Dátová veda v sebe istým spôsobom spája štatistiku, informatiku, výpočtovú techniku, komunikáciu, manažment a sociológiu s cieľom skúmania a študovania dát. [5] Jednou z najčastejších je definícia „Dátová veda je inter-disciplinárnym odborom, ktorý využíva vedecké metódy, procesy, algoritmy a systémy pre získavanie znalostí a poznatkov z dát či už v štruktúrovanej alebo neštruktúrovanej podobe.“ [5]

Dátová veda tiež úzko súvisí s data miningom, strojovým učením a s pojmom „big data“.

Opierajúc sa o túto definíciu je možné povedať, že aj vizualizácia a analýza dát sú súčasťou dátovej vedy.

4.2 Dáta v podniku

Dáta sú pre fungovanie organizácie rozhodujúce – sú základom pre rozhodovanie a pomocou nich podnik koordinuje svoju činnosť. Ak dáta nie sú spravované s rovnakou pozornosťou ako ostatné podnikové aktíva môže dochádzať k menej vhodným rozhodnutiam a znižovaniu efektivity. [6] Jedným z rozhodujúcich faktorov, ktoré ovplyvňujú podnikové procesy sú aj kvalita a spracovanie dát.

Kvalita dát je zásadná nie len pre rozhrania systémov, ale aj aby osoby zodpovedné za rozhodovanie mohli rozhodovať na základe spoľahlivých dát. [6] Kvalita dát je definovaná stupňom, ktorým dané dáta zodpovedajú príslušným charakteristikám. V širšom zmysle je kvalita „vhodnosťou pre daný účel“. Kvalita dát je veľmi širokou témou, ktorá zahŕňa oblasti ako kompletnosť, presnosť, konzistenciu a včasnosť. [6]

Podľa [7] pri spracovávaní dát alebo pri dátovej analýze sú tomuto procesu priradené všetky koncepty a nástroje primárne sa zaoberajúce hodnotením a analýzou dát.

Podľa [8] analyzovanie a využívanie dát môže prispieť k informovanému rozhodovaniu v rôznych fázach celého výrobného procesu.

Dáta sú pre fungovanie organizácie rozhodujúce – sú základom pre rozhodovanie a pomocou nich podnik koordinuje svoju činnosť. Ak dáta nie sú spravované s rovnakou pozornosťou ako ostatné podnikové aktíva môže dochádzať k menej vhodným rozhodnutiam a znižovaniu efektivity. [6]

4.3 Data mining

Pojem Data mining, tiež známy ako „knowledge discovery in databases“ sa po prvýkrát objavil koncom osemdesiatych rokov. V súčasnosti je jednou z najvyspelejších výskumných oblastí v rámci vedy o dátach. [9] Data mining je cieľovo orientované modelovanie dát a jeho techniky slúžia na získavanie informácií, vzťahov, trendov a ďalších ešte neobjavených informácií z veľkého počtu interných databáz. [9]

Chen [10] hovorí, že data mining znamená použitie existujúceho systému manažmentu databáz na získavanie dotazov a tvorenie reportov. Kombinujú sa s multidimenzionálnou analýzou, štatistickou analýzou a aplikovaním online analytického spracovania dát (OLAP) na získanie relevantnej analýzy pre rozhodovanie.

V podstate data mining znamená objavovanie bezprecedentných implicitných poznatkov v databáze. Všetky tieto metódy sú metódami extrakcie užitočných informácií z databáz.

Aplikačný výskum data miningu sa týka vývoja rôznych data miningových modelov aplikovaných do rôznych odvetví. [10] Medzi typické oblasti patria: analýza a prognóza trhu, analýza a prognóza predaja veľkých supermarketov, predajné kanály a analýza cien, a ďalšie. [9] Široké využívanie modelov data miningu je aj v oblasti manažmentu zákazníkov či marketingu (napr. predikcia vplyvu reklamy na produkciu).

4.4 Typy dát

Pre správnu analýzu dát je veľmi dôležité dobre poznať typ dát, s ktorými pracujeme. Dátové typy sú dôležitým konceptom, pretože štatistické metódy môžu byť použité len pre niektoré z nich. Spojité dáta sa musia analyzovať iným spôsobom než kategorické dáta, v opačnom prípade by sme dospeli k zlému výsledku analýzy. Preto nám poznanie dátových typov, s ktorými nakladáme umožňuje vybrať správnu metódu analýzy.

Z jedného pohľadu je možné dáta deliť na štruktúrované, semi-štruktúrované a neštruktúrované dáta. [11]

Nie všetky dáta sú si rovné, napríklad dáta získané zo sociálnych sietí sú úplne odlišné od dát generovaných systémami dodávateľských reťazcov. Niektoré dáta sú štruktúrované (napr. v relačných databázach), ale prevažná väčšina dát je neštruktúrovaných. Spôsob akým sú dáta zbierané, spracovávané a analyzované záleží od ich formátu.

Štruktúrované dáta sú vysoko organizované a formátované do podoby, vďaka ktorej sa jednoducho vyhľadávajú v relačnej databáze. Neštruktúrované dáta nemajú žiaden preddefinovaný formát ani spôsob organizácie, čo značne sťažuje ich zber, spracovanie a analýzu. Okrem odlišného spôsobu získavania, spracovávania a analýzy, sa štruktúrované a neštruktúrované dáta budú nachádzať v kompletne odlišných databázach. [12]

Štruktúrované dáta [12] – tieto dáta sú najčastejšie kategorizované ako kvantitatívne dáta. Sú to dáta, ktoré sa jednoducho zapisujú do polí a stĺpcov v relačných databázach a tabuľkách. Štruktúrované dáta spôsobili revolúciu v systémoch založených

na papierových dokumentoch, o ktoré sa spoločnosti opierali pri chode ich podnikania. Zatiaľ čo sú štruktúrované dáta stále užitočné, postupne sa všetky spoločnosti snažia rozložiť a spracovať neštruktúrované dáta pre budúce príležitosti.

Príkladom štruktúrovaných dát môžu byť napríklad: meno, dátum, adresa, čísla kreditných kariet, informácie o akciách (stock information), geolokácia, a ďalšie.

Štruktúrované dáta sú vysoko organizované a jednoducho zrozumiteľné pre strojový jazyk. Práca s relačnými tabuľkami umožňuje relatívne rýchle vyhľadanie, input a manipuláciu s obsiahnutými dátami, čo je najatraktívnejšou vlastnosťou štruktúrovaných dát. Programovací jazyk používaný na správu štruktúrovaných údajov sa nazýva štruktúrovaný dotazovací jazyk, známy tiež ako SQL (structured query language).

Neštruktúrované dáta [12] – sú najčastejšie kategorizované ako kvalitatívne dáta a nemôžu byť spracované a analyzované pomocou konvenčných nástrojov a metód.

Neštruktúrovanými dátami sú napríklad: text, video, audio, mobilná aktivita, aktivita na sociálnych sieťach, snímky zo satelitov, zábery z bezpečnostných systémov a mnoho ďalších. Môžu byť získavané z textových súborov, prezentácií, sociálnych sietí, emailov, webových stránok, komunikácií ako sú čety, nástroje na podnikovú kolaboráciu, posielanie správ mobilmi a dáta z multimédií (videá, zvukové súbory, digitálne fotografie a pod.).

Tieto dáta sa dajú len veľmi ťažko rozložiť, pretože nemajú žiaden preddefinovaný model, čo znamená, že ich nie je možné usporiadať do relačnej databázy. Namiesto toho sú na správu neštruktúrovaných dát najvhodnejšie nerelačné alebo NoSQL databázy (Not-only-SQL).

Viac ako 80% dát generovaných v dnešnej dobe je považovaných za neštruktúrované a toto číslo bude stúpať s významom Internetu vecí. Porozumieť neštruktúrovaným dátam nie je ľahká úloha. Vyžaduje si to pokročilú analýzu a vysoký úroveň technickej expertízy, čo môže byť veľmi nákladné pre mnohé spoločnosti.

Tí, ktorí dokážu efektívne využívať neštruktúrované dáta získavajú veľkú konkurenčnú výhodu. Zatiaľ čo štruktúrované dáta poskytujú pohľad na zákazníkov z vtáčej perspektívy, neštruktúrované dáta nám môžu poskytnúť hlbšie pochopenie

zámerov a správania sa zákazníkov. Napríklad, data miningové techniky aplikované na neštruktúrované dáta môžu firmám pomôcť pochopiť nákupné návyky a načasovanie, vzory v nákupoch, vnímanie konkrétneho produktu a mnoho ďalších vecí.

Neštruktúrované dáta sú tiež kľúčom k prediktívnym analytickým softvérom. Napríklad, dátové senzory pripevnené k industriálnym strojom môžu včas upozorniť výrobcov na netypickú aktivitu. Vďaka takejto informácií môže byť údržba vykonaná predtým ako príde k nákladnému poškodeniu stroja.

Pološtruktúrované dáta – sú dáta, ktoré uchovávajú interné značky a značky, ktoré identifikujú samostatné dátové prvky, čo umožňuje zoskupenie informácií a vytváranie hierarchie. Dokumenty aj databázy môžu byť čiastočne štruktúrované. Tento typ sa často využíva v podnikovej praxi a napríklad e-mail je veľmi častým príkladom pološtruktúrovaného dátového typu (obsahuje tzv. metadáta, ktoré umožňujú vyhľadávanie podľa kľúčových slov). Pološtruktúrované dáta sa používajú na prenos dát na webe prostredníctvom značkovacích jazykov ako je napríklad XML, sú používané aj nerelačné databázy.

V súčasnosti firmy používajú často kombináciu všetkých vyššie spomenutých typov dát, pričom databázy postupne dosahujú niekoľko terabytov. Z pojmu veľkých dátových skladov sa postupne udomácnil pojem „big data“, ktorý neznamená iba to, že databáza je veľká (z hľadiska bytov), ale pojem big data je charakterizovaný troma atribútmi objemom, rýchlosťou a rôznorodosťou (volume, velocity a variety). Pre to, aby bolo možné hovoriť o big data musí platiť, že majú veľký objem (high volume), veľkú rýchlosť (high velocity) a veľkú rôznorodosť (high variety). To znamená, že je ich veľké množstvo, ktoré je ťažko spracovateľné konvenčným spôsobom, dáta sa rýchlo menia a sú rôznych foriem. Podľa [13] „Big data je pojem, ktorý sa používa na popis údajov s veľkým objemom, vysokou rýchlosťou a / alebo vysokou rozmanitosťou; vyžaduje nové technológie a techniky na ich zachytenie, uloženie a analýzu; a používa sa na zdokonalenie rozhodovania, pre vytváranie, poskytovanie poznatkov a na podporu a optimalizácia procesov.“

Big data sú často získavané zo sociálnych sietí, rôznych senzorov inteligentných strojov, GPS údajov, rôznych textových súborov. Na analýzu a prácu s big data sú používané iné techniky a nástroje ako sú používané pri štruktúrovaných dátach. [14]

4.4.1 Štatistické typy dát

Väčšina dát patrí do jednej z dvoch skupín: kategorické dáta alebo numerické dáta. Prvou skupinou sú **kategorické dáta** známe tiež pod názvom kvalitatívne dáta, ktoré predstavujú charakteristiky ako napríklad pohlavie osoby, rodné mesto alebo typy filmov, ktoré sa danej osobe páčia. Kategorické dáta môžu nadobúdať numerické hodnoty (napríklad 1 pre mužov a 2 pre ženu), ktoré však nemajú matematický význam. [15]

Tieto dáta ďalej delíme na:

Nominálne dáta – nominálne hodnoty reprezentujú diskrétne jednotky a využívajú sa na označovanie premenných, ktoré nemajú žiadnu nominálnu hodnotu. Nominálne dáta nemajú žiadne poradie, a preto môžeme ich poradie zamieňať bez straty významu.

Informácie z týchto dát môžeme získať pomocou meraní:

- Frekvencie
- Proporcií
- Percent

Pre jednoduchú vizualizáciu tohto typu dát môžeme použiť koláčový alebo stĺpcový graf.

Ordinálne dáta – ordinálne hodnoty reprezentujú diskrétne hodnoty a ich špecifické poradie. Sú preto veľmi podobné nominálnym dátam, avšak pre ordinálne dáta znamená zmena poradia tiež zmenu alebo stratu významu, čo je hlavným obmedzením ordinálnych dát. Vďaka týmto ich vlastnostiam sú ordinálne dáta najčastejšie využívané na meranie nenumerických vlastností, akými môžu byť napríklad šťastie alebo spokojnosť zákazníkov.

Pri narábaní s ordinálnymi dátami môžeme použiť rovnaké metódy pre vizualizáciu ako pri nominálnych dátach (koláčový graf a stĺpcový graf), ale tiež máme k dispozícií niekoľko dodatočných spôsobov ich sumarizácie, ktorými sú percentily, medián, funkcia mode a medzi-kvartilový rozptyl.

Druhou skupinou dát sú **numerické dáta**. Tieto ďalej delíme na:

Diskrétne dáta – o diskrétnych dátach je možné hovoriť, keď sa jedná o odlišné a oddelené hodnoty dát. Tento typ dát nie je možné zmerať, ale je možné dáta tohto typu spočítať. V podstate sa jedná o dáta, ktoré môžeme zaradiť do klasifikácie. Jedným z typických príkladov je tiež počet padnutých „hláv" zo 100 hodov mincou.

Vizualizácia takýchto dát sa realizuje prostredníctvom niektorých typov grafov ako napríklad: stĺpcový a zložený stĺpcový graf, pruhový a zložený pruhový graf, radarový graf. [16] Každý z týchto grafov má svoje výhody, nevýhody a presný účel použitia podľa toho, či ide o jednodimenzionálne alebo multidimenzionálne dáta [17]

Spojité dáta – reprezentujú hlavne merania, ktoré prebiehajú napr. v čase, a preto ich hodnoty nemôžu byť diskrétne vypočítané. Spojité dáta môžeme následne rozdeliť na niekoľko ďalších typov [18]:

Intervalové dáta – intervalové hodnoty predstavujú zoradené jednotky s rovnakými odstupmi. Preto môžeme hovoriť o intervalových dátach, keď máme premennú, ktorá obsahuje numerické hodnoty, ktoré sú zoradené (postupnosti) a u ktorých poznáme presné rozdiely medzi jednotlivými hodnotami (napríklad „Teplota: -10, -5, 0, 5, 10"). Problémom práce s intervalovými dátami môže byť skutočnosť, že tieto dáta nemajú „pravú nulu". Vo vyššie uvedenom príklade to znamená, že neexistuje hodnota „žiadna teplota". Intervalové dáta môžeme sčítať i odčítať, ale nemôžeme ich násobiť, deliť ani vypočítať pomer. Pretože tieto dáta nemajú žiadnu „pravú nulu", nie je možné na nich aplikovať veľkú časť popisnej a inferenčnej štatistiky.

Pomerové dáta – pomerové hodnoty sú tiež zoradené jednotky s rovnakými rozdielmi. Od intervalových dát sa líšia hlavne tým, že majú absolútnu nulu (napríklad váha nebo dĺžka).

Na spracovanie spojitých dát môžeme použiť väčšinu metód. Môžeme ich sumarizovať pomocou percentilu, mediánu, medzi-kvartilového rozptylu, funkcie mode, štandardného rozptylu a rozsahu.

Na vizualizáciu spojitých dát môžeme využiť histogram alebo krabicový graf. Pomocou histogramu môžeme overiť centrálnu tendenciu, variabilitu, modalitu.

Histogram však neukáže prítomnosť odľahlých hodnôt a preto je využívaný aj krabicový graf.

Toto je jedna z možných klasifikácií dát, nie je však jediná, keďže dáta môžeme klasifikovať na základe rôznych vlastností do rôznych skupín a reálne neexistuje žiadny fixný a „najsprávnejší“ spôsob klasifikácie dát.

Vyššie uvedené dáta môžeme prehľadnejšie kategorizovať pomocou nasledujúcej tabuľky Tabuľka 1, kde je uvedený aj typ vizualizácie prislúchajúci danému typu dát.

Tabuľka 1 Typy dát a ich vizualizácia, spracované podľa [19]

Typ dát	Príklad	Náhľad	typ (vizualizácie)
Kategorické	nenumerické dáta ako typy kníh/autorov	porovnanie, proporcie	Stĺpcový graf, pruhový graf, „bullet graph“, koláčový, skladaný stĺpcový graf, skladaná plocha, skladaná 100% plocha a „treemap“
Jednorozmerné	jedno-číselné premenné, napr. cena knihy	distribúcia, proporcie, frekvencie	Histogram, graf hustoty a box plot
Geopriestorové	špecifická poloha označená zemepisnou šírkou a dĺžkou, regiónom, mestom, štátom alebo hranicami	lokácia, porovnanie, trendy	kartogram, bublinový graf, mapa s vyznačenými bodmi, mapa spojení

Viacrozmerné	2 alebo viac číselných premenných, napr. váha, výška a IQ	vzťahy, proporcie, porovnanie	Bodový graf, matica bodových grafov, bublinový graf, rovnobežné súradnice, radarový graf, „bullet graph“ a teplotná mapa (heatmap).
Časové rady	roky, mesiace, dni, hodiny, minúty, sekundy alebo dátumy	trendy, porovnanie, cykly	čiarový, „sparkline“, plocha, „stream graph“, bublinový graf, skladaná plocha, stĺpcový graf.
Text	jednotlivé slová alebo vety ako napr. kľúčové slová z recenzií	porovnanie, "pocit", frekvencia	„Word cloud“, proporčný plošný graf, histogram, stĺpcový graf.
matrice susednosti	kto koho kontaktuje alebo kto sa s kým pozná (v sieti)	prepojenie, vzťahy, centralita, interakcie	Orientovaný alebo neorientovaný sieťový diagram

4.5 Data warehouse

Dátový sklad je úložisko všetkých dát, ktoré môžu byť potenciálne využité pre podporu obchodného rozhodovania, zhromaždené z rôznych zdrojov do spoločného centrálneho formátu. Dátový sklad je navrhnutý tak, aby obsahoval viac informácií ako informačné systémy (pretože je schopný ukladať historické údaje na sledovanie trendov v čase) a je vyladený tak, aby umožňoval čo najefektívnejšiu distribúciu týchto informácií „správnym ľuďom“. [6]

Dátové sklady sú určené na zhromažďovanie dát, aby ich bolo možné prepojiť a získať z nich rôzne náhľady (napríklad viacdimenzionálne). Sklad poskytuje túto integráciu

a umožňuje porovnávať dátové hodnoty so súborom kritérií kvality údajov, ktoré má organizácia definované ako súčasť dátovej architektúry. Dátový sklad je teda podľa definície centrálnym úložiskom informácií v organizácii, ktoré podporujú proces rozhodovania. [6]

5. Význam vizualizácie dát

Dáta v praxi zohrávajú kľúčovú úlohu pri podnikových procesoch, v manažérskych činnostiach, ale aj v každodennom živote pri riešení životných situácií. Dáta nielen opisujú prostredie, v ktorom sa vykonávajú procesy, ale sú základom získavania informácií pre rozhodovanie. Manažérske rozhodovanie v praxi je potrebné vykonávať rýchlo na základe získaných informácií a tak nie je možné používať vyhľadávanie v rozsiahlych tabuľkách, databázach ani zložité nástroje BI, ktoré nemajú vhodné zobrazovacie nástroje pre získané informácie (napr. LISPminer, MatLab a pod.)

Množstvo informácií získaných prostredníctvom zraku je omnoho väčšie ako množstvo informácií získaných iným spôsobom. Vizualizácia dát je tak využitím prirodzených ľudských schopností na zlepšenie spracovania dát, zefektívnenie organizácie a lepšiu interpretáciu získaných informácií (poznatkov) z dát. [20]

5.1 Dôvody vizualizácie dát

V dnešnej dobe sme všetci zahltení dátami, ktoré je potrebné vyjadriť nejakým spôsobom, bez ohľadu na (pracovnú) oblasť. Úlohou vizualizácie dát je lepšie vyjadriť podnikové informácie získané z dát pomocou kombinácie grafov. Firmy sa postupne preorientovávajú z tradičného procesného manažmentu na manažment založený na dátach. Vizualizácia dát môže pomôcť analytikom komplexnejšie porozumieť dátam a získať pohľady, ktoré majú väčšiu cenu z aj obchodného hľadiska. [21]

V súčasnosti firmy narábajú s veľkým objemom dát (z logistických systémov, zber dát z GPS, výroba Just-in-Time, dáta o zákazníkoch zo systémov CRM a pod.), ktoré sa menia veľkou rýchlosťou. Napríklad CRM systémy zbierajú dáta o zákazníkoch v elektronickom podnikaní a obchodoch a sú často zbierané len v digitálnej forme. Ďalším odvetvím, ktoré narába s veľkým objemom dát je tiež zdravotníctvo. Dáta sú statickou zložkou informačných systémov a ich významom je popísať oblasť, v ktorej sa deje nejaká činnosť

(procesy, rozhodovacie úlohy, a pod.). V zdravotníctve ide o citlivé údaje, ktoré sú ukladané v rôznych formách (od textových súborov, cez obrázky z RTG až po videá, resp. multimediálne dáta) a získavanie informácií z týchto dát musí byť vysoko efektívne.

Dáta sú dôležitým vstupom pre manažérske činnosti a rôzne úlohy ako plánovanie, predvídanie, monitoring, hodnotenie, priraďovanie a podobne. Väčšina týchto procesov je založená na subjektívnom rozhodovaní, ktoré používa ako vstupy práve dáta.

Prostredníctvom BI a techník data miningu je možné z dát získavať nielen relevantné informácie pre ďalšie činnosti a rozhodovacie procesy v praxi, ale je taktiež možné na ich základe zostaviť prediktívne modely s využitím do budúcnosti, alebo klasifikačné modely (napr. klasifikácia zákazníkov, dodávateľov, či marketingových nástrojov, a pod.).

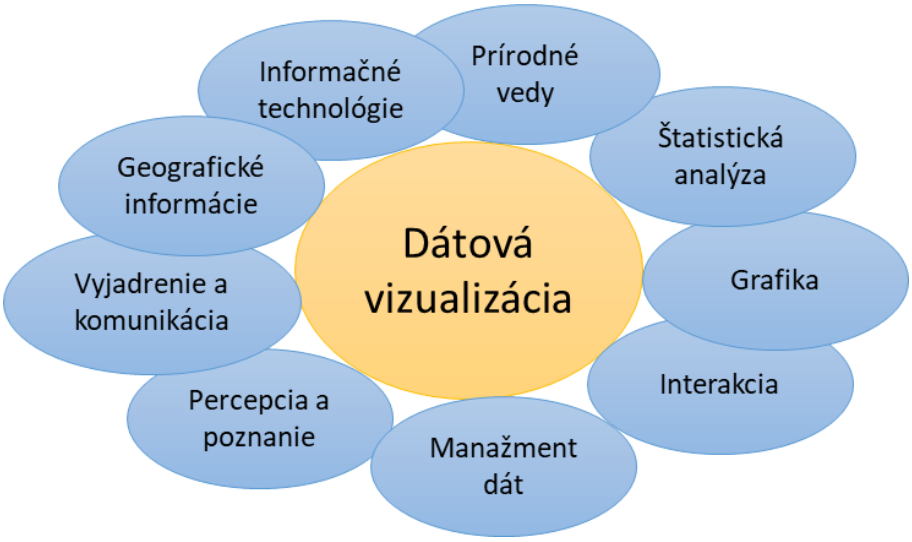
5.2 Vizualizácia dát

Vo svete dátovej vedy je v súčasnosti jednoznačne vizualizácia dát jedným z najpoužívanejších pojmov. Z tohto dôvodu je priblíženiu a objasneniu pojmu vizualizácie dát venovaná táto podkapitola.

Vizualizácia rozlišuje tri hlavné kategórie (typy), za ktoré sú považované vedecká vizualizácia, informatická vizualizácia a vizuálna analýza. Dátová vizualizácia je kombináciou týchto troch spomínaných smerov a je novým východiskovým bodom pre oblasť vizuálneho výskumu. [20]

Generalizovaná dátová vizualizácia zahŕňa rôzne disciplíny ako napríklad informačné technológie, prírodné vedy, štatistickú analýzu, grafiku a geografické informácie a schematicky tieto oblasti sú zachytené na obrázku 2 [20].

Vizualizácia ponúka vizuálnu a interaktívnu reprezentáciu dát, čo umožňuje používateľom lepšie porozumieť daným dátam. [10]



Obrázok 2 Dátová vizualizácia, spracované podľa [20]

Vizualizácia dát je jednou z metód explorácie, zobrazenia a vyjadrenia dát pomocou vizuálnej komunikácie. Vizuálne sprostredkováva rôzne druhy dát a informácií, ktoré sú základom dizajnu. S pokrokom technológií sa vizualizácia dát stala rozmanitejšou a bohatšou. [9]

Dôvody vizualizácie dát vysvetľujú autori [22] aj takto: „Vizualizovanie dát znamená organizovanie informácií na základe vhodného priestorového umiestnenia a ďalších princípov vizuálneho vnímania, ktoré podporujú utváranie percepčných záverov. Pre ľudí je pomerne jednoduché vyvodiť percepčné závery, keďže ich vizuálne vnímanie je lepšie (v porovnaní s rýchlymi programovateľnými algoritmami) a transformácia dát v ľudskej pamäti je neuveriteľne rýchla. Vizualizácia teda zvyšuje schopnosť hľadať a rozpoznávať, čím značne zvyšuje schopnosť priradenia významu [22]“

Card a kol. [23] uvádzajú, že vizualizácia môže redukovať čas potrebný na vyhľadanie dát a môže efektívne odhaliť vzory v dátach.

Vizualizácia môže efektívne kombinovať strojovú a ľudskú inteligenciu na získanie náhľadu z dát za účelom podpory informovaného rozhodovania v komplikovaných scenároch. [8]

Analyzovanie dát pomocou vizuálnych záverov je jednoduchšie a kognitívne menej náročné ako pohľad na nespracované dáta, pretože umožňuje identifikáciu blízkosti, podobnosti, kontinuity a záveru. [22]

Zásadným cieľom vizualizácie je vytvoriť špecifický náhľad alebo vykonať špecifickú úlohu zvýraznením jednotlivých vlastností data setu. Náhľadmi môže byť objavovanie trendov, korelácie, asociácie, zhluky (clusters), udalosti, ktoré umožňujú vytvorenie alebo potvrdenie hypotézy, ako aj prezentácia informácií konkrétnemu publiku rozprávaním presvedčivého príbehu podporovaného dátami na účely rozhodovania. [22]

Podľa štúdie Vicenteho [23] je vizualizácia dát medziodborová disciplína, ktorá využíva obrovskú komunikačnú silu obrázkov na poskytnutie zrozumiteľného vysvetlenia vzťahu medzi dátami. V dobe takzvaných Big Data vzrástol význam vizualizácie, keďže v každom okamihu je generované obrovské množstvo nespracovaných dát.

Pre vytvorenie alebo výber vhodných vizualizácií sú podľa Brehmera a Munznera [22] rozhodujúce tri etapy: 1) kódovanie (výber a návrh vhodných vizuálnych foriem), 2) manipulácia (umožňuje používateľovi narábať s údajmi) a 3) uvedenie (umožňuje používateľovi pridať ďalšie dáta a uložiť výsledky).

Vizualizácia sa stala dôležitým nástrojom na vysvetlenie a pochopenie veľkých a komplexných dát v mnohých oblastiach. [8]

V klasifikácií rozdeľujeme vizualizácie na základe dvoch znakov: typu dát, ktoré je možné v navrhovanej vizualizácii reprezentovať (hierarchická vizualizácia a viac-atribútová vizualizácia), a základné rozloženie (polárna vizualizácia alebo vizualizácia založená na Karteziánskej sústave súradníc). [22]

Funkcia vizualizácie dát sa odzrkadľuje najmä v dvoch aspektoch: prvým je zobrazenie dát a druhým business analýza. [21]

5.3 Súčasný stav využitia vizualizácie v praxi

Cieľom tejto podkapitoly nie je mapovať ani komentovať pandemickú situáciu v daných štátoch ani vo svete. Cieľom je poukázať na vzrast významu vizualizácie dát aj pre širokú verejnosť a ilustrovať niektoré z dobrých praktík vizualizácie ako aj niekoľkých spôsobov, ktoré môžu viesť k nepresnej interpretácií vizualizovaných dát.

V súčasnej dobe sa do povedomia širokej verejnosti dostala vizualizácia dát aj vďaka pandémií nového koronavírusu Covid-19, ktorý sa po prvýkrát vyskytol koncom roka 2019 v Číne. Potreba sledovať neustále sa meniace údaje o výskyte tohto vírusu,

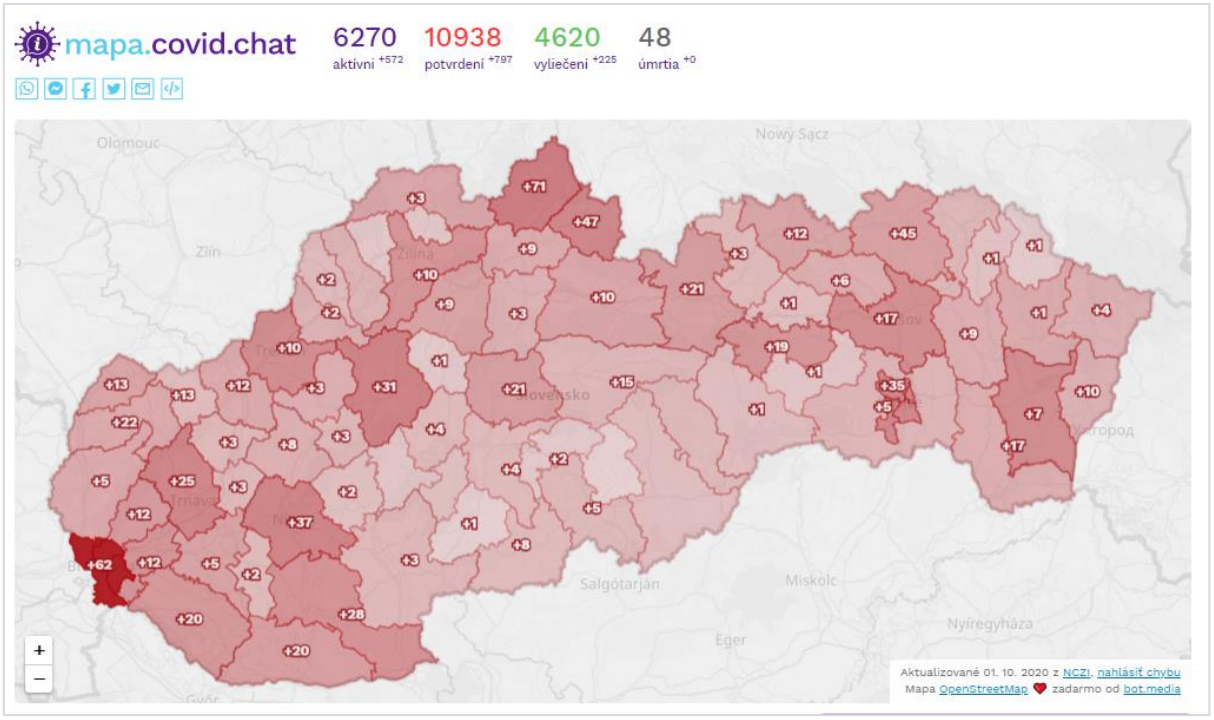
infikovaných, ale aj vyliečených pacientoch priviedla mnohé tímy na myšlienku vizualizácie už verejne dostupných dát, keďže práve pomocou vizualizácie sú tieto dáta jednoduchšie pochopiteľné aj pre laika, alebo človeka, ktorý nemá čas alebo chuť vyhľadávať dáta v rôznych tabuľkách a dokumentoch.

Každá krajina zbiera a uverejňuje svoje oficiálne dáta, na základe ktorých je postavená prevažná väčšina všetkých vizualizácií pre daný štát. Spravidla sú oficiálne štatistiky presné a často aktualizované, na rozdiel od niektorých neoficiálnych štatistík. Mnoho neoficiálnych vizualizácií je vytvorených rôznymi mediálnymi skupinami a spoločnosťami.

Množstvo dát, ktoré boli zbierané počas vrcholiaceho obdobia pandémie bolo obrovské, čo prispelo k tomu, že okrem výborných a prehľadných vizualizácií sme mali možnosť vidieť aj niektoré menej prehľadné alebo zriedka aktualizované vizualizácie.

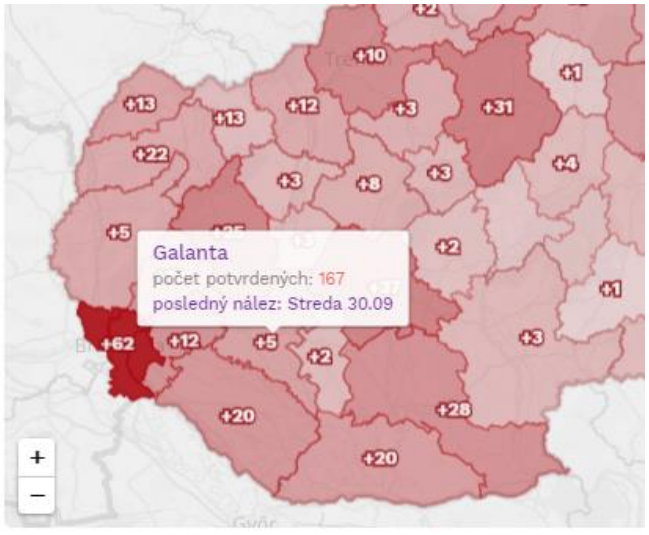
Okrem prehľadnosti a aktuálnosti dát pre dané vizualizácie je dôležité aj uvedenie kontextu, z ktorého vychádzali autori pri tvorbe vizualizácie. V niektorých prípadoch samotná vizualizácia môže byť dobre spracovaná, ale daný web neposkytuje dostatočné množstvo informácií pre pochopenie vizualizácie, čo môže viesť k rôznym interpretáciám danej vizualizácie.

Jednou z takýchto vizualizácií je stránka mapa.covid.chat. [24] Na tejto stránke je dostupná vizualizácia aktuálneho počtu pozitívne testovaných pacientov, vyliečených pacientov a úmrtí na území Slovenskej Republiky.



Obrázok 3 Vizualizácia mapa.covid.chat [24]

Vo vrchnej časti obrázku 3 je vidieť čísla pre počet aktívnych prípadov, počet potvrdených a vyliečených prípadov a počet úmrtí. Ďalej je tu mapa Slovenska s rôznymi číslami pre dané okresy zafarbené rôznymi odtieňmi červenej farby. Bez znalosti situácie, alebo aspoň približného poznania dát, z ktorých vizualizácia vychádza by bolo pomerne ťažké pochopiť, čo dané čísla vyjadrujú.



Obrázok 4 Mapa detail [24]

Na obrázku 4 môžeme vidieť, že po umiestnení kurzoru na daný okres sa zobrazia podrobnejšie informácie ako názov okresu (Galanta), počet potvrdených prípadov (167) a posledný nález v danom okrese (v stredu 30.09.2020). Stále však nie je zrozumiteľné, čo znamenajú jednotlivé plusové čísla na mape.

Avšak poznajúc situáciu v danej krajine a aspoň približné počty pacientov v danom období, sa dá jednoducho

identifikovať, že plusové čísla reprezentujú nové pozitívne prípady v danom okrese za daný deň.

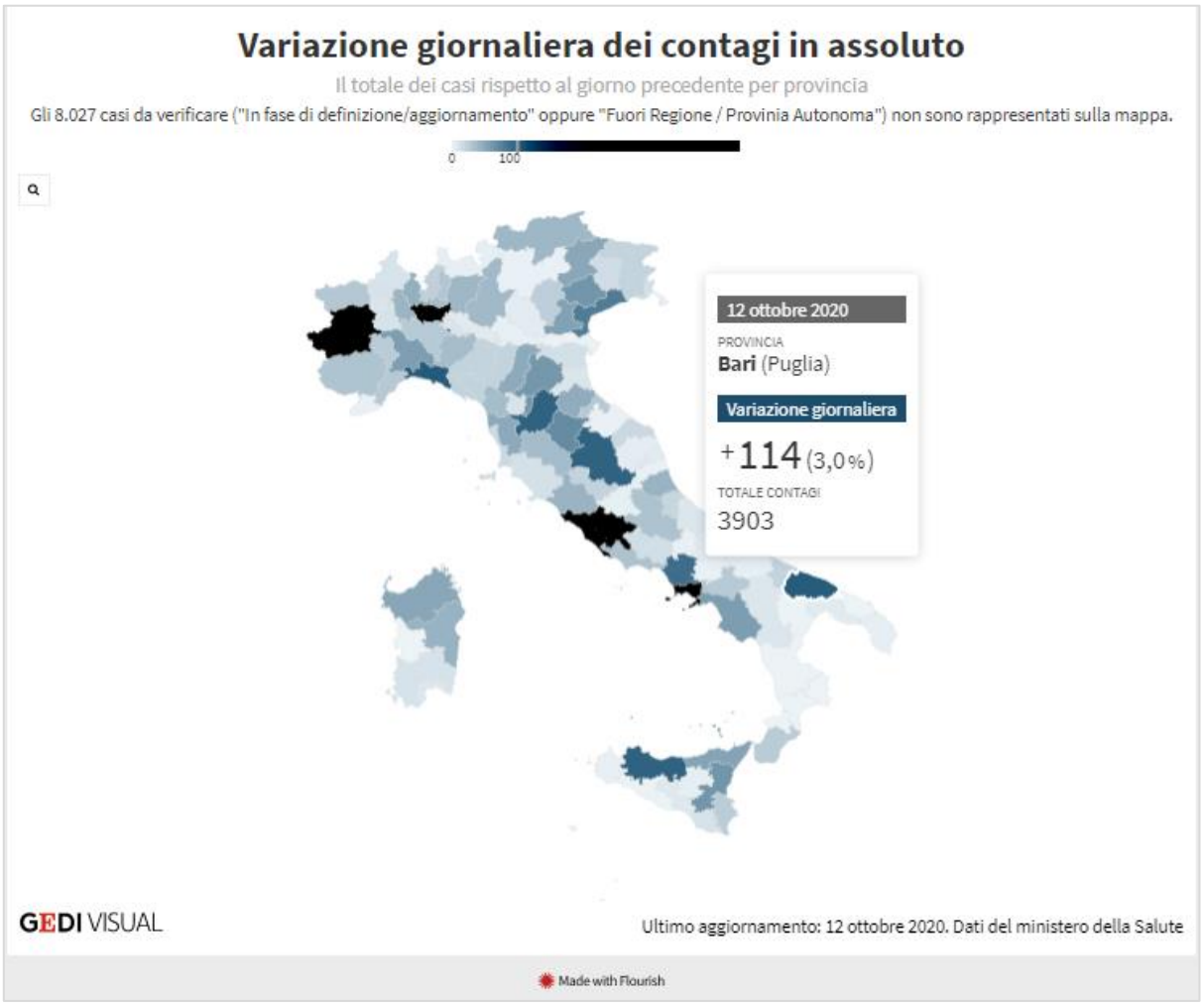
Z tejto krátkej ukážky je možné uzavrieť, že dobrá vizualizácia by mala byť nielen založená na dôverných dátach, na ich prehľadnom ich zobrazení, ale aj brať do úvahy publikum, pre ktoré je určená a tomu prispôsobiť poskytnutý kontext a vysvetlenie. Hlavne v prípade, ak daná vizualizácia nebude osobne prezentovaná publiku a teda nebude možnosť ju dodatočne okomentovať. Preto by mali vizualizácie vždy obsahovať aj prehľadnú a kompletnú legendu alebo popis jednotlivých údajov zobrazených vo vizualizácii.

Ďalším pomerne častým spôsobom, ktorým je možné skresliť informácie je prezentovanie celkových čísel bez kontextu počtu obyvateľov pre daný okres alebo štát. Zobrazenie týmto spôsobom vynecháva fakt, že dané číslo pozitívnych prípadov môže byť relatívne malé alebo veľké v závislosti na veľkosti populácie zobrazovanej geografickej oblasti. Najmä zo začiatku bolo časté uvádzanie nových prípadov týmto spôsobom. Postupne však prevažná väčšina zdrojov prešla na uvádzanie počtov nových prípadov prepočítaných na 100 tisíc obyvateľov. Zjednotením mierky sa teda uľahčilo porovnanie stavu aktuálnych prípadov v jednotlivých regiónoch, keďže pomerové údaje berú do úvahy aj veľkosť populácie a teda poskytnuté informácie sú objektívnejšie.

Na druhej strane je publikovaných aj mnoho dobre spracovaných a vhodných vizualizácií. S postupom času práve takýchto vizualizácií pribúda. Ako príklad uvádzame vizualizáciu dennej variácie prípadov v jednotlivých okresoch Talianska. [25] Táto vizualizácia je spracovaná spoločnosťou Gedi Visual [25] a zverejnená na web stránke spolu s ďalšími vizualizáciami postupu pandémie v Taliansku, medzi inými napríklad absolútna variácia pozitívnych prípadov od začiatku sledovania vírusu alebo vzťah medzi novými prípadmi a osobami, ktoré neboli nikdy predtým testované.

Napríklad Obrázok 5 ukazuje (odhliadnuc od jazyka vizualizácie) dodržanie konvencie, kde s tmavnúcou farbou okresov pribúdajú aj pozitívne prípady za deň. Po umiestnení kurzoru na vybraný okres sa zobrazí tabuľka s aktuálnym dátumom, názvom okresu s názvom regiónu, do ktorého daný okres spadá. Zobrazené sú tri číselné údaje: denná variácia pozitívnych prípadov (+114), percentuálny podiel nových pozitívnych

prípadov na celkovom počte prípadov v okrese (3,0%) a posledným číselným údajom je celkový počet pozitívnych prípadov v danom okrese (3903).



Obrázok 5 Denná variácia prípadov v Taliansku [25]

Na tejto stránke sú k dispozícií aj ďalšie veľmi dobre spracovaná vizualizácie zamerané na iné údaje ohľadom vývoja pandémie na území v Taliansku. Každá vizualizácia má okrem legendy aj krátky výstižný popis a umiestnenie viacerých vizualizácií na jednu stránku pomáha dokresliť kompletný a objektívnejší obraz sledovaného javu.

6. Nástroje na exploráciu dát

Aby bolo možné robiť kvalitné rozhodnutia nad dátami, je potrebné správne pochopiť a analyzovať dáta. Ľudská schopnosť založená na skúsenostiach a ľudské

vnímanie je jedinečné v tom, že ľudia dokážu hľadať vzory v dátach aj keď nie sú úplné. Avšak pri veľkom množstve dát to človek nezvláda. Preto potrebuje IT nástroje na prieskum dát.

„Prieskum dát pomáha analytikovi zvoliť najvhodnejší nástroj na spracovanie a analýzu dát a využíva vrodenú ľudskú schopnosť rozpoznávať v dátach vzory, ktoré analytické nástroje nemusia zachytiť.“ [26]

Explorácia teda prieskum dát je počiatočným krokom analýzy dát, v ktorom je neštruktúrovaným spôsobom skúmaný rozsiahly data set v snahe odhaliť počiatočné vzory, charakteristiky a body záujmu. [27] Cieľom tohto procesu nie je uskutočniť hĺbkovú analýzu, ale pomôcť odhaliť dôležité trendy a hlavné body na následné podrobnejšie preskúmanie. [27] Tento proces uľahčuje hĺbkovú analýzu, keďže môže pomôcť zamerať sa na určité skúmanie a začať proces eliminácie irelevantných bodov a vyhľadávaní, ktoré by skončili bez výsledku. [27] Hlavným prínosom je zoznámenie sa s existujúcimi dátami, čo umožňuje jednoduchšie nájsť presnejšie odpovede.

V explorácii dát môže byť využitá kombinácia manuálnych metód a automatizovaných nástrojov. Táto kombinácia je veľmi prínosná, keďže každý z prístupov skúma tie isté dáta z iného uhľa pohľadu. Manuálna analýza pomáha používateľovi zoznámiť sa s informáciami a môže poukázať na prítomnosť všeobecných trendov. Automatizované nástroje sú vynikajúce na vyradenie menej použiteľných dát, reorganizáciu dát do skupín, ktoré uľahčujú analýzu. [27]

Pre bežných používateľov v podniku predstavuje explorácia dát spôsob získavania znalostí z organizačných dát, čo im môže pomôcť lepšie pochopiť situáciu, odpovedať na konkrétne podnikové otázky a podporiť rozhodovací proces. [26]

Explorácia a následná analýza dát sú základom pre rozhodovanie a riadenie procesov. Pre exploráciu dát sú často využívané aj vizualizačné nástroje, keďže vizuálne spracované dáta sú istým spôsobom prehľadnejšie a niektoré ich charakteristiky sa dajú týmto spôsobom veľmi jednoducho identifikovať. Vizualizáciou dát je teda možné získať náhľady na dáta, o ktoré je možné oprieť rozhodovanie.

Pre podniky je dôležité využívať rozhodovanie na základe reálnych dát, čo prináša výsledky „šité na mieru“ danému podniku. Jednou z hlavných výhod rozhodovania na

základe dát je zníženie nákladov, keďže okrem iného je pomocou analýzy možné presnejšie určiť dopady jednotlivých rozhodnutí bez potreby ich praktického zavedenia do podniku.

6.1 Business Intelligence

Pojem Business Intelligence (BI) je už ustáleným pojmom ako v praxi tak aj vo výskume. BI zahŕňa prístupy ako zber dát, skladovanie, spracovanie, analýza a prezentovanie podnikových dát. [7]

Hlavným cieľom Business Intelligence je podpora rozhodovania, ktorá má podniku pomôcť robiť lepšie rozhodnutia. [28]

Tento pojem zaviedol analytik skupiny Gartner, Howard Dresner, v polovici 90. rokov a definoval ho ako súhrnný pojem pre koncepty a metódy pre podporu rozhodovania prostredníctvom informačnej analýzy, dodania a spracovania. Odvtedy sa pojem BI rozšíril aj do obchodnej praxe a vedy, ale stále existujú nezhody v chápaní tohto pojmu, čo vedie k rôznym definíciám. [7]

Jednou z najbežnejších definícií BI je že sú to informácie, ktoré podporujú rozhodovanie v podniku, doručené správnym ľuďom v správnom formáte a čase. [6] „V správnom čase“ znamená mať dané informácie k dispozícií v čase rozhodovania, ale aj mať informáciu v správnej hodnote. [6] „V správnom formáte“ zahŕňa aj spôsob prezentácie dát, napr. niekedy potrebuje používateľ vidieť trendy na vysokej úrovni, inokedy mu postačujú informácie na nižšej úrovni. [6] a napokon „správnym ľudom“ zabezpečuje to, aby sa potrebné informácie distribuovali ľuďom, ktorí dané informácie potrebujú na realizovanie rozhodnutia. Je preto kľúčové dobre poznať jednotlivé skupiny používateľov a rôzne úrovne zodpovednosti, ktoré nesú. [6]

Na inom mieste autori [29] uvádzajú, že BI systém je možné definovať aj ako integrovanú sadu nástrojov, technológií a programovaných produktov, ktoré sa používajú na zhromažďovanie, integráciu, analýzu a na sprístupnenie dát. V ďalšej definícií je BI definovaná ako architektúra a kolekcia integrovaných prevádzkových aplikácií a databáz slúžiacich aj pre podporu rozhodovania, ktoré poskytujú podnikovej komunite prístup k podnikovým dátam. [29] Jednou z ďalších možností ako definovať BI je, že je to proces zberu veľkého množstva dát, analýza týchto dát a prezentovanie

reportov na vysokej úrovni, ktoré kondenzujú podstatu dát pre vytvorenie základu podnikových akcií umožňujúc manažmentu uskutočňovať zásadné každodenné rozhodnutia. [29]

Zjednodušene povedané, hlavná úloha BI systémov zahŕňa inteligentnú exploráciu, integráciu, agregáciu a multidimenzionálnu analýzu dát pochádzajúcich z rôznych zdrojov. Z tohto dôvodu BI platformy ponúkajú kompletné zbierky nástrojov pre tvorbu, vývoj, podporu a údržbu BI aplikácií. [29]

V podnikaní je dôležité mať postačujúce informácie vysokej kvality a kľúčové ukazovatele výkonnosti (KPIs – Key Performance Indicators) ako základ pre rozhodovanie. Používanie Business Intelligence má tri hlavné ciele: vylepšenie rozhodovacieho základu, zvyšovanie transparentnosti podnikových akcií a poukázanie na vzťahy medzi „osamelými“ informáciami. [7]

BI umožňuje podnikom vylepšovať vývoj nových produktov poskytovaním náhľadov v oblastiach ako je napríklad dynamika zákazníkov a poskytovaním konzistentného obrazu o vnútornom stave a operáciách. [29]

Využívaním BI systémov sú podniky podporované v tvorbe transparentných a inteligentných procesov a podnikových dát. Zamestnancom je umožnené vykonávať lepšie rozhodnutia, rýchlejšie dosahovať požadované výsledky a kontinuálne ich rozvíjať. Ďalšou výhodou BI systémov je, že vďaka nim môžu podniky zlepšovať vzťah medzi zákazníkmi a dodávateľmi, znižovať náklady, minimalizovať risky a zároveň zvyšovať pridanú hodnotu. [7]

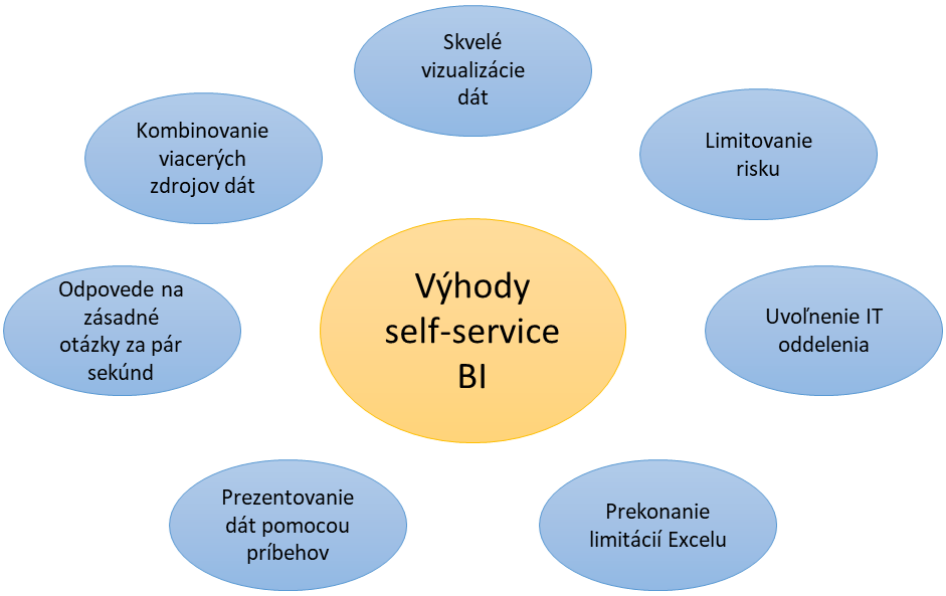
Pre dosiahnutie úspechu v BI musí podnik pestovať kultúru spolupráce naprieč celou organizáciou tak, aby všetci pracovali na dosiahnutí strategickej vízie a dobre ju pochopili. [29]

V dnešnej dobe existujú rôzne implementácie BI medzi inými napríklad BI suites (EBIS), ktoré zvyčajne poskytujú škálovateľnosť BI a dosahujú nie len na interných používateľov, ale aj na kľúčových zákazníkov, dodávateľov a niekedy až na širokú verejnosť. Ďalšími sú napríklad nástroje na dotazovanie a vykazovanie, pokročilé BI nástroje – najmä OLAP (On-Line Analytical Processing) a BI platformy pre vývoj BI aplikácií. [29]

Business inteligencia (BI), poskytovanie informácií na podporu rozhodovacieho procesu, je dosiahnuteľná, iba ak je postavená na údajoch zaručenej kvality, ktoré sú pre podnik relevantné. Aby to bolo možné, je nevyhnutný úspešný framework architektúry dát. [6]

6.2 Business Intelligence self-service tools

Existuje mnoho nástrojov a spôsobov ako vizualizovať dáta. Jedným z komplexnejších typov nástrojov sú práve self-service BI nástroje a keďže práve týmto je práca venovaná, v tejto časti práce sú bližšie popísané rozdiely medzi self-service a inými typmi nástrojov.



Obrázok 6 Výhody BI self-service nástrojov, spracovné podľa [30]

Obrázkom 6 sú ilustrované výhody, ktoré self-service BI nástroje prinášajú. Základnou výhodou, ktorú je možné odvodiť aj z názvu týchto nástrojov, je samostatnosť používateľa pri ich používaní. Po pripojení zdroja dát môže používateľ sám získať náhľady zo skúmaných dát bez pomoci od špecializovaného IT tímu alebo pracovníka.

Hlavnou výhodou je možnosť získať odpovede ohľadom dát v priebehu niekoľkých minút a uvoľnenie IT oddelenia, keďže vďaka self-service nástrojom pracovníci IT oddelenia nemusia priamo pristupovať k databázam a vytvárať dátové modely pre všetky požiadavky.

Ďalšími výhodami sú tiež možnosť kombinácie rôznych zdrojov dát a prekročenie limitov, s ktorými je možné sa stretnúť pri vizualizovaní v Exceli. Hlavnými benefitmi sú však najmä možnosť vytvoriť výborné a pútavé vizualizácie, vytvorenie dashboardu a možnosť prezentovať dáta pomocou príbehov vytvorených práve pomocou self-service BI nástrojov.

V tabuľke č. 2 je možné vidieť krátke zhrnutie základných rozdielov medzi tradičnými a self-service BI nástrojmi.

Tabuľka 2 Tradičné vs self-service nástroje [31]

Tradičné BI nástroje	Self-service BI nástroje
používateľ zadá požiadavku na IT odborníka na vytvorenie reportu alebo dashboardu	IT tím nastaví nástroj a poskytne prístup používateľom
odborník extrahuje a spracuje relevantné dáta a vytvorí dátový model	používateľ pristupuje priamo k dátam, filtruje ich za účelom vytvorenia dátového modelu
používateľ schváli report alebo dashboard, prípadne zadá požiadavky na zmenu	používateľ vylepšuje a upravuje vygenerované reporty, aby vyhovovali všetkým požiadavkám

Self-service business intelligence označuje procesy, nástroje a softvér používaný firmami, aby umožnili používateľom samostatne vybrať, filtrovať, porovnávať a analyzovať dáta bez zásahu špecializovaného a pokročilého IT tréningu. [30]

Cieľom self-service nástrojov je poskytnúť používateľom väčšiu voľnosť a zodpovednosť, keďže základom je osamostatnenie a nezávislosť používateľov pri využívaní informácií vo firme, či spoločnosti. Sebestačnosť používateľov pri narábaní s dátami vedie k decentralizácií business inteligencie v podniku. [32]

Self-service nástroje umožňujú používateľom prístup a prácu s dátami podniku aj bez hlbšej znalosti a skúseností so štatistickou analýzou, BI alebo data miningom. [33]

Podobne ako pri iných typoch nástrojov na analýzu a vizualizáciu, ani pri self-service nástrojoch neexistuje univerzálne riešenie, ktoré by maximálne spĺňalo všetky požiadavky každej úlohy vo všetkých podnikoch. Používateľom s rôznymi používateľskými rolami umožňuje vykonávať odlišné úlohy podľa typu pridelenej role. [32]

Na to, aby sme overili či je daný nástroj skutočne self-service nástrojom vhodným aj pre iných ako technických používateľov sa môžeme zamerať na niekoľko otázok:

- Dajú sa dáta jednoducho spájať? Je táto funkcionalita vbudovaná v nástroji alebo je potrebné ju vyvinúť na mieru?
- Koľko skriptovania bude potrebné na prípravu dát pre analýzu? Je na to potrebná veľmi pokročilá znalosť SQL?
- Budú obchodné oddelenia sebestačné alebo sa budú musieť stále obracať a spoliehať na profesionálne služby dodávateľa daného nástroja? [34]

Ak je odpoveď na všetky tri z týchto otázok kladná, je možné povedať, že sa jedná o takzvaný self-service business intelligence nástroj.

6.3 Tradičné BI nástroje vs. self-service nástroje

Ako už bolo vyššie spomenuté, medzi tradičnými a self-service nástrojmi existujú značné rozdiely, ktoré budú v tejto kapitole bližšie rozobrané.

Tradičné BI nástroje vyžadujú komplexné IT prostredie, priestor na skladovanie dát a takmer nepretržitú angažovanosť IT tímu.

Tradičná BI implementácia je komplexná a náročná na zdroje, zatiaľ čo self-service BI je nástroj okamžite pripravený na použitie. Je však možné nájsť viacero rozdielov ku každému z prístupov. [31]

Tabuľka 3 Porovnanie BI nástrojov, spracované podľa [31]

	Tradičné BI	Self-service BI
IT setup	Prevažne riadený IT s takmer nepretržitou angažovanosťou IT odborníkov a odborníkov na dáta. Rozvoj s viacerými komponentami, kde si každý komponent vyžaduje špecializovaný personál na implementáciu a údržbu.	IT odborníci sú potrební na počiatočnú implementáciu a následne sa venujú požiadavkám na infraštruktúru. Vyžaduje podstatne menší počet IT personálu na údržbu.
Agilita	Prístup vyhradený pre IT personál a odborníkov na dáta. Príležitosti trhu môžu ostať zacyklené na týždne až mesiace v požiadavkách a reportoch.	Používatelia môžu robiť analýzu dát a generovať reporty v reálnom čase. Môžu testovať dáta a korelácie vďaka vytváraniu dátových modelov tzv. "za pochodu".
Typ dát	Potreba štruktúrovať dáta pred ich použitím	Využitie dát v rôznych formátoch a z rôznych zdrojov.
Typ reportovania	Zamerané na zodpovedanie otázok z minulosti alebo súčasnosti. Obmedzené poskytovanie reportov na požiadanie.	Okrem reportingu z minulosti poskytuje tiež prediktívny a normatívny reporting. Rôzne možnosti reportingu na požiadanie.
Správa dát	Silná angažovanosť IT personálu a odborníkov na dáta pri čistení, správnom ukladaní a bezpečnosti dát, celkové riešenie správy dát.	Potreba politiky správy dát na definovanie procesov pre čistenie a skladovanie dát, rôzne oprávnenia pre prístup k dátam.

Aj z tohto krátkeho porovnania v tabuľke 3 je možné vidieť, že každý z dvoch porovnávaných typov BI nástrojov má svoje výhody a určite aj svoje využitie. S postupom doby, pokrokom technológií a rapídnym nárastom objemu produkovaných dát sa však rýchlo dostávajú do popredia práve self-service nástroje, keďže vďaka nim má prístup k analýze a vizualizácií dát väčšie množstvo ľudí a podnikov a teda je možné viac uplatňovať rozhodovanie na základe dát.

7. Praktická časť - CRISP

Táto kapitola je venovaná praktickej časti práce, kde boli vybrané a porovnané dva self-service nástroje. Postup vypracovania praktickej časti čiastočne sleduje metodológiu CRISP.

7.1 Atribúty self-service nástrojov BI

Pre tak komplexné nástroje akými sú práve self-service BI nástroje je možné zvoliť celú radu kritérií pre porovnanie. Napríklad spoločnosť Gartner [35] hodnotí nástroje s ohľadom na 15 vybraných kritérií. Pre potreby tejto práce boli však kritériá zosumarizované a vybraté 3 kritériá, ktoré spolu poskytujú dostatočný rozsah porovnania väčšiny základných funkcií vybraných nástrojov.

Jedným z prvých krokov každej analýzy dát je príprava a čistenie dát, preto je prvým kritériom pre porovnanie práve dostupnosť prípravy a čistenia dát vo vybraných nástrojoch. Porovnanie je zamerané hlavne na spôsob, akým čistenie a príprava prebiehajú, či je možné dáta upraviť priamo v nástroji, alebo či je potrebná úprava pomocou iného nástroja alebo aplikácie.

Ďalším kritériom pre porovnanie je rozmanitosť ponúkaných typov vizualizácií a grafov pre rôzne typy dát. V tejto oblasti sú predpokladané iba minimálne rozdiely, keďže vizualizácie a grafy sú jednou zo základných zložiek self-service BI nástrojov.

Záverečným kritériom pre porovnanie je dostupných analytických nástrojov a konkrétna ponuka štatistických ukazovateľov a metód, ktoré sú v nástrojoch použiteľné. Pri hodnotení tohto kritéria sú brané do úvahy nielen počet dostupných metód a ukazovateľov, ale aj presnosť (napríklad pri prognóze budúceho vývoja) a dostupnosť vysvetlenia daných metód.

7.2 Výber vhodných nástrojov na komparáciu podľa zvolených atribútov

Existuje mnoho rôznych BI self-service nástrojov a žiaden z nich nie je možné univerzálne vyhlásiť za najlepší. Každá firma má špecifické požiadavky na nástroj, ktorý by chcela využiť alebo využíva a teda potrebám každej firmy najviac vyhovuje iný nástroj. Aj z tohto dôvodu je možné nájsť na internete mnoho hodnotení a odporúčaní typu „10 najlepších BI nástrojov“. Nie všetky sú však dostatočne objektívne a preto bol za relevantný a overený zdroj hodnotení vybraný zoznam najlepších self-service nástrojov zostavený spoločnosťou Gartner. [35]

Gartner vydáva každoročne takzvaný Magický kvadrant analytických a Business Intelligence platforiem (zobrazený na obrázku 7). Ako vyplýva z názvu, nejedná sa o rebríček, ale o umiestnenie najznámejších BI self-service nástrojov do 4 kvadrantov podľa schopnosti výkonu a kompletnosti vízie. Všetky vybrané nástroje sú teda rozdelené na „niche players“, „visionaires“, „challengers“ a „leaders“.

Podľa Gartneru sa už ABI platformy (Application Binary Interface) nerozlišujú primárne na základe ich vizualizačných vlastností, pretože tieto sa stali ich neodmysliteľnou súčasťou. Diferenciácia sa presunula na rozšírenú analytiku a integrovanú podporu podnikového reportingu. Gartner taktiež uvádza zoznam 15 kritických oblastí, podľa ktorých hodnotí spôsobilosť jednotlivých platforiem. Týmito oblasťami sú: bezpečnosť, ovládateľnosť, cloud, prepojiteľnosť dátových zdrojov, príprava dát, komplexnosť modelu, katalogizácia, automatizované výstupy (Automated insights), pokročilá analytika, vizualizácia dát, dopytovanie v prirodzenom jazyku (natural language query), „data storytelling“, vstavaná analýza (embedded analytics), natural language generation (NGL) a reporting. [35]

Obrázok 7 Magický kvadrant od spoločnosti Gartner [35]

Pre finálne porovnanie boli vybrané dva nástroje z kategórie lídrov. Výber oboch nástrojov z rovnakej kategórie zaručuje objektívne porovnanie, keďže o všetkých nástrojoch v danej kategórií je možné predpokladať, že sa nachádzajú na veľmi podobnej úrovni zrelosti produktu a rozšírenosti na trhu.

7.2.1 Microsoft Power BI

Ako prvý bol vybraný Microsoft Power BI z kvadrantu lídrov, ktorý sa nachádza priamo na pozícií lídra. Túto pozíciu sa mu darí udržať nielen vďaka komplexnému a vizionárskemu produktu, ale aj vďaka širokému dosahu na trhu prostredníctvom Microsoft Office. [35]

Microsoft Power BI ponúka prípravu dát, objavovanie znalostí na základe vizualizácií, interaktívne dashboardy a rozšírenú analytiku. Power BI je dostupný v SaaS verzii bežiacej na Azure cloud alebo v on-premises verzii na Power BI Report serveri. Power BI Desktop môže byť použitý ako samostatný personálny nástroj pre analýzu. Inštalácia Desktop verzie je však potrebná v prípade, ak experti zložito spájajú rozličné zdroje, ktoré si vyžadujú zapojenie/využitie lokálnych zdrojov dát. Microsoft týždenne

aktualizuje cloudové služby, pre ktoré za rok 2019 vydal stovky vylepšených funkcií ako napríklad prepojenie s LinkedIn alebo vylepšené geografické mapovanie. [35]

7.2.2 Tableau

Pre porovnanie s MS Power BI bol vybraný nástroj Tableau, keďže sa jedná o nástroj podobnej vyspelosti a popularity.

Tableau je taktiež Gartnerom zaradené do kvadrantu Lídrov. Ponúka vizuálne založenú exploráciu, ktorá sprístupňuje používateľom prípravu, analýzu a prezentáciu dát. Tableau má veľmi silný marketing a je rozšíreným produktom na trhu.

V poslednej dobe Tableau značne rozšírilo ponuku svojich produktov, najmä v oblasti analýzy a riadenia. Boli zavedené „Ask Data“ a „Explain Data“, ktoré poskytujú dotazovanie v prirodzenom jazyku a automatické prehľady (insights). Taktiež boli pridané serverové add-ons, ktoré umožňujú management serveru, migráciu obsahu a optimalizáciu rozloženia pracovnej záťaže. Tableau tiež presunulo značnú časť svojich zákazníkov na cloud pomocou Tableau Online. [35]

7.3 Použitie metodiky CRISP DM

Táto podkapitola je venovaná samotnému praktickému porovnaniu vybraných nástrojov, výberu data setu vhodného pre ilustráciu funkcionality nástrojov a podrobnejšiemu popisu funkcií ponúkaných vybranými nástrojmi. Ako bolo uvedené v predchádzajúcej podkapitole, vybrané boli nástroje Tableau a Microsoft Power BI, konkrétne boli využité Desktop verzie oboch nástrojov.

Najprv bude v rámci fázy Pochopenie dát z metodiky CRISP popísaný postup a uvažovanie pri hľadaní a výbere vhodného data setu, na ktorom by bolo možné dostatočne predviesť všetky základné a podstatné funkcie vybraných self-service nástrojov.

Ďalej bude popísaný konkrétny spôsob a nástroj (v prípade Microsoft Power BI je súčasťou nástroja), v ktorom môže byť zdroj dát upravený a očistený od chybných hodnôt tak, aby bol pripravený na efektívne využitie pre nasledujúcu analýzu a vizualizáciu. Tento krok zodpovedá druhej fáze metodiky CRSIP a to Príprave dát.

Následne bude popísaný spôsob tvorby vizualizácií v jednotlivých nástrojoch a taktiež budú porovnané a stručne popísané analytické funkcie ponúkané každým z nástrojov.

Na záver podkapitoly budú zhodnotené rozdiely medzi vybranými self-service nástrojmi, čo zodpovedá fáze Vyhodnotenia.

7.3.1 Pochopenie dát

Keďže porovnanie vybraných self-service nástrojov nemá reálny cieľ nasadenia zmien do podniku, ktoré by priniesli pozitívne zmeny, bola vynechaná aj fáza Pochopenia cieľa, ktorá by v ideálnom prípade použitia metodiky CRISP otvárala celý projekt.

Na vytvorenie vhodných vizualizácií a dobrej analýzy dát je potrebné dobre poznať a pochopiť spracovávané dáta, preto je prvou a dôležitou fázou práve Pochopenie dát, ktoré v tomto prípade zahŕňa aj popis požiadaviek na finálny data set, keďže dáta nie sú limitované žiadnym podnikom.

Na to, aby mohlo byť porovnanie uskutočnené bolo potrebné nájsť a vybrať vhodný data set. Pri hľadaní vhodného data setu bolo potrebné vziať do úvahy mnoho aspektov, ktorých naplnenie daným data setom by umožnilo čo najlepšie predviesť a ilustrovať možnosti využitia oboch nástrojov v praxi. Medzi tieto aspekty patria okrem iných napríklad: množstvo a kvalita dát obsiahnutých v data sete, vhodnosť témy dát, dobrá spracovateľnosť a vizualizácia dát.

Je dôležité, aby set obsahoval dostatočné množstvo dát, ideálne niekoľko tisíc záznamov, na získanie reprezentatívnej vzorky. Vďaka tomu je možné objektívnejšie vypočítať štatistické ukazovatele a vyvodiť závery, ktoré sa viac približujú k realite v danom sektore alebo oblasti.

Pri hľadaní vhodného data setu bola tiež braná do úvahy kompletnosť zozbieraných dát, keďže chýbajúce hodnoty negatívne ovplyvňujú spracovateľnosť a dôveryhodnosť data setu. Z toho dôvodu bolo potrebné vybrať data set, v ktorom aj po odstránení záznamov s prázdnymi alebo neplatnými hodnotami ostane dostatočné množstvo ostatných údajov na vytvorenie zaujímavej a objektívnej analýzy.

Vybrané self-service BI nástroje sú primárne vyvíjané za účelom efektívnej analýzy a vizualizácie podnikových dát. Na podnikových dátach je teda najlepšie vidieť všetky

funkcie a možnosti daných nástrojov. Z pochopiteľných dôvodov, medzi ktoré patria v prvom rade najmä GDPR a obchodné tajomstvo, však podnikové dáta nie sú voľne dostupné v dostatočnom rozsahu.

Data set bol hľadaný na stránke kaggle.com, ktorá ponúka po registrácií možnosť publikovať vlastné data sety, prehliadať data sety zverejnené inými používateľmi alebo prispieť k riešeniu úloh zadaných používateľmi portálu. Zdroje dát môžu byť zverejnené v rôznych formátoch ako napríklad čiarkami oddelených textových súboroch typu .csv, JSON, SQLITE, alebo v iných formátoch. Najčastejšie publikovaným formátom zverejňovaných data setov je práve formát .csv, v ktorom je aj data set použitý pre účely tejto práce.

Napokon bol zvolený data set zaoberajúci sa rezerváciami hotelov. [36] Tento data set bol pôvodne zverejnený ako súčasť vedeckého článku tímu portugalských autorov, ktorý bol uverejnený spoločnosťou Elsevier v časopise s názvom Data in Brief. [37] V tomto data sete nájdeme údaje o týchto faktoch:

- o aký typ hotela sa jedná,
- kedy bola rezervácia vytvorená,
- počet dospelých osôb a detí pre danú rezerváciu,
- zvolený typ stravovania,
- či sa jedná o nového zákazníka,
- či bola daná rezervácia zrušená,
- rezervovaný typ izby a reálne pridelený typ izby,
- počet zmien prevedených pre jednu rezerváciu,
- počet požadovaných parkovacích miest,
- počet špeciálnych požiadaviek pre každú rezerváciu, a ďalšie.

Data set obsahuje cez 110 tisíc záznamov o rezerváciách a vyžadoval si určitú mieru čistenia a úprav pre využitie v tejto práci. Prevažná väčšina záznamov bola kompletná, no niektoré záznamy obsahovali prázdne hodnoty alebo chybne zapísané hodnoty v niektorých poliach. Pole s číslom firmy bolo kompletne odstránené, keďže obsahovalo viac ako polovicu prázdnych hodnôt a zvyšné hodnoty tiež nemali žiadnu výpovednú

hodnotu, keďže kvôli GDPR boli názvy firiem konvertované iba na číselný údaj, ktorý neposkytuje žiadne ďalšie informácie o danej firme.

Dáta sú obsiahnuté v textovom .csv súbore, ktorého totožné kópie boli použité pre Tableau ako aj pre Microsoft Power BI. Tento postup bol zvolený z toho dôvodu, že každý zo self-service nástrojov poskytuje iné možnosti a postupy spracovania a čistenia dát pred vizualizáciou.

Pred samotným spracovaním v self-service nástrojoch boli prevedené menšie úpravy v pôvodnom textovom .csv súbore pomocou Excelu. V súbore bol dátum plánovaného príchodu každej rezervácie rozdelený na tri samostatné polia – deň, mesiac a rok. Tableau ani Microsoft Power BI však nie sú schopné pracovať s týmito dátami ako s dátumom, ak sa nenachádzajú v jednom poli. Tableau ani Microsoft Power BI neposkytujú možnosť zlúčenia viacerých stĺpcov so zachovaním hodnôt zo všetkých stĺpcov, preto boli tieto dáta spojené v Exceli pomocou vzorca CONCAT(), čím sa zmenšil počet potrebných polí z troch na jedno pole pre dátum, z ktorého sú nástroje efektívne schopné rozoznať a pracovať s dátumom aj s jednotlivými rokmi, mesiacmi alebo dňami.

Po zhliadnutí možností úpravy dát v jednotlivých self-service nástrojoch, boli v Exceli ešte pridané polia „totalGuests“ pre počet všetkých hostí v jednej rezervácií a „underage“ pre počet maloletých hostí, keďže v pôvodnom súbore boli maloletí hostia rozdelení na deti a batoľatá, ale pre potreby vizualizácie bolo vhodnejšie mať tieto dve polia spočítané do jedného. Ostatné úpravy čistenia a prípravy dát boli prevedené v jednotlivých nástrojoch, tj. v Tableau a Microsoft Power BI.

Nástroj Tableau Prep, ktorý bude bližšie popísaný v nasledujúcej podkapitole, obsahuje veľmi užitočnú funkciu pre lepšie pochopenie dát. Okrem bežného náhľadu záznamov jednotlivých polí, je možné zobraziť mini vizualizácie záznamov v podobe pruhových grafov jednotlivých hodnôt daného poľa. Veľmi to sprehľadňuje spracovávané dáta, na prvý pohľad je možné vidieť, ktoré polia obsahujú nadmerné množstvo prázdnych hodnôt a teda je možné ich hneď odstrániť. MS Power BI podobnú vizualizáciu neponúka.

7.3.2 Príprava dát

Táto kapitola je venovaná ďalšej fáze z metodiky CRISP a to Príprave dát, ktorá zahŕňa všetky aktivity spojené s čistením a prípravou dát do finálnej podoby data setu. V tejto práci prebehla táto fáza dvakrát, keďže jedným z cieľov porovnania je aj porovnať celý proces v každom z vybraných self-service nástrojov. Túto fázu bolo potrebné vykonať dvakrát z totožného počiatočného data setu, aby bolo možné objektívne porovnať rozhranie aj ponúkané funkcie čistenia a prípravy dát v oboch nástrojoch. Ako prvý bude opísaný proces čistenia a rozhranie nástroja Tableau a následne popísané rovnaké úpravy čistenia a prípravy data setu v nástroji MS Power BI.

Tableau

Samotný nástroj Tableau ponúka niekoľko možností prípravy dát ako súčasť nástroja. V tejto časti je možné prepojiť zdroje dát a vybrať a spojiť dáta z rôznych zdrojov alebo jednotlivých tabuliek zo zdrojových súborov. Ďalej je možné premenovať polia (stĺpce) tabuliek, zmeniť dátové typy, vytvoriť vypočítané pole a rozdeliť hodnoty. Nie je však možné vykonať niektoré úpravy ako napríklad vypustenie nepotrebných stĺpcov (polí).

Na tieto a ďalšie komplikovanejšie zmeny v príprave a čistení dát poskytuje Tableau samostatný nástroj s názvom Tableau Prep, špeciálne vytvorený za účelom prípravy a čistenia dát. Tento nástroj je automaticky k dispozícii všetkým používateľom s Tableau licenciou teda vrátane študentov využívajúcich Tableau na rok zadarmo. Okrem funkcií, ktoré ponúka Tableau Desktop je tu dostupné veľké množstvo ďalších funkcií. V nástroji je možné vytvoriť „flow“ (tok), ktorý vizuálne mapuje každý krok spracovania dát. Je možné spojiť rôzne množstvo zdrojov dát v ľubovoľnom bode toku. Všetky zmeny sú prevedené až po spustení toku, ktorý zmeny priamo prevedie do výstupného súboru. V Tableau Prep je možné pridať kroky so špecifickými úlohami pre čistenie, agregáciu dát, „pivot“, tj. zamenenie stĺpcov a riadkov, spojenie dát, skript alebo output. Každý typ kroku je odlíšený inou farbou. Tiež je možné spojiť niekoľko rôznych tokov.

Obrázok 8 Tok v Tableau Prep [vlastné spracovanie]

Na obrázku 8 je možné vidieť malú ilustráciu ako vyzerá vytvorený tok prípravy dát. Keďže použitý zdroj dát obsahoval veľmi málo hodnôt, ktoré by si vyžadovali čistenie, tok obsahuje iba dva kroky a výstup. V prvom kroku boli špecifikované dátové typy niektorých polí (hlavne dátumu a skratky krajín), ktoré automatická identifikácia určila nepresne. V druhom kroku boli odstránené polia s vysokým počtom hodnôt „null“, ktoré by z tohto dôvodu boli nepoužiteľné. Tiež boli odstránené tri polia, ktoré obsahovali hodnoty pre deň, mesiac a rok rezervácie zvlášť, ale boli nahradené jedným poľom so štandardným formátom dátumu v predchádzajúcej úprave data setu pomocou Excelu.

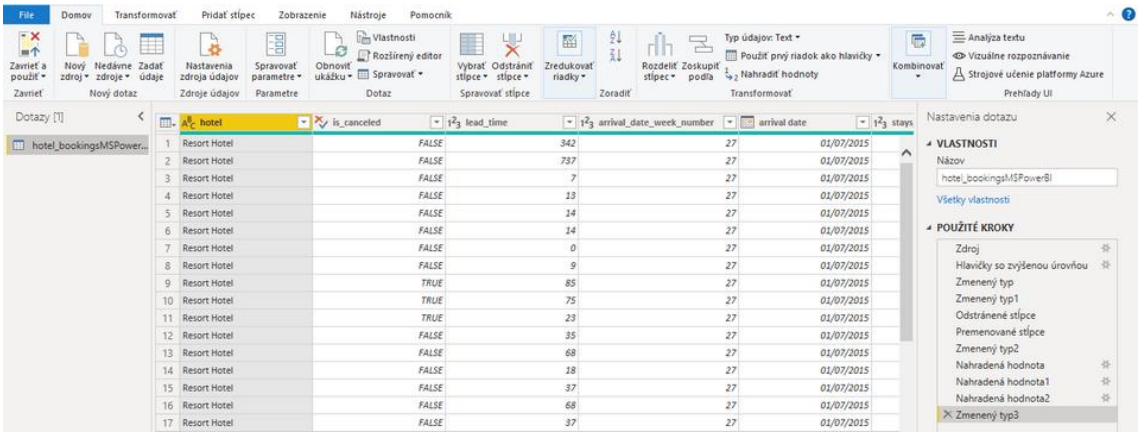
Počas spracovávania vizualizácie bolo zistené, že Tableau je schopné generovať a umiestniť krajiny podľa trojpísmenových skratiek vytvorených na základe ISO 3155–3:2013, ale nie je schopné ich rozdeliť na skupiny podľa kontinentov. Z tohto dôvodu bolo pridané ďalšie pole, ktoré explicitne priraďuje ku každej krajine kontinent, na ktorom sa nachádza. Toto rozdelenie umožnilo porovnať množstvo hostí na jednotlivých kontinentoch a zároveň umožnilo bližšie zameranie sa na Európu odkiaľ pochádza veľká časť zozbieraných dát.

Power BI

Ako už bolo vyššie spomenuté, aj pre túto vizualizáciu bola použitá identická kópia zdroja dát ako pre prácu s nástrojom Tableau a to verzia originálneho data setu po upravení polí pre dátum a pridaní polí pre celkový počet hostí v rezervácií a pre počet maloletých hostí. Podobne ako v nástroji Tableau je možné prepojiť veľké množstvo zdrojov dát od statických súborov formátu .csv, tabuliek z Excelu, XML, JSON alebo pdf až po databázy (SQL Server, IBM, Oracle, MySQL, Access, Amazon Redshift, a ďalšie), rôzne online služby alebo priamo prepojiť web stránky cez URL adresu. Microsoft Power BI tiež poskytuje vlastnú Power Platform, z ktorej je možné prepojiť data sety, toky dát (dataflows), a spoločné dátové služby. Na túto platformu je tiež možné publikovať používateľom vytvorené reporty jedným klikom priamo z domácej obrazovky nástroja Power BI.

Na rozdiel od Tableau, Microsoft Power BI obsahuje zabudovaný editor dát, v ktorom je možné dáta očistiť a transformovať do finálnej podoby. Po prepojení zdroja dát je možné priamo prepnúť do nového okna editora dát. Tento editor poskytuje rozhranie

vizuálne veľmi podobné klasickému rozhraniu, aké poskytuje Microsoft Excel. Toto rozhranie je zobrazené na obrázku 9 a bude bližšie popísané aj s vysvetlením jeho funkcií.



Obrázok 9 Rozhranie Power BI editora dát [vlastné spracovanie]

V ľavej časti sa nachádza záložka, na ktorej sú vypísané názvy tabuliek, z ktorých pochádzajú aktuálne upravované dáta. Uprostred okna je zobrazený náhľad dát z vybranej tabuľky, ktorá bola zvolená v ľavej časti obrazovky. Podobne ako v klasickej verzii Excelu sú dáta zobrazené v tabuľke s názvami v hlavičke v prvom riadku. Pri názve každého poľa (stĺpca) je naznačený typ dát, ktoré toto pole obsahuje. Toto označenie typov dát napomáha k lepšej prehľadnosti najmä pri práci so zdrojmi dát obsahujúcimi veľké množstvo rôznych typov dátových polí. Na druhej strane názvu je podobne ako v Exceli tlačidlo, ktorým je možné zobraziť možnosti filtrovania a usporiadania dát v danom poli. Na rozdiel od filtrovania v Exceli, kde sú nepotrebné odfiltrované dáta ukryté, Power BI tieto nepotrebné dáta vymaže.

Na pravej strane obrazovky je možné zobraziť záložku s nastaveniami dotazu/tabuľky. Ďalej je tu možné vidieť prehľad všetkých aplikovaných krokov na úpravu a čistenie dát. Z tohto zobrazeného zoznamu krokov je možné jednotlivé kroky zrušiť a niektoré typy krokov ako napríklad filter určitých dát, je možné upraviť. Vymazanie kroku, po ktorom boli následne vykonané ďalšie kroky môže narušiť alebo ovplyvniť celý proces úpravy dát.

V hornej časti obrazovky sa nachádza lišta s prevažnou väčšinou všetkých ponúkaných funkcií. Pre účely tejto práce nie je potrebné zaoberať sa každou funkciou zvlášť a preto budú popísané v skupinách podľa toho, čomu sa jednotlivé funkcie venujú.

Z tejto časti používateľského rozhrania je možné pristupovať okrem iných aj k týmto funkciám:

- aplikovanie zmien a uzavretie editora, ktoré sú jednými z najdôležitejších funkcií tejto časti rozhrania. Na rozdiel od iných editorov je nevyhnutné vykonané zmeny aplikovať a až po aplikovaní sa prevedené zmeny zobrazia aj mimo editora, aplikovaním zmien z editora nie je ovplyvnení pôvodný zdroj dát.
- prepojenie ďalších zdrojov dát,
- upravenie súčasných prepojení dátových zdrojov,
- spravovanie jednotlivých stĺpcov a riadkov dát,
- možnosť rozdelenia hodnôt do skupín, podľa nastaviteľných kritérií,
- možnosť spustenia skriptov programovacieho jazyka R alebo jazyka Python,
- možnosť nahradenia hodnoty vybraných záznamov, a ďalšie.

Podobne ako pri spracovávaní dát v nástroji Tableau Prep, boli odstránené nepotrebné polia, ktoré obsahovali príliš veľa prázdnych hodnôt, alebo neboli dostatočne dobre využiteľné pri analýze v kombinácií s inými dátami. Na rozdiel od Tableau, Power BI do inej miery správne identifikoval dátové typy jednotlivých polí.

V MS Power BI je možné priamo pole celých čísel obsahujúce iba hodnoty 0 a 1 transformovať na pole typu pravda/nepravda (boolean). MS Power BI mal však mierne problémy s automatickou identifikáciou iných dátových typov. Z pôvodného súboru .csv, kde boli typy všetkých polí nastavené na „Všeobecné“, ich Power BI prevzal, ale identifikoval všetky ako text aj v prípade kedy dané pole obsahovalo čiste číselné hodnoty. Pre tieto dátové polia, obsahujúce číselné hodnoty alebo dátum, bolo preto nutné upraviť ich dátový typ manuálne. Pri manuálnej zmene typu dát došlo k chybám, pretože Power BI nekonvertuje automaticky iné ako číselné znaky zadané do číselného poľa na prázdne hodnoty. Bolo teda potrebné manuálne filtrovať nečíselné hodnoty a nahradiť ich prázdnou hodnotou „null“ a až po týchto krokoch bolo možné dané dátové pole úspešne konvertovať na typ celých čísel. Ako posledná bola vykonaná zmena typu poľa s dátumami, taktiež z pôvodne identifikovaného textového typu.

Rovnako ako Tableau ani MS Power BI nie je schopný identifikovať kontinenty jednotlivých krajín, ale je schopný všetky krajiny zobraziť na mape na základe ich trojpísmenových skratiek. Rozdelenie na kontinenty bolo preto pridané manuálne do pôvodného súboru rovnako ako pri príprave dát v Tableau.

Po vykonaní týchto krokov boli zmeny aplikované a spracovanie dát pokračovalo v samotnom MS Power BI mimo editačného okna.

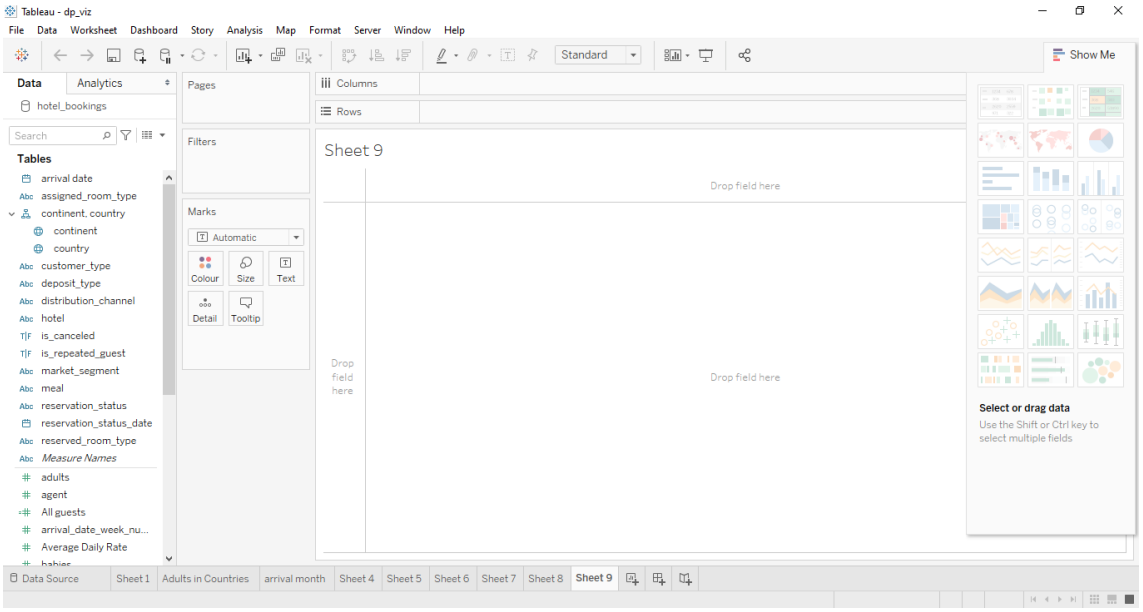
7.3.3 Modelovanie

Podľa metodiky CRISP je v tejto fáze vytvorený model. Súčasťou modelovania je aj vizualizácia a analýza skúmaných dát, v ktorej je možné využiť aj štatistické modely ako napríklad časové rady alebo zhlukovanie. Preto v tejto podkapitole bude popísaný spôsob vizualizácie v jednotlivých nástrojoch a následne budú popísané analytické funkcie ponúkané porovnávanými nástrojmi.

Vizualizácia v Tableau

V tejto podkapitole bude bližšie popísaný proces vizualizácie v nástroji Tableau. Bude tu priblížené používateľské rozhranie, ktoré Tableau ponúka ako aj výber grafov, ktoré sú v tomto nástroji k dispozícií.

Pri práci s akýmkoľvek nástrojom je potrebné sa najprv oboznámiť s prostredím nástroja. Za týmto účelom poskytuje Tableau širokú škálu e-learningových materiálov od stručných videí, ktoré zbežne prezentujú základné funkcionality nástroja po podrobné školiace materiály zložené z inštruktážnych videí a zdrojov dát ku cvičeniam ako aj pracovné zošity s vypracovanými problémami z inštruktážnych videí.



Obrázok 10 Rozhranie nástroja Tableau Desktop [vlastné spracovanie]

Obrázok 10 zobrazuje rozhranie nástroja Tableau Desktop. Pre vytvorenie vizualizácie je potrebné prepojiť Tableau a zdroj dát, ktorý môže byť buď statický súbor napríklad tabuľka v Exceli, textový súbor, pdf súbor a ďalšie. Alternatívne, môže byť použitý dynamický zdroj dát zo servera, kde je možné pracovať s automatickým obnovovaním dát v reálnom čase, tak ako sa menia v databázy na serveri. Dostupné sú prepojenia so všetkými bežnými servermi ako napríklad MySQL, Microsoft SQL Server, Oracle alebo Amazon Redshift, ale dostupné je aj množstvo iných serverov a online zdrojov.

Na ľavej strane sa nachádza záložka s názvami polí, ktoré Tableau automaticky identifikuje z dátového zdroja, ale používateľ ich môže modifikovať pri prepájaní dát alebo aj neskôr. Tableau rozdeľuje dáta na dimenzie a miery (Dimensions and Measures). Miery sú čisto číselné dáta a je možné doplniť rôzne vypočítané polia s použitím už existujúcich polí a hodnôt. V dimenziách sú potom obsiahnuté ostatné dátové typy ako text, dátum, boolean, alebo geografická poloha. Na rozdiel od mier, je možné dimenzie usporiadať do hierarchií jednoduchým presunutím názvu jednej dimenzie na názov inej dimenzie. Takto vytvorené hierarchie je následne možné použiť v grafoch. Napríklad dátum samotný funguje v Tableau ako hierarchia. V predvolenom nastavení sú dáta zobrazené v grafe rozdelené podľa rokov, ale keďže sa jedná o hierarchiu, môže byť rozbalená na štvrťroky, následne na mesiace alebo aj na

jednotlivé dni. Týmto spôsobom je možné v grafe rozbaliť všetky stupne každej hierarchie. Ak však niektorý stupeň hierarchie nie je potrebný, alebo zahlcuje prehľadnosť vizualizácie, je možné daný stupeň zmazať z grafu bez ovplyvnenia zobrazenia ostatných stupňov danej hierarchie.

V tej istej časti rozhrania sa po kliknutí na tlačidlo „Analytics“ zobrazí ponuka dostupných štatistických metód, ktoré budú bližšie popísané v jednej z nasledujúcich podkapitol.

Najväčšia plocha je venovaná na zobrazenie priamo danej vizualizácie, ktorú je možné jednoducho vytvoriť pomocou pretiahnutia názvu polí do tejto plochy. Tableau automaticky identifikuje, ktorý graf je najčastejšie používaný pre zobrazenie daného typu dát a tento typ grafu následne implementuje. Napríklad, v prípade číselných dát Tableau vytvorí stĺpcový graf pre dané pole dát; pre geografické údaje (najčastejšie v podobe názvov krajín alebo regiónov) je vytvorená mapa s označením krajín, ktoré sa nachádzajú v dátovom poli zdrojového súboru. Samozrejme je možné použiť iný typ grafu podľa preferencií používateľa. Všetky grafy je možné vizuálne prispôsobiť cieľu, ktorý chce používateľ podložiť danou vizualizáciou. Je teda možné zmeniť použité farby, veľkosti typ a veľkosť písma pre grafy aj pre nadpisy, atď.

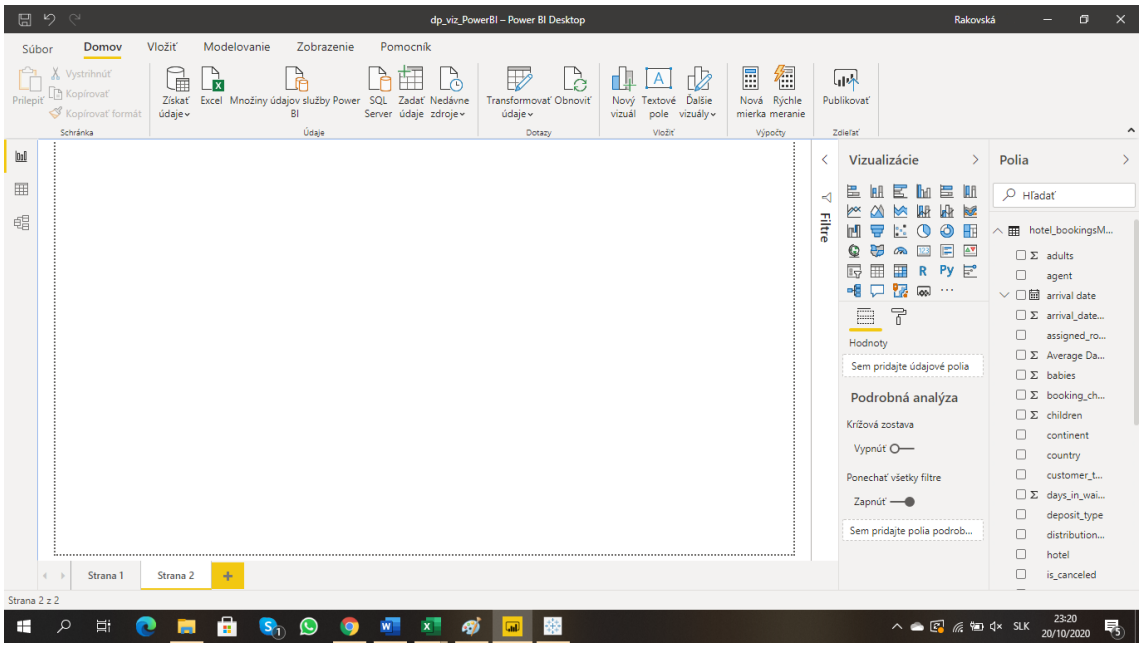
Na pravej strane rozhrania je umiestnená záložka „Show me“, ktorá je v základnom nastavení zobrazená v zrolovanej podobe. „Show me“ je špeciálnou funkciou nástroja Tableau Desktop, ktorá prispieva k tomu, že tento nástoj je možné považovať za self-service nástroj. Po označení názvov polí v ich zozname táto funkcia vyznačí vhodné grafy pre vizualizáciu vybraných dát a na základe overených vizualizačných praktík označí najvhodnejší graf pomocou oranžového orámovania. Táto funkcia je veľkou výhodou pre rýchle tvorenie grafov, ktoré môže byť veľmi efektívne využité aj v prostredí podnikov hlavne zamestnancami, ktorí sa bežne nevenujú analýze dát, ale z nejakého dôvodu potrebujú svoje dokumenty podložiť práve vizualizáciou dát, ktorá často oveľa efektívnejšie vysvetlí alebo dokreslí prezentovanú problematiku.

Vizualizácia v Microsoft Power BI

V tejto podkapitole bude bližšie popísané prostredie nástroja Microsoft Power BI, spôsob, akým je v ňom možné tvoriť vizualizácie ako aj základné funkcie tohto nástroja.

Na obrázku 11 je zobrazené kompletné používateľské rozhranie nástroja Microsoft Power BI. Toto rozhranie je usporiadané veľmi podobne ako vyššie opísané rozhranie editora nástroja Power BI.

Na ľavej strane sú zobrazené tri ikony, pomocou ktorých je možné prepínať medzi zobrazením plochy pre tvorbu vizualizácií, tabuľky spracovávaných dát a dátovým modelom týchto dát.



Obrázok 11 Rozhranie nástroja MS Power BI [vlastné spracovanie]

V pravej polovici rozhrania môžu byť zobrazené tri záložky pre filtre, vizualizácie a polia. V záložke pre vizualizácie je možné vybrať typ grafu, upravovať už existujúce grafy (napríklad dosadiť do grafu iné dáta, alebo hodnoty, podľa potreby vymeniť osi alebo zmeniť farbu a formát grafu). Z tejto záložky sú tiež dostupné funkcie analýzy dát, ktoré budú bližšie popísané v nasledujúcej podkapitole. V poslednej tretej záložke je vypísaný zoznam názvov polí dát, kde iné ako textové dáta sú označené malou ikonou kalendára pre dátum alebo symbolom sumy pre číselné dáta.

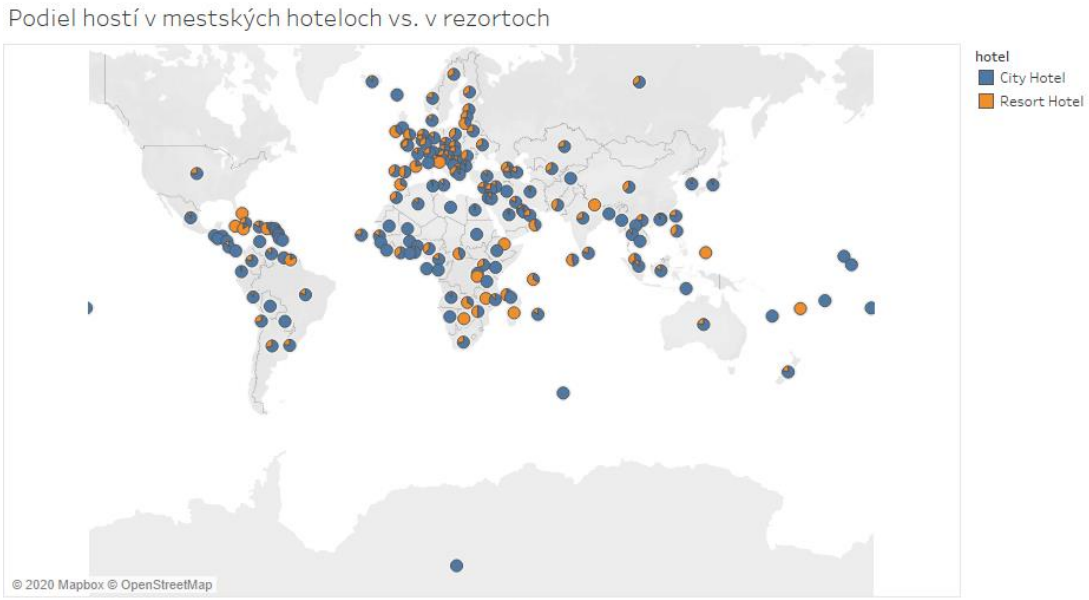
Rovnako ako v rozhraní editora Power BI, aj v tomto rozhraní je v hornej lište dostupných mnoho funkcií. Medzi tie základné patrí napríklad prepojenie dátových zdrojov, vloženie nových vizuálov alebo textu, modelovanie, cez ktoré sú dostupné výpočty nových polí, zobrazenie, ktoré umožňuje zmeniť estetickú stránku tvorených dashboardov a pomocník.

7.3.3.1 Porovnanie výslednej vizualizácie

V nasledujúcej časti budú bližšie popísané ukážky z vytvorených vizualizácií a prípadne aj proces ich tvorby, ak si vyžadoval špeciálne kroky alebo úpravy oproti štandardnému postupu, ktorým je označenie skúmaných polí a zvolenie jedného z navrhovaných grafov.

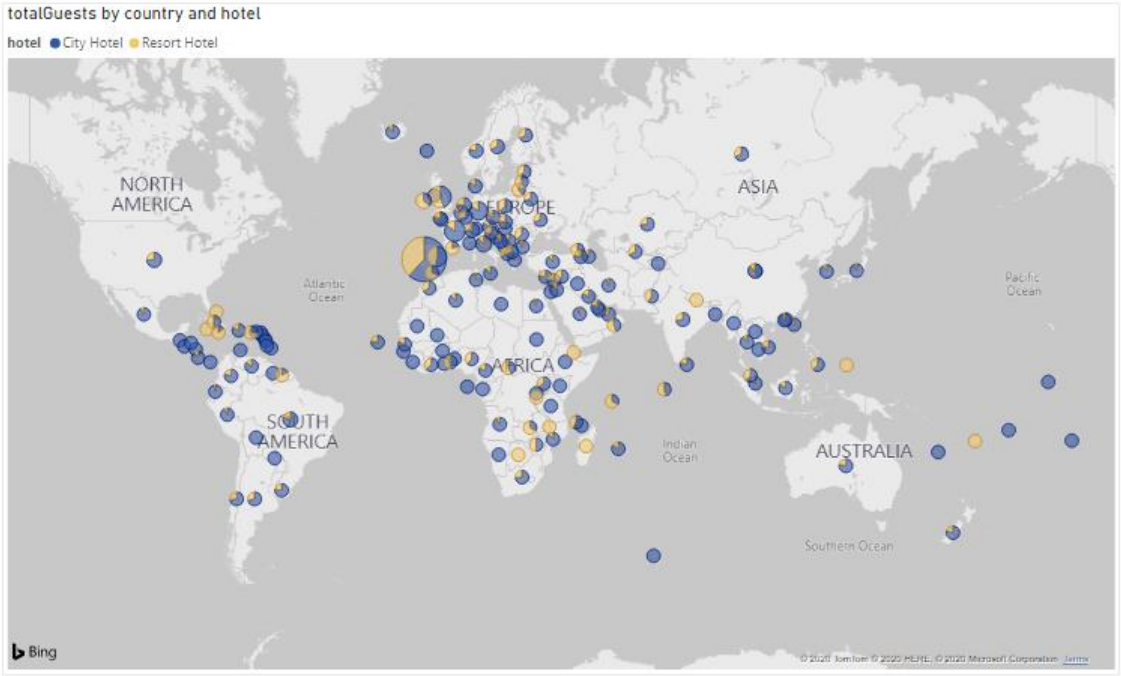
Ako bolo už vyššie spomenuté skúmaný data set obsahuje približne 110 tisíc záznamov o rezerváciách v dvoch typoch hotelov po celom svete. Na obrázkoch 12 a 13 je zobrazená mapa s označením krajín, v ktorých sa nachádzajú hotely z použitého data setu. Kvôli GDPR boli odstránené všetky údaje, ktoré by mohli identifikovať konkrétny hotel alebo hosťa. Z tohto dôvodu boli ponechané iba dva údaje o samotných hoteloch a to názov krajiny, v ktorej sa daný hotel nachádza a typ hotela. Konkrétnejšie, boli zbierané údaje iba z dvoch typov hotelov: z rezortov a z mestských hotelov. Každá krajina je označená malým koláčovým grafom pre ilustráciu podielu využitia mestských hotelov a rezortov. Mestské hotely sú reprezentované modrou farbou a rezorty oranžovou farbou. Zvolená bola mierne netypická kombinácia zobrazenia mnohých menších koláčových grafov na mape, keďže týmto spôsobom je možné zobraziť dve skutočnosti vyplývajúce z dát a vizualizácia aj napriek tomu zostane dostatočne prehľadná. Cieľom tejto vizualizácie je uviesť obecenstvo do skúmanej problematiky a ilustrovať nepomer využitia hotelov, kde v prevažnej väčšine krajín značne prevláda využitie mestských hotelov nad rezortmi.

Na obrázku 12 je zobrazená vyššie popísaná vizualizácia spracovaná v nástroji Tableau. V tomto nástroji je jednoduché vytvoriť vizualizácie vďaka fukncii „drag and drop“. Stačí teda želanú položku potiahnuť do plochy na tvorbu vizualizácií, týmto spôsobom je však možné aj vložiť jednotlivé dátové polia do stĺpcov alebo riadkov tvoreného grafu, zmeniť farbu alebo typ značiek v grafe. Napríklad v obrázku 12 je týmto spôsobom možné zmeniť značky na mape z malých koláčových grafov na kruhy rôznej veľkosti v závislosti od počtu hostí, alebo ponechať jednoduché značky bez farebného rozlíšenia, ktoré zobrazia detaily v tabuľke po presunutí myši na vybranú značku.



Obrázok 13 Podiel hostí v hoteloch, Tableau [vlastné spracovanie]

Na obrázku 13 je zobrazená vyššie popísaná vizualizácia zrealizovaná pomocou nástroja Power BI. Je tu možné vidieť, že Power BI automaticky priraďuje značkám na mape aj veľkosť v závislosti od počtu hostí v danej krajine.



Obrázok 12 Podiel hostí v hoteloch, Power BI [vlastné spracovanie]

Power BI tiež ponúka „drag and drop“ funkciu pre tvorbu grafov rovnako ako Tableau, ale vzhľad jednotlivých značiek na mape je nutné zmeniť v menu, v ktorom sa upravuje aj celkový vzhľad grafu.

Tieto grafy boli zvolené iba pre krátku ilustráciu tvorby grafov v oboch nástrojoch. V nasledujúcej tabuľke 4 je možné vidieť porovnanie ponúkaných grafov a vizualizácií v jednotlivých nástrojoch ako aj tých, ktoré sú spoločné pre oba nástroje.

Tabuľka 4 Porovnanie dostupných grafov [vlastné spracovanie]

spoločné	Tableau	Power BI
tabuľka	teplotná mapa (heat map)	skupinový pruhový a stĺpcový graf
skladaný stĺpcový graf	highlight map	100% skladaný stĺpcový graf
čiarový graf	pruhový graf	skladaný plošný graf
plošný graf	stĺpcové grafy vedľa seba	pásový graf
koláčový graf	zobrazenie bodov hodnôt (circle views)	vodopádový graf
prstencový graf	circle views vedľa seba	lievik (funnel)
treemap	boxplot	mierka
bodový graf	Gantt	vizuál R skriptu
mapa	bullet graph	vizuály v jazyku Python
kartogram	bublinový graf	dekompozičný strom
histogram		
kombinácia stĺpce+čiarový graf		

7.3.3.2 Analytické funkcie

Aj keď sa porovnávané nástroje radia medzi najvyspelejšie na trhu, disponujú odlišnými funkciami na spracovanie analýzy dát. Tabuľka 5 poskytuje prehľad obsiahnutých ukazovateľov a štatistických metód v jednotlivých nástrojoch.

Tabuľka 5 Porovnanie analytických funkcií [vlastné spracovanie]

		Tableau	Power BI
línia konštanty		✓	✓
línia trendu		✓	✓
	lineárny trend	✓	✗
	logaritmický trend	✓	✗
	exponenciálny trend	✓	✗
	power trend	✓	✗
	polynomický trend	✓	✗

minimum		v referenčnej línii	✓
maximum		v referenčnej línii	✓
línia priemeru		✓	✓
línia mediánu		✓ aj kvartily	✓
percentil		✗	✓
symmetry shading		✗	✓
predikcia		✓	✓
box plot		✓	✗
súčty (totals)		✓	✓
cluster		✓	✓
referenčná línia		✓	✗
refeferenčné pásmo		✓	✗
distribučné pásmo		✓	✗
pomerová línia		✗	✓
histogram		✓	✓
odľahlé hodnoty		Nie automaticky	Dodatočný balíček
skupiny		✓	✓

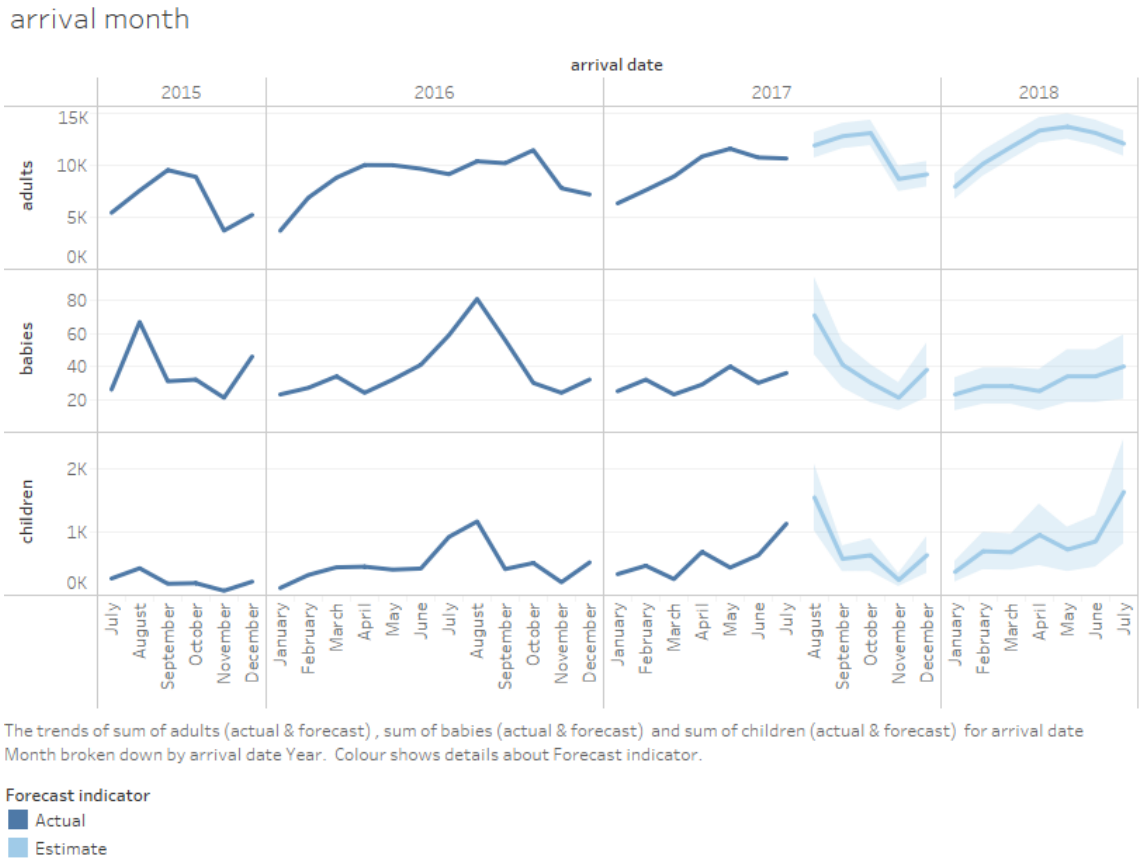
Jedným zo zásadnejších nedostatkov MS Power BI je nemožnosť využitia iného ako lineárneho trendu pre analýzu dát. Lineárny trend je možné aplikovať na čiarový, plošný a na stĺpcový graf. Na rozdiel od MS Power BI, Tableau ponúka širšiu škálu trendov, ktoré je možné aplikovať na viac druhov grafov.

Ďalším z rozdielov v tabuľke sú línie minima a maxima, referenčná línia a medián. V týchto prípadoch je nutné doplniť tabuľku o vysvetlenie. V tabuľke je možné vidieť rozdiely v ponuke týchto funkcií v jednotlivých nástrojoch. Oba nástroje však poskytujú línie minima aj maxima, ale v Tableau nie sú ponúkané ako samostatné línie pod týmto názvom, ale je možné ich zobraziť pomocou nastavenia referenčnej línie na maximum, resp. minimum. Platí to aj naopak, MS Power BI nedisponuje funkciou referenčnej línie, ale všetky jej funkcie (minimum, maximum, konštanta) sú nahraditeľné inými samostatnými funkciami. Tableau tiež ponúka spolu so zobrazením mediánu aj zobrazenie kvartilov, ktoré MS Power BI neponúka.

Ani jeden z nástrojov neponúka automatické identifikovanie odľahlých hodnôt, v MS Power BI je však možné využívať túto funkciu po inštalácií balíčka Outliers Detection.

Jedným z mála výraznejších rozdielov pri tvorbe vizualizácií a pri testovaní analytických funkcií nástrojov je rozdielny výstup predikcie.

Súčasťou tvorby vizualizácií bolo zvolené aj vytvorenie čiarového grafu, ktorý sleduje počet dospelých hostí, detí a batoliat v dané dni počas sledovaného obdobia, tj. od júla 2015 do júla 2017.

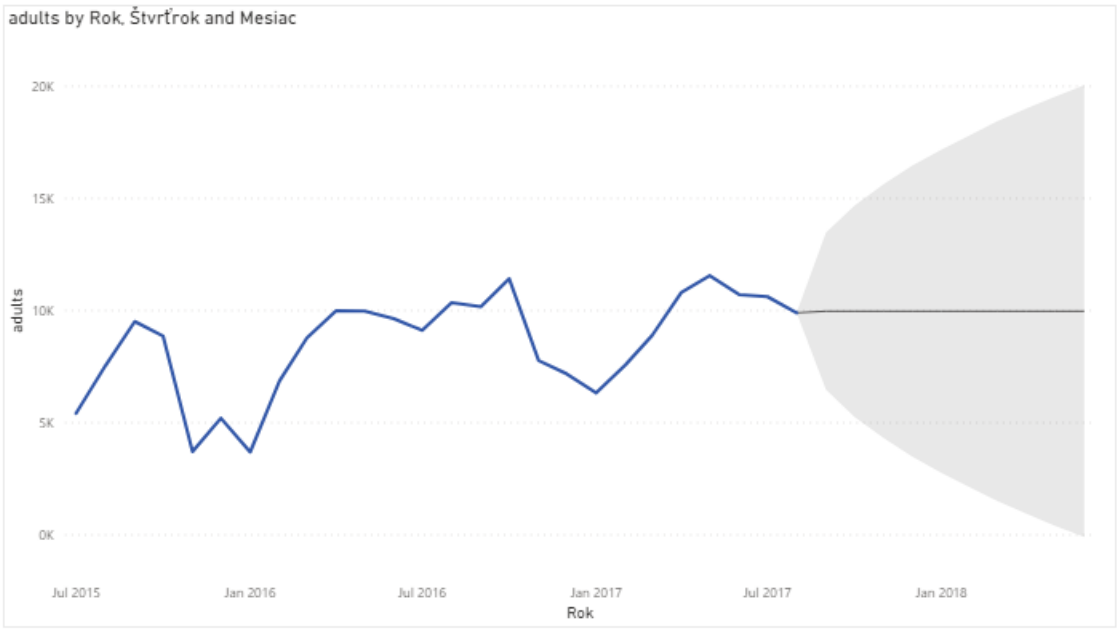


Obrázok 14 Predikcia príchodu hostí, Tableau [vlastné spracovanie]

Na obrázku 14 je výstup predikcie z nástroja Tableau. Tableau používa na predikciu časových radov exponenciálne vyrovnávanie a automaticky nastaví sezónnosť, ak je prítomná v skúmaných dátach.

V grafe je zobrazený počet hostí, ktorí prišli do hotela v daný mesiac a predikcia tohto počtu na nasledujúcich 12 mesiacov. Ak pôvodný čiarový graf obsahuje viac ako jednu dimenziu (nečíselné pole dát), Tableau vytvorí graf pre každú z nich a takisto postupuje aj pri výpočte a zobrazení predikcií. V použitom príklade je teda možné vidieť grafy pre dospelých hostí, deti a novorodencov v jednej vizualizácii, čo uľahčuje porovnanie ich vývoja. Táto konkrétna predikcia je na nasledujúcich 12 mesiacov s 95% intervalom spoľahlivosti, agregovaná po mesiacoch.

Power BI tiež používa exponenciálne vyrovnávanie, konkrétne pre sezónne dáta využíva algoritmus ETS AAA a pre dáta bez sezónnosti algoritmus ETS AAN. Power BI vyberie jeden z týchto algoritmov automaticky na základe analýzy historických dát čiarového grafu, pre ktorý je potrebné vytvoriť predikciu.



Obrázok 15 Predikcia príchodu hostí, Power BI [vlastné spracovanie]

Predikcia na obrázku 15 je vytvorené v nástroji Power BI, taktiež na nasledujúcich 12 mesiacov s 95% intervalom spoľahlivosti a automaticky detegovanou sezónnosťou. Power BI ponúka vytvorenie predikcie pre maximálne jednu dimenziu v jednej vizualizácií.

Pri porovnaní týchto dvoch predikcií je možné vidieť niekoľko rozdielov. Základným rozdielom je, akým spôsobom sú predikcie zobrazené. Tableau ponúka konkrétne čísla odhadu pre každý mesiac so zobrazeným intervalom spoľahlivosti, zatiaľ čo Power BI označí plochu, ktorá predstavuje hodnoty, ktoré môže nadobudnúť počet hostí v mesiaci, na ktorý je počítaná predikcia. Týmto spôsobom zobrazenia Power BI ponúka oveľa menej presnú predikciu oproti predikcii v nástroji Tableau.

Ďalším rozdielom je, že Tableau je schopné poskytnúť predikciu pre viacero veličín v rámci jednej vizualizácie a Power BI počíta predikcie pre maximálne jednu veličinu na jednu vizualizáciu. Nástroj Power BI totiž neumožňuje zobrazenie viacerých sledovaných polí (dospelí, deti, batoľatá) a zároveň vypočítať a vpísať do toho istého grafu aj

prognózu pre všetky vybrané polia dát. Preto bol vytvorený jeden graf so zobrazenými všetkými tromi dátovými poľami a druhý graf, ktorý poskytuje prognózu počtu prichádzajúcich dospelých hostí na nasledujúci rok.

7.3.4 Vyhodnotenie

V tejto kapitole budú zhrnuté výsledky porovnávania self-service nástrojov Tableau Desktop a Microsoft Power BI Desktop. Taktiež budú popísané odlišné postupy, alebo prípadné komplikácie, ktoré nastali pri praktickom využití vybraných self-service nástrojov.

7.3.4.1 Rozdiely v čistení a príprave dát

Celá táto podkapitola je venovaná rozdielom v čistení a príprave dát, keďže týchto rozdielov bolo viacero a z veľkej časti ovplyvňujú celkovú použiteľnosť oboch nástrojov.

Prvým základným rozdielom je rozsah v akom je možné dáta očistiť a pripraviť pre analýzu priamo vo vybranom nástroji. Microsoft Power BI poskytuje kompletné prostredie potrebné pre čistenie dát v zabudovanom editore. Nástroj Tableau zvolil iný prístup a zabudované ponúka hlavné funkcie ako zmenu typu dát, alebo premenovanie jednotlivých dátových polí. Pre zložitejšie úpravy ako rôzne delenie do skupín a agregovanie aj pre pohodlnejšie spájanie väčšieho počtu zdrojov dát je k dispozícií nástroj Tableau Prep, ktorý je k dispozícií pre všetkých používateľov Tableau. Potreba stiahnuť tento ďalší nástroj je miernou komplikáciou pri čistení dát, ale veľkou výhodou je vizuálne zmapovanie všetkých krokov a prepojení zdrojov dát. V istých prípadoch tiež môže byť výhodou, že po spustení „data flow“ v Tableau Prep je možné výstup uložiť ako nový súbor v Tableau formáte alebo v iných formátoch, napr. vo formáte vhodnom pre Excel.

Ďalší zásadný rozdiel je automatické identifikovanie typov dát obsiahnutých v jednotlivých poliach. Ako bolo spomenuté v predchádzajúcich kapitolách, všetky dáta boli do oboch nástrojov importované z identického textového súboru formátu .csv, v ktorom boli všetky polia nastavené na typ „všeobecné“. Tableau automaticky a korektne identifikovalo textové polia, číselné polia aj polia s dátumami. Na druhej

strane, MS Power BI automaticky označil všetky polia za textové, čo si vyžadovalo následnú manuálnu zmenu typu polí.

Nástroje sa tiež líšia v manipulácií s hodnotami iného typu ako je nastavený typ dátového poľa, v ktorom sa nachádzajú. V číselnom poli sa nachádzalo niekoľko hodnôt „NA“ (not applicable/ not available), čo je zaužívanou skratkou pre údaje, ktoré nie sú k dispozícií. Tableau túto skratku automaticky identifikovalo a nahradilo prázdnou hodnotou „null“. MS Power BI túto skratku nebolo schopné identifikovať a preto vypísalo chybové hlásenie o zistení nekonzistentných záznamov v danom poli. Na nápravu bolo potrené chybné hodnoty vyhľadať a nahradiť a následne manuálne zmeniť dátový typ poľa.

7.3.4.2 Ostatné rozdiely

1. Prvým rozdielom je poňatie výslednej vizualizácie, keďže každý z nástrojov poskytuje odlišné možnosti tvorby vizualizácií. V nástroji Tableau sú k dispozícií tri rôzne typy hárkov pre tvorbu vizualizácie. Prvým je hárok pre základnú vizualizáciu. V tomto type hárku môže byť umiestnená iba jeden graf s legendou. Druhým typom je dashboard. Dashboard slúži pre usporiadanie vybraných viacerých vizualizácií, tak aby boli k dispozícií na jednom hárku, resp. na jednej pracovnej ploche. Jedinečným typom hárku pre Tableau je „príbeh“. Ako názov napovedá, tu je možné vytvoriť z vybraných grafov príbeh, ktorý sa skladá z jednotlivých krokov. V každom kroku môže byť umiestnený iný graf, alebo ten istý graf s dodatočným komentárom, filtrom dát, atď. Tento spôsob je výborný na prezentovanie a vysvetlenie myšlienkového procesu pri analyzovaní dát a súvislosti medzi jednotlivými grafmi. Je možné publikovať všetky hárky, alebo iba príbeh, resp. vybraný dashboard.

Na rozdiel od Tableau, Power BI ponúka iba jeden typ hárku a to dashboard. Dashboard funguje na rovnakom princípe ako dashboard v Tableau, ale je samozrejme možné využiť ho aj na prezentáciu jediného grafu.

2. Oba nástroje ponúkajú širokú škálu zdrojov dát, ktoré je možné prepojiť a čerpať z nich dáta na spracovanie. V Microsoft Power BI je možné pracovať aj s dátami priamo z web stránky. Na takéto prepojenie stačí zadať URL a Power BI sám automaticky deteguje použiteľné dáta zo stránky. Na to, aby bol Power BI schopný detegovať dáta

musí stránka obsahovať aspoň jednu tabuľku. Ak Power BI nájde na stránke tabuľky, zobrazí ich a je možné vybrať všetky, s ktorými chce následne používateľ pracovať. Tieto vybrané tabuľky sú následne prenesené do editora a ďalej je s nimi možné pracovať ako s dátami so všetkých ostatných zdrojov.

V nástroji Tableau nie je možné prepojiť web stránku ako zdroj dát iba pomocou URL ani iným spôsobom. Dáta musia byť buď na niektorom z dostupných serverov alebo v statickom súbore.

3. Po dokončení práce s vizualizáciami je možné ich uložiť lokálne alebo v oboch nástrojoch je možnosť publikovať vytvorené vizualizácie online. Z Power BI je možné publikovať priamo do online Power BI workspace. Na využívanie tejto služby je potrebné, aby mal používateľ aj všetci jeho kolegovia, ktorí budú publikované vizualizácie prezerať, Power BI Pro licencie.

V nástroji Tableau sú dve možnosti publikovania online. Prvou možnosťou je publikovanie na Tableau Online. V Tableau Online má používateľ k dispozícií vlastný súkromný workspace, z ktorého môže zdieľať svoje dashboardy a vizualizácie napríklad so svojimi kolegami. Ďalšou možnosťou je publikovať vytvorené vizualizácie na Tableau Server. Rozdiel medzi týmito dvomi možnosťami je, že Tableau Online funguje plne na cloudovom úložisku, ale z toho dôvodu má obmedzenú kapacitu 10GB na jedného používateľa. Tableau Server nie je cloudový a plne záleží od toho ako si ho zriadi daná firma, ktorá ho chce využívať.

4. Posledným rozdielom, ktorý do veľkej miery prispieva ku komfortu najmä používateľa začiatočníka je dostupnosť prehľadnej dokumentácie a tutoriálov k nástrojom. Každý z vybraných nástrojov poskytuje web stránky s určitým množstvom materiálov a návodov pre rôzne úrovne používateľov, od začiatočníkov až po skúsených používateľov. Keďže e-learning vybraných nástrojov je spracovaný iným systémom, hlavné rozdiely sú uvedené v nasledujúcej tabuľke 6.

Tabuľka 6 Porovnanie e-learningových materiálov [vlastné spracovanie]

e-learning	TB	MS Power BI
Forma	samostatný portál	dokumentácia na webe MS
Jazyk	angličtina	slovenčina - nekompletná / angličtina
Videá	tutoriály	ukážka funkcie (2-5min)
Rozdelenie	"career paths" aj jednotlivé témy samostatne	vždy viacero funkcií v jednej téme
Komunita	komunita používateľov	dostupná na inej stránke ako dokumentácia
súbor so zdrojom dát	✓	✗
prepis videa v pdf	✓	✗
výsledné vizualizácie z videa	✓	✗

Pre používateľa, ktorý nikdy nepracoval s nástrojmi tohto typu môže byť teda hlavným problémom pri nástroji MS Power BI roztrúsenosť materiálov ako aj chýbajúce vysvetlenia, kedy a prečo určité metódy využiť, keďže tutoriály v dokumentácií vysvetľujú iba ako nástroj používať (kam kliknúť pre zobrazenie danej funkcie, a pod.). V originálnych dokumentoch a tutoriáloch je tiež zložité nájsť vysvetlenie k jednej konkrétnej funkcii, keďže sú vždy v jednej téme spracované viaceré funkcie a často krát z názvu témy nie je zrejmé, o ktoré funkcie sa jedná.

Originálne videá od MS je možné považovať za ukážku toho, čo MS Power BI dokáže a jeho jednotlivých funkcií bez detailného vysvetlenia ich využitia. Ak používateľ potrebuje vysvetlenia k jednotlivým funkciám, musí ich hľadať na iných stránkach a od iných autorov. V tomto prípade je z veľkej miery MS suplovaný rôznymi videami s tutoriálmi a vysvetleniami na YouTube, ktoré však nemusia byť od expertov danej oblasti.

Na druhej strane pri používaní nástroja Tableau môže byť najväčšou prekážkou jazyková bariéra, keďže všetky materiály sú dostupné iba v angličtine a nie v slovenčine.

Je však možné predpokladať, že používateľ, ktorý potrebuje alebo chce pracovať s nástrojom tohto typu ovláda aspoň základnú úroveň angličtiny.

Dostupnosť všetkých materiálov na jednom mieste aj s pomerne precíznymi vysvetleniami k jednotlivým funkciám nástroja je pre používateľov veľkým uľahčením a konkurenčnou výhodou.

8. Záver

V práci boli porovnávané dva vybrané self-service nástroje pre analýzu a vizualizáciu dát. Pre toto porovnanie boli vybrané nástroje Tableau a MS Power BI, na základe ich porovnateľných vlastností a funkcionalít .

V úvodnej časti práce bola spracovaná rešerš a analýza dostupných zdrojov ku základným pojmom týkajúcim sa témy analýzy a vizualizácie dát. Konkrétne na témy: dátová veda, dáta v podniku, data mining, typy dát, data warehouse, vizualizácia, business intelligence a BI self-service nástroje. Boli tu tiež vysvetlené rozdiely medzi tradičnými a self-service BI nástrojmi. V rámci tejto teoretickej časti práce bol tiež vysvetlený význam vizualizácie dát a súčasný stav využitia vizualizácie v praxi, keďže vizualizácia dát sa stáva veľmi rozšírenou a uznávanou podporou rozhodovania na základe dát.

Cieľom praktickej časti bolo vybrať porovnať dva self-service BI nástroje, ktoré boli vybrané na základe hodnotenia zostaveného spoločnosťou Gartner. Vybrané boli nástroje Tableau a MS Power BI. Praktická časť bola spracovaná na základe CRISP metodiky a teda nasledovala jej poradie fáz.

Pre lepšie a objektívnejšie porovnanie týchto nástroj bol zvolený postup porovnania funkcií na zhodnom data sete. Z tohto dôvodu bolo potrebné zorientovať sa v ponuke voľne dostupných data setov, určiť kritériá a vybrať konkrétny data set. Keďže self-service nástroje sú primárne budované na analýzu a vizualizáciu podnikových dát, ktoré však podliehajú GDPR a obchodnému tajomstvu, nakoniec bol vybraný data set zaoberajúci sa rezerváciami hotelov so záznamami z krajín po celom svete.

Nasledujúc fázy metodiky CRISP, najprv bol bližšie popísaný vybraný data set v rámci fázy Pochopenie dát. Následne boli v rámci fázy Príprava dát, dáta očistené od chybných a nepotrebných hodnôt a záznamov, aby boli čo najefektívnejšie využiteľné pri analýze.

Keďže sa jedná o veľmi vyspelé produkty so širokým spektrom používateľov, naplnil sa predpoklad toho, že vo funkcionalite nástrojov budú iba minimálne rozdiely. MS Power BI chýbajú niektoré analytické funkcie (napr. logaritmický a exponenciálny trend, krabicový graf), ktoré Tableau poskytuje. V Power BI je však možné tieto funkcie doplniť

pomocou integrácie skriptov jazyka R alebo Python, čo pre bežného používateľa nie je vhodné riešenie.

Z hľadiska celkovej používateľskej prívetivosti je Tableau o niečo vhodnejšie, keďže oproti Power BI je menej potrebné preklikávať medzi rôznymi oknami a záložkami, viac funkcií je zobrazených pomocou ikon a viac aspektov tvorenej vizualizácie je možné zmeniť jednoducho pomocou „drag and drop“ funkcie.

V konečnom dôsledku nie je možné jasné povedať, či je niektorý z nástrojov jednoznačne lepší ako ten druhý. Preto by používateľ pri výbere mal zvážiť napríklad aj tieto kritériá: úroveň svojej pokročilosti v danej oblasti a poskytovateľa iných nástrojov v podniku, potrebu zapojenia mnohých zdrojov dát zároveň a náročnosť čistenia a prípravy dát. MS Power BI má zabezpečenú hladkú integráciu naprieč svojimi nástrojmi a eliminuje sa potreba vytvárania ďalšieho používateľského účtu, ak spoločnosť daného užívateľa využíva aj iné produkty od firmy Microsoft a používateľ nepotrebuje vykonať zložité čistenie a prípravu dát. Ak však používateľ potrebuje zapojiť mnoho zdrojov dát zároveň a vykonať komplikovanejšie úpravy pri čistení a príprave dát, je možné danému používateľovi odporučiť Tableau.

9. Zdroje

[1] J. Smetana, „Metodika CRISP-DM ako proces získavania znalostí z databáz,“ [Online]. Available: http://smartvia.sk/metodika-crisp-dm-ako-proces-ziskavania-znalosti-z-databaz/. [Cit. 29 Október 2020].

[2] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, R. Wirth, „CRISP-DM 1.0,“ SPSS, [Online]. Available: https://the-modeling-agency.com/crisp-dm.pdf. [Cit. 31 Október 2020].

[3] J. Paralič, „Objavovanie znalostí,“ [Online]. Available: https://www.researchgate.net/publication/228710132_Objavovanie_znalosti_v_databazach. [Cit. 9 November 2020].

[4] J. Piaček, M. Kravčík, „Komparácia,“ [Online]. Available: http://dai.fmph.uniba.sk/~filit/fvk/komparacia.html. [Cit. 9 November 2020].

[5] L. Cao, „What Is Data Science. In: Data Science Thinking,“ Springer, [Online]. Available: https://doi.org/10.1007/978-3-319-95092-1_2. [Cit. 30 Október 2020].

[6] J. Jordan, C. Ellen, „Business need, data and business,“ *Journal of Digital Asset Management,* %1. vyd.5, pp. 10-20, 2009.

[7] O. Azeroual, H. Theel, „The Effects of Using Business Intelligence Systems on an Excellence Management and Decision-Making Process by Start-Up Companies: A Case Study,“ *International Journal of Management Science and Business Administration,* %1. vyd.3, pp. 30-40, 2018.

[8] F. Zhou, X. Lin, C. Liu, Y. Zhao, P. Xu, R. Liu, T. Xue, L. Ren, „A survey of visualization for smart manufacturing,“ Springer, [Online]. Available: https://doi.org/10.1007/s12650-018-0530-2. [Cit. 8 September 2020].

[9] E. Qi, X. Yang, Z. Wang, „Data mining and visualization of data-driven news in the era of big data,“ Springer, [Online]. Available: https://doi.org/10.1007/s10586-017-1348-8. [Cit. 08 September 2020].

[10] N. Chen, W. Liu, R. Bai, A. Chen , „Application of computational intelligence technologies in emergency management: a literature review,“ Springer, [Online]. Available: https://doi.org/10.1007/s10462-017-9589-8. [Cit. 9 September 2020].

[11] C. Taylor, „Structured vs. Unstructured Data,“ Datamation, [Online]. Available: https://www.datamation.com/big-data/structured-vs-unstructured-data.html. [Cit. 26 Január 2020].

[12] D. Pickell, „Structured vs Unstructured Data – What's the Difference?,“ Learning Hub, [Online]. Available: https://learn.g2.com/structured-vs-unstructured-data. [Cit. 10 August 2020].

[13] H. Watson, „Big Data Analytics: Concepts, Technology, and Applications,“ [Online]. Available: https://www.researchgate.net/publication/331945370_Update_Tutorial_Big_Data_Analytics_Concepts_Technology_and_Applications. [Cit. 9 November 2020].

[14] R. Suja, „Big Data,“ [Online]. Available: http://www.infostat.sk/web2015/sk/_publikacie/Big_Data.pdf. [Cit. 9 November 2020].

[15] N. Donges, „Data Types in Statistics,“ towards data science, [Online]. Available: https://towardsdatascience.com/data-types-in-statistics-347e152e8bee. [Cit. 10 August 2020].

[16] V. Rovnik, „Best Charts to Show Discrete Data,“ [Online]. Available: https://www.webdatarocks.com/blog/best-charts-discrete-data/. [Cit. 25 Január 2020].

[17] D. Sarkar, „The Art of Effective Visualization of Multi-dimensional Data,“ [Online]. Available: https://towardsdatascience.com/the-art-of-effective-visualization-of-multi-dimensional-data-6c7202990c57. [Cit. 25 Január 2020].

[18] iSixSigma, „CONTINUOUS DATA,“ [Online]. Available: https://www.isixsigma.com/dictionary/continuous-data/. [Cit. 26 Január 2020].

[19] K. Sosulski, „Data Visualization Made Simple: Insights Into Becoming Visual,“ New York, Routledge, 2020.

[20] L. Chou, „9 Data Visualization Tools That You Cannot Miss in 2019,“ towards data science, [Online]. Available: https://towardsdatascience.com/9-data-visualization-tools-that-you-cannot-miss-in-2019-3ff23222a927. [Cit. 10 August 2020].

[21] L. Chou, „How Can Beginners Design Cool Data Visualizations?,“ towards data science, [Online]. Available: https://towardsdatascience.com/how-can-beginners-design-cool-data-visualizations-d413ee288671. [Cit. 11 August 2020].

[22] L. Perkhofer, C. Walchshofer, P. Hofer, „Does design matter when visualizing Big Data?,“ Springer, [Online]. Available: https://doi.org/10.1007/s00187-020-00294-0. [Cit. 09 September 2020].

[23] K.-W. Su, C.-L. Liu, Y.-W. Wang, „A principle of designing infographic for visualization representation of tourism social big data,“ Springer, [Online]. Available: https://doi.org/10.1007/s12652-018-1104-9. [Cit. 08 September 2020].

[24] bot.media, s. r. o., „mapa.covid.chat,“ [Online]. Available: https://mapa.covid.chat/. [Cit. 1 Október 2020].

[25] GEDI Visual, „Coronavirus, la situazione in Italia,“ [Online]. Available: https://lab.gedidigital.it/gedi-visual/2020/coronavirus-i-contagi-in-italia/. [Cit. 12 Október 2020].

[26] QlikQ, „Data Exploration,“ [Online]. Available: https://www.qlik.com/us/data-analytics/data-exploration. [Cit. 27 Október 2020].

[27] Sisense, „What is Data Exploration?,“ [Online]. Available: https://www.sisense.com/glossary/data-exploration/. [Cit. 27 Október 2020].

[28] OLAP, „What is Business Intelligence (BI)?,“ OLAP.com, [Online]. Available: https://olap.com/learn-bi-olap/olap-bi-definitions/business-intelligence/. [Cit. 11 August 2020].

[29] I. A. Jamaludin, Z. Mansor, „Review on Business Intelligence (BI) Success Determinants in Project Implementation,“ *International Journal of Computer Applications,* %1. vyd.8, pp. 24-27, 2011.

[30] Datapine, „Self-Service BI Tools,“ [Online]. Available: https://www.datapine.com/articles/self-service-bi-tools. [Cit. 10 August 2020].

[31] R. Bhandari, „Traditional vs. Self-Service BI: Analytics Alternatives Explained,“ Software Advice, [Online]. Available: https://www.softwareadvice.com/resources/traditional-bi-vs-self-service/. [Cit. 10 August 2020].

[32] bi-survey.com, „Self-Service BI: An Overview,“ [Online]. Available: https://bi-survey.com/self-service-bi. [Cit. 10 August 2020].

[33] TechTarget, „self-service business intelligence (BI),“ [Online]. Available: https://searchbusinessanalytics.techtarget.com/definition/self-service-business-intelligence-BI. [Cit. 11 August 2020].

[34] Sisense, „Self-Service Business Intelligence,“ [Online]. Available: https://www.sisense.com/glossary/self-service-bi/. [Cit. 11 August 2020].

[35] J. Richardson, R. Sallam, K. Schlegel, A. Kronz, J. Sun, „Magic Quadrant for Analytics and Business Intelligence Platforms,“ Gartner, [Online]. Available: https://www.gartner.com/doc/reprints?id=1-1XYUYQ3I&ct=191219&st=sb. [Cit. 15 August 2020].

[36] J. Mostipak, „Hotel booking demand,“ [Online]. Available: https://www.kaggle.com/jessemostipak/hotel-booking-demand. [Cit. 1 September 2020].

[37] N. Antonio, A. de Almeida, L. Nunes, „Hotel booking demand datasets,“ Elsevier, [Online]. Available: https://doi.org/10.1016/j.dib.2018.11.126. [Cit. 10 September 2020].

[38] Tableau, „What is business intelligence? Your guide to BI and why it matters,“ [Online]. Available: https://www.tableau.com/learn/articles/business-intelligence. [Cit. 11 August 2020].

10. Prílohy

Kópia zadania práca

UNIVERZITA HRADEC KRÁLOVÉ
Fakulta informatiky a managementu
Akademický rok: 2018/2019

Studijní program: Systémové inženýrství a informatika
Forma studia: Prezenční
Obor/kombinace: Informační management (im2-p)

Podklad pro zadání DIPLOMOVÉ práce studenta

Jméno a příjmení:	**Bc. Natália Rakovská**
Osobní číslo:	**I1800751**
Adresa:	Palachova 1129, Hradec Králové – Nový Hradec Králové, 50012 Hradec Králové 12, Česká republika
Téma práce:	Porovnání kombinovaných nástrojů pro vizualizaci a analýzu dat
Téma práce anglicky:	Evaluation and comparison of Self-Service BI Tools
Vedoucí práce:	prof. RNDr. Hana Skalská, CSc. Katedra informatiky a kvantitativních metod

Zásady pro vypracování:

1.Typy dat a možnosti jejich vizualizace
2. Nástroje na exploraci dat
3. Stanovení kriterií pro porovnání
4. Volba reprezentantů k porovnání, zdůvodnění volby
5. Praktické porovnání
6. Porovnání nástrojů

Seznam doporučené literatury:

PYLE, Dorian. *Data preparation for data mining*. MAKRIDAKIS, Spyros, Steven C. WHEELWRIGHT a Rob J. HYNDMAN. *Forecasting: Methods and Applications*.

Podpis studenta: Rakovská Datum: 16.11.2020

Podpis vedoucího práce: Datum: