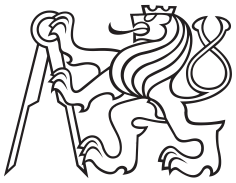
Bakalářská práce



České
vysoké
učení technické
v Praze

F3

Fakulta elektrotechnická
Katedra počítačů

Návrh a vytvoření úvodního kurzu datových analýz

Oleksandra Tahirbekova

Školitel: Ing. Pavel Náplava, Ph.D.
Obor: Softwarové inženýrství a technologie
Leden 2023

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

I. OSOBNÍ A STUDIJNÍ ÚDAJE

Příjmení: **Tahirbekova** Jméno: **Oleksandra** Osobní číslo: **477718**

Fakulta/ústav: **Fakulta elektrotechnická**

Zadávající katedra/ústav: **Katedra počítačů**

Studijní program: **Softwarové inženýrství a technologie**

II. ÚDAJE K BAKALÁŘSKÉ PRÁCI

Název bakalářské práce:

Návrh a vytvoření úvodního kurzu datových analýz

Název bakalářské práce anglicky:

Design and creation of data analysis introduction course

Pokyny pro vypracování:

Analyzujte problematiku "datových analýz". S vedoucím práce navrhněte strukturu a vytvořte úvodní kurz datových analýz pro začátečníky. Postupujte následovně:
1) Definujte pojem "datová analýza" a popište problematiku datových analýz.
2) Prozkoumejte metody, postupy a nástroje, které se v této oblasti používají, primárně se zaměřte na využití jazyk Python.
3) Podívejte se na existující kurzy, které se touto problematikou zabývají. Zaměřte se na uživatele, kteří se znalostmi blíží studentům oboru SIT.
4) Na základě předchozích analýz a společně s vedoucím práce navrhněte a vytvořte praktický kurz, jehož součástí budou materiály pro seznámení s problematikou (literatura), praktické úkoly (datové množiny, konkrétní úkoly) a průběžné nebo alespoň závěrečný test, ověřující znalosti.
5) Navržený kurz uživatelsky ověřte na vybrané skupině studentů, případně dalších potenciálních zájemcích o kurz.
6) Navrhněte, jakým způsobem lze zařadit (integrovat) vytvořený kurz do stávající výuky programu SIT.

Seznam doporučené literatury:

1. Hans Rosling, Ola Rosling, Anna Roslingová Rönnlundová, Faktomluva: Deset důvodů, proč se mýlíme v pohledu na svět – a proč jsou věci lepší, než vypadají, Jan Melvil Publishing, 2018
2. William McKinney: Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython, 2017
3. Anil Maheshwari: Data Analytics Made Accessible, 2014
4. Jacqueline Kazil, Katharine Jarmul: Data Wrangling with Python, 2016

Jméno a pracoviště vedoucí(ho) bakalářské práce:

Ing. Pavel Náplava, Ph.D. Centrum znalostního managementu FEL

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

Datum zadání bakalářské práce: **30.09.2022** Termín odevzdání bakalářské práce: **10.01.2023**

Platnost zadání bakalářské práce: **24.09.2024**

Ing. Pavel Náplava, Ph.D. podpis vedoucí(ho) práce	podpis vedoucí(ho) ústavu/katedry	prof. Mgr. Petr Páta, Ph.D. podpis děkana(ky)

III. PŘEVZETÍ ZADÁNÍ

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

. _____
Datum převzetí zadání

_____
Podpis studentky

Poděkování

Chtěla bych poděkovat především svému vedoucímu, panu Ing. Pavlu Náplavovi, Ph.D. za podporu, trpělivost a pomoc při vypracovávání této práce. Dále také děkuji mámě a kamarádům, kteří mě během celého studia neustále podporovali.

Prohlášení

Prohlašuji, že jsem tuto práci vypracovala samostatně za použití uvedené literatury a zdrojů v souladu s Metodickým pokynem č. 1/20.

V Praze, dne 10. ledna 2023

Abstrakt

Jelikož v dnešní době je datová analýza nezbytnou součástí života každého člověka, zvlášť v IT sféře, vypracovala jsem tuto bakalářskou práci, jejíž hlavním cílem je seznámení se s problematikou datových analýz, prozkoumání postupů a nástrojů, které se v ní používají, a následný návrh praktického úvodního kurzu pro začátečníky. Při vypracování daného kurzu byl kladen důraz na práci v Pythonu, jakožto hlavního jazyka pro manipulaci s daty. Kurz je určen pro zájemce, kteří se svými znalostmi blíží ke studentům starších ročníků studijního programu Softwarové inženýrství a technologie.

Klíčová slova: data, datová analýza, notebook, dataset, kurz, Python.

Školitel: Ing. Pavel Náplava, Ph.D.

Abstract

Since nowadays data analysis is a necessary part of every person's life, especially in the IT sphere, I developed this bachelor's thesis, the main goal of which is to explore the topic of data analysis, to examine the methods and tools used in it, and subsequently to design a practical introductory course for beginners. While developing the given course, I put emphasis on working in Python as the main language for data manipulation. The course is intended for those whose knowledge is close to students in the advanced years of the Software Engineering and Technology study programme.

Keywords: data, data analysis, notebook, dataset, course, Python.

Obsah

Kapitola 1

Slovník zkratek a pojmů

Pojem	Vysvětlení
SIT	Softwarové inženýrství a technologie, studijní program na FELu.
FEL	Fakulta elektrotechnická.
Data mining	Neboli dolování dat. Je to analytická metodologie získávání netriviálních skrytých a potenciálně užitečných informací z dat.
UI	User interface, neboli uživatelské rozhraní.
OOP	Objektově orientované programování.
CRISP-DM	Z anglického Cross-Industry Standard Process for Data Mining. Je to proces zkoumání dat.
API	Z anglického application programming interface, neboli rozhraní pro programování aplikací.
RDF	Z anglického Resource Description Framework, neboli systém popisu zdrojů.
DBMS	Z anglického database management system, neboli Systém řízení báze dat.
SW	Software.

Kapitola 2

Úvod

2.1 Představení a cíle práce

Analýza dat je v současné době jednou z nejpopulárnějších a rychle se rozvíjejících oblastí na světě. Právě z tohoto důvodu je základní porozumění datové analýze a její aplikaci v moderním světě nezbytnou znalostí pro spoustu povolání v IT sféře. Vzhledem k tomuto cíli byla také vytvořena daná bakalářská práce - pro seznámení se a základní porozumění této důležité a obrovské oblasti. Především se jednalo o mou vlastní zkušenost s datovou analýzou, ale postupem času z toho vznikl kurz, který by měl pomoci všem lidem, kteří s tímto oborem nemají žádnou předchozí zkoušenost.

V průběhu práce se toto téma ukázalo jako velice obšírné a rozmanité. Proto i když jsem se v rámci teoretické části bakalářské práce snažila obecně ukázat více nástrojů, které se používají při práci s analýzou dat, v praktické části jsem se soustředila výhradně na práci v Pythonu, jakožto hlavního jazyka pro práci s daty; a na odpovídající knihovny, které tuto práci nejen zjednodušují, ale také poskytují určitá rozhraní, která se v dané sféře považují za standard. Celá praktická část je koncipována jako jeden projekt, ve kterém každý předchozí krok navazuje na ten další, aby studenti meli možnost projít celým procesem od začátku do konce a během něj narazit na kritické a zákeřné body ve všech fázích při práci s daty.

Nakonec jsem během cesty psaní bakalářské práce od úplné nuly do dokončení této práce měla 2 hlavní cíle:

1. prozkoumat a zanalyzovat problematiku "datových analýz", včetně metod, postupů a nástrojů, které se pro ni používají;
2. navrhnout praktický kurz pro začátečníky, včetně zkoušky na konci.

2.2 Čím se tento kurz liší od ostatních?

Předtím, než jsem začala pracovat na této bakalářské práci, položila jsem si otázku "Čím můj kurz bude lepší než ostatní, které už existují? Proč by studenti měli zvolit právě můj kurz?". Protože jsem věděla, že bez tohoto osobně neuvidím smysl ve své práci. Abych si na to dokázala odpovědět, zvolila jsem 2 různé výzkumné cesty:

1. První spočívala v tom, že jsem nastudovala mnoho materiálů na internetu: jak menších tutoriálů, řešících menší problémy zvlášť, tak i větších videokurzů z Youtube a Udemy, které obsahovaly teoretickou i praktickou část. Přesně tato cesta mi pomohla v tom, abych se seznámila s problematikou datové analýzy a abych začala řešit nějaké praktické úlohy, které by mně otevřely různé cesty pro pokračování. Jenže klíčové slovo v předchozí větě bylo "nějaké". Drtivá většina úloh, se kterými jsem se setkala, mně moc nedávala smysl, protože jsem věděla, že nejsou realistické, nejsou z praxe, z reálného světa. Co to přesně znamenalo?
    - Zaprvé, ve většině úloh byl vynechán krok získání dat. Autoři jiných kurzů často na začátku poskytují již hotový dataset a začínají rovnou manipulací s daty. Proto studenti nemají možnost narazit na problémy, se kterými se setkávají datoví analytyci na začátku projektu.
    - Další problém spočívá v tom, že předem připravené datasety nebo nasimulované získání dat vedlo k tomu, že se ve většině případů student nesetká s nutností detailního čištění a transformace dat, protože "získaná"data už budou čistá a připravená. To znamená, že se student zase nenaučí řešit skutečné problémy.
    - Zatřetí, většina kurzů ve svých úlohách řešila jenom happy day scénáře. Případy, kdy něco mohlo selhat, se studentovi skoro neukazovaly. Osobně pro mě to pak byl obrovský problém, protože většinu času, jak i v klasickém kódování, jsem pak trávila právě ošetřováním výjimek a řešením chyb. A proto tento happy day postup během studia považuji za opravdu nešťastný.

   Na jednu stranu všechny body, které jsem uvedla výše, můžeme považovat za plus, když mluvíme o úvodním kurzu. Protože se informace pro

začátečníky mají sdílet postupně, krok za krokem, od těch nejvíce triviálních po pokročilejší, aby jich v jeden okamžik nepřibylo najendou příliš hodně a aby se pak neztratila motivace pokračovat ve studiu. Na druhou stranu ale tím, že student dostane již hezky předem předpřipravená data, nenaučí se tomu nejdůležitějšímu, co podle mého názoru má předat učitel studentovi - přemýšlet.

2. Druhá výzkumná cesta spočívala v tom, že jsem během výměnného pobytu na korejské univerzitě Sungkyunkwan University absolvovala předmět "Introduction to Big Data Analytics". Bylo to udělané z toho důvodu, že mě zajímal přístup k výuce podobného předmětu právě v akademické sféře. A obrovským bonusem pro mě bylo to, že se akademická sféra v Koreji lišila docela dost od té, na kterou jsem byla zvyklá já. Díky tomu jsem měla možnost navnímat a porovnat kardinálně různé přístupy, abych si odnesla z každého to nejlepší.

   Nakonec jsem tam narazila na více problémů, které vedly k samým zmatkům mezi studenty:

   - Snažili se nám ukázat manipulace s daty jak v Pythonu, tak i v programovacím jazyku R během jednoho semestru, s tím, že mezi nimi nebyl vysvětlen žádný rozdíl. Také chyběly příklady a situace, ukazující, kde a z jakého důvodu který jazyk je lepší k využití.
   - Příklady na probírané algoritmy byly udělané nešťastně z toho důvodu, že neukazovaly výhody a smysl jejich využití.
   - Pár týdnů byly zaměřeny na práci s databázemi, protože studenti na SKKU s tím neměli žádné předchozí zkušenosti. A po tom ohromném množství informací, co jsme dostali, hodně studentů na pořádné pochopení tohoto tématu už nemělo prostor.
   - Ukazovali nám dokonce i Hadoop a Spark, které podle mého názoru nebyly pořádně zaintegrované do předchozí části semestru, protože na ni nijak nenavazovaly a nebyly logicky spojeny.

   Nehledě na to, že tento semestr mne vůbec nepomohl s inspirací pro svůj kurz, byl pro mě velice důležitý, protože mi ukázal, čeho nechci dosáhnout pomocí svého kurzu - nechci zmást studenty. Mým úkolem je naučit lidi, nikoliv je odradit od studia. Pro mě to znamenalo, že mám:

   - na začátku definovat cílovou skupinu s konkrétními vstupními znalostmi;
   - omezit množství probíraných témat;
   - vytvořit takové praktické úlohy, které budou logicky navazovat na sebe.

Všechny výše uvedené informace jsem vzala v potaz pro vypracování svého kurzu pro začátečníky, a věřím, že se mi to úspěšně podařilo.

2.3 Pro koho bude tento kurz nejvíce přínosný?

Většina vzdělávacích IT kurzů má pro své zájemce vstupní požadavky, aby kurz byl co nejvíce přínosný pro vybranou cílovou skupinu a ušetřil čas těm, kdo má jinou úroveň znalostí - buď už nižší nebo vyšší. Můj kurz také není výjimkou. Ačkoli je určen pro začátečníky, jedná se o začátečníky právě v datové analýze, nikoli v IT světě. Proto pro vyšší efektivitu a splnění očekávání studentů bych ráda tuto cílovou skupinu určila na začátku. Tak si pojďme říct, co přesně bych doporučila zájemcům před začátkem tohoto kurzu:

1. Mít předchozí zkušenost s programovacím jazykem Python (nebo jiným OOP jazykem). Z toho důvodu, že celá praktická část je postavena na Pythonu, toto je báze, bez které by bylo maximálně neefektivní začínat kurz.
2. Pro získání a extrakci dat:
   - být povědomý s HTML nebo podobným jazykem (XML, XHTML apod.) a CSS;
   - umět se vyznat v regulárních výrazech (regex);
   - mít základní znalosti Selenia.
3. Absolvovat kurz statistiky, jelikož se bude hodně pracovat se základními statistickými pojmy a metodami, například histogram, boxplot, matice korelace a různá rozdělení.
4. Umět pracovat s databázemi a navrhnout vlastní datový model. Toto se bude hodit pro efektivnější ukládání dat.

Velkou výhodou by bylo, kdyby měl student povědomí se způsoby a technikami business analýzy (definice problému, stanovení cílů atd.). Toto téma se v rámci kuzru probírat nebude, ale pro budoucí interakce s datovou analýzou to může být velice vhodnou a důležitou znalostí.

2.4 Zařazení do programu SIT

Studijní program SIT byl vždy docela široko zaměřen, a proto mají studenti po jeho absolvování přehled o více oblastech v IT světě. Z toho důvodu podle mého názoru taková rychle se rozvíjící oblast jako datová analýza mu do značné míry ve svém studijním plánu chyběla.

A s tím, jak jsem v moment napsání této bakalářské práce studentkou SITu, a před nástupem na FEL jsem neměla žádné zkušenosti s programováním a vším souvisejícím, tak jsem se rozhodla, že by pro moje cíle bylo ideální se soustředit na studenty, kteří se svými znalostmi blíží právě ke studentům 2-3. ročníku SITu. Je to zaprvé z toho důvodu, že já jsem měla přesně tuto úroveň znalostí. A zadruhé jsem během práce zjistila, že datová analýza je velice obšírné téma, a proto by nebylo vhodné, aby se studenti museli naučit ještě něco navíc, kromě obrovského množství informací, které budou vysvětleny a probrány v mém kurzu. Jinak by se pro ně stal kurz nepřehledným a nezvládnutelným, zvlášť s tím, že jsem se v daném výukovém kurzu snažila ukázat a probrat co nejvíce postupů, technologií a zákeřných momentů v této oblasti.

Proto všechny požadavky a znalosti, které jsem uvedla v předchozí podkapitole, studenti SITu ideálně naplňují. Pojďme to ověřit:

1. První bod naplňuje předmět B6B36ZAL - Základy algoritmizace, který je ideální pro začátečníky v programování, protože je orientován na práci v Pythonu a ukazuje celou bázi, nutnou pro jakékoli manipulace s daty. Více pokročilé funkce (například lambda funkce), použité v kurzu, se také probírají v rámci předmětu B0B36PJV - Programování v JAVA.
2. Druhý bod se probírá ve více předmětech:
   - HTML a CSS jsou ukázány v předmětu B6B39ZWA - Základy webových aplikací, dokonce do větší míry detailů, než je třeba pro absolvování mého kurzu;
   - na regex studenti v průběhu studia narazí v různých předmětech 2. a 3. semestru, například v již výše zmíněných B6B36ZAL a B0B36PJV;
   - Selenium se probírá v rámci B6B36TS1 - Testování software. A pro to, aby ho student byl schopën využít pro získání dat, není potřebné mít další znalosti.

3. Ve čtvrtém semestru studenti SITu mají celý předmět B6B01PST - Pravděpodobnost a statistika, který poskytuje veškeré informace, využitelné pro analýzu dat v odpovídající části bakalářské práce.

4. Pracovat s databázemi a navrhnout vlastní datový model studenty naučí předmět B0B36DBS - Databázové systémy, který je základem pro efektivní a správné ukládání dat a s nimi následnou práci.

5. Problematikou business analýzy se obsáhle zabývá sada dalších předmětů: B6B36ZPR - Základy projektového řízení, B6B36SMP - Sběr a modelování požadavků, B6B36NSS - Návrh softwarových systémů a B6B16INS - Informační systémy.

Všechny výše uvedené předměty studenti SITu mají v prvním a druhém ročníku studia, což znamená, že by v případě zájmu o datovou analýzou šel ideálně zařadit daný kurz do studijního programu SIT ve 4. nebo 5. semestru studia. Tím pádem by studenti mohli využít svých čerstvých znalostí z minulých předmětů a navázat na další velké zajímavé téma.

2.5 Jak je kurz koncipován?

Předtím, než se pustíme do hlavní části mé bakalářské práce, je důležité ukázat, jak je celá práce koncipována, jak do toho zapadá kurz pro začátečníky a jak jej třeba vnímat.

Zásadní věc, kterou je třeba si ujasnit, je to, že celá praktická část mé bakalářské práce je kurz sám o sobě. S tím, jak analýza dat je pouze sled kroků, z nichž každý hraje klíčovou roli pro ty následující, jsem vytvořila daný kurz, aby každá další fáze (v praktické části - notebook) navazovala na fázi předchozí. To znamená, že tento proces je podobný řetězu po sobě jdoucích, vzájemně propojených fází, proto budeme neustále pracovat s výstupy z předchozích kroků.

A aby celá struktura a postup při průchodu práce byly 100% jasné, uvádím níže krátké shrnutí každé ze zbylých kapitol:

1. Kapitola 3 je čistě teoretická a slouží jako úvod do problematiky a motivace se datovou analýzou zabývat.

2. Kapitola 4 je více prakticky zaměřena a je v roli můstků mezi kapitolou 3 a 5, protože připravuje čtenáře na čistě praktickou část:

   - V sekci 4.1. se ukazuje proces CRISP-DM, který se považuje za standard v oblasti datové analýzy. Jsou tam definovány všechny klasické kroky a úkoly analytických projektů, ale bez hlubších detailů a vysvětlení.
   - Sekce 4.2. navazuje na tu předchozí, ale už je zaměřena právě na postup analýzy v praktické části mého kurzu. To znamená, že některé body z CRISP-DM tam jsou upraveny a/nebo vynechány z logických, tam popsaných důvodů. Navíc každý krok je podrobně vysvětlen z hlediska motivace se tím zabývat, podrobnějšího popisu postupů jednotlivých fází, a problémů, na které se dá narazit.

     Každá jednotlivá podsekce (výjimka 4.2.1) odpovídá pak stejnojmenné sekci z praktické části a následně i příslušnému notebooku. Například: máme sekci *4.2.2 Získání a extrakce dat*. Na ni logicky bude navazovat *5.1 Notebook 1 - Získání a extrakce dat*, a pak i samotný notebook číslo 1.
   - Sekce 4.3 krátce představuje několik nástrojů, které se používají pro analýzu dat, s tím, že ukazuje (ne)výhody a možnosti každého z nich. Nakonec je ale důraz kladen na práci v Pythonu, a proto se právě této části (včetně nápomocných pro Python nástrojů) věnuje nejvíc času.

3. V kapitole 5 se z praktického hlediska popisují prodělané kroky v noteboocích, hlavním úkolem kterých je na konci projektu zkusit předpovědět cenu pronájmu bytů. Kapitola je rozdělená na sekce, které jsou 1:1 namapované na notebooky. V každé z nich jsou uvedeny i použité důležité funkce a/nebo knihovny. Není třeba se bát toho, že tam není moc detailů, protože všechny kroky do detailů jsou pak okomentovány v noteboocích spolu s kódem.

4. Kapitola 6 slouží k výsledkům kurzu: k otestování práce na vybrané skupině studentů a ke zkoušce. Důležitým je také bod o tom, jak by studenti mohli pokračovat v daném projektu dál, aby dosáhli lepších výsledků.

5. Kapitola 7 už je jen závěr bakalářské práce, který se nezvtahuje ke kurzu.

Z toho vyplývá, že kurzem nakonec vyšla nejen praktická část mé práce, ale skoro celá bakalářská práce. Sice se přímo praktické části týkají pouze kapitoly 4, 5 a samotné notebooky, ale kapitoly 2, 3 a 6 jsou také velice užitečné pro zájemce o kurz nebo případné studenty.

Kapitola 3

Obecné informace

Předtím než se ponoříme do praktického světa datové analýzy, seznámíme se s novými technologiemi a budeme řešit reálné, zajímavé problémy, je vhodné si nejdříve ujasnit, co vůbec pojem datové analýzy znamená, proč je potřeba se jí zabývat a jaký prospěch nám může přinést.

3.1 Co jsou data, neboli jsou data a informace to samé?

Téměř každý den se člověk setkává s pojmy data a informace a oba v dnešní době zní hodně synonymicky a často se zaměňují. Nicméně není tomu tak.

Co jsou tedy data? Akademická odpověď na tuto otázku by byla taková: **data** jsou sbírkou obecných, nefiltrovaných detailů ve formě textů, obrázků, symbolů, popisů nebo jednoduchých pozorování událostí nebo věcí, které lze analyzovat. Jednoduše řečeno, data jsou cokoli, co je nám dáno (ne nadarmo to slovo pochází z latinského slova datum, v překladu - něco dáné). Celý svět se totiž skládá z dat, která jsou generována vším kolem nás a včetně nás, ať už je v konkrétní situaci potřebujeme nebo ne. Naším úkolem je jen pochopit, co s nimi musíme udělat, aby se pro nás stala užitečná.

Jak bylo uvedeno výše, data mohou být daná ve zcela odlišných formátech, všechny ale lze rozdělit do 2 kategorií:

- kvantitativní - ukazují množství něčeho, jsou mnohem jednodušší pro další zpracování.
- kvalitativní - popisují vlastnost nebo kvalitu, pak je práce s nimi mnohem obtížnější.

Informace je ale to, co získáme po zpracování, interpretaci a uspořádání nezpracovaných dat. Informace je kus dat, který má nějaký význam - je výsledkem měření, pozorování, zkušeností. Informace většinou mají interpretaci či vysvětlení, tj. mají tzv. kontext. Vně kontextu není informace odlišitelná od dat. informace, na rozdíl od dat, mají význam samy o sobě. Informace jsou proto daty, která člověk nějakým způsobem zpracoval - rozřadil, vyčistil, rozdělil či předělal.

Aby byl rozdíl jasnější a nezůstali bychom pouze v teoretické rovině, pojďme si tento materiál pro lepší pochopení rozebrat a sjednotit na jednoduchém příkladu:

- Představme si, že máme soubor chaoticky exportovaných dat, například z profilu zákazníka z internetového obchodu: *Dlouhá, 1 dítě, Zuzana, 4, 11200, Krátká, 512, Praha, 3, 15.9.2010*. Proč jsou data exportována v tomto formátu bez popisu každého pole? Je tam vůbec pole? Je tam několik logických prvků? To může být například problémem špatné extrakce dat z databáze nebo z webových stránek, anebo tento kus dat bez kontextu, v němž byla nalezena, nedává smysl.

  Vidíme, že soubor je nestrukturovaný a nicneříkající. Tudíž jako zpracovatelé těchto dat nevíme, jak s tím máme pracovat a co provést. Kdybychom potřebovali z těchto dat dostat takovou informaci, jako je například adresa, potřebujeme vědět, jak vypadá textový zápis adresy a co musí obsahovat. Dokonce i skutečnost, že v tomto souboru musíme najít adresu, již hovoří o tom, o jaký kontext jde - o hledání adresy.

  Je docela jasné, že v tomto úkolu nebudeme potřebovat datum narození a informaci o dětech. Takže teď nam zbyla: *Dlouhá, Zuzana, 4, 11200, Krátká, 512, Praha, 3.*

  A teď, když zkusíme dát dohromady ostatní údaje, tak narazíme na to, že data byla od začátku velice nakvalitně zpracována (exportována), protože bez dalšího kontextu nám vůbec není jasné, co z údajů *Dlouhá, Krátká* je ulice a co je příjmení. Také nevíme, co znamenají čísla *3, 4*, jestli to je patro, číslo orientační nebo něco dalšího.

V tuto chvíli pořád pracujeme pouze s daty, nikoliv informacemi - daný textový řetězec zatím nemá vypovídající hodnotu. Zároveň ale už víme, že musí obsahovat adresu. To znamená, že teď máme několik možností, jak by tato adresa mohla vypadat. Část z nich je uvedena níže v tabulce 3.1. Bohužel bez dalšího kontextu nejsme schopni se dozvědět, co z ní je ta správná varianta. Z toho vyplývá docela častý problém v analýze dat - špatná interpretace dat.

Zuzana Dlouhá Krátká 512 11200 Praha	Zuzana Krátká Dlouhá 512/3 11200 Praha	Zuzana Krátká Dlouhá 4 11200 Praha

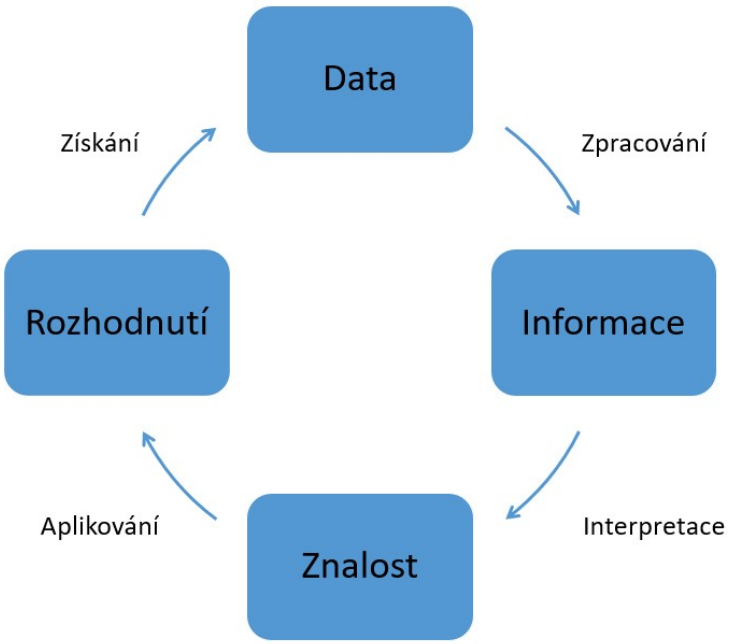
Tabulka 3.1: Příklad špatné interpretace dat.

Informací se stane právě kus dat (jedna z položek tabulky), u něhož víme, že reprezentuje adresu. Informace je tedy soubor dat, který má interpretaci a má hodnotu (dá se s ním dále pracovat). Například tato informace může posloužit k sestavení seznamu adres k zaslání letáků se slevonými kupóny. Data se ale v surovém stavu nedají takto použít.

- Je však nutné zavést i pojmy kvalita dat a kvalita informací. Kvalitou dat se rozumí zdroj, který tato data generuje. Kvalitní data jsou vytvořena dokonalým procesem - nelze považovat za data výstup z mikrofónu, ve kterém není slyšet hlas nahrávajícího. Nekvalitní data jsou většinou zašuměná. Kvalití informace je ale výsledkem zpracování dat.

  Vysvětlím to na příkladu "ze života". Chtěli bychom upéct dort a máme k dispozici vejce a mouku. Mouka a vejce je analogií pro data - obě suroviny se k nám dostaly, nevytvořili jsme si je sami. Dort je ale výsledkem vykonání postupů dle receptu, proto můžeme považovat dort za informaci - na základě "dat"jsme schopni vytvořit něco odvozeného. Kvalitními daty se rozumí takové suroviny, které nejsou ničím poškozeny - v našem případě, že vejce jsou čerstvá a mouka je správného druhu. Kvalitní informace je tedy správně upečený dort. Je zřejmé, že pomocí prošlých vájec či nesprávného druhu mouky není možné dle stejné receptury upéct chutný dort. Stejně tak to funguje i s daty a také informacemi - pokud data nejsou kvalitní, je nemožné z nich získat kvalitní informaci. Ale je potřeba pamatovat i na to, že pokud data kvalitní jsou (tzn. že máme čerstvá vejce a dobrou mouku), tak to neznamená, že dostaneme kavlitni informaci (vždy můžeme v troubě nastavit špatnou teplotu a dort shoří). I toto je úkolem analýzy dat - zajistit, že na konci budeme mít kvalitní informaci.

Na přikladech výše jsem dopodrobna vysvětlila pojmy dat a informací, které jsou základními kameny v tzv. životním cyklu dat, znázorněném na obrázku 3.1.



Obrázek 3.1: Životní cyklus dat

Nejdřív se z dat vygeneruje informace pomocí čištění, řazení a kategorizace. Dále z informací vznikne znalostní báze - soubor informací, který jako celek dává smysl a tuto informaci popisuje, vytváří její rozdělení. Znalostí může být seznam výšek všech žáků ve třídě, zatímco informace je pouze jedna spočtená výška a datum je skutečnost, že člověk je nějak vysoký. Posledním prvkem je rozhodnutí - výsledek naplnění znalostní báze. Data jsou sbírána a zpracována za nějakým účelem - analýzy, optimalizace, potvrzení hypotézy atd. Tento účel definuje i rozhodnutí. V příkladu se žáky tímto rozhodnutím může být nákup desek, které odpovídají výšce dětí ve třídě, aby se jim lépe sedělo během výuky.

Tento cyklus je ve velké míře inspirován DIKW pyramidou, viz [i-S] a [Aya19]. Největším rozdílem mezi nimi je to, že jsem změnila pyramidu právě na cyklus, protože mým úkolem bylo ukázat čtenářům, že se jedná o proces, který nikdy nekončí. Pyramida na můj pohled není nejšťastnější volbou pro reprezentaci daného úkolu, protože z ní vyplývá, že se jedné o jednosměrný a jednorázový proces. Není ale tomu tak. Protože jakmile uděláme nějaké rozhodnutí (získáme moudrost), tak zase začneme získávat nová data pro další úkol nebo rozhodnutí, v procesu nikoliv nekončíme.

3.2 Co je datová analýza?

Když půjdeme do etymologie samotného slova "analýza", tak uvidíme, že k nám přišlo ze starověkého Řecka a znamená „rozmotat“, „osvobodit“. V současné době ale neexistuje žádná konkrétní definice pojmu “datová analýza”, protože jej lze definovat hodně sice podobnými, ale pořád o něco odlišnými způsoby. Několik z nich je uvedeno níže:

1. Daniel Burrus, obchodní konzultant a mluvčí pro obchod a inovace, o analýze dat říká [Bur15]: „Analytika dat je věda o získávání vzorců, trendů a použitelných informací z velkých souborů dat. Mnoho z nich pomůže lidem pracovat chytřeji a rychleji, protože máme data o všem, co se děje.“
2. Stephan Kudyba ve své knize „Big Data, Mining, and Analytics“ píše [Kud14], že analýza dat je proces kontroly, čištění, transformace a modelování dat s cílem objevit užitečné informace, vyvodit závěry a podpořit rozhodování.
3. V obchodním prostředí se v poslední době stala velmi populární definice analýzy dat od Maria Faria, viceprezidenta společnosti Gartner [Far15]: "Analýza dat je transformace dat do závěrů, na základě kterých se budou rozhodovat a uskutečňovat akce za pomoci lidí, procesů a technologií. Pro každého obchodního lídra nebo top manažera je podle této definice důležité hledat skryté vzorce a získávat nové znalosti v analýze dat".

Výše uvedené interpretace tohoto pojmu lze zjednodušit a zobecnit pomocí jedné věty: Analýza dat je práce s daty za účelem sbírání užitečných informací, které pak lze použít k přijímání informovaných rozhodnutí.

3.3 Proč se datovou analýzou zabývat?

3.3.1 Digitalizace světa

Proces digitalizace je často označován jako digitální transformace a hluboce mění podobu dnešního podnikání a ovlivňuje společnosti v každém odvětví a spotřebitele po celém světě. Digitální transformace není o vývoji zařízení (i když se vyvíjet budou), jde o integraci dat do všeho, co děláme. Svět řízený daty bude vždy zapnutý, bude vždy sledovat, pozorovat a naslouchat, protože se bude stále učit. To, co vnímáme jako náhodnost, bude ohraničeno vzorci normality pomocí sofistikovaných algoritmů umělé inteligence, které přinesou budoucnost novými a personalizovanými způsoby.

Data jsou jádrem digitální transformace. Spotřebitelé jsou závislí na datech a jejich větším množství v reálném čase. Dnes více než 5 miliard spotřebitelů komunikuje s daty každý den – do roku 2025 to bude 6 miliard, neboli 75 procent světové populace. V roce 2025 bude mít každý připojený člověk alespoň jednu datovou interakci každých 18 sekund [DR18].

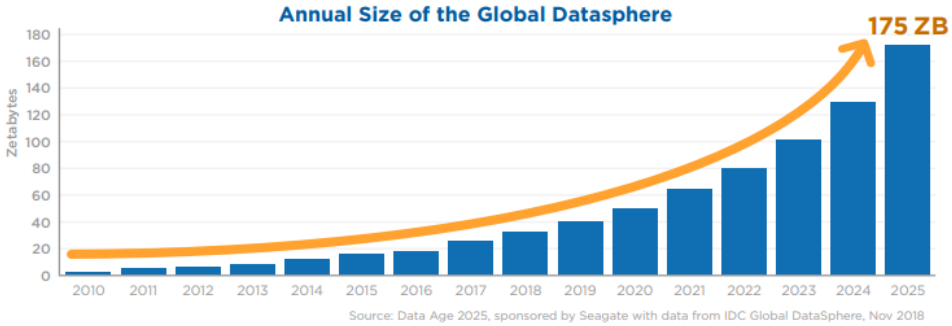
3.3.2 Využití dat dnes mění způsob, jakým žijeme

Podniky v průmyslových odvětvích po celém světě využívají data k transformaci, aby se staly agilnějšími, zlepšily zákaznickou zkušenost, zavedly nové obchodní modely a vyvinuly nové zdroje konkurenční výhody.

Spotřebitelé žijí ve stále více digitálním světě a jsou závislí na online a mobilních kanálech, aby se mohli spojit s přáteli a rodinou, získat přístup ke zboží a službám a provozovat téměř každý aspekt svého života, a to i ve spánku.

Velká část dnešní ekonomiky závisí na datech a tato závislost se v budoucnu ještě zvýší, protože společnosti budou získávat, katalogizovat a vydělávat na datech v každém kroku svého dodavatelského řetězce; podniky shromažďují obrovské množství údajů o zákaznících, aby zajistily vyšší úroveň personalizace; a spotřebitelé integrují sociální média, zábavu, cloudová úložiště a personalizované služby v reálném čase do svých životních proudů.

Důsledkem tohoto rostoucího spoléhání se na data bude nikdy nekončící rozšiřování velikosti globální datové sféry. International Data Corporation předpovídá [DR18], že globální datová sféra vzroste z 33 zettabytů z roku 2018 na 175 zettabytů do roku 2025, viz Obrázek 3.2 pro lepší přehlednost.



Obrázek 3.2: Roční velikost globální datové sféry. Zdroj [DR18]

Je docela těžké si představit, jak obrovské je toto číslo (jeden zettabajt odpovídá bilionu gigabajtů), ale kdybychom mohli stáhnout celou globální datovou sféru roku 2025 s průměrnou rychlostí 25 Mb/s, pak by to jedné osobě trvalo 1,8 miliardy let, nebo pokud by to dělal paralelně každý člověk na světě bez přestávky, pak bychom to mohli udělat za 81 dní.

3.3.3 K čemu to je prospěšné?

Analýza dat hraje klíčovou roli ve všech obchodně orientovaných a dokonce i vládních operacích, od lepšího porozumění publiku a zákazníkům až po předpovídání přírodních a sociálních katastrof a vývoj umělé inteligence. A s exponenciálně rostoucím množstvím dat, která se hromadí v reálném čase, není důvod, proč bychom z nich nemohli udělat konkurenční výhodu:

- Pokud mluvíme o podnikání, tak začneme u zákazníků, kteří jsou pravděpodobně nejdůležitějším prvkem v každém podnikání. Pomocí analýzy je možné získat vizi všech aspektů souvisejících se zákazníky; pochopit, které kanály používají ke komunikaci, jejich demografické údaje, zájmy, zvyky, nákupní chování a další. Analytika dat může například pomoci bance personalizovat zkušenosti zákazníků, zdravotnickému systému předpovědět budoucí potřeby zdravotní péče nebo zábavní společnosti vytvořit nový streamovací hit.
- Z dlouhodobého hlediska datová analýza pomůže podpořit úspěch i marketingových strategií, umožní identifikovat nové potenciální zákazníky a zabrání plýtvání zdrojů na cílení na nesprávné lidi nebo odesílání

nesprávných zpráv. Spokojenost zákazníků bude možné sledovat pomocí analýzy recenzí klientů.

- Z pohledu managementu je také možné těžit z analýzy dat, protože správná analýza a správné závěry z ní pomohou činit obchodní rozhodnutí na základě faktů, nikoli prosté intuice. Může to přispět k porozumění toho, kam investovat kapitál, zjistit příležitosti k růstu, předvídat příjmy nebo řešit problémy dříve, než se nastanou. Pak to usnadní a urychlí prezentaci dat interaktivním způsobem různým zúčastněným stranám.

Tím, že datová analýza poskytuje užitečné informace a statistiky, obecně pomáhá snižovat rizika spojená s rozhodováním. Rozhodování založená na datech, jsou zpravidla důkladnější a přehlednější pro všechny strany. Datová analýza je tedy nedílnou součástí podniků ve snaze zefektivnit stávající procesy a kampaně, stejně tak jako otevřít nové příležitosti a umět je monitorovat a správně vyhodnocovat.

Kapitola 4

Rešerše

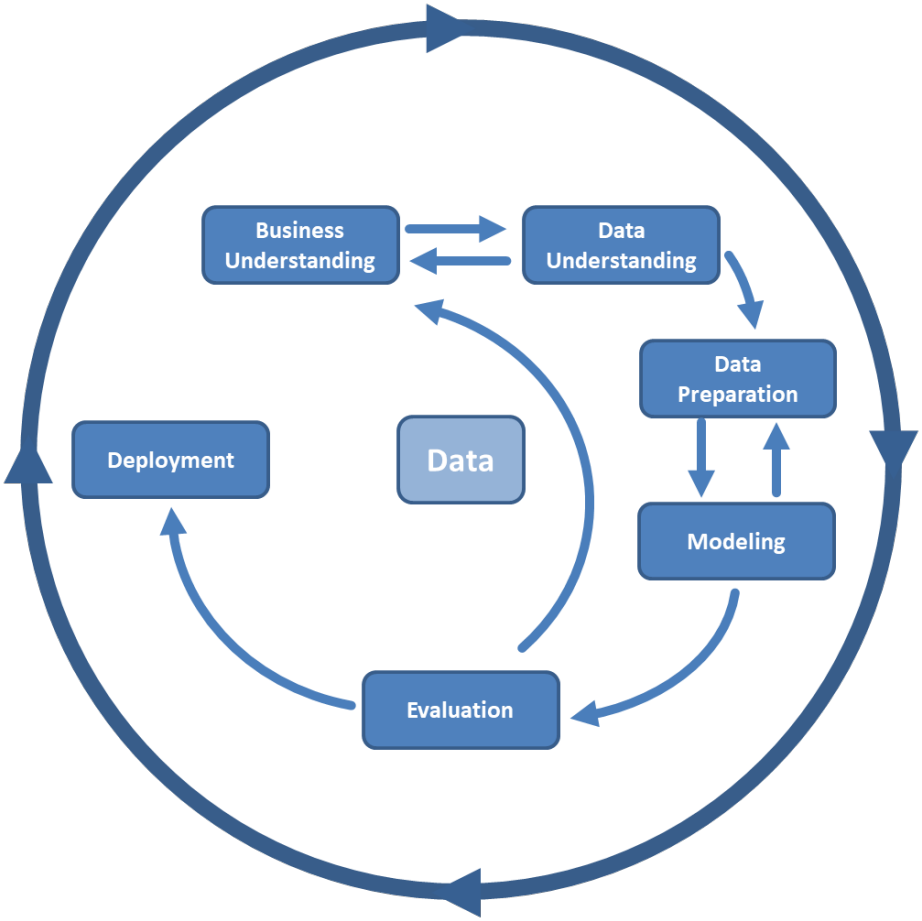
4.1 Postup při analýze dat

Analýza dat je pouze sled kroků, z nichž každý hraje klíčovou roli pro ty následující. Tento proces je podobný řetězu po sobě jdoucích, vzájemně propojených fází. Je to ale komplexní proces, ve kterém projekty zahrnují různé zainteresované strany, zdroje dat a cíle. Ve snaze udržet pořádek byly vytvořeny různé osnovy a metodiky, které mají pomoci novým i zkušeným datovým vědcům se způsoby organizace a strukturou jejich práce. Těchto metodik je docela dost a abychom teď nešli do hloubky v rozdílech každé z nich (pro zájemce doporučuji tento článek [All]), tak bych ráda ukázala jen jednu z nich, která se považuje za standard v oblasti datové analýzy a zároveň je jednou z nejvíce populárních a používanějších v současné době, viz [Nic22].

Cílem následující podsekce je seznamit čtenáře a ukázat mu obecný princip fungování modelu CRISP-DM, včetně všech fází a nutných úkolů pro jejich dokončení. Nebudeme však dopodrobna probírat každý krok (pro této účely můžu doporučit [DTL05]), jelikož se většina z nich ukáže v podkapitole 4.2.

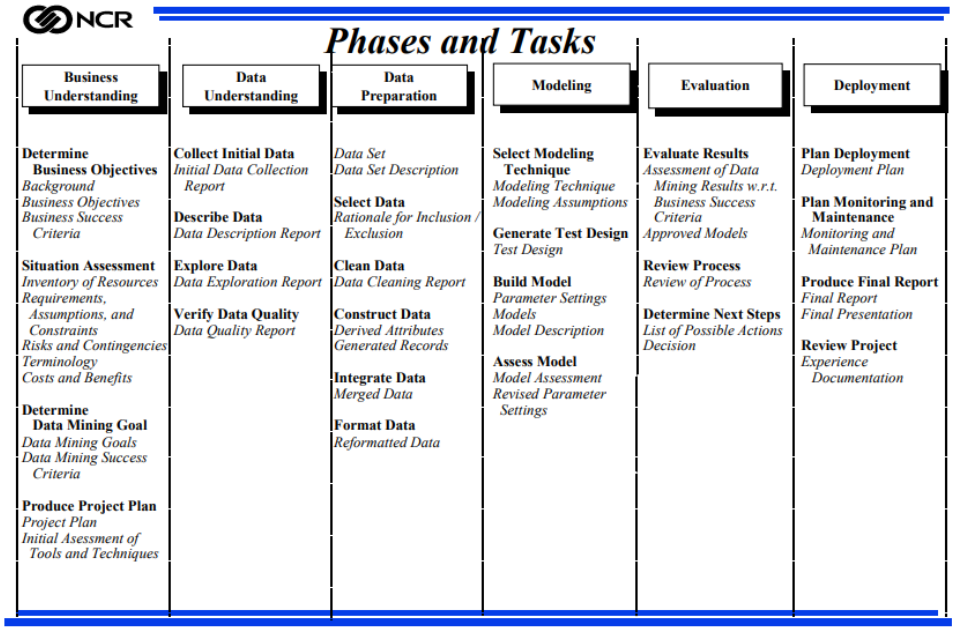
4.1.1 CRISP-DM

CRISP-DM (z anglického Cross-Industry Standard Process for Data Mining) je meziodvětvový standardní proces zkoumání dat. Tento proces popisuje životní cyklus data miningu, který se skládá z 6 fází: od definice problému z obchodního hlediska až po nasazení technického řešení. Posloupnost fází není striktně definována, přechody se mohou opakovat z iterace do iterace, viz Obrázek 4.1. Každá následující fáze je ale obvykle silně ovlivněna výstupy vytvořenými v rámci fáze předchozí.



Obrázek 4.1: Grafické zobrazení modelu CRISP-DM. Zdroj [Hei20]

Všechny fáze CRISP-DM jsou rozděleny do úkolů, na konci každé z nich musí být dosaženo konkrétního výsledku, viz Obrázek 4.2. Podívejme se na tyto fáze životního cyklu dolování dat podrobněji:



Obrázek 4.2: Fáze a úkoly v modelu CRISP-DM. Zdroj [Bro99]

1. **Obchodní porozumění (Business Understanding)** – stanovení cílů projektu a obchodních požadavků. Poté jsou tyto znalosti převedeny do prohlášení o problému dolování dat a předběžného plánu pro dosažení cílů projektu. Úkoly fáze obchodního porozumění jsou následující:
   - Definovat obchodní cíle.
   - Vyhodnotit situaci.
   - Definovat cíle pro analýzu dat.
   - Vytvořit plán projektu.
2. **Porozumění datům (Data Understanding)** - sběr dat a poznávání informací, identifikace problémů s kvalitou dat (chyby nebo opomenutí hodnot). Je tedy třeba pochopit, jaké informace jsou k dispozici, pokusit se najít zajímavé soubory dat nebo vytvořit hypotézy o přítomnosti skrytých vzorců v nich. Úkoly fáze porozumění datům, viz [Cha00]:
   - Nasbírat vstupní data.
   - Popsat data.
   - Prozkoumat data.
   - Ověřit kvalitu dat.

3. **Příprava dat (Data Preparation)** - získání finálního datasetu z výchozích multiformátových dat, který bude použit při modelování. Úkoly této fáze lze provádět mnohokrát bez předem určeného pořadí:
   - Vyčistit data.
   - Odvodit chybějící data.
   - Sloučit data.
   - Převést data do správného formátu.
4. **Modelování (Modeling)** - v této fázi se na data aplikují různé techniky modelování, sestavují se modely a jejich parametry se upravují na optimální hodnoty. Úkoly fáze modelování dle [DTL05] jsou uvedeny dále:
   - Vybrat a aplikovat vhodné modelovací techniky.
   - Kalibrovat parametry vybraného modelu (data mining algoritmu) za účelem jeho optimálního nastavení a získání relevantních výsledků.
   - Mít na paměti, že často k řešení jednoho data miningového problému je možné využít několik rozdílných technik a modelů. Obecně se doporučuje využít více různých technik a jejich výsledky kombinovat.
   - Z předchozích kroků je možné, že vyplyne potřeba vrátit se zpět k fázi přípravy dat a jejich modifikaci tak, aby bylo pracováno s co nejvhodnějšími daty přizpůsobenými konkrétní zvolené data miningové technice.
5. **Evaluace (Evaluation)** - rozbor kvantitativních charakteristik kvality modelu, potvrzení či vyvrácení skutečnosti, že díky zkonstruovanému modelu bylo dosaženo všech obchodních cílů. Hlavním cílem etapy je najít důležité obchodní úkoly, kterým nebyla věnována náležitá pozornost. Úkoly hodnotící fáze jsou následující:
   - Vyhodnotit výsledky.
   - Provést revizi procesu.
   - Určit další postup.
6. **Nasazení (Deployment)** - v závislosti na požadavcích může být fáze nasazení jednoduchá (vypracování závěrečné zprávy) nebo složitá, například automatizace procesu analýzy dat pro řešení obchodních problémů. Nasazení je obvykle implementace přijatých modelů v aplikační oblasti. Úkoly fáze nasazení jsou uvedeny dále:
   - Naplánovat nasazení.
   - Naplánovat podporu a monitorování nasazeného řešení.
   - Udělat závěrečnou zprávu.

- Provést revizi projektu.

Petr Berka [Ber03] uvádí, že nejdůležitější fáze je porozumění problému, která zabere 20 % času, ale má 80% významu. Časově nejnáročnější je fáze přípravy dat, která zabírá 80 % času s 20% významem, přičemž modelování a analýza zkoumaných dat zabere 5 % času a má 5% významu.

4.2 Postup v mém kurzu

V předchozí podsekci byl uveden klasický standardní přístup k datové analýze ve větších projektech, od samého začátku do konce. V mém kurzu ale některé kroky budou vynechány z kapacitních (není třeba se snažit do kuzru pro začátečníky umístiti příliš hodně informací) a logických (nejsme ve firmě s N systémy, kam bude třeba řešení nasazovat) důvodů. Zůstal tam ale nejdůležitější princip: každá fáze bude navazovat na tu předchozí, což znamená, že v každém dalším notebooku budeme pracovat s výstupy z předchozího. Proto jsem pro naše účely klasický model CRISP-DM trochu zjednodušila, zpřehlednila a níže uvedla postup, kterého se budeme držet během celé praktické části:

1. Definice problému.
2. Získání a extrakce dat.
3. Příprava dat - čištění a transformace dat.
4. Zkoumání, generování a vizualizace dat.
5. Predikce na základě získaných dat.

Každý z výše uvedených kroků bude podrobněji vysvětlen v následujících sekcích a následně předveden v praktické části bakalářské práce.

4.2.1 Definice problému

Proces analýzy dat začíná dlouho předtím, než jsou shromážděna nezpracovaná data. V první fázi práce s daty musíme pochopit, proč potřebujeme data shromažďovat a analyzovat a také jaká data potřebujeme. Stanovení cílů a předběžných hypotéz o datech pak bude základem našeho projektu.

Účelem této fáze, nutné pro to, aby se analýza zaměřila na dosažení požadovaných výsledků, je:

1. definovat cíle organizace,
2. zhodnotit současnou situaci,
3. definovat cíle analýzy dat.

V rámci této bakalářské práce se z kapacitních důvodů nebudou dopodrobna probírat výše uvedené jednotlivé kroky. Zájemcům o tyto informace bych doporučila natstudovat každý krok této fáze samostatně. Studenti SITu, jak jsem zmiňovala dřív, měli možnost se s touto problematikou seznámit v rámci spousty předmětů B6B36ZPR, B6B36SMP, B6B36NSS a B6B16INS, a proto se od nich již očekávají odpovídající znalosti.

Definice problému navíc slouží k tomu, aby byla vytvořena jakási koncepce (formálněji konceptuální model) řešeného problému, byly zjištěny zúčastnené entity a určeny zdroje, odkud data, nutná pro nalezení řešení, budou pocházet. Problémem se často stává existující neefektivita podniku, pálčivý a stále nevyřešený problém, který neumožňuje další rozvoj organizace. Může se stát i opačný jev - nasbíraná data se stanou podkladem pro tzv. exploratorní analýzu, tj. hledání souvislostí a generování nových znalostí (predikce) na základě již existujicích podkladů. Proto cílem analýzy dat může být jak optimalizace, tak i nalezení nových příležitostí.

4.2.2 Získání a extrakce dat

Jakmile je problém identifikován, prvním krokem při provádění analýzy je získání dat. Tento krok je často v literatuře opomíjen a autoři rovnou začínají tím, jak začít s daty manipulovat. Jenže v praxi to vypadá jinak: zdroj dat občas není znám, ale je znám problém. Úkolem datového analytika je tedy zjistit, odkud potřebné zdroje lze získat a zda je to vůbec možné. V praktické části svého projektu ukazuji, že například problém predikce ceny nájmu nemovitosti má jasně definovanou otázku - "Jaká je na základě inzerce nejpravděpodobnější cena měsíčního nájmu daného bytu?". Máme tedy určit, kam bychom se mohli podívat po relevantních zdrojích.

Internet je dobré místo, kde začít hledat data. Ale většinu z nich není snadné jen tak dostat. Ne všechna data jsou uložena ve strukturovaném tvaru buďto v souboru nebo databázi. Kdyby tomu tak bylo, stačilo by zjistit datové schéma patřičného souboru (např. JSON schéma odpovědí webového serveru nebo API, RDF schéma datové sady, atd.) nebo nastudovat dotazovací jazyk příslušné databáze (např. MySQL pro relační DBMS, SPARQL pro RDF databáze), což by nám umožnilo s daným uložištěm komunikovat.

Problém nastane ve chvíli, kdy nic takového nemáme k dispozici. Většinou se to stane v okamžiku, kdy se snažíme zorientovat se v možných relevantních datových zdrojích: monitorovací systém na výrobní lince, čidla a senzory u robotů, logy u SW, stránky na internetu. Relevantní data mohou být totiž přímo před našima očima na nějakém webovém portálu, který nemá aplikační rozhraní (API) pro strojové zpracování dat. Proto je v praxi běžné, že si analytik musí umět poradit s daty ve formátu HTML/XML nebo v nestrukturovaném textu (čímž jsou v podstatě aplikační logy).

Jelikož získávání dat je občas nezbytným úkonem, rozhodla jsem se této části datové analýzy věnovat větší pozornost a ukázat posluchačům, jak lze efektivně a jednoduše získávat relevantní data a pak je převádět do tvaru, kterým pak mnozí autoři začínají první kapitolu o datové analýze, tj. do tzv. datasetu. Tento přístup má i tu výhodu, že zohledňuje i reálnou složku každodenního života datového analytika - práce se skutečnými daty. Mnohé kurzy a tutoriály zkoumají datasety, které jsou buď předem záměrně připraveny pro didaktické účely nebo které jsou syntezovány tak, aby na nich šel ukázat studovaný jev, tj. je dopředu známo, že řešení problému, o němž jsem diskutovala v předchozí kapitole, existuje. Z mého pohledu takový přístup odtrhuje analytika od reality a nepřipravuje ho na výzvy, se kterými se může setkat v praxi. Toto jsou důvody pro to, jakým způsobem jsem koncipovala praktickou část předmětu, zaměřenou na práci s programovacím jazykem Python. Praktickou část jsem vypracovala tak, že jsem navrhla end-to-end postup (neboli pipeline), kterým

by měli projít zájemci o tuto disciplínu.

Pro data obsažená v HTML formátu nebo v podobné struktuře (XML, XHTML apod.) jsem využila několik technik parsingu, které vyžadují alespoň základní znalosti výše zmíněného Selenia, se kterým měli studenti možnost seznámit se v předmětu B6B36TS1, a HTML a CSS v předmětu B6B39ZWA. Celý proces probíhá následovně:

1. Pomocí Selenia jsem schopna procházet webové stránky a tak ukládat jejich HTML a CSS obsahy.
2. Dále za pomoci široce používaného nástroje BeautifulSoup jsem schopna procházet HTML/XML stromovou strukturu dokumentu a vyhledávat na základě identifikátorů a vlastností (přesněji atributů) prvků stromu potřebná data, např. text v seznamech, záznamy v tabulkách, tagy a klíčová slova u článků atd.
3. Získaná data uložím do uložiště v jednom z několika známých formátů na základě vhodnosti pro daný týp dat - například JSON slovník, CSV tabulka, kolekce tabulek, graf atd.

Po dokončení vyhledávání budeme mít seznam údajů, se kterými již lze pracovat a které je třeba analyzovat.

Během získání dat je potřeba dbát na kvalitu získáváných dat, protože většinou tento proces není ani rychlý, ani levný, ani opakovatelný. Představme si sběr dat pomocí měření teploty pacientů po podání nějakého experimentálního léku. Kdybychom se dopustili stejné chyby u každého pacienta (např. měřili nefunkčním nebo nekalibrovaným měřícím přístrojem), bylo by nutné pokus opakovat, což v případě léků není bezpečné. Podobné důvody platí u čistě IT projektů - analyzovaná data mohou být výstupem z velmi náročného a dlouhého zpracování, které je velmi nákladné. Navíc zpracování velkého množství dat je často zdlouhavé a každá odhalená chyba v dalších krocích může vést k prodlévám v realizaci projektu. Nedostatečná pozornost věnovaná volbě dat tedy povede k tomu, že data nebo modely nebudou pro zkoumaný problém vhodné a celou analýzu včetně sběru dat bude potřeba provádět znova.

Více k tomu v podkapitole 5.1.

4.2.3 Příprava dat - čištění a transformace dat

Příprava dat mi na začátku ze všech fází analýzy přišla jako nejméně problematický krok, ale ve skutečnosti vyžaduje obrovské množství úsilí a času. Data jsou často sbírána z různých zdrojů, z nichž každý je může nabízet ve své vlastní formě nebo formátu, a proto je musíme připravit pro proces analýzy, tj. provést unifikaci.

Příprava dat je logickým krokem poté, co jsme dokázali data najít a prohlásili je za relevantní. Je důležité pochopit, co do tohoto procesu vstupuje a co musí vystupovat. Jak jsem psala na začátku, na pojem *analýza* lze nahlížet různými způsoby, proto je potřeba pečlivě rozmyslet nejen název oboru, ale i jednotlivé kroky. Proto zkusme odpovědět na otázku, co máme v současné chvíli - máme většinou textové hodnoty, které jsme našli nebo obdrželi z datového zdroje. Umíme je interpretovat? Umíme říct, zda jsou připraveny pro další analýzu? Co vlastně další analýza bude vyžadovat? Jsou to otevřené otázky a není jednoduché na ně hledat odpovědi. Stejně tak jako není jednoduché navrhnout správný proces zpracování dat. Abych byla konkrétní, tak pokud se například jedná o záznam, který by měl reflektovat cenu nebo výměru, ale je uložen jako textový řetězec, nebudu schopna v dalších krocích s těmito daty pracovat. Proto proces přípravy dat slouží k tomu, aby data byla použitelná v dalších krocích analýzy.

Proces přípravy dat zahrnuje následující kroky: čištění, normalizaci a transformaci záznamů do "optimalizovaného"souboru dat. Obvykle se jedná o tabulkovou formu, která se pro analýzu dat jeví jako přirozená - v tabulkách sloupce slouží k popisu nějaké charakteristiky, vlastnosti, zatímco řádky většinou znázorňují jednu konkrétní datovou entitu. Například ve své praktické části jsem každý byt reprezentovala jako záznam v tabulce, zatímco každý sloupec znamenal například typ budovy, užitečnou plochu nebo počet pokojů.

Čištění dat je cíleno na záznamy, které nemají správné kódování, mají zbytečné řídicí znaky atd. Normalizace dat zahrnuje převod dat na stejný formát, například stejný formát pro datumy, částky atd. Tranformací dat je převod na jiný datový typ, doplnění dat o chybějící hodnoty.

V daném kroku je důležité ověřit správnost předchozího kroku - získání dat. Velmi často se stane to, že z nějakého výsledku této fáze vyplyne, že data jsou vygenerována nebo získána s nepřesnostmi, které pak budou ovlivňovat další fáze. V praktické části svého projektu jsem na takový problém zrovna narazila během studia kompletně nevyplněných sloupců u stažených detailů inzerátů.

Mnoho problémů ale může vzniknout kvůli neplatným, nejednoznačným nebo chybějícím hodnotám, duplicitním polím nebo datům, která se nevejdou do platného rozsahu. Tyto problémy můžou způsobit nekonzistence v budoucích fázích.

Více k tomu v podkapitole 5.2.

4.2.4 Zkoumání, generování a vizualizace dat

Zkoumání dat je analýza dat v grafické nebo statistické reprezentaci za účelem nalezení vzorců nebo vztahů, které se pak dají využít pro splnění konkrétního cíle. Vizualizace je jedním z nejlepších nástrojů pro zvýraznění a znázornění takových vzorů. V posledních letech se vizualizace dat vyvinula natolik, že se stala samostatnou disciplínou. Četné technologie a nástroje, z nichž jedním z nejpopulárnějších je Tableau (a PowerBI), se v dnešní době používají výhradně pro grafické zobrazení informace (ať už jde o grafy, matice, tabulky, dashboardy nebo něco dalšího).

Nicméně než se člověk pustí do samotné vizualizace, je potřeba se zamyslet nad tím, zda těmto datům opravdu rozumí. Nemalý přehled jsme mohli již získat v předchozích fázích, nicméně je nesmírně důležité si uvědomit, že většina dosud provedených úkonů byla připravná - data jsme nasbírali, ztransformovali, znormalizovali, připravili je pro strojové zpracování. Pořád těmto datům ale nerozumíme! Nevíme, které vlastnosti jsou popsány diskrétními veličinami, které naopak spojitými. Nevíme, zda v datech nejsou vychýlené hodnoty (anglicky *outliers*), které budou vadit při predikci a analýze závislosti a dokonce i visualizaci. Nevíme dále, jak moc máme prázdných hodnot a co s nimi máme dělat (zda potřebujeme tyto hodnoty vyhodit anebo provést tzv. *augmentaci* nebo *imputation* dat na základě existujících vzorků).

Tato část zpracování dat se v literatuře uvádí jako Explorative Data Analysis (EDA) (česky *explorativní datová analýza*). Cílem této analýzy je pochopit samotná data, najít v nich souvislosti, doplnit chybějící hodnoty nebo je dopočítat.

EDA se zabývá jak zkoumáním jedné proměnné s cílem pochopit, jak je rozložená, jaká jsou její minima a maxima, průměr, medián atd.; tak i zkoumáním vztahů mezi různými proměnnými. Jedním z nejzásadnějších kroků je například výpočet korelace mezi všemi spojitými náhodnými veličinami s cílem pochopit jejich vzájemné vztahy a možné vlivy. V případě diskrétních

veličin se zajímáme o jejich možné hodnoty, statistiky hodnot a vlivy hodnot na ostatní proměnné. V tomto bude hodně nápomocná vizualizace, která umožňuje názornější zobrazení nejen vztahů, ale i míry propojenosti dvou či více proměnných (např. pomocí heatmap). Vizualizace slouží k interpretaci statistik, proto jde ruku v ruce s EDA.

Zkoumání dat se skládá z předběžné studie, která je nezbytná k pochopení typu a významu sebraných informací. Spolu s informacemi sebranými při definici problému tato kategorizace určuje, která metoda analýzy dat je pro definici modelů nejvhodnější. Pro některé metody je například žádoucí, aby vstupní data splňovala nějaké jisté předpoklady (u lineální regrese vstupní data mají být normálně rozložena), proto je potřeba tyto předpoklady potvrdit pomocí vizualizačních nástrojů nebo testování hypotéz (nad rámec daného předmětu).

Více k tomu v podkapitole 5.3.

4.2.5 Predikce na základě získaných dat

Prediktivní analytika je proces analýzy dat, který se používá k vytvoření nebo nalezení vhodného statistického modelu k předpovědi pravděpodobnosti výsledku. Cílem prediktivního modelování je odpovědět na tuto otázku: „Co se na základě známého chování v minulosti s největší pravděpodobností stane v budoucnu?"

Klasifikace a regrese jsou dva hlavní predikční úlohy. Klasifikace je proces hledání nebo objevování modelu (funkce), který pomáhá při oddělování dat do více kategorií. V klasifikaci jsou data zařazena do různých kategorií na základě některých parametrů, ovlivňujících tzv. závislou proměnnou. Úkolem navrhovatele klasifikačního algoritmu je zjistit, jaké známé proměnné můžou co nejpřesněji určit ty neznámé. Regresní analýza je statistický model, který se používá k predikci spojitých dat namísto diskrétních. Tato metoda modeluje spojité funkce, proto je občas zvána generativní metoda.

Oba typy predikcí jsou ovšem závislé jednak na vstupních datech a jednak na vnitřních parametrech použité metody. Oba tyto faktory hrají kritickou roli v budoucím úspěchu či neúspěchu predikce. Prediktivní modely jsou ve své podstatě navrženy tak, aby byly schopny popsat neznámý statistický model, nebo dokonce parametry známého statistického modelu (viz. metoda maximální věrohodnosti). Tyto parametry zjistí na základě výběru z procesu,

který se chová podle hledaného statistického modelu. Proto je kritické mít o tomto modelu co nejvíce vypovídající datový vzorek. Vypovídací hodnotu má jak velikost datového vzorku (čím více dat máme, tím obecně lépe jsme schopní odhadnout neznámé parametry), tak i kvalita jednotlivých částí každého vzorku. Vzorek či měření se obecně skládá z několika částí, kterým se říká *příznaky* neboli anglicky features. Tyto příznaky popisují vzorek a přidávají mu obrysy, na jejichž základě jsme schopni určit další charakteristiky vzorku. Například na základě informace o počtu pokojů v inzerovaném bytě a blízkosti k MHD jsme shopni lépe určit hodnotu této nemovitosti. Proto je důležité mít datový model, který dobře určuje hledanou hodnotu. U predikce touto hodnotou může být např. cena pronájmu či měsíční náklady.

Datový model je předmětem detailní analýzy, kterou provádí datový analytik na základě dat, která má k dispozici, společně s expertem v řešené doméně. V případě prodeje či pronájmu nemovitosti může jít o reáltního makléře. Expert určí příznaky, které jsou rozhodující v otázce predikované hodnoty - ceny, pak datový analytik zajistí dostupnost určených dat a jejich kvalitu. Kvalita dat v otázce predikce hraje nejdůležitější roli, proto je povinností datového analytika průzkum všech významných proměnných a zajištění jejich kvality, čimž je standardizace dat, čištění dat, normalizace dat apod. Více k tomu v podkapitole 5.4.

Kromě transformace existujících dat je rovněž potřeba zajistit jejich přítomnost ve všech vzorcích - byt s chybějící výměrou nebude vhodným vzorkem pro přesný výpočet prodejní ceny. Chybějící hodnoty v réalných úlohách oproti těm, které jsou v hojném počtu k dispozici na internetu s pečlivě předpřipravenými datasety, představují obrovskou výzvu. Chybějící hodnoty se mají detailně studovat a u každé takové proměnné se musí rozhodnout, jakým způsobem ji dopočítat nebo vůbec. Například, když chybí jen pár hodnot u nějaké proměnné, chybějící hodnoty lze nahradit agregovanou hodnotou, vypočtenou z těch přítomných - např. průměr nebo medián, nebo poddrobit důslednějšímu zkoumání. Pokud ale chybí většina hodnot u néjaké proměnné, není možné ji dopočítat na základě jen pár existujících - dopouštěli bychom se velké chyby. Úplně chybějící hodnoty u nějaké proměnné lze vyřešit lepším měřením či odstraněním nejpravděpodobněji nějaké chyby při sběru dat. V nejnepříznivějším případě je potřeba se rozhodnout, zda takovou proměnnou nevynechat z datasetu.

Než se analytik začne zabývat samotnou predikcí, musí mít k dispozici datový vzorek, u kterého si je jistý, že má v sobě potřebnou kvalitu. Pokud do prediktivního modelu vstupuje nekvalitní vzorek, sotva se model natrénuje přesně. Kvalitní datový vzor tvoří většinu úspěchu budoucího modelu. Nekvalitní dataset má za následek model s velkou chybovostí nebo vychýleností (např. predikuje stále pouze dražší nemovistosti).

Predikce je založená na principu trénování modelu strojového účení. Těchto modelů je obrovské množství. Každý ale má své silné a slabé stránky. Každý má své předpoklady, které klade na data a na datový a statistický model hledané hodnoty, které je potřeba splnit, aby výstup z modelu byt věrohodný - viz. lineární regrese [Fre05]. Některé modely si lépe poradí s různými typy vstupních dat - viz. rozhodovací stromy. Některé modely mají pouze jedno unikátní řešení - viz. SVM [Fre05], některé z nich mají nekonečně mnoho - viz. perceptron [Fre05]. Každý z nich je založen na vlastním matematickém modelu a principu. Z tohoto důvodu zde neuvádím všechny a nejdu dopodrobna. Nejdůležitější je za pomocí kvalitního datového vzorku vyzkoušet si ty, pro které platí předpoklady kladené na data, a vybrat si ten model, který vykazuje lepší chování.

Je zvykem u predikce rozdělit existující datový vzorek na část trénovací a testovací. Trénovací vzorek vstupuje do trénování modelu, testovací vzorek slouží k nezávislému ověření funkčnosti modelu na datech, která předtím neviděl. Porovnání modelů se provádí na základě společné metriky, která je pro danou úlohu nejvhodnější. Například u úlohy predikce numerické hodnoty to může být RMSE - root mean squared error - průměrná hodnota absolutní hodnoty odchylky predikované hodnoty od té pravé.

Více k tomu v podkapitole 5.4.

4.3 Jaké nástroje se používají pro analýzu dat?

Po tom, co jsme si ukázali postup při analýze dat, bylo by vhodné říct, jaké nástroje můžou plnit každou jednotlivou fázi. S tím, jak je datová analýza obrovský obor, a jak moc problémů může řešit, bylo vyvinuto opravdu velké množství nástrojů, aplikací a jazyků, které nyní používají odborníci denně, aby dosáhli požadovaného výsledku. V rámci tohoto kurzu by nebylo možné je všechny ukázat, proto pojďme probrat ty 3 nejčastější, se kterými se setká začátečník v této oblasti: Python, Excel a PowerBI/Tableau.

Python je plnohodnotný programovací jazyk, který se používá v mnoha různých odvětvích. Jedno z nejpopulárnějších využití má Python právě v oblasti datové analýzy a strojového učení. Python je jednoduchý, interpretovatelný jazyk s podporou objektově orientovaného programování. Učicí křivka v Pythonu je velmi strmá v porovnání s jazyky jako je Java či C++. Python jako nástroj pro analýzu dat poskytuje téměř neomezené možnosti - umí zpracovávat tabulky, dobře si poradí s velkými datovými zdroji, dá se v něm vytvořit

vizualizace a reporty. Spolu s analytickou částí, v níž jsou většinou poptávány kroky jako získání, extrakce, normalizace a transformace dat, poskytuje jednoduché nástroje na vyvtoření webových služeb a API pro snadnější integraci. Veškerý tento arzenál je ale potřeba vždy naprogramovat ručně. Holý Python nemá grafické rozhraní a pro potřebnou analýzu dat ji potřebuje naprogramovat od začátku do konce. Proto není určený pro každého, kdo se chce zabývat analýzou dat, jelikož vyžaduje alespoň základní znalosti programování a datových struktur. Více informace o výhodách, nevýhodách a použití Pythonu a příslušných knihoven uvádí ve svých knihách Wes McKinney, viz [McK17], a Jacqueline Kazil s Katharine Jarmul, viz [Jac16].

Excel je tradičním nástrojem pro práci s tabulkami. Ve chvíli, kdy se zpracovávají data ve formátu tabulek, je tento nástroj velmi oblíbený u datových analytiků. Nicméně má omezené možnosti z pohledu získání a čištění dat - Excel je aplikace s uživatelským rozhraním, do které uživatel nahraje soubor, jejž chce analyzovat. Excel se zpravidla používá ve chvíli, kdy data jsou ve správnem formátu a kdy nemají chybějící hodnoty. Excel je mnohem lepší nástroj na rychlou exploratórní analýzu dat a na spočtení základních statistik a ověření hypotéz. Další silnou stránkou jsou tzv. kontingenční tabulky - agregované tabulky, umožňující jiný pohled na data, viz. Obrázek 4.3. Excel sice disponuje možností psaní kódu ve Visual Basicu, není to v současné době standardem a využívá se velice zřídka.

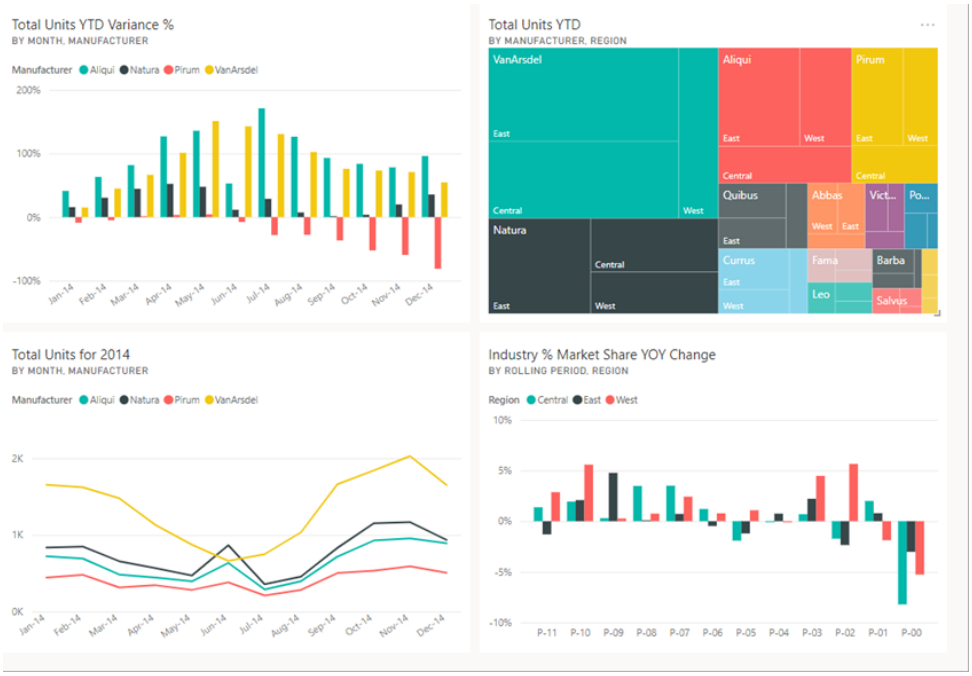
E1 =ZÍSKATKONTDATA("Prodej",A3)

	A	B	C	D	E
1	Oblast	Sever		Celkový prodej	55 340 Kč
2					
3	Součet prodejů		Produkt		
4	Měsíc	Prodejce	Nápoje	Plodiny	lkový součet
5	Led	Karásek	5 310 Kč	4 060 Kč	9 370 Kč
6	Led	Chvojková	880 Kč	990 Kč	1 870 Kč
7	Souhrn za leden		6 190 Kč	5 050 Kč	11 240 Kč
8	Úno	Karásek	2 310 Kč	8 800 Kč	11 110 Kč
9	Úno	Chvojková	1 320 Kč	2 910 Kč	4 230 Kč
10	Souhrn za únor		3 630 Kč	11 710 Kč	15 340 Kč
11	Bře	Karásek	9 620 Kč	3 090 Kč	12 710 Kč
12	Bře	Chvojková	9 150 Kč	6 900 Kč	16 050 Kč
13	Souhrn za březen		18 770 Kč	9 990 Kč	28 760 Kč
14	Celkový součet		28 590 Kč	26 750 Kč	55 340 Kč

Obrázek 4.3: Ukázka kontingenční tabulky v Excelu. V tomto příkladu vrátí funkce =ZÍSKATKONTDATA("Prodej";A3) celkovou částku prodeje z kontingenční tabulky. Zdroj [Sup22]

PowerBI je obdobou Excelu a spíše jeho vylepšením vůči netechnickým lidem. Díky tomu, že má přívětivější uživatelské rozhraní, jednodušší vizualizační funkce a snadné generování reportů, je tento nástroj navržen pro business uživatele. PowerBI je cloudová služba, která má spoustu možností pro zís-

kání dat - má širokou škálu nástrojů a služeb, odkud si může stáhnout data. Dále nemá problém s většími datasety oproti Excelu a je určen zejména pro vytvoření tzv. dashboardů (přehledů) pro manažery a pro tzv. decision makery - pro ty, kteří konzumují data a potřebují je vidět ve formátu, který jim poskytuje nejvíce informací. Proto je v PowerBI kladen důraz na snadné vytvoření visualizací, rychlou konzumaci dat a dostupnost reportů či dashboardů jak z klasických obrazovek, tak i z mobilu. Navíc prezentace dat formou PowerBI (příklad je uveden na Obrázku 4.4) je interaktivní - uživatel je schopen interagovat s reportem a přizpůsobovat ho, zatímco v Excelu je nutný zásah kvalifikovanějšího člověka.



Obrázek 4.4: Ukázka PowerBI rozhraní. Zdroj [Lea21]. Pro představení více grafů umožňujících přehlednou prezentací informací je určen zdroj [Lea22]

Nakonec PowerBI vyžaduje nejméně technických dovedností pro ovládání ze všech tří nástrojů. Excel je výbornou aplikací pro práci s tabulkami a manipulaci s nimi, zatímco Python je plnohodnotný švýcarský nůž, který dokáže přizpusobit proces analýzy dat přesně dle potřeb uživatele, ale za cenu hlubších znalostí programování a vizualizace.

Když se podíváme na Obrázek 4.5, který shrnuje možnosti každého z výše uvedených nástrojů, tak uvidíme, že naše požadavky nejlépe splňuje Python, což znamená, že je nejvíce univerzální volbou pro tento kurz. To je jeden z důvodů, proč se v praktické části bude jednat pouze o práci s Pythonem a s příslušnými knihovnami. Druhým důvodem je to, že Python je bází, standardem pro datovou analýzou, a proto se v kurzu pro začátečníky probírat musí.

	Excel	Python	Tableau/Power BI
Získání dat	Ne	Ano	Ne
Čištění a transformace	Ano	Ano	Částečně
Zkoumání, generování a vizualizace	Částečně	Ano	Částečně
Predikce	Ne	Ano	Ano

Obrázek 4.5: Tabulka ukazující možnosti různých nástrojů pro datovou analýzu.

Důležité je ale zmínit, že stejně jako i v jakémkoliv jiném IT odvětví, programovací jazyk sám o sobě toho moc neudělá. Pro smysluplné úkoly a jejich řešení je vždy třeba mít pod rukou vhodné rozhraní. V případě Pythonu je většina lidí zvyklá na vývojové prostředí PyCharm od JetBrains, ale v datové analýze je zvykem používat Jupyter Notebook, který vždy jde spolu s Anacondou. Co to ale je, k čemu a proč se používá, bude vysvětleno v následujících podsekcích: 4.3.1 a 4.3.2

4.3.1 Anaconda

Analýza dat je oborem, v němž existuje obrovské množství různých knihoven a nástrojů pro práci s daty. Spousta z nich má velké komunity vývojářů a velkou historii změn a novinek. Každá datová úloha je iterativním procesem, který se stoupajícím množstvím požadavků zpravidla zvýšuje množství nástrojů, které se používají pro splnění těchto požadavků. V praxi to znamená, že se na řešení různých úkolů používají různé knihovny a frameworky. Proto pro každý projekt v ideálním případě by mělo existovat izolované pracovní prostředí s jen nutnými knihovnami a moduly. Tato prostředí by měla udžovat závislosti mezi moduly, platné jen pro daný projekt, stejně tak jako jejich verze. Fixace přesných verzí v ekosystému jazyka Python je klíčové, jelikož novější verze knihoven často rozbíjí fungování kódů, psaných pro verze starší, a to z důvodu přejmenování funkcí, změn jejich parametrů či jejich zániku.

Dalším důvodem pro izolaci prostředí je fakt, že každý projekt vyžaduje jinou sadu knihoven, častokrát i jiné verze stejných knihoven, což není možné udržet v rámci globálně nainstalovaného Pythonu. Dále je nutné podotknout, že datová analýza se ve dnešní době provádí na různých místech - v cloudu, na vestavných zařízeních, na různých operačních systémech a různých architekturách, proto omezení arzenálu knihoven a striknì dodržení jejich verzí je pro datovou analýzu klíčové.

Právě z výše zmíněných důvodů v ekosystému Pythonu existuje manažer prostředí - **Anaconda** (nebo **conda**), viz [anaom]. Conda se používá pro naplnění potřeb z předchozích dvou odstavců - je nástrojem, který umožňuje vytvoření a použití několika na sobě nezávislých Python prostředí s různými knihovnami a moduly. Conda se hojně využívá pro souběžnou práci na několika projektech, z nichž každý má jiné požadavky. Conda umožňuje nejen instalaci různých verzí balíčků, ale i různých verzí Pythonu, např. v robotice se stále aktivně používá Python verze 2, který již oficiálně není podporován.

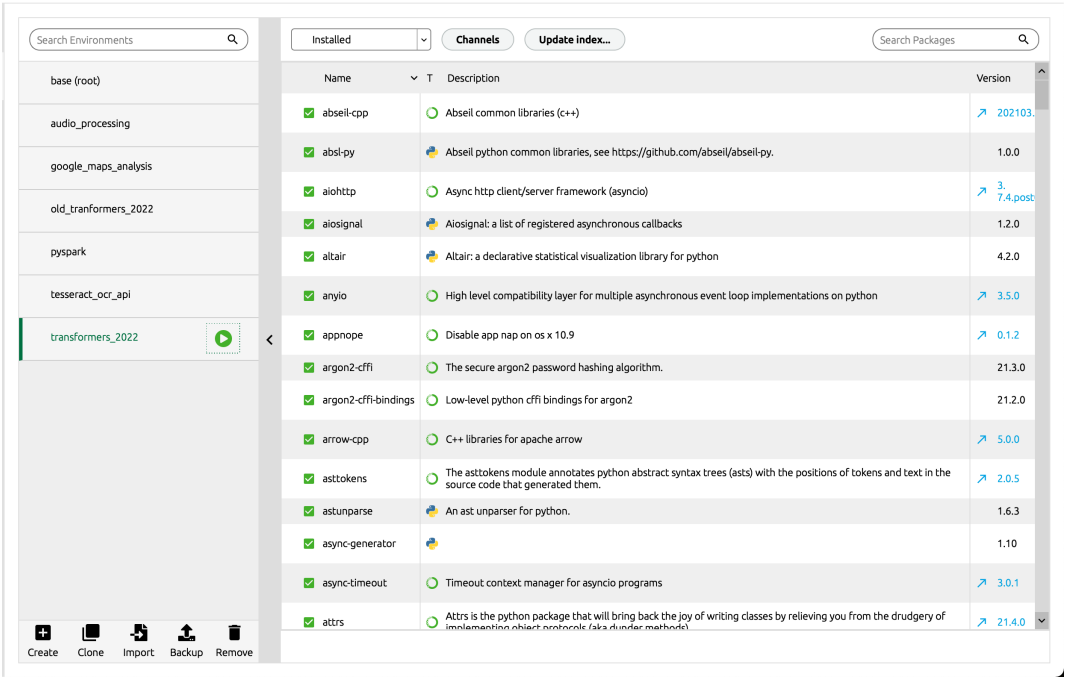
Conda funguje následujícím způsobem:

1. Pro každé prostředí se vytvoří zvláštní adresář, v němž je uložen spouštěč interpretátoru Pythonu spolu s příslušnými moduly.
2. Jakmile se uživatel rozhodne použít pro práci s projektem dané prostředí, Anaconda ho navede na správný spouštěč Pythonu, který následně zidentifikuje jemu patřící moduly a knihovny.
3. Tento spouštěč neví o existenci paralelních Python spouštěčů, proto to pro zdrojové kódy a grafická vývojová prostředí nevytváří zmatek.

Anaconda a conda jsou synonyma. Jediný rozdíl spočívá v tom, že Anaconda nabízí pohodlné uživatelské rozhraní, které slouží pro přehled všech prostředí a manipulaci s nimi, viz Obrázek 4.6. Pod kapotou ale běží conda, proto tato slova lze považovat za stejná. Jinými slovy, Anaconda = conda + UI.

Pro více informací ohledně práce s Anacondou nemohu doporučit nic lepšího než oficiální dokumentaci, viz [anaor].

Anaconda prostředí jsou jednoduše přenositelná na jiná zařízení, což je výhodné pro kontinuální a plynulou práci se zdrojovými kódy.



Obrázek 4.6: Příklad uživatelského rozhraní Anaconda. Vlevo je seznam existujících prostředí, vpravo je přehled instalovaných Python modulů, které jsou součástí tohoto prostředí. Toto rozhraní nabízí kompletní správu prostředí.

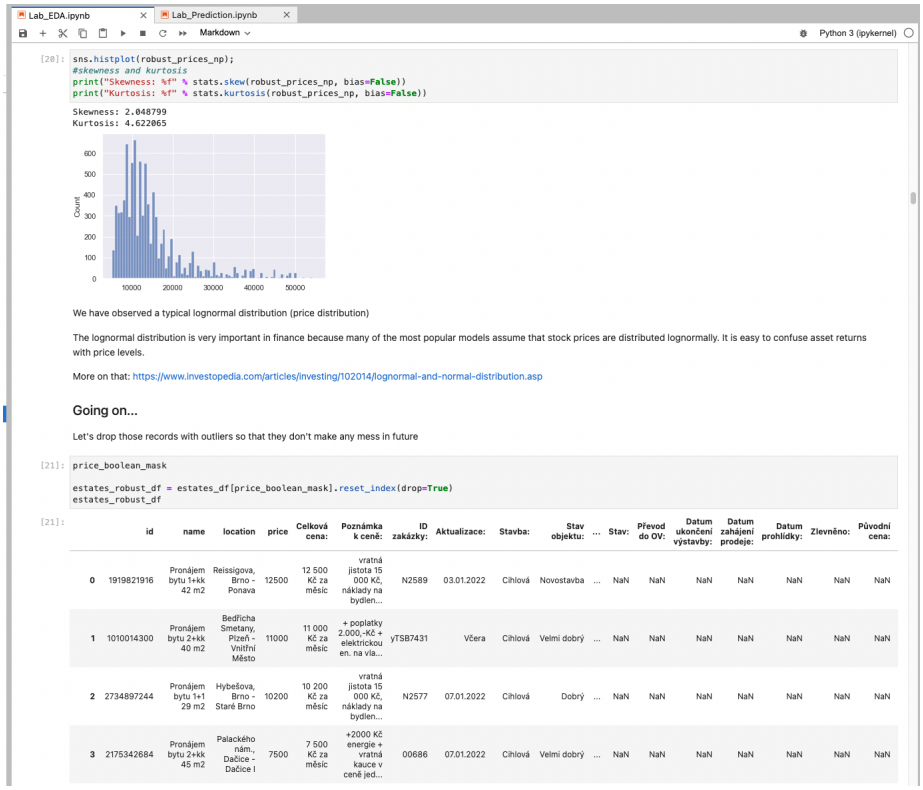
4.3.2 Jupyter Notebook

Co se práce se zdrojovými kódy ale týče, datoví analytici většinou pracují iterativně a protože hodně zkoumají data a testují hypotézy, nemají v oblibě standardní vývojová prostředí jako PyCharm nebo VSCode. Tyto IDE zpravidla spouštějí kód jako sadu příkazů a pak skončí. Pro datovou analýzu je ale pohodlnější tesnější interakce s daty, proto v oboru je zvykem používání jiného nástroje - a tím je Jupyter Notebook nebo Jupyter Lab.

Jupyter Notebook, viz [juprg], je webové rozhraní, redaktor kódu, který se zakládá na principu tzv. sessions, neboli globálního stavu Python interpretátoru, kdy se s tímto stavem dá interagovat, například zadefinovat proměnnou, provést změnu a pak se dotázat na aktuální hodnotu dané proměnné. Tento princip je velice podobný principu debuggeru, když se kód spouští krok za krokem a výsledek jednotlivých kroků je uložen do globalního stavu aktuálního běhu a je dostupný ke čtení a editaci. Jupyter Notebook je přesně tímto nástrojem, který umožňuje psát kód, spouštět ho a pak dále doplňovat kód bez nutnosti spouštění ještě jednou. Je založen na principu pracovních sešitů (tzv. notebooků) a tzv. buněk (anglicky cells), které tvoří jednotlivý blok kódu. Notebook jako takový je jinak formatovaný zdrojový kód psaný v Pythonu. Jednotlivé buňky v noteboocích lze spustit zvlášť. Veškerý výstup se zobrazí pod patřičnou buňkou. Výsledek spuštění kódu v buňce se uloží do globalního stavu a vzniklé proměnné budou dostupné pro čtení a editaci v jiných buňkách. Tímto způsobem je analytik schopen pracovat postupně a rychle se dostat k obsahu tabulek, souborů, rychle vytvořit visualizaci a navrhnout postupy a porovnat je. Příklad takového prostředí je zobrazen na obrázku 4.7.

Notebooky tvoří základ analytického postupu a zároveň jej dokumentují. Notebooky lze současně považovat jak za výsledek práce analytika ve formě zdrojového kódu, tak i reportu, který lze prezentovat samostatně a který je navíc interaktivní pro potřebu rychlé změny. Proto jsou notebooky tak populární, že se staly standardem v analýze dat a v současné době valná většina analytických knihoven je optimalizována pro práci v noteboocích, např. visualizační knihovny neukládají výsledky do disku, ale rovnou je zobrazují jako výstup z notebooku.

I když je práce v Jupyteru docela intuitivní, pro ty, kteří se s tím nikdy v životě nepotkali, bych velice doporučila podívat se na již hotové návody pro používání daného nástroje. Může to ušetřit hodně času hledáním a poznáváním důležitých funkcionalit a možností Jupyteru. Pro tyto účely mohu nabídnout videotutorial od Corey Schafer, viz [jupWk] nebo standardní textový návod od Dataquest [jupal].



	id	name	location	price	Celková cena:	Poznámka k ceně:	ID zakázky:	Aktualizace:	Stavba:	Stav objektu:	...	Stav:	Převod do OV:	Datum ukončení výstavby:	Datum zahájení prodeje:	Datum prohlídky:	Zlevněno:	Původní cena:
0	1919821916	Pronájem bytu 1+kk 42 m2	Reissigova, Brno - Ponava	12500	12 500 Kč za měsíc	vratná jistota 15 000 Kč, náklady na bydlen...	N2589	03.01.2022	Cihlová	Novostavba	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	1010014300	Pronájem bytu 2+kk 40 m2	Bedřicha Smetany, Plzeň - Vnitřní Město	11000	11 000 Kč za měsíc	+ poplatky 2.000,-Kč + elektrickou en. na vla...	yTSB7431	Včera	Cihlová	Velmi dobrý	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	2734897244	Pronájem bytu 1+1 29 m2	Hybešova, Brno - Staré Brno	10200	10 200 Kč za měsíc	vratná jistota 15 000 Kč, náklady na bydlen...	N2577	07.01.2022	Cihlová	Dobrý	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	2175342684	Pronájem bytu 2+kk 45 m2	Palackého nám., Dačice - Dačice I	7500	7 500 Kč za měsíc	+2000 Kč energie + vratná kauce v ceně jed...	00686	07.01.2022	Cihlová	Velmi dobrý	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Obrázek 4.7: Příklad části notebooku, v níž jsou vidět buňky (jsou na šedivějším pozadí, očislována 20, 21) s kódy a výstupy vykonání těchto příkazů - vypočtené statistiky, nakreslený graf, vytištěná tabulka.

Kapitola 5

Praktická část

Jak jsem už několikrát zmínila, praktická část mé bakalářské práce je zaměřena výhradně na práci s daty v Pythonu a její výstup se skládá ze 4 notebooků (viz. Gitlab [1] - jen pro přidané uživatele, nebo soubory v zip archivu), které odpovídají výše popsaným krokům postupu při analýze dat. V každém notebooku pro lepší pochopení jeho fungování budou popsány jednotlivé nejdůležitější kroky, které se tam odehrávají. Pro budoucí možnost využití notebooků v roli studijního materiálu jsem je psala v anglickém jazyce, abych neomezovala cílovou skupinu jenom na česky mluvící lidi. Před spouštěním projektu doporučuji přečíst *README.txt* a *requirements.txt* a nastavit podle ně pracovní prostředí.

Definici problému, jakožto prvnímu kroku při analýze, nebudeme v rámci této práce věnovat velké množství času kvůli důvodům, které jsem už zmiňovala v odpovídající kapitole, a proto místo něj jen krátce zopakuji cíl, kterého chceme dosáhnout. Naším úkolem je odpovědět na otázku "Jaká je nejpravděpodobnější cena měsíčního nájmu daného bytu na základě inzerce?".

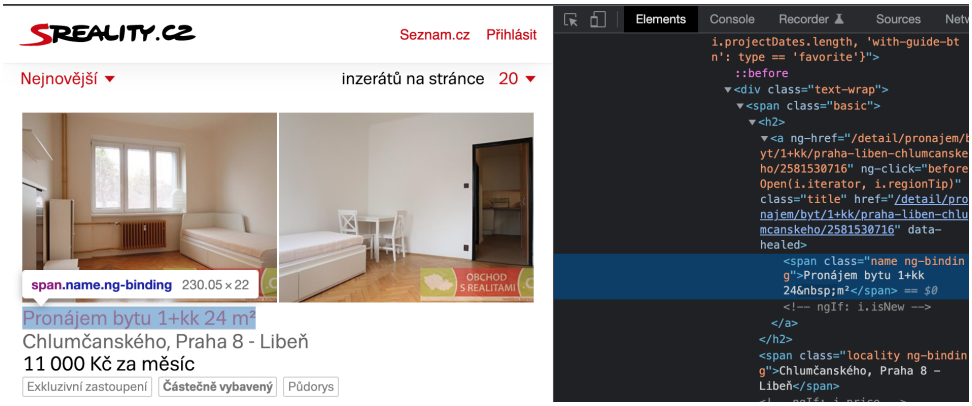
Pro vyřešení tohotu problému jsem využila stránky Sreality z toho důvodu, že jsou o něco lépe koncipovány než například Bezrealitky a podobné zdroje; a s ohledem na to, že stále naším cílem je se zaměřit na úroveň znalostí studentů SITu, tak nechceme probírat ty nejtěžší příklady na prvních etapách seznámení se s datovou analýzou.

[1] Celý projekt je dostupný skrz odkaz - https://gitlab.fel.cvut.cz/tahirole/semestralni-projekt.

5.1 Notebook 1 - Získání a extrakce dat

Notebook přidané uživatelé najdou na Gitlabu [1], viz notebook *1-Data-Acquisition-and-Extraction*. Pro ostatní je dostupný v zip archivu.

Naším úkolem v tomto notebooku je stáhnout co nejvíce informací o každém bytě vystavěném na pronájem. Celý proces vytěžování spočívá v tom, že otevíráme vybrané stránky [2] se seznamem všech dostupných inzerátů, pak zkoumáme, jakým způsobem se dá chytře získat odkaz na každý z nich. Toto zkoumání spočívá v tom, že si v prohlížeči otevřeme okno s HTML kódem a nalezneme příslušný kontejner a jeho identifikátor, který pak využíjeme k extrakci dat, viz Obrázek 5.1, a potom otevíráme každou stránku s inzerátem a stahujeme HTML soubor a id každého inzerátu na lokální dataset. Všechny úkony se musí provést automaticky, proto klíčovou roli v tomto notebooku hraje nástroj Selenium, který jsem používala přes Python rozhraní.



Obrázek 5.1: Proces zkoumání HTML šablony

Zde je docela důležité zmínit, že veškerý proces je velice časově náročný (v našem případě to bylo kolem 15 000 stránek a trvalo to několik hodin), a je nutné s tím počítat. Také bych chtěla ještě jednou zdůraznit, že pro další kroky je velice důležité celý proces projít postupně (a na malém vzorku, např. na pár stránkách) a ošetřit případné chyby, které se s největší pravděpodobností tak či onak vyskytnou.

[1] Všechny notebooky použité v daném projektu se nachází na mém GitLabu, a to na adrese https://gitlab.fel.cvut.cz/tahirole/semestralni-projekt.

[2] SReality.cz - inzeráty s nemovitostmi - https://www.sreality.cz/.

Důležité funkce/knihovny, použité v tomto notebooku:

1. Python funkce pro zpracování seznamů: list(map), lambda funkce, reduce, filter,
2. Python funkce pro práci s URL adresami: urllib.parse a urllib.urljoin,
3. BeautifulSoup - nástroj pro exktrakci dat z HTML/XML souborů.

Jelikož proces stahování velkého množství stránek je dlouhý, notebook obsahuje spoustu duplicitního kódu, který jsem přizpůsobila tak, abych v případě výskytu nějaké chyby nemusela začínat od začátku a mohla pokračovat.

V případě, že se HTML struktura stránek změní, což je běžná praktika, jsem všechny stažené inzeráty uložila v příloze, aby studenti mohli pokračovat v procesu bez nutnosti přizpůsobovat kód novému HTML stromu.

5.2 Notebook 2 - Příprava dat - čištění a transformace dat

Notebook přidané uživatelé najdou na Gitlabu [1], viz notebook *2-Data-Preparation-and-Data-Transformation*. Pro ostatní je dostupný v zip archivu.

Cílem tohoto notebooku je práce s již staženými HTML soubory za účelem převodu neupravených dat do lepšího a pohodlnějšího formátu pro další zpracování formátu, například do tabulky.

Po první části prodělané práce máme představu o tom, že se každý HTML soubor skládá ze tří struktur, které nás budou zajímat:

1. metadata každého inzerátu - popis vlastnosti každé nemovitosti (výměra, podlaží, počet bazénů atd.),
2. textový popis inzerátu - důvod je ten, že občas užitečná data o nemovitosti jsou obsažena jen ve slovním popisu, nikoliv v tabulce,

[1] Všechny notebooky použité v daném projektu se nachází na mém GitLabu, a to na adrese https://gitlab.fel.cvut.cz/tahirole/semestralni-projekt.

3. seznam služeb v okolí bytu a jejich vzdálenost (zastávka MHD, restaurace, škola atd.).

Pomocí knihovny BeautifulSoup budeme vytěžovat každou z nich, což je docela podrobně po krocích popsáno v samotném notebooku. Metadata každého inzerátu se uloží společně do jedné tabulky, služby v okolí bytu pak budou uloženy ve zvláštních tabulkách, zvláštních pro každý inzerát, a textový popis se uloží v jednom hromadném JSON souboru.

Dalším krokem bylo to, že jsem musela vymyslet, jakým způsobem chci chytře ukazovat všechna data o vlastnostech bytů ze všech inzerátů najednou, abych se pokaždé při porovnání nějakých hodnot nemusela dívat do více tabulek. Zde se konceptuálně důležitým řešením, jak jsem již naznačila v odstavci výše, objevila možnost otočit původně navrženou tabulku o 90 stupňů, aby to nebyla sturktura vždy o dvou sloupcích, která se prakticky moc nedá využít, ale opravdová, smysl dávající tabulka, kde sloupce popisují nějaký konkrétní typ metadat nebo vlastnost (jméno, cena atd.), řádky popisují samotné byty neboli zvláštní inzeráty. Tabulku se službami bohužel do toho nešlo chytře připojit, protože to je komplexnější struktura, která se skládá z unikátních informací pro každý byt, proto jsem je uložila zvlášť pro každý byt. Každá tabulka je rozlišitelná pomocí identifikátoru inzerátu.

Důležité funkce/knihovny, použité v tomto notebooku:

1. Pathlib - Python knihovna pro práci se soubory a se souborovým systémem,
2. BeautifulSoup - extrakce dat z HTML struktur,
3. Pandas - knihovna pro práci s daty, která jsou reprezentována jako tabulka. Transformace, čištění a normalizace dat.

Výstupem ze druhého notebooku je předzpracovaný a připravený dataset k použití, uložený na lokálním souborovém systému.

5.3 Notebook 3 - Zkoumání, generování a vizualizace dat

Notebook přidané uživatelé najdou na Gitlabu [1], viz notebook *3-Data-exploration-generation-and-visualisation*. Pro ostatní je dostupný v zip archivu.

Nyní, po prodělané práci, je třeba zjistit, o jaká data se v tabulkách konkrétně jedná. Představme si, že výstupy z druhého notebooku, tj. všechny připravené tabulky teď vidíme poprvé v životě. Potřebujeme se s daty nejprve seznámit a porozumět jim.

V tomto notebooku následně provedeme základní kroky exploratorní analýzy dat - EDA. Samozřejmě se podíváme na názvy sloupců v tabulce, abychom alespoň přibližně pochopili, o co se jedná; pak pomocí knihovny Pandas můžeme jednoduše zjistit počet nenulových hodnot v každém z nich, dozvíme se o datových typech a spoustu jiných věcí. Po tom, co už alespoň trochu máme přehled o tabulce, vrátíme se ke hlavnímu cíli provedení celé analýzy a pochopíme, že máme začít pracovat s cenou, o tu přece máme největší zájem. Práce s cenou jako s datovým typem *string* (což je teď formát ceny) nám moc nepomůže v predikci cen bytů. Musíme umět pracovat s cenou jako s numerickou hodnotou, nikoliv řetězcem, a proto podle návodu, který je popsán v samotném notebooku, převedeme tento sloupec na typ *int*.

První krok máme za sebou. Dále nás zajímá rozložení cen všech bytů, a proto zkusíme spočítat a zobrazit histogram. Narážíme ale na další problémy, jako jsou outliery neboli odlehlé hodnoty, které pomocí klasických statistických metod, jako je např. interval spolehlivosti, musíme odstranit. Potom převedeme rozdělení ceny na normální rozdělení (původně to bylo logaritmicko-normální rozdělení), které nám v budoucnu pomůže s předpovídáním ceny. Téměř identický proces uděláme i s dalšími položkami, například s počtem pokojů nebo výměrou.

Notebook navíc poskytuje hodně jiných ukázek, zajímavých a důležitých momentů, na které jsem narazila při práci s tímto datasetem. Jsou tam popsány jednotlivé metody a je také představena práce s datem a časem, s diskrétními veličinami, s vykreslováním různých závislostních grafů a statistik atd.

[1] Všechny notebooky použité v daném projektu se nachází na mém GitLabu, a to na adrese https://gitlab.fel.cvut.cz/tahirole/semestralni-projekt.

Důležité funkce/knihovny, použité v tomto notebooku: všechny, uvedené v předchozích 2 noteboocích, a k tomu navíc:

1. numpy - Python knihovna pro práci s numerickými daty a výpočet statistik,
2. seaborn - Python knihovna pro vizualizaci,
3. datetime - standardní Python knihovna pro práci s datumy a časy, poskytuje funkcionalitu pro operaci s časovými datovými položkami,
4. re - standardní Python knihovna pro práci s regulárními výrazy.

5.4 Notebook 4 - Predikce na základě získaných dat

Notebook přidané uživatelé najdou na Gitlabu [1], viz notebook *4-Prediction-Basing-On-Obtained-Data*. Pro ostatní je dostupný v zip archivu.

Ve fázi EDA jsme zjistili, co se v datech nachází, v jakém jsou data vztahu vůči zkoumané proměnné (cena nemovitosti), provedli základní analýzu a standardizaci proměnných. Nyní se logicky zabýváme dalším krokem analýzy dat, kterým je predikce. Predikce si klade za cíl najít vztah mezi proměnnými, které můžeme získat pomocí měření či parsingu, a hodnotou, o kterou se zajímáme. V daném případě by nás zajímala predikce ceny pronájmu nemovistosti na základě jejich parametrů. Parametry nemovistosti jsme v detailu studovali v předchozích noteboocích. Celkově máme následující datové zdroje:

1. Hlavní tabulku se základními parametry - výměra, podlaží, dispozice atd.
2. Vedlejší JSON databázi s popisy každé nemovitosti.
3. Tabulku se vzdálenostmi k nejbližším objektům občanské vybavenosti - kavárnám, restauracím, metru, škole atd.

[1] Všechny notebooky použité v daném projektu se nachází na mém GitLabu, a to na adrese https://gitlab.fel.cvut.cz/tahirole/semestralni-projekt.

Pro účely predikce je potřeba si pamatovat, že metodicky existují dva typy proměnných - nezávislé a závislé. Závislá proměnná je ta, kterou se pokoušíme predikovat. Nezávislé proměnné jsou ty, které máme na vstupu: vyplněný formulář či měření. Nezávislá proměnná je takto pojmenována přesně z důvodu, že ji nemůžeme ovlivnit - ty jsou pouze dané. V úloze predikce je potřeba prozkoumat nezávislé proměnné do detailu a připravit je pro vstup do prediktivního modelu.

Příprava nezávislých proměnných se jinak nazývá feature engineering. Je to zčásti podobný process jako EDA, ale zde se snažíme o to, aby proměnné před vstupem do modelu nesly nějakou užitečnou informaci. Feature engineering je nejnáročnější process, protože zahrnuje:

1. Výběr relevantních proměnných a odstranění těch, které nemají na výsledek žádný vliv.
2. Správu proměnných s chybějícími záznamy: NaN hodnoty. Rozhodnutí, zda tyto hodnoty dopočítat (a jak) nebo tuto proměnnou zcela vynechat.
3. Extrakce dat z proměnných, které nejsou ve "správném"formátu.

Výběr relevantních proměnných je závislý na problému, který řešíme. Například, je zcela evidentní, že čím je nemovitost větší, tím je zpravidla vyšší její cena. Proto v případě, kdy nám v datasetu chybí výměra, je potřeba tento údaj nejspíše doplnit. Je potřeba se rovnou rozhodnout, které proměnné lze vyloučit, abychom nezesložiťovali model - toto rozhodnutí je opět předmětem domluvy s doménovým expertem a může být podpořeno ve fázi EDA, např. skutečnost, že daný byt má blízko ke kinu, nemá vliv na cenu, proto tento údaj lze zcela opustit.

Proměnné s chybějícími záznamy jsou nejnáročnější částí - je potřeba tyto hodnoty buď dopočítat (vzít průměr nebo jinou náhradu za NaN hodnotu), nebo celou proměnnou vyhodit. Například, když je 90 procent sloupce tvořeno NaN hodnotami, je obecně lepší tuto proměnnou vynechat než s ní pracovat dál. Pokud je celý sloupec tvořen NaN hodnotami, je to signál, že tento sloupec byl špatně vytvořen - viz. sloupec "Výtah"v notebooku. Spolu s některými dalšími sloupci naznačuje, že jsme museli extrahovat data z HTML stránek sofistikovaněji (některá data byla reprezentována obrázkem, nikoliv textem). Občas cbybějící hodnoty lze jednoduše dopočítat: v našem případě NaN hodnota u proměnné "Balkon"znamená, že balkón není, proto ji zde lze nahradit nulou. Nicméně ne vždy je to možné. Pro predikci je důležité mít všechny hodnoty k dispozici, proto analýza chybějících hodnot je nutná.

Narovnání formatu dat je rovněž velmi důležité. U predikce rozlišujeme dva typy proměnných - kategorické a numerické. Kategorické proměnné jsou např. "Výtah", která má pouze dvě hodnoty: ano nebo ne. Takovým proměnným říkáme binární. Další takovou proměnnou je "Energetická náročnost budovy"- třída A, B, C atd. Kategorické proměnné nejsou navzájem měřitelné, proto je potřeba je převést do tzv. one-hot encodingu pomocí metody *get_dummies* z balíčku Pandas. Tím vytvoříme několik nových sloupců, které budou signalizovat to, jaká hodnota této proměnné je pro tento řádek správná. Například, když proměnná má 5 různých hodnot, vznikne 5 binárních sloupců pro každou z nich. Je potřeba ale u tohoto převodu dávat pozor na počet vzniklých unikátních hodnot: například u proměnné "Energetická náročnost budovy"na začátku máme mnohem více než 7 hodnot, přitom logicky jich můžeme mít pouze 7, proto mnohé proměnné musí být dodatečně předzpracovány a narovnány.

Jednodušším typem proměnných je numerický typ. Ten je dobře měřitelný, dvě hodnoty dané proměnné lze porovnávat. Proto je tento typ jednodušší na zpracování. Existují však situace, kdy v datech máme numerické proměnné v nesprávném formátu, např. proměnná "Podlaží", která je v našem datasetu reprezentována jako string a má doprovodný text, např. "4. podlaží z celkových 10". Tuto proměnnou v notebooku převádíme na numerický typ jednoduchou manipulací se stringem. Dále vytvaříme novou numerickou proměnnou "Počet podlaží", kterou naplníme hodnotami ze stejné vstupní proměnné. Tímto nám vzniknou dvě nové numerické proměnné z původní kategorické. V notebooku převádíme spoustu kategorických proměnných na numerický typ.

Ve chvíli, kdy máme vybrány proměnné, které použijeme pro predikci, musíme připravit dva datasety - trénovací a testovací. Data z trénovacího datasetu budou použitá k nalezení vztahu mezi závislou a nezávislými proměnnými. Data z testovacího datasetu budou sloužit k ověření, že náš model je skutečně schopen predikce. Navíc pomocí stejného testovacího datasetu jsme schopni porovnávat různé prediktivní modely.

Pro predikci jsem zvolila několik různých modelů, které jsou implementovány v balíčku sklearn:

1. LinearRegression - základní algoritmus lineární regrese založený na metodě nejmenších čtverců.
2. KNeighborsRegressor - regresní model založený na jednoduchém klasifikačním modelu k nejbližším sousedům.
3. DecisionTreeRegressor - regresní model založený na robustním klasifikačním modelu rozhodovacího stromu.

4. Ridge - lineární regrese doplněná o regularizační prvek, který má za úkol zmenšovat velikost parametrů regresního modelu.
5. Lasso - podobná technika jako Ridge, velikost parametrů se ale nezmenšuje, ale probíhá tzv. feature selection - model nachází parametry, které nemají vliv na predikci a nastavuje příslušné váhy v modelu na nulu. Je to taktéž regularizační metoda jako Ridge.
6. MLPRegresson - jednoduchá neuronová síť založená na algoritmu perceptron.

V balíčku sklearn je implementován jednoduchý postup trénování a porovnání modelu. Každý zvolený model má stejné rozhraní na trénování. Ke každému modelu jsme spočítali dvě základní metriky (viz [Aks]): MSE (mean square error) a RMSE (root mean square error): první metrika ukazuje, jak se predikce liší z pohledu modelu: je to průměrný rozdíl kvadrátů predikované a skutečné hodnoty, zatímco RMSE ukazuje průměrný rozdíl ve stejných jednotkách, které predikujeme.

Další výhodou je to, že jsme schopni nahlédnout na parametry natrénovaného modelu. Například u LinearRegression modelu vidíme, jak špatně jsme natrénovali model - některé parametry jsou o mnoho řádů větší než ostatní a model se nechová stabilně. Důvodem je nesplnění přísných předpokladů klasické lineární regrese: standardizace vstupních proměnných vůči sobě (všechny by měly mít průměr 0 a rozptyl 1). U Lasso ale vidíme, jak regularizace pomáhá zjednodušit model a vynechat proměnné, které nemají vliv na predikovanou hodnotu. Proto vidíme, že model se chová stabilněji a vykazuje uspokojivou přesnost.

Kapitola 6

Ukončení kurzu

6.1 Otestování práce na vybrané skupině studentů

Nejlepším způsobem, jak ohodnotit kvalitu jakéhokoli studijního programu nebo kurzu, je zpětná vazba studentů, kteří kurzem už prošli. Právě proto je nezbytnou součástí mé bakalářské práce její otestování na vybrané skupině studentů.

Pro tento účel jsem oslovila několik studentů (nakonec mi bylo ochotno pomoci 5 studentů), spadajících do cílové skupiny tohoto výukového kurzu, aby ho vyzkoušeli na sobě a nechali mi zpětnou vazbu v daném dotazníku. Tento dotazník [1] obsahuje 2 sekce:

1. Vstupní informace o zájemci o kurz. Aby hodnocení bylo víc objektivní, je třeba se na začátku dozvědět znalosti studentů předtím, než začali s osvojením daného kurzu.
2. Hodnocení kurzu. Zde pomocí otázek různého typu (jak otevřených, tak i uzavřených) studenti, co prošli kurzem, mohli anonymně sdělit svou zpětnou vazbu.

[1] Dotazník lze najít na adrese https://forms.gle/Pip7AzvRBxuGYvaJ9

6.1.1 Výsledky

Všechny výsledky jsou dostupné ve dvou přílohách, doporučuji je si přečíst před pokračováním do dalších sekcí:

- *Hodnocení dle studentů.xlsx* - zde jsou uvedena individuální hodnocení každého studenta ve formátu tabulky.
- *Hodnocení dle otázek.docx* - zde je uveden souhrn odpovědí dle otázek.

První sekce otázek

Jak jsem už zmínila výše, první část otázek se týkala zjišťování vstupních informací o studentech. Níže uvádím souhrn odpovědí včetně mých komentářů:

1. S tím, jak jsem se snažila oslovit právě studenty, spadající do cílové skupiny daného kurzu, tak je vidět, že vstupní požadavky všichni plní více než na polovinu. Průměrně to je dokonce více než 8 z 10 bodů.
2. Zkušenosti s datovou analýzou většina studentů měla minimální, tj. na úrovni menší orientace v problematice. Každý ale alespoň trochu věděl, co to je za oblast IT a obecně čím se zabývá.

   Máme ale i jednoho více zkušeného člověka, který se v této sféře už nějakou dobu pohyboval. Sice to není úplně začátečník, ale může se to hodit pro další názor ohledně této práce.

   Studenti ohodnotili svoji předchozí zkušenost s datovou analýzou průměrně na 4/10 body.
3. A protože kurz je spojen se studijním programem SIT, tak mě zajímalo, kolik z oslovených studentů bylo nebo jsou studenty tohoto programu, abych pak mohla ověřit, jestli se mi správně podařilo nastavit úroveň složitosti kurzu a jestli by ho šlo do SITu zařadit. Nakonec máme 3 studenty SITu z celkově 5 oslovených studentů.

S ohledem na výše uvedené informace, považuji úkol zvolení správné oslovené skupiny studenty za úspěšně splněný.

Druhá sekce otázek

Výsledky druhé části otázek, na které studenti odpovídali již po ukončení kurzu, jsou následující:

1. Všichni studenti kurz dokončili, tzn. že jejich hodnocení je kompletní.
2. Srozumitelnost kurzu všichni hodnotí kladně - průměrně to je 7/10. Je zde ale důležité uvést, že 10 znamená to, že studenti během nastudování nevyužívali žádné další zdroje informace, aby porozuměli celé probírané látce. Toto ale nebylo mým cílem, protože by se jinak kurz považoval za příliš jednoduchý.

   Všichni studenti můj názor potvrdili a do poznámky uvedli, že pro dohledávání dalších informací (například pro dokumentaci jednotlivých knihoven) použili internet.
3. Přínosnost kurzu, k mému velkému potěšení, byla ohodnocena velice kladně. 4/5 studentů dali maximální počet bodů s komentářem, že si z kurzu odnesli hodně. S tím, jak byl poslední student více zkušený, tak je samozřejmostí, že minimálně část probírané látky pro něj byla opakováním, a proto přínosnost pro sebe ohodnotil 5/10. Ocenil ale užitečnost příkladů pro začátečníky.

   Průměrně je to tím pádem 9/10.
4. Odpovědi na otázku "Co se Vám na kurzu líbilo?"byly dost rozsáhlé. Jako plusy celé práce byly dohromady vyzdvihnuty následující body:

   - Kurz je jedním velkým projektem, kde následující fáze vždycky navazují na předchozí.
   - Omezenost kurzu pouze na Python.
   - Návaznost na znalosti ze studijního programu SIT.
   - Postupnost v ponoření se do procesu datové analýzy: od obecného popisu problematiky na začátku do nejmenších komentářů na konci.
   - Dostatek doplňujících materiálů.
   - Složitost kurzu.
   - Jasné příklady ze života.
   - Určené vstupní požadavky na začátku kurzu.
   - Dostatek komentářů.
   - Čitelnost notebooků.

5. Připomínek ohledně toho, co se studentům na kurzu nelíbilo, bylo rozhodně méně:

   - Jedné studentce úroveň složitost kurzu nepřišla nastavena ideálně. Podle jejího názoru byl o něco složitější než čekala.
     Na to bych ráda reagovala tím, že souhlasím s tím, že se zcela bázové příklady v praktické části neprobíraly. Nebylo to ale mým cílem, protože jsem chtěla ukázat víc reálnou datovou analýzu s více zákeřnými body. Ale stejně jsem ukázala a použila všechny bázové funkce, které by začátečníci museli umět - jen ve větším rozsahu a na skutečném příkladu. Navíc k tomu v minulé otázce jsem naopak dostala kladné hodnocení ohledně složitosti. Proto si myslím, že porovnání studentky s jazykovou znalostí A2 je docela přesné. I když A2 je stále začátečník, takže toto plní účel kurzu.
   - Dále jednomu studentovi v kurzu chybělo zpracování obrázků. S tím také souhlasím, toto je velice zajímavý úkol, ale kvůli tomu, že je dost rozsáhlý, tak jsem to z kapacitních důvodů vynechala.
   - Stejný komentář bych nechala i ohledně metody shlukování.
   - Dalším bodem byla přítomnost zkoušky na konci kurzu. Ale s ohledem na to, že byla určena pouze pro to, aby studenti pochopili své silné a slabé stránky na konci, tak tento bod nepovažuji za závažný.
   - Poslední připomínka se týkala nedostatku komentářů v noteboocích. Se studentkou jsem se na konci spojila, abych upřesnila konkrétní místa, kde toho bylo málo. Nakonec jsem do 3. a 4. notebooků přidala pár dalších upřesňujících komentářů.

6. Doporučení pro zlepšení kurzu byly následující:

   - Přidat ke kurzu vyučujícího jako ve škole.
     Nemůžu s tím nesouhlasit, protože také považuji učitele za neoddělitelnou část výuky. Ale bohužel to nemohlo být vyřešeno v rámci dané bakalářské práce.
   - Udělat kurz na PowerBI.
     Je možné, že tento nápad využije někdo další.
   - Ukázat zpracování obrázků a textového popisu inzerátu, včetně následující změny chování predikce.
     K tomu jsem se už vyjadřovala v předchozí otázce - bohužel jsem to z kapacitních důvodů vynechala, ale je to opravdu dobrý nápad.

Na závěr bych ráda řekla, že s ohledem na všechny komentáře od studentů, kterým velice děkuji za ochotu mi pomoci, si myslím, že se mi úspěšně podařilo naplnit cíle daného kurzu. Je samozřejmostí, že tam zůstaly body, které se dají zlepšit, ale dosáhnout ideálu v datové analýze je skoro nemožné, nebo by to byl velice časově náročný proces. Avšak tyto body jsou podrobněji rozebrány v sekci 6.3.

6.2 Zkouška

Pro ověření znalostí studentů na konci kurzu jsem vytvořila zkoušku, jejíž součástí jsou otázky ze všech fází daného kurzu. Zkouška je zcela volitelná a určená pouze pro to, aby studenti věděli, co si z probírané látky pamatuji a co by bylo lepší přečíst a/nebo probrat ještě jednou. Proto nebyla nutnou součástí ani pro těch 5 studentů, kteří byli ochotni vyzkoušet kurz na vlastní kůži, i když k ní měli přístup a většina z nich to nakonec ústně prošla. Zkouška se skládá z 16 otázek s otevřenými odpověďmi. V příloze pro tento účel jsou uvedené 2 dokumenty:

- první obsahuje pouze otázky - *Zkouška - otázky.docx*;
- druhý obsahuje i odpovědi - *Zkouška - odpovědi.docx*.

6.3 Jak bych pokračovala dále?

Během kurzu pro začátečníky jsem se snažila ukázat všechny nejdůležitější kroky v datové analýze. Je jasné, že jsme neprobrali úplně všechno, co bylo možné, a že kurz není 100 % kompletní, a proto pro nejzvědavější studenty a zájemce o kurz mohu doporučit ještě několik dalších kroků v rámci samostudia pro zlepšení výsledků predikce ceny:

1. Zpracovat obrázky (symboly) v tabulce v popisu k inzerátu. Tím jsou myšleny fajfky a křížky u takových parametrů, jako je například parkování nebo výtah. Proto by bylo nutné na základě pevné logiky změnit symboly na "Ano/Ne"a zařadit to do prediktivního modelu, aby v odpovídajících sloupcích už nebyly neznámé hodnoty.
2. Zpracovat textový popis všech inzerátů. Jak jsme už zjistili v praktické části, popis některých parametrů v tabulce chybí, a místo toho je uváděn normálně v textovém popisu. Tím pádem bychom také mohli zredukovat množství chybějících parametrů a ještě víc zlepšit predikci cen.
3. Zajímavým úkolem by bylo také vytvořit a následně ověřit testovací hypotézy. Například "Cena stejného bytu v Praze je o 5 tisíc dražší než ve zbylé části České republiky"nebo "Absence balkonů nesnižuje průměrnou cenu bytu", aby studenti ještě jednou uviděli aplikaci statistiky v analýze dat.

Kapitola 7

Závěr

7.1 Obecně

Hlavním cílem mé bakalářské práce bylo zcela od nuly seznámit se s problematikou datové analýzy, udělat rešerši a prozkoumat různé postupy a nástroje, které se v ní používají. Následně jsem měla navrhnout praktický kurz pro začátečníky, jehož součástí jsou materiály pro seznámení s problematikou (literatura), praktické úkoly (datové množiny, konkrétní úkoly) a závěrečný test, ověřující znalosti. Na konci bylo mým úkolem uživatelsky ověřit navržený kurz na vybrané skupině studentů a navrhnout, jakým způsobem lze zařadit vytvořený kurz do stávající výuky programu SIT.

Z toho vyplývá, že první část mé práce je čistě teoretická a vysvětluje pojem datové analýzy, motivaci se jí zabývat a popisuje jednotlivé kroky při analýze dat. 4 z 5 kroků jsou následně velice podrobně rozebrány v druhé, praktické části pomocí speciálně vytvořených notebooků (viz. Gitlab - jen pro přidané uživatele, nebo soubory v zip archivu). Při jejich vytváření jsem záměrně nevzala již předpřipravené datasety nebo úlohy například z Kagglu, protože jsem chtěla zkusit a pak i ukázat budoucím studentům reálné problémy, na které se dá narazit při řešení skutečného úkolu. Proto jsem si zvolila zadání, které reflektuje hodně aspektů a rizik práce s daty, jež jsou podrobně probrána v noteboocích.

Kurzem nakonec vyšla nejen praktická část mé práce, ale skoro celá bakalářská práce. Sice se přímo praktické části týkají pouze kapitoly 4, 5 a samotné notebooky, ale kapitoly 2, 3 a 6 jsou také velice užitečné pro zájemce o kurz nebo případné studenty.

Na konci, abych měla možnost vyhodnotit kvalitu svého studijního kurzu, jsem uvedla výsledky otestování své práce na potenciálních zájemcích. Ověřilo se, že nakonec studenti s mým kurzem byli dost spokojení a ohodnotili ho velice kladně, i když samozřejmě měli několik připomínek pro možné zlepšení. Díky tomu byla přidána další sekce mě bakalářské práce, a to o tom, jak se kurz dá doplnit, aby se zlepšily výsledky prediktivního modelu.

Velkou vyhodou je i to, že se mi během vypracování projektu podařilo využít i dovednosti studentů studijního programu Softwarové inženýrství a technologie z minulých semestrů a předmětů, například:

1. nástroje pro testování SW *Selenium* z předmětu B6B36TS1;
2. jazyk HTML a práce s regulárními výrazy z B6B39ZWA;
3. statistické metody jako histogram, boxplot, matice korelace, různá rozdělení z B6B01PST;
4. obecná ponětí, techniky datového návrhu a práce s relačními daty z B0B36DBS.

Proto podle mého názoru se mi podařilo naplnit všechny pokyny pro vypracování dané bakalářské práce a mohu ji považovat za kompletní.

7.2 Vyhodnocení práce

Kvůli tomu, že jsem se předtím s datovou analýzou nikdy v životě nepotkala, tak bych pro sebe tuto bakalářskou práci ohodnotila jako velice časově náročnou. Během jejího vypracování jsem si přečetla spoustu článků o problematice datové analýzy jako takové, ve většině případů, samozřejmě, v angličtině, ale pro větší přehled jsem také zkusila češtinu a ruštinu. Kdyby byla nutnost uvést je všechny, tak bych s tím měla problém, protože toho bylo opravdu hodně. Místo toho bych chtěla zmínit jen 2 články, které považuji pro začátečníky za velmi užitečné a doporučené: "Analytics is not storytelling...", viz [Cas19], a "What makes a data analyst excellent?", viz [Cas20]. Autorem obou je Cassie Kozyrkov, Chief Decision Scientist at Google.

Co se týče praktické části mého projektu, tak jsem se, jak bylo vidět, zaměřila výhradně na práci s Pythonem. Implementační část také nikdy nebyla mojí silnou stránkou, proto pro to pro mne byla velká výzva. Předtím, než jsem začala pracovat na svém vlastním projektu, jsem:

1. prošla několik ukázkových příkladů z Kagglu [1], například "House Prices" (viz [Ped17]), který ačkoli vypadá podobně tomu mému, ale v podstatě neřeší ani polovinu těch věcí, které jsem probrala ve svých noteboocích;
2. se naučila základům práce v Jupyteru a s Anacondou;
3. přečetla ostatní pomocné materiály. Tady chci zdůraznit jen ty, ze kterých jsem načerpala mnoho znalostí a pak je ihned využila při vypracování notebooků:
   - Python and Data Analysis, viz [Nat20].
   - Jupyter tips and tricks, viz [Chr18].
   - Jupyter formatting cheatsheet, viz [Ing17].
   - Pandas tutorial - základy, viz [Pan22].
   - Pandas tutorial - práce s datovými tabulkami, viz [Kar22].
   - Python map/list/filter/reduce funkce, viz [Har21].
   - Python lambda functions, viz [Gun21].
   - Práce se souborovým systémem v Pythonu, viz [Pyt21].
   - Dokumentace BeautifulSoup - extrakce dat z HTML/XML souborů, viz [Leo20].
   - Anaconda - správa Python prostředí, viz [Pan16].

[1] Kaggle lze najít na adrese https://www.kaggle.com/.

Čas strávený nad vypracováním praktické části své bakalářské práce zhruba odpovídá následujícím hodnotám (ale ještě jednou chci připomenout, že nejsem vývojářský typ člověka, a proto si myslím, že by tato hodnota nejspíš byla menší v případě, že by to dělal jiný začátečník):

1. Práce s vlastními notebooky - 200 hodin:
   - první - 35 hodin;
   - druhý - 50 hodin;
   - třetí - 65 hodin;
   - čtvrtý - 50 hodin.
2. Napsání smysluplných komentářů pro začatečníky v noteboocích - 20 hodin.

Celkově to je zhruba 220 hodin. Čas, strávený na to, abych se sama s tímto tématem seznámila a abych vymyslela koncepci kurzu, nebudu schopna uvést. Takže bych na závěr řekla, že tato práce pro mě ani zdaleka nebyla jednoduchá, ale velice zajímavá a přínosná. Samotná příprava příkladů mi nepřišla jako extrémně složitý úkol z toho hlediska, že se strukturou notebooků dalo inspirovat na Kagglu, ale vzhledem k tomu, že s tím, jak jsem příklady vytvářela, jsem se zároveň i učila, jak to všechno funguje, tak bych práci pro sebe opravdu ohodnotila jako náročnou.

Příloha A

Literatura

[Aks] Akshita Chugh, *MAE, MSE, RMSE, Coefficient of Determination, Adjusted R Squared — Which Metric is Better?*, Medium.

[All] Data Science Process Alliance, *Data Science Methodologies and Frameworks Guide.*

[anaom] *Anaconda*, dostupné z https://www.anaconda.com.

[anaor] *Anaconda.Documentation*, dostupné z https://docs.anaconda.com/navigator/.

[Aya19] Ayan Brahmachary, ITIL Foundation , *DIKW Model: Explaining the DIKW Pyramid or DIKW Hierarchy.*

[Ber03] Petr Berka, *Dobývání znalostí z databází*, Academia, 2003, p. 28.

[Bro99] Meta S. Brown, *The CRISP-DM User Guide*, Wayback Machine, 1999, p. 14.

[Bur15] Daniel Burrus, *Why Your Company Needs Data Analytics.*

[Cas19] Cassie Kozyrkov, *Analytics is not storytelling... On the nature of analytics, part 1 of 2.*

[Cas20] ———, *What makes a data analyst excellent? On the nature of analytics, part 2 of 2.*

[Cha00] Pete Chapman, *CRISP-DM 1.0: Step-by-step data mining guide*, The Modeling Agency - online, 1999, 2000, p. 18.

[Chr18] Christoph Rieke, *Jupyter Tips and Tricks*, Medium, 2018.

[DR18] John Rydning David Reinsel, John Gantz, *The digitization of the world from edge to core.*

[DTL05] Chantal D. Larose Daniel T. Larose, *Discovering Knowledge in Data: An Introduction to Data Mining*, John Wiley Sons, 2005, p. 7.

[Far15] Mario Faria, *The rise of people management analytics.*

[Fre05] David Freedman, *Statistical models: Theory and practice*, 01 2005.

[Gun21] Gunjan Goyal, *Anonymous or Lambda Functions in Python: A Beginner's Guide!*, Analytics Vidhya, 2021.

[Har21] Harsh Dhamecha, *An Explanation to Python's Lambda, Map, Filter and Reduce*, Analytics Vidhya, 2021.

[Hei20] Florian Heinrichs, *Using crisp-dm to grow as data scientist.*

[i-S] i-SCOOP, *The DIKW model for knowledge management and data value extraction.*

[Ing17] Inge Halilovic, *Markdown for Jupyter notebooks cheatsheet*, Medium, 2017.

[Jac16] Jacqueline Kazil, Katharine Jarmul, *Data Wrangling with Python*, O'Reilly Media, 2016, pp. 4–7.

[jupWk] *Jupyter Notebook Tutorial: Introduction, Setup, and Walkthrough*, dostupné z https://www.youtube.com/watch?v=HW29067qVWk.

[jupal] *How to use jupyter notebook: A beginner's tutorial*, dostupné z https://www.dataquest.io/blog/jupyter-notebook-tutorial/.

[juprg] *Jupyter*, dostupné z https://jupyter.org/.

[Kar22] Karlijn Willems, *Pandas Tutorial: DataFrames in Python*, Datacamp, 2022.

[Kud14] Stephan Kudyba, *Big Data, Mining, and Analytics. Components of Strategic Decision Making.*

[Lea21] Learn Microsoft, *Dashboards for business users of the Power BI service.*

[Lea22] ———, *Visualization types in Power BI.*

[Leo20] Leonard Richardson, *Beautiful Soup Documentation*, Crummy, 2020.

[McK17] William McKinney, *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*, O'Reilly Media, 2017, pp. 2–8.

[Nat20] Natassha Selvaraj, *A Beginner's Guide to Data Analysis in Python*, Towards Data Science, 2020.

[Nic22] Nick Hotz, Data Science Process Alliance, *What is CRISP DM?*

[Pan16] Pankaj Mathur, *What is Anaconda and Why should I bother about it?*, 2016.

[Pan22] Pandas. User Guide, *10 minutes to pandas*, 2022.

[Ped17] Pedro Marcelino, *Comprehensive data exploration with Python. House Prices - Advanced Regression Techniques*, Kaggle, 2017.

[Pyt21] Python, *pathlib — Object-oriented filesystem paths*, 2021.

[Sup22] Support Microsoft, *ZÍSKATKONTDATA (funkce)*.