EKONOMICKÁ UNIVERZITA V BRATISLAVE

FAKULTA HOSPODÁRSKEJ INFORMATIKY

Evidenčné číslo: 1003006/B/2020/36089192888434692

BIG DATA A ICH SPRACOVANIE

Bakalárska práca

2020 **Richard Šteiner**

EKONOMICKÁ UNIVERZITA V BRATISLAVE

FAKULTA HOSPODÁRSKEJ INFORMATIKY

BIG DATA A ICH SPRACOVANIE

Bakalárska práca

Študijný program:	Manažérske rozhodovanie
Študijný odbor:	Kvantitatívne metódy v ekonómií
Školiace pracovisko:	Katedra matematiky a aktuárstva
Vedúci záverečnej práce:	Mgr. Andrea Kaderová, PhD.

Bratislava 2020 **Richard Šteiner**

Čestné vyhlásenie

Čestne vyhlasujem, že záverečnú prácu som vypracoval samostatne a že som uviedol všetku použitú literatúru.

Dátum: **.....**

podpis študenta

ABSTRAKT

ŠTEINER, Richard: *Big Data a ich spracovanie*. – Ekonomická univerzita v Bratislave. Fakulta hospodárskej informatiky; Katedra matematiky a aktuárstva. – Vedúci záverečnej práce: Mgr. Andrea Kaderová. - Bratislava: FHI EU, 2020, 43s.

Cieľom bakalárskej práce je objasniť v dnešnej dobe tak dôležitú prácu s dátami, spracovanie týchto dát, ich následnú analýzu. Výška celkových maloobchodných tržieb odráža úroveň kúpyschopnosti obyvateľstva a ich rast, naznačuje vyššiu spotrebu obyvateľstva. Práve preto by mal byť tento makroekonomický ukazovateľ starostlivo sledovaný ako jeden z hlavných indikátorov ďalšieho vývoja ekonomiky. Preto považujeme za dôležité analyzovať a prognózovať vývoj maloobchodných tržieb, aby tvorcovia hospodárskej politiky mali k dispozícii aj kvantitatívny nástroj pri rozhodovaní. Práca je rozdelená do štyroch kapitol. Obsahuje 6 grafov, 3 tabuľky a 19 obrázkov. V prvej kapitole stručne popíšeme všeobecné poznatky o Big Data, ich charakteristiku, vlastnosti, postup spracovania a spôsoby akými sa v súčasnosti Big Data využívajú. V ďalšej časti sa nachádza hlavný cieľ a čiastkové ciele pre dosiahnutie hlavného cieľa. Tretia kapitola je zameraná na metodológiu a metódy použité pre splnenie cieľov. Máme tu opísane základné charakteristiky a modely, ktoré sme využili pri analýze a následnom prognózovaní. V poslednej kapitole aplikujeme metódy z predošlej kapitoly a analyzujeme údaje, kde našou hlavnou úlohou je prognóza priemerných denných maloobchodných tržieb v USA na nasledujúci rok. Prognózu porovnáme so skutočným stavom za prvé 4 mesiace roku 2020.

Kľúčové slová: Big Data, spracovanie dát, jazyk R, prognóza

ABSTRACT

ŠTEINER, Richard: Big Data and its processing – University of Economics in Bratislava. The Faculty of Economic Informatics; Department of Mathematics and actuary. - Supervisor: Mgr. Andrea Kaderová PhD. - Bratislava: FHI EU, 2020, 43s.

The aim of the bachelor thesis is to explain the importance of working with data, processing of this data and subsequnet analysis. The level of total US retail sales reflects the level of the purchasing power of population and its growth idicates higher comsumption of the population. That is why this macroeconomic indicator should be carefully monitored as one of the main indicators of the further development of economy. Therefore, we consider it importatnt to analyze and forecast the development of retail sales, so that the architects of economic policies could use the results of that analysis as quantitative instruments for making decisions. The thesis is divided into 4 chapters. It contains 6 graphs, 3 tables and 19 images. In the first chapter we briefly describe general knowledge about Big Data, their characteristics, properties, steps to process it and ways in which Big Data is currently used. In the next section is the main goal and partial goals for achieving the main goal. The third chapter focuses on the methodology and methods used to meet the target. We have described the basic characteristics and models that we used in the analysis and subsequent forecasting. In the last chapter, we apply the methods from the previous chapter and analyze the data, where our main task is to forecast the average daily retail sales in teh US for the following year 2020. We will compare the forecast with the actual situation for first 4 months of 2020.

Key words: Big Data, data processing, R language, forecast

Obsah

ÚVOD

Témou bakalárskej práce sú „Big Data a ich spracovanie.“ Podľa výpočtov spoločnosti IBM bolo v roku 2012 denne vygenerovaných 2,5 exabajtov nových dát, pričom predpoklady hovoria, že tento objem každým rokom narastá. Dáta sú generované z rôznych zdrojov, či už sa jedná o servery, osobné počítače, mobilné telefóny, rôzne digitálne senzory, alebo zariadenia označované ako internet vecí.

Bakalárska práca je rozdelená do štyroch kapitol. V úvodnej kapitole bakalárskej práce sme sa venovali všeobecným poznatkom o danej problematike. Rozoberáme súčasný stav, definujeme a charakterizujeme Big Data a tiež sa venujeme krokom, ktoré sú potrebné pre spracovanie týchto dát. Na záver kapitoly uvedieme odvetvia, v ktorých sa analýzy veľkých dát naplno využívajú. Taktiež uvedieme niekoľko praktických príkladov pre lepšiu predstavu, ako Big Data prospievajú modernej spoločnosti.

V druhej kapitole sme určili ciel práce a čiastkové ciele pre realizáciu hlavného cieľa, čo je analýza vývoja maloobchodných tržieb v Spojených štátoch amerických a následná predikcia na budúci rok. Výslednú predikciu sme porovnali so skutočným stavom za prvé 4 mesiace roku 2020. Na predikciu sme využili programovací jazyk R. Tretia kapitola sa zameriava na postupy akými dáta spracovávame a analyzujeme. Popíšeme modely, ktoré sme testovali na prognózovanie a postupy výberu toho najlepšieho modelu.

Záverečná kapitola bola zacielená na analýzu a predikciu priemerných denných maloobchodných tržieb v Spojených štátoch americký v programovacom jazyku R. Najprv sme dáta zozbierali, spracovali a upravili do nami želanej podoby. Analyzovali sme ich a následne sme aplikovali vybrané modely popísané v tretej kapitole. Na základe výsledkov predbežných analýz sme vybrali model, ktorý najlepšie opisuje nami vybrané dáta a použili sme ho na predikciu priemerných denných maloobchodných tržieb na nasledujúci rok 2020. Predikované dáta sme porovnali so skutočným stavom prvých 4 mesiacov roku 2020.

1 Súčasný stav riešenej problematiky a charakteristika Big Data

V súčasnom digitálom svete sa význam a objem dát neustále rozširuje. Výskum z roku 2014 od IDC (International Data Corporation) predpokladá, že celkový objem dát, ktorý v roku 2013 tvoril 4,4 zettabajtov, sa každé dva roky zdvojnásobí do roku 2020 na 44 zettabajtov, čo predstavuje 44 biliónov gigabajtov dát [8].

Nárast dát je badateľný v mnohých oblastiach digitálneho sveta, či sa jedná o informačné zdroje, sociálne siete a médiá alebo internet vecí. Súčasný trend predpokladá dramatický nárast dát pripadajúci na jednotlivca v spoločnosti, ale o mnoho väčší objem dát sa očakáva zo strany strojov v podobe logov, GPS záznamov z mobilov alebo dopravných prostriedkov, zaznamenávanie a ukladanie obchodných transakcii a podobne.

Štátne inštitúcie taktiež generujú veľké množstvo dát, ktoré sú väčšinou verejnosti nedostupné. Tieto dáta zatiaľ ležia nevyužité čakajúc na svoj čas. Spracovávanie týchto dát môže prinášať prospech a pomáhať nielen firmám v súkromnom sektore ale aj obyčajným občanom a samozrejme predovšetkým štátnym inštitúciám v rozhodovaní o veľkých projektoch a investíciách [21].

Podľa prieskumu Centra pre digitálne vedenie štátu z roku 2018 sa 43 z 50 štátov USA vyjadrilo, že hľadajú pracovnú silu v oblasti business intelligence a analýzy dát [22].

Prognózovanie je bežnou štatistickou úlohou v podnikaní, kde pomáha v rozhodovaní ohľadom plánovania výroby, prepravy alebo personálu a poskytuje návod na dlhodobé strategické plánovanie. Podnikateľské predikcie sa však často robia zle a zamieňajú sa s plánovaním a cieľmi. Prognózovanie by malo byť neoddeliteľnou súčasťou rozhodovacích činností manažmentu, pretože môže hrať dôležitú úlohu v mnohých oblastiach organizácie. Moderné spoločnosti vyžadujú krátkodobé, strednodobé a dlhodobé predpovede v závislosti od konkrétneho použitia [12].

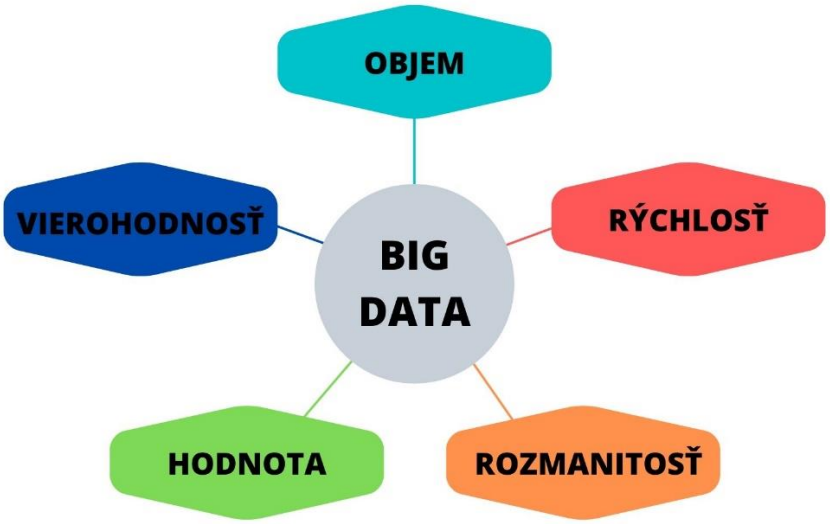
1.1 Charakteristika Big Data

V roku 2005 Roger Magoulas definoval pojem Big Data ako veľké množstvo dát, ktoré nie je možné spracovať pomocou bežných nástrojov a techník, kvôli ich veľkosti a komplexnosti [1].

Neexistuje presná definícia, ale všetky majú niečo spoločné a to, že Big data je pojem pre súbor dát, ktorých rozsah je natoľko veľký, že je ťažké ich spracovať v rozumnom čase tradičnými databázovými nástrojmi alebo aplikáciami.

Big Data sú charakterizované najčastejšie tromi hlavnými znakmi „3V“ volume (objem), velocity (rýchlosť) a variety (rozmanitosť) a ďalšími 2 znaky value (hodnota) a veracity (vierohodnosť).

Obrázok č. 1: Základné vlastnosti Big Data



Zdroj: Vlastné spracovanie

Okrem týchto základných znakov sa uvádzajú aj iné charakteristiky ako viscosity (zložitosť), variability (premenlivosť), volatility (trvanlivosť), viability (použiteľnosť) a validity (platnosť).

1.1.1 Objem

Už z názvu Big Data vyplýva, že sa jedná o veľké množstvo dát, ktoré sa denne generuje z rôznych zdrojov ako sú sociálne média, stroje, obchodné procesy, ľudská interakcia atď. Taktiež ak konkrétne údaje chceme považovať za Big Data alebo nie, závisí to od objemu údajov ktoré máme k dispozícii. Preto práve objem je jednou z charakteristík, ktorú je potrebné zohľadniť pri riešení veľkých dát. Teraz sú ľudia prepojení viac ako

kedykoľvek predtým a toto prepojenie vedie k väčšiemu počtu dátových zdrojov, čo následne vedie k tomu, že množstvo dát neustále rastie [2].

V súčasnej dobe sa množstvo dát rýchlo rozširuje a v blízkej budúcnosti bude aj naďalej zaznamenaný ich exponenciálny nárast [13].

1.1.2 Rýchlosť

Táto charakteristika odkazuje na frekvenciu s akou sú dáta tvorené , uchovávané a zdieľané v reálnom čase. Napríklad senzory v moderných autách, ktoré monitorujú a odosielajú rôzne aspekty výkonu vozidla, potrebujú byť spracované v reálnom čase. Dáta sú generované rýchlo a tým pádom musia byť spracované rovnakou rýchlosťou [6].

Rast počtu digitálnych zariadení, ako sú smartfóny, vedie k rýchlemu vytváraniu dát a zväčšuje potrebu ich spracovania v reálnom čase [7].

1.1.3 Rozmanitosť

Spracovávané dáta pochádzajú z rôznych zdrojov, obsahujú rôzne typy dát a sú ukladané v rôznych formátoch ako sú napríklad dokumenty, databázy alebo dáta o polohe GPS. Big Data obsahujú heterogénny formát dát. Existujú štruktúrované, neštruktúrované a semi-štrukturované formáty dát. [6]

Neštruktúrované dáta predstavujú viac ako 80% dát v spoločnostiach a ich zdrojom sú obvykle sociálne siete. Ako príklad neštruktúrovaných dát môžeme uviesť videá, filmy, fotografie, finančné transakcie, e-maily alebo meteorologické záznamy [11].

Pokiaľ rôznorodosť dátových zdrojov rastie, rastie aj zložitosť získavania hodnoty z týchto dát. Ľudia nemôžu zvládať takéto zaťaženie, a preto je možné v takomto prípade použiť techniku „deep learning“ (hlboké učenie). Siete pre hlboké učenie môžu zaistiť používanie rôznych dátových formátov pre získavanie hodnoty [2].

1.1.4 Hodnota

Ďalším charakteristickým znakom je hodnota, ktorá rozširuje pôvodné hlavné znaky „3V“. Nejde o vlastnosť, ktorú majú len Big Data, ale ide o všeobecnú vlastnosť toho, že dáta, ktoré spracovávame by mali mať pre nás hodnotu a mali by byť prínosom. Je dôležité aby analýza Big Data smerovala k výsledkom, ktoré organizácii prinesú hodnotu [6].

Pre lepšiu predstavu, ako sa z Big Data získava hodnota, spoločnosť McKinsey & Company spravila výskum v oblasti zdravotnej starostlivosti a maloobchodnom sektore v Spojených štátoch amerických, v správe verejného sektoru Európskej únie, v globálnom výrobnom sektore a globálnych osobných dátach o polohe. Pomocou tohto výskumu zdôraznili, že Big Data majú obrovský význam pre hospodárstvo, zlepšujú produktivitu a konkurencieschopnosť podnikov i verejného sektoru a dokonca prinášajú výhody aj pre spotrebiteľov.

1.1.5 Vierohodnosť

Ak teda chceme získať hodnotu z dát, ktoré zbierame, musia tieto dáta pochádzať z dôveryhodného zdroja. Získané dáta môžu byť neúplné, nejasné a nekonzistentné, preto pri analýze a porovnávaní treba zistiť, ktoré dáta sú vierohodné a ktoré nie.

IBM pridalo vierohodnosť do konceptu „3V“, čo ukazuje na nespoľahlivosť súvisiacu s niektorými zdrojmi dát. Napríklad z komunikácie na sociálnych sieťach, kde pocity zákazníkov obsahujú dôležité informácie, ale sú neisté kvôli ľudskému úsudku [7].

1.1.6 Ďalšie charakteristiky

Okrem základných charakteristík „3V“, hodnoty a vierohodnosti boli následne pridané nové charakteristiky:

- **Zložitosť**
  Komplexná správa obrovského a rôznorodého dátového súboru je zložitá kvôli korelácii a vzájomným závislostiam v štruktúre Big Data [10].

- **Premenlivosť**
  Často sa stáva, že rýchlosť Big Data nie je konzistentná kvôli skutočnosti, že tieto dáta sú generované prostredníctvom obrovského množstva zdrojov [10].

- **Trvanlivosť**
  Táto charakteristika ukazuje, že Big Data majú obmedzenú dobu platnosti a uchovávania. Pre plné využitie potenciálu Big Data by sa malo určiť, kedy tieto dáta už nie sú relevantné a nemôžu byť použité pre analýzu [10].

- **Použiteľnosť**

  Pre analýzu Big Data by mali byť vybrané určité atribúty, ktoré s najväčšou pravdepodobnosťou môžu predpovedať najvýznamnejšie faktory, ktoré ovplyvňujú podnikové výsledky. Pokiaľ sa pomocou štatistických testov a výpočtov zistí, že medzi týmito premennými je výrazná korelácia, procesy generovania a zhromažďovania týchto dát by mali byť zlepšené, čo ďalej umožní sledovanie objavovaných súvislostí [10]

- **Platnosť**

  Aj napriek tomu, že dátový súbor môže byť vierohodný, môže sa stať, že mu zle porozumieme a tým pádom môže byť neplatný. Platnosť zdrojov produkujúcich Big Data a analýzy by mali byť exaktné, pokiaľ tieto výsledky budú následne použité pri procese rozhodovania [10].

1.2 Spracovanie Big Data

Údaje sú zhromažďované v obrovskom rozsahu a poháňajú väčšinu aspektov modernej spoločnosti, vrátane mobilných služieb, maloobchodu, výroby, finančných služieb, biologických a fyzikálnych vied. Rozhodnutia, ktoré sa predtým zakladali na dohadoch alebo starostlivo skonštruovaných modeloch, sa teraz môžu robiť na základe samotných údajov. Spracovanie Big Data prebieha nasledovne

1.2.1 Zhromažďovanie dát

Zber údajov je prvým krokom v procese spracovania dát. V tejto fáze sa dáta nespracovávajú pretože pred spracovaním sa musí určiť, ktoré dáta sa budú zbierať a ako sa následne budú nahrávať. Údaje sa získavajú z rôznych dostupných zdrojov. Proces zhromažďovania musí zabezpečiť, aby zozbierané údaje boli presne definované a dôveryhodné. Zhromaždené údaje budú neskôr použité ako informácie a práve preto musia byť v čo najvyššej kvalite. Všetky tieto faktory budú mať výrazný vplyv na konečný výstup [18].

Väčšina údajov nie je predmetom záujmu a je možné ich filtrovať. Jednou z výziev je definovať tieto filtre takým spôsobom, aby sa zredukovalo celkové množstvo údajov a ostali iba relevantné údaje, s ktorými sa bude pracovať. Eliminácia nepotrebných dát pomôže k zrýchleniu nasledujúcich fáz. Ďalšou výzvou je automatické generovanie

správnych metadát, ktoré popisujú, čo za údaje sú nahrané, ako sú nahrané a aké parametre majú nahrané dáta [16].

1.2.2 Príprava a čistenie údajov

Často sa stáva, že zozbierané údaje nie sú vo formáte, ktorý je vhodný na analýzu. Príprava predstavuje manipuláciu s dátami do formy, ktorá je vhodná pre ďalšie spracovanie. V tejto fáze sa „surové“ údaje vyčistia a dôkladne kontrolujú, či neobsahujú chyby alebo nespracované dáta. Analyzovať dáta, ktoré neboli starostlivo skontrolované, môže viesť k veľmi zavádzajúcim výsledkom, ktoré sú silne závislé od kvality vstupných údajov. Zlá dátová kvalita či rôzna dátová komplexnosť, spôsobená neštruktúrovanými dátami a rozličnými formátmi, robí túto fázu časovo najnáročnejšiu [16].

1.2.3 Dátová integrácia

Vzhľadom na rôznorodosť údajov, nestačí ich iba zaznamenať a vložiť do databázy. Ak sa v úložisku nachádza iba hromada množín údajov, je nepravdepodobné, že by niekto iný mohol opätovne tieto dáta použiť. Pre efektívnu analýzu dát je potrebné aby sa overené údaje pretransformovali do strojom čitateľnej formy, ktorá je zrozumiteľná pre počítače a je automaticky riešiteľná [9].

Na spracovanie údajov je potrebné veľké množstvo výpočtového výkonu, takže sa väčšina údajov musí riadiť formálnou a prísnou syntaxou. Kvôli nákladom sa mnohé podniky uchyľujú k outsourcingu tohto procesu.

1.2.4 Príprava a čistenie údajov

Počas tejto fázy sa vložené údaje spracovávajú pomocou rôznych algoritmov strojového učenia a umelej inteligencie. Cieľom je vytvoriť výstup respektíve interpretáciu údajov.

Metódy pre spracovanie a dolovanie Big Data sa zásadne líšia od tradičnej štatistickej analýzy používanej na malej vzorke. Big Data sú často chaotické, dynamické, heterogénne, vzájomne prepojené a nedôveryhodné. Avšak aj napriek tomu sú tieto chaotické dáta hodnotnejšie ako malé vzorky, pretože všeobecné štatistiky získané z bežných vzorcov a korelačnej analýzy zvyčajne prevažujú nad individuálnymi odchýlkami a často odhaľujú spoľahlivejšie skryté vzorce a vedomosti [9].

1.2.5 Interpretácia

Najpodstatnejším krokom spracovania Big Data je interpretácia. Analyzovať Big Data je jedna vec, ale ak používateľ nevie správne pochopiť výsledky tak to neprinesie žiadnu hodnotu. Dátový analytik musí výsledky analýz spracovať do takej podoby aby boli zrozumiteľné aj pre ostatných riadiacich pracovníkov spoločnosti. Výsledky sú prezentované väčšinou vo forme grafov, videí, audia, obrázkov alebo obyčajného krátkeho textu [16].

1.3 Využitie Big Data

Big Data sa používajú najmä v IT sektore, v ktorom spoločnosti ako Google a Microsoft analyzujú rôzne dáta, aby spravili určité rozhodnutia, ktoré majú neskôr dopad na súčasné a budúce technológie. Avšak Big Data používajú aj iné spoločnosti ako napríklad Amazon alebo Netflix, ktoré na základe dát, odporúčajú svojim zákazníkom - používateľom určité položky alebo filmy, ktoré by sa im mohli páčiť.

1.3.1 Financie

Banky a ostatné spoločnosti poskytujúce finančné služby zhromažďujú obrovské dáta o svojich klientoch. Počas rokov premenili tieto dáta na konkurenčnú výhodu. Medzi najväčšie problémy tohto odvetvia patria - podvody s kreditnými kartami a cennými papiermi, vykazovanie podnikových úverových rizík alebo archivácia audítorských záznamov [17].

Vlády prijímajú zákony, ktoré priamo ovplyvňujú toto odvetvie. Americká komisia pre cenné papiere a burzy (SEC) využíva big data na monitorovanie finančných trhov. Používajú nástroje na sieťovú analýzu a spracovanie prirodzeného jazyka aby zachytili nezákonnú obchodnú aktivitu na týchto trhoch.

Banky získavajú informácie o svojich klientoch procesom nazývaným „KYC“ (Know your customer). Je to proces overovania totožnosti jednotlivých klientov. Spoliehajú sa na získané údaje aby dokázali napríklad zabrániť praniu špinavých peňazí, krádeži identity alebo iným finančným podvodom. Analýza údajov môže spoločnostiam pomôcť pri identifikácii možných podvodov. Jeden prediktívny model strojového učenia vytvorený spoločnosťou QuntumBlack už v prvom týždni používania zistil podvodné transakcie v hodnote 100 000 amerických dolárov [19].

Bank of America vytvorila virtuálneho asistenta s názvom Erica, ktorý využíva prediktívnu analýzu a nástroje na spracovanie prirodzeného jazyka, aby zákazníkom pomohla zobraziť históriu bankových transakcií alebo informácie o blížiacich sa platbách. Zástupcovia Bank of America tvrdia, že asistent nakoniec preštuduje zvyky ľudí a poskytne príslušné finančné poradenstvo.

1.3.2 Priemysel

Digitálna revolúcia zmenila aj výrobný priemysel. Výrobcovia hľadajú nové spôsoby, ako využiť dáta, ktoré generujú, napríklad na zlepšenie prevádzkovej efektívnosti, zefektívnenie obchodných procesov a odhalenie poznatkov, ktoré prinesú zisk a rast. Big data môžu pomôcť predpovedať zlyhanie strojového zariadenia. Potenciálne zlyhania je možné zistiť analýzou údajov, ktoré sú rôzneho formátu, ako sú vek zariadenia, model zariadenia, značka, údaje zo senzorov, chybové hlásenia, teplota motora a ďalšie. Vďaka týmto údajom môžu výrobcovia maximalizovať bezpečnosť a čas prevádzkyschopnosti zariadenia a tým pádom aj optimalizovať náklady na údržbu. Prevádzková efektivita je jednou z oblastí, v ktorých majú big data najväčší vplyv na ziskovosť. Vďaka údajom sa môžu výrobné procesy analyzovať, hodnotiť a podniky môžu aktívne reagovať na spätnú väzbu od zákazníkov a predvídať budúce požiadavky. Optimalizácia výrobných liniek môže znížiť náklady a zvýšiť príjmy. Veľké údaje môžu výrobcom pomôcť pochopiť tok výrobkov cez ich výrobné linky a zistiť, ktoré oblasti môžu byť prínosom. Analýza údajov odhalí, ktoré kroky vedú k predĺženiu výrobného procesu a ktoré oblasti spôsobujú oneskorenie [17].

Dnes prakticky už každá firma dokáže sledovať tisíce parametrov vo svojom výrobnom procese. Analýzou enormného množstva dát, ktoré stále pribúdajú, dokáže firma predikovať potrebu servisných zásahov a preventívnych opatrení, čo má za následok šetrenie pri prípadných neočakávaných odstávkach z dôvodu porúch [4].

1.3.3 Zdravotníctvo

Zdravotnícke organizácie používajú dáta na rôzne účely od zlepšenia ziskovosti po pomoc pri záchrane životov. Zdravotnícke spoločnosti, nemocnice a vedci zhromažďujú obrovské množstvo údajov. Tieto dáta sú veľmi užitočné, avšak nie vtedy ak sú navzájom izolované. Big data môžu hrať dôležitú úlohu v genomickom výskume. Vedci môžu pomocou údajov identifikovať gény chorôb, aby pomohli pacientom určiť zdravotné

problémy, ktorým môžu v budúcnosti čeliť. Výsledky môžu dokonca zdravotníckym organizáciám umožniť navrhnúť špecializovanú liečbu [17].

Nositeľné sledovacie zariadenia prenášajú údaje lekárom a informujú ich, či pacienti užívali lieky alebo či sa liečia podľa stanoveného plánu. Zozbierané údaje v priebehu času poskytnú lekárom komplexný pohľad na pacientov zdravotný stav a ponúkajú podrobnejšie informácie ako obyčajné osobné návštevy. Okrem iného, dáta a analýzy pomáhajú nemocniciam skrátiť čakaciu dobu a zlepšiť starostlivosť. Niektoré platformy skúmajú údaje hromadne, nachádzajú v nich vzorce a následne dávajú odporúčania na dosiahnutie pokroku [19].

1.3.4 Maloobchod

Konkurencia v maloobchode je veľmi veľká. Spoločnosti sa snažia byť v niečom iné, výnimočné, aby sa odlíšili od tých ostatných. Veľké dáta sa používajú vo všetkých fázach maloobchodného procesu - od vývoja produktov po predikciu budúceho dopytu až po optimalizáciu online stránky a kamenného obchodu. Analýza dát poskytuje informácie potrebné na udržanie spokojnosti zákazníkov. Maloobchodníci využívajú big data a hľadajú nové spôsoby, ako si získať a udržať zákazníkov. Veľké dáta môžu pomôcť predvídať dopyt zákazníkov. Klasifikáciou kľúčových atribútov minulých a súčasných produktov a následným modelovaním vzťahu medzi týmito atribútmi, môžu vytvárať prediktívne modely pre nové produkty a služby. Spoločnosti musia analyzovať veľké množstvo údajov, rôzneho formátu, na základe ktorých potom vytvoria segmenty, podľa správania sa zákazníka [17].

Každý zákazník má pre spoločnosť hodnotu. Avšak niektorí prinášajú vyššiu hodnotu ako iní. Big data poskytujú informácie o správaní a výdavkoch zákazníkov. Spoločnosti musia analyzovať veľké množstvo údajov o zákazníckych transakciách a vytvoriť sofistikované modely, ktoré skúmajú minulé správanie a predpovedajú budúce kroky. Akonáhle poznajú tých najlepších zákazníkov, marketing ich môže zacieliť špeciálnymi ponukami [17].

Spoločnosti v tomto sektore pracujú s rôznymi druhmi dát. Môžu to byť údaje od zákazníkov prostredníctvom vernostných kariet, ekonomické a demografické údaje, sociálne siete a web. Vďaka týmto informáciám potom môžu napr. vytvárať obchodné stratégie, predpovedať dopyt, upravovať ceny, identifikovať potenciálnych zákazníkov a udržať si súčasných [5].

1.3.5 Štát

Vládne inštitúcie v USA čelia niekoľkým technickým a manažérskym výzvam, pokiaľ ide o dátovú analýzu. Občania budú pravdepodobne ochotnejší poskytovať citlivé a užitočné údaje vládam, ak tieto údaje využijú pre zvýšenie kvality života a zlepšenie verejných služieb. Vlády musia určiť, aké druhy údajov skutočne potrebujú. Pri mnohých analytických úlohách, napríklad pri meraní efektívnosti programu, ktorý chcú zaviesť, by interné údaje mohli byť všetko čo je potrebné. Ale pri úlohách ako odhaľovanie podvodov alebo projekcia výnosov sú externé dáta nevyhnutné pre spracovanie analýz [23].

Big Data môžu byť užitočné pri správe prírodných zdrojov ako sú pôda, lesy alebo voda. Napríklad agentúra pre ochranu životného prostredia vedie databázu ohľadom regulovaných objektov, ktorá sa nazýva ECHO. Táto databáza poskytuje integrované údaje o zhruba 800 000 objektoch ohľadom splnenia a dodržovania enviromentálnych regulácii [24].

Spojené štáty americké majú jeden z najmodernejších systémov správy daní pre rôzne sektory hospodárstva. Avšak aj napriek tomu, podľa vládnej agentúry IRS, ktorá spracováva dane, štát príde zhruba o 300 miliárd USD kvôli chybám v zdaňovaní alebo pre podvodné praktiky. IRS využíva Big Data na boj proti týmto problémom. Robo-audity spracovávajú daňové priznania a porovnávajú ich z údajmi z tretích strán ako sú poskytovatelia platobných kariet, internetové platobné systémy, sociálne média, e-mail a mnoho ďalších. Zber a analýza týchto údajov umožňujú IRS vytvárať a sledovať unikátne atribúty týkajúce sa finančného správania, jednoduchšie vymáhať dane a bojovať proti problémom [20].

2 Cieľ práce a metódy skúmania

Cieľom práce je analyzovať a prognózovať vývoj denných maloobchodných tržieb v Spojených štátoch amerických na základe vybraných modelov v programovacom jazyku R. Maloobchodné tržby predstavujú jeden z významných makroekonomických ukazovateľ, ktorý vypovedá o výkonnosti ekonomiky a sú dôležitým faktorom pre rast ekonomiky.

Pre potreby práce musíme byť oboznámený so základnými princípmi štatistky a algebry. Na spracovanie použijeme programovací jazyk R, ktorý je určený pre štatistickú analýzu a grafické zobrazenie dát. Vzhľadom na to, že je open source, má za sebou mimoriadne veľkú komunitu ľudí, ktorí ho neustále aktualizujú o najmodernejšie techniky. Pre spracovanie problematiky budeme potrebovať balík s názvom fpp2, ktorý obsahuje funkcie pre prognózovanie a grafické zobrazenie dát. Programovací jazyk R sa stal štandardom v mnohých oblastiach štatistiky.

Pomocou vybraných metód analyzujeme vybrané dáta a na základe komparácie vybraných ukazovateľov, vyberieme tú najvhodnejšiu metódu pre následnú prognózu. Aby sme splnili hlavný cieľ práce, predstavíme čiastkové ciele, vďaka ktorým bližšie špecifikujeme našu problematiku:

- Transformácia a úprava dát;
- Analýza vývoja denných maloobchodných tržieb v USA od začiatku roka 1992 do konca roka 2019;
- Testovanie vybraných modelov a výber najvhodnejšieho modelu;
- Určenie krátkodobej prognózy vývoja priemerných denných maloobchodných tržieb v USA pomocou vybraného modelu a porovnanie so skutočným stavom za prvé 4 mesiace roku 2020

3 Metodika práce a metódy skúmania

Pred samotnou analýzou, je dôležité zhromaždiť informácie o danej problematike. Táto kapitola bakalárskej práce sa zameriava na teoretický opis základných vlastností časových radov a metodiku, ktorú sme využili pri analýze dát.

3.1 Transformácia a úprava dát

Úprava historických údajov môže často viesť k jednoduchšej prognóze. Zaoberáme sa dvomi úpravami: úprava kalendára a úprava inflácie. Účelom týchto úprav a transformácii je zjednodušiť vzorce v historických údajoch odstránením známych zdrojov variácie alebo zvýšením konzistentnosti modelu v celom súbore dát. Jednoduchšie vzorce vedú k presnejším prognózam.

Kalendár – vzhľadom na to, že základné údaje predstavujú mesačné maloobchodné tržby, medzi mesiacmi budú rozdiely len kvôli rozdielnemu počtu dní, preto tieto dáta upravíme do podoby priemerných denných maloobchodných tržieb.

Inflácia – dáta sú ovplyvňované hodnotou peňazí a práve preto ich musíme ešte pred samotným modelovaním upraviť. Finančné časové rady sa zvyčajne upravujú tak, aby sa všetky hodnoty uvádzali v peňažných hodnotách za konkrétny rok.

3.2 Analýza zložiek časového radu

Časový rad, môžeme chápať ako súbor pozorovaní, ktoré sú chronologicky zoradené vo vzostupnom poradí. Pomocou časového radu teda zaznamenávame pohyb skúmanej premennej. Vychádzame z rozdelenia časového radu na štyri formy pohybu, a to trendový, sezónny, cyklický a náhodný. V praktických úlohách sa najčastejšie stretávame s potrebou oddeliť a vyčistiť trendový a sezónny výkyv a na základe takto upraveného modelu vypočítať prognózu.

Trend – ak hodnoty pozorovaní rastú alebo klesajú s časom, tak hovoríme o pôsobení trendovej zložky. Ak hodnoty pozorovaní rastú alebo klesajú s časom, tak hovoríme o pôsobení trendovej zložky. Trend v časových radoch vyjadruje hlavnú dlhodobú tendenciu vo vývoji radu v čase.

Sezónna zložka - ide o opakovanú krátkodobú fluktuáciu hodnôt pozorovanej premennej s periódou jeden rok alebo kratšou.

Zatiaľ čo modely exponenciálneho vyhladzovania sú založené na opise trendu a sezónnosti údajov, cieľom modelov ARIMA je opísať autokorelácie v údajoch. Práve preto táto metóda vyžaduje stacionárne časové rady.

Stacionarita – je vlastnosť časového radu, ktorá u hodnôt vyvoláva tendenciu vracať sa ku konštante. Stacionaritu môžeme chápať v dvoch zmysloch a to:

- V striktnom zmysle môžeme požadovať nemennosť skúmanej náhodnej premennej v čase, bez ohľadu na to, aký si z nej vyberieme časový úsek.
- V slabom zmysle ako nemennosť základných štatistických mier skúmanej náhodnej premennej v čase, teda jej strednej hodnoty rozptylu a kovariancie.

Ekonomické časové rady často obsahujú trend a preto bežným javom je nestacionarita procesu vzhľadom na priemer. Preto sa musí časový rad hodnôt transformovať. Ak trend časového radu je lineárny, potom jednoduché prvé diferencie proces stacionarizujú [14].

Poznáme niekoľko spôsobov transformácie, pre naše potreby práce použijeme **1. diferenciu** ako obvyklý spôsob pre získanie stacionárneho radu. Diferencovanie je výpočet zmeny medzi po sebe idúcimi pozorovaniami v pôvodnom časovom rade, a môže byť zapísané ako:

$$y'_t = y_t - y_{t-1}$$

Kde: y'_t je rozdiel medzi závislými premennými v čase t

y_t je hodnota premennej v čase

y_{t-1} je hodnota premennej posunutá o jedno obdobie

t sa nesmie rovnať 0, pretože neexistuje diferencia pre 1. pozorovanie y'_1

3.3 Predikčné metódy a ich porovnanie

Niektoré predikčné metódy sú prekvapivo jednoduché a účinné pri tvorení predikcie. Pre potrebu bakalárskej práce je nutné opísať modely, ktoré budeme zvažovať pre prognózu časového radu.

Naivný model

Naivné modely vyjadrujú hypotézy o vzťahu medzi dvoma hodnotami tej istej premennej, ktorú sledujeme v obdobiach bezprostredne nasledujúcich po sebe. Na rozdiel

od iných prognostických techník ich môžeme použiť pre tvorbu prognózy aj vtedy, keď disponujeme malým rozsahom empirických pozorovaní.

Ak hodnoty premennej v čase rastú, tak môžeme použiť iný typ naivého modelu – tzv. **sezónny naivný model**. Tento typ určuje budúcu hodnotu pomocou poslednej pozorovanej hodnote z tej istej sezóny, v našom prípade z toho istého mesiaca predchádzajúceho roka. Model môžeme zapísať v tvare:

$$\hat{y}_{T+h|T} = y_{T+h-m\,(k+1\,)}$$

kde $\hat{y}_{T+h|T}$ - skratka pre odhad y_{T+h} na źaklade dát y_1,...,y_T

m je sezónna prióda

k je počet pozorovaní v prognózovanom období

Model vychádza z predpokladu, že rovnaký prírastok ako zaznamenávame v súčasnom období, zaznamenávame aj v období prognózovanom [12].

Exponenciálne vyrovnávanie

Ďalšou metóda ktorá by mohla byť vhodná na analýzu údajov je metóda exponenciálneho vyrovnávania.

Exponenciálne vyrovnávanie patrí medzi tzv. adaptívne modely, ktorých hlavnou vlastnosťou je, že najnovšie pozorovania časového radu považujeme za najdôležitejšie pre vytvorenie prognózy. Zdôraznenie najneskorších hodnôt premennej je možné dosiahnuť voľbou odlišných váh jednotlivých pozorovaní tak, aby tie najnovšie údaje mali najväčšiu váhu [15].

$$\hat{y}_{T+h|T} = \propto y_T + \propto (1-\propto) y_{T-1} + \propto (1-\propto)^2 y_{T-2}$$

α – vyrovnávacia konštanta, kde $0 \leq \alpha \leq 0$

Ak sa vyrovnávacia konštanta blíži nule, väčšia váha sa pripíše pozorovaniam v minulosti, a ak sa konštanta blíži k jednotke viac váhy sa prisudzuje nedávnym pozorovaniam.

Táto metóda sa využíva najmä na krátkodobé predikcie. Jej výhodou je, že sa hodí na predikciu dát, pri ktorých nie je jasné či obsahujú trendovú alebo sezónnu zložku.

Box-Jenkinsova metodológia ARIMA modelov

Boxova-Jenkinsonova metodológia je východiskom pre modelovanie nestacionárnych časových radov a sezónnych časových radov. V súčasnosti sa táto technika používa ako nový smer dynamického modelovania ekonomických javov.

Nevyhnutnou podmienkou pre využitie metód je stacionarita časového radu. Exponenciálne vyhladzovacie modely a modely ARIMA (model autoregresných integrovaných kĺzavých priemerov) sú dva najpoužívanejšie prístupy k prognózovaniu časových radov. Zatiaľ čo modely exponenciálneho vyhladzovania sú založené na opise trendu a sezónnosti údajov, cieľom modelov ARIMA je opísať autokoreláciu v údajoch [12]. ARIMA model sa skladá z 3 procesov:

Autoregresný proces (AR)

Vyjadruje modelovanie vývoja hodnôt časového radu na základe časovo oneskorených hodnôt toho istého časového radu, teda závislosť časového radu od jeho vlastných oneskorených hodnôt. Pracuje so závislosťou medzi pôvodnými hodnotami a oneskorenými hodnotami časového radu. Proces nám pomáha modelovať lineárnu závislosť hodnôt časového radu y_t od p oneskorených hodnôt.

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + e_t$$

kde e_t je náhodná premenná

Proces kĺzavých priemerov (MA)

Vyjadruje lineárnu závislosť hodnôt časového radu y_t od q oneskorených hodnôt náhodných šokov e_t. Podstatou kĺzavých priemerov je popísať vývoj hodnôt časového radu pomocou časového radu náhodných šokov.

$$y_t = c + e_t + \phi_1 e_{t-1} + \phi_2 e_{t-2} + \cdots + \phi_q e_{t-q}$$

Proces Integrácie (I)

Ak skombinujeme diferencovanie s autoregresným modelom a modelom kĺzavých priemerov dostaneme nesezónny ARIMA model. Pri tvorbe tohto modelu je dôležité nestacionárny časový rad transformovať na stacionárny pomocou diferencií. To znamená,

že pôvodný rad nahradíme radom vytvoreným z jeho diferencií prvého, druhého alebo vyššieho rádu. Tieto časove rady reprezentujú tzv. integrované procesy.

$$y'_t = c + \phi_1 y'_{t-1} + \cdots + \phi_p y'_{t-p} + \theta_1 e_{t-1} + \cdots + \theta_q e_{t-q} + e_t$$

Porovnanie prognóz

Presnosť použitých prognostických metód overujeme pomocou dvoch vybraných charakteristík.

Smerodajná odchýlka chýb prognóz

Je často používaným meradlom rozdielu medzi prognózovanými a skutočnými hodnotami časového radu. Používa sa pri rozhodovaní o výbere najvhodnejšieho modelu, ktorý najlepšie opisuje vybraný časový rad.

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)}$$

Vyberáme model, ktorý bude mať najmenšiu hodnotu tejto odchýlky.

Autokorelačná funkcia

Typickým znakom časových radov je silná korelovanosť. Autokorelačná funkcia určuje mieru lineárnej závislosti časovo posunutých veličín y_t a y_{t-k}. Koeficienty aurokorelácie reziduí sú definované vzťahom

$$r_k = \frac{\sum_{t=k+1}^{T}(y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum_{t=1}^{T}(y_t - \bar{y})^2}$$

Kde T je dĺžka časového radu

Grafické zobrazenie autokorelačnej funkcie sa nazýva **korelogram**. Tento graf nachádza uplatnenie pri modelovaní časových radov a pri ich dekompozícii. Jednotlivé stĺpce autokorelačnej funkcie zobrazujú autokorelačné koeficienty, ktoré vyjadrujú silu lineárnej závislosti medzi hodnotami časového radu. Cieľom je aby tieto stĺpce neprekročili určitú hodnotu.

4 Výsledky práce

4.1 Údajová základňa

Dáta, ktoré použijeme na analýzu čerpáme z databázy FRED-u (Federal Reserve Bank of St. Louis – Federálna rezervná banka v St. Louis). Údaje o maloobchodných tržbách v USA sú zaznamenávané v mesačných intervaloch od roku 1992 do decembra 2019 v nominálnych amerických dolároch.

Graf č. 1: Maloobchodné tržby v USA od roku 1992 do 2019

Zdroj: www.fred.stlouisfed.org, 2020

Avšak predtým ako začneme analyzovať údaje, upravíme ich aby sme mohli výsledky jednoduchšie interpretovať. Stiahnuté dáta sú v nominálnych dolároch, ale vzhľadom na to, že peniaze sa znehodnocujú infláciou, upravíme tieto dáta o infláciu. Bez tejto úpravy by sme predikovali zmenu tržieb a zároveň zemenu inflácie v jednom. Preto upravíme dáta tak, aby sme ich mali v súčasnej reálnej hodnote USD. Ako bázickú hodnotu sme vybrali hodnotu CPI v decembri 2019. Následne každý mesiac vydelíme počtom dní, čím získame hodnoty priemerných denných tržieb v danom mesiaci.

Tabuľka č.1: Časť údajovej základne od roku 1992 do roku 2019 v MS Excel

Dátum	Tržby v mil. USD (nominálne)	CPI	index	Tržby v mil. USD (12/2020)	Počet dní v mesiaci	Priemerné denné tržby v mil. USD
1992-01-01	130683	138,300	0,535125598	244209,9584	31	7877,740595
1992-02-01	131244	138,600	0,536286391	244727,4483	29	8438,877528
1992-03-01	142488	139,100	0,538221046	264738,8114	31	8539,96166
1992-04-01	147175	139,400	0,539381839	272858,6492	30	9095,288307

Zdroj: Vlastné výpočty v MS Excel.

Stĺpec A – dátum, B – Tržby v miliónoch v nominálnych USD, C – CPI (Index spotrebiteľských cien), D – index (CPI v decembri 2019 ÷ CPI v danom mesiaci), E – Mesačné tržby v miliónoch USD upravené o infláciu, F – počet dní v mesiaci, G – Priemerné denné tržby v miliónoch USD upravené o infláciu.

Na analýzu použijeme priemerné denné tržby v USA, čiže výsledok predikcie budú priemerné tržby za jeden deň v mesiaci.

4.2 Integrácia a spracovanie dát

Pre spracovanie údajov sme použili programovací jazyk R, ktorý postačuje na analýzu dát a ich grafické zobrazenie. Prvý krok bude vloženie dát do R pomocou funkcie read_excel() a následné zobrazenie a kontrola údajov cez funkciu view()

Obrázok č.2: Vloženie dát pomocou funkcie read_excel() a zobrazenie dát cez view()

```
priemerne_denne_maloobchodne_trzby <- read_excel("D:/Rko/data/priemerne_denne_maloobchodne_trzby.xlsx")
View(priemerne_denne_maloobchodne_trzby)
```

Zdroj: Vlastné spracovanie v programovacom jazyku R.

Tabuľka č.2: Priemerné denné tržby v miliónoch USD v daný mesiac

	Dátum	Priemerné denné tržby v mil. USD
1	1992-01-01	7877.741
2	1992-02-01	8438.878
3	1992-03-01	8539.962
4	1992-04-01	9095.288
5	1992-05-01	9095.997
6	1992-06-01	9337.250
7	1992-07-01	9054.055
8	1992-08-01	9028.250
9	1992-09-01	9045.723
10	1992-10-01	9177.477
11	1992-11-01	9386.191
12	1992-12-01	11210.411

Zdroj: Vlastné spracovanie v programovacom jazyku R.

Následne načítame balíček, ktorý obsahuje funkcie potrebné pre analýzu, grafické zobrazenie a prognózu na nasledujúci rok.

Obrázok č. 3: Načítanie balíka fpp2 pomocou funkcie library()

```
library(fpp2)
```

Zdroj: Vlastné spracovanie v programe R.

V ďalšom kroku definujeme premenné a ukážeme, že sa jedná o časový rad, pretože hodnoty sú zaznamenávane a pozorované postupne v čase. Y je premenná, „ts“ predstavuje funkciu časového radu a v hranatej zátvorke určíme, že 2. stĺpec dátového súboru je ten, ktorý nás zaujíma. Ďalej definujeme začiatok časového radu, to je rok 1992 začínajúci prvým mesiacom a frekvenciu v akej sa dáta ukladajú, čiže 12, lebo je 12 mesiacov v roku.

Obrázok č. 4: Funkcia ts() pre definovanie časového radu

```
Y <- ts(priemerne_denne_maloobchodne_trzby[,2],start=c(1992,1),frequency = 12)
```

Zdroj: Vlastné spracovanie v programovacom jazyku R

Pred samotnou predikciou treba najprv spraviť predbežnú analýzu správania sa časového radu. Pre časové rady je grafická analýza najužitočnejšia. Vytvoríme graf, ktorý bude podobný ako na obrázku č.2, avšak nepoužijeme pôvodné údaje, ktoré predstavujú celkové mesačné maloobchodné tržby, ale naše upravené údaje ktoré zobrazujú priemerné denné malobochodné tržby v danom mesiaci.

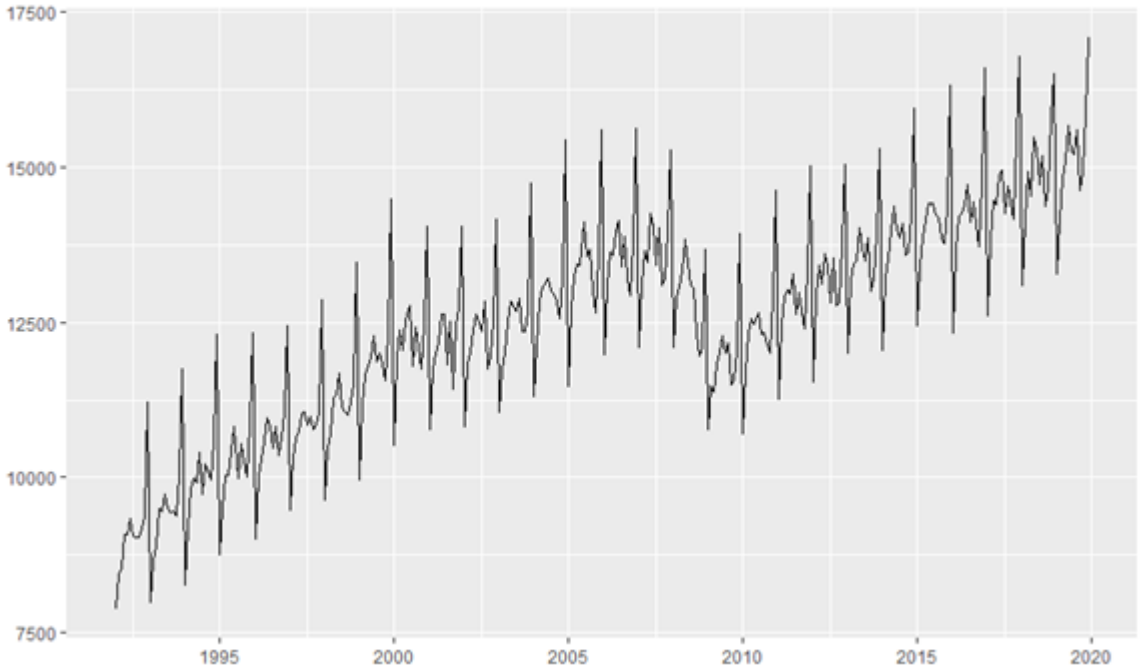
Autoplot() je funkcia pre zobrazenie grafu. V zátvorke je premenná, ktorú sme definovali v predošlom kroku.

Obrázok č. 5: Funkcia autoplot() pre zobrazenie grafu

```
autoplot(Y)
```

Zdroj: Vlastné spracovanie v programovacom jazyku R.

Graf č. 2: Priemerné denné maloobchodné tržby v mil. USD



Zdroj: Vlastné spracovanie v programovacom jazyku R.

Graf je podobný ako graf č.1 aj napriek tomu, že sme dáta upravili o infláciu. Môžeme vidieť podobný rastúci trend, keďže celkové maloobchodné tržby rastú, samozrejme okrem rokov 2008 a 2009 v období krízy. Pravdepodobne časový rad obsahuje aj sezónnu zložku, pretože môžeme pozorovať vysoké výkyvy hodnôt, kedy tržby prudko rastú a následne prudko klesajú. Vzhľadom na to, že tieto dáta majú rastúci trend a pravdepodobne obsahujú sezónnu zložku, transformujeme tieto dáta na staciónarne.

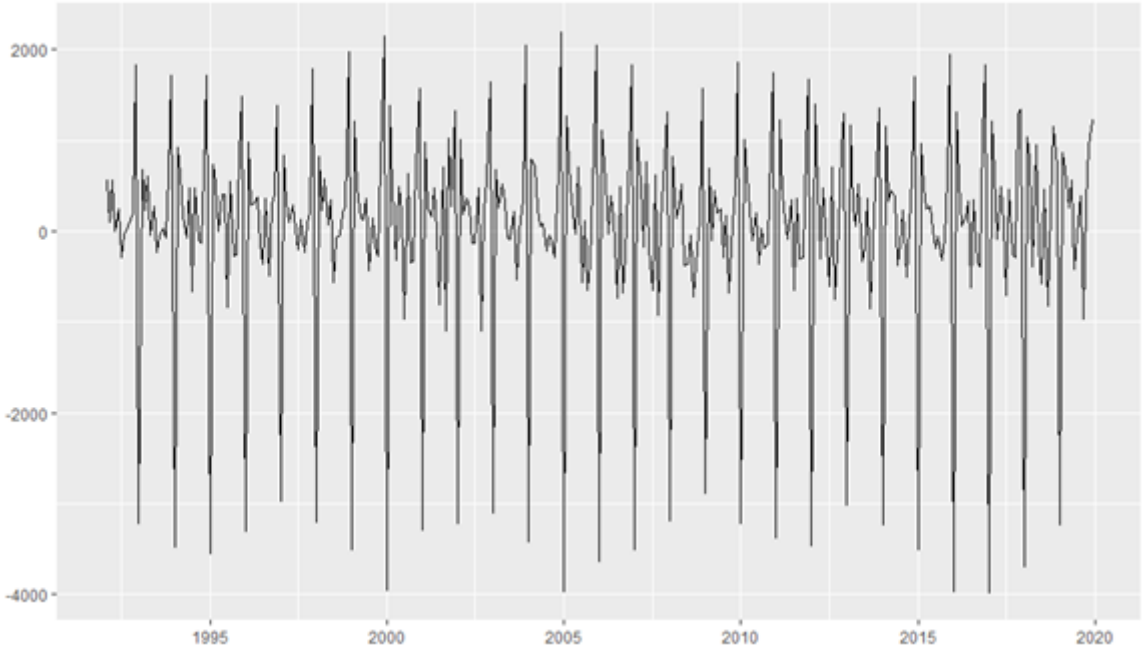
Prvou diferenciou sa časový rad stane stacionárnym. Namiesto pôvodných dát budeme teda pracovať s dátami, ktoré predstavjú medzimesačnú zmenu priemerných denných tržieb. Prvá diferencia bude teda rozdiel medzi priemernými dennými tržbami v januári 1992 a februári 1992 atď. Na zobrazenie grafu opäť použijeme funkciu autoplot().

Obrázok č. 6: funkcia diff() pre výpočet diferencie a zobrazenie grafu autoplot()

```
DY <- diff(Y)

autoplot(DY)
```

Zdroj: Vlastne spracovanie v programovacom jazyku R

Graf č. 3: Zmeny priemerných denných maloobchodných tržieb v mil. USD



Zdroj: Vlastné spracovanie v programovacom jazyku R

Týmto procesom sme sa zbavili trendu a na grafe č. 3 môžeme vidieť veľké kolísanie hodnôt a ďalšou analýzou zistíme, či sa tieto výkyvy pravidelne opakujú alebo sú náhodné.

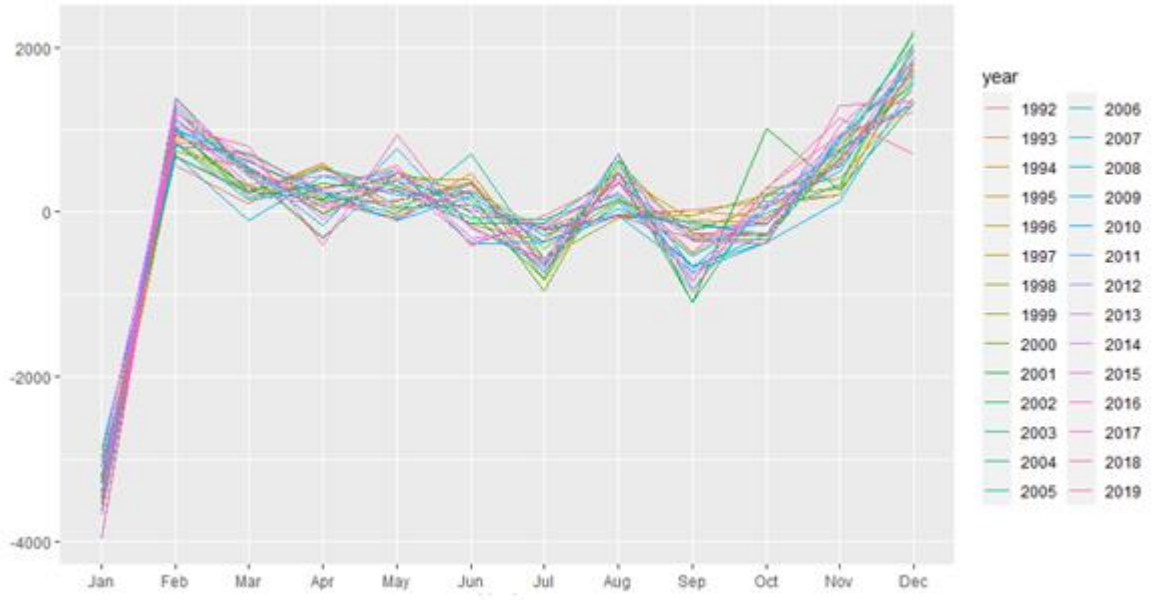
Pomocou funkcie ggseasonplot() dokážeme celý časový rad rozdeliť na rovnako dlhé časti, v tomto prípade jeden rok. Môžeme vidieť ako sa každý jeden rok správala zmena priemerných denných tržieb, pretože každý rok je reprezentovaný inou farebnou linkou na grafe č. 4 nižšie.

Obrázok č. 7: Funkcia ggseasonplot() pre zobrazenie grafu č.4

```
ggseasonplot(DY)
```

Zdroj: Vlastné spracovanie v porgramovacom jazyku R

Graf č. 4: Zmena priemerných denných maloobchodných tržieb podľa rokov v mil. USD



Zdroj: Vlastné spracovanie v programovacom jazyku R

Zmena výšky priemerných denných tržieb z decembra na január vyzerá byť vždy negatívna. Zmena sa každým rokom líši ale možeme povedať, že táto zmena bola počas sledovaného obdobia vždy negatívna. Taktiež môžeme pozorovať pravidelný nárast od novembra do decembra, čo naznačuje sezónne opakovania sa určitých javov.

Ďalší spôsob ako môžeme odhaliť sezónnosť je pomocou funkcie ggsubseriesplot().

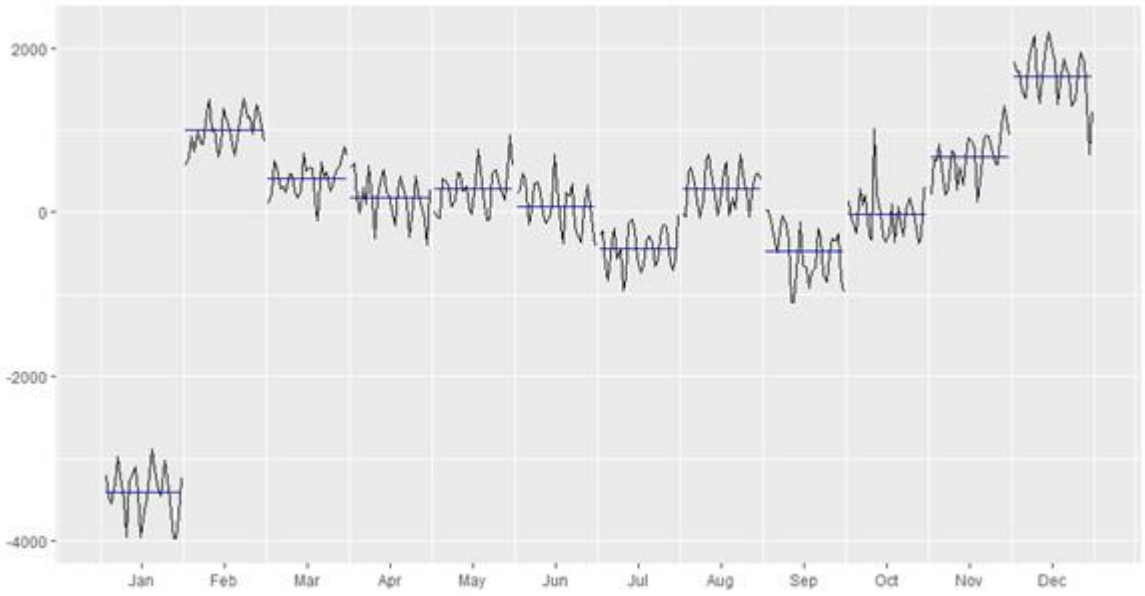
Obrázok č. 8: Funkcia ggsubseriesplot() pre zobrazenie grafu č. 5

```
ggsubseriesplot(DY)
```

Zdroj: Vlastné spracovanie v programovacom jazyku R

Na grafe č. 5 nižšie, čierna čiara spája zmeny v jednotlivých mesiacoch počas celého sledovaného obdobia (napr. zemny v decembri 1992,1993,1994...2019) a modrá čiara reprezentuje priemer daných hodnôt v mesiaci. Vďaka tejto funkcí môžeme pozorovať, že priemerná zmena v januároch jednotlivých rokov je hlboko negatívna a naopak priemerná zmena vo februári a decembri je vždy pozitívna.

Graf č. 5: Zmena priemerných denných maloobchodných tržieb podľa mesiacov



Zdroj: Vlastné spracovanie v porgramovacom jazyku R

Na základe grafickej analýzy dát môžeme tvrdiť, že denné priemrné maloobchodné tržby v USA sa systematicky zvyšujú, teda majú rastúci trend. Okrem rastúceho trendu vidno aj cyklické zmeny, ktoré sa pravidelne opakujú – tzv. sezónnosť. Pre odstránenie trendu sme využili prvú diferenciu, ktorá tiež obsahuje sezónnu zložku. Vzhľadom na to, že dáta majú trend a sezónnosť, musíme tieto javy brať do úvahy prí výbere predikčných modelov.

4.3 Aplikácia modelov a ich porovnanie

Sezónny naivný model

Prvý model ktorý budeme testovať je sezónny naivný model. V jazyku R použijeme funkciu snaive(), summary() pre zobrazenie zhrnutia, kde je uvedený typ modelu, aký sme použili a taktiež štandardná odchýlka a checkresiduals() pre grafické zobrazenie.

Obrázok č. 9: Funkcia snaive() pre naivný sezónny model

```
model1 <- snaive(DY)
summary(model1)
checkresiduals(model1)
```

Zdroj: Vlastné spracovanie v programovacom jazyku R

Kvalita modelu sa jednoducho overuje podľa veľkosti reziduálnej odchýlky. Je to rozdiel medzi skutočnou hodnotou v čase a prognózovanou hodnotou v tom istom čase.

$$e_t = Y_t - \hat{Y}_t$$

Symbol e označuje tzv. rezíduum. Čím je hodnota reziduí bližšie nule, tým presnejší model sa podarilo získať.

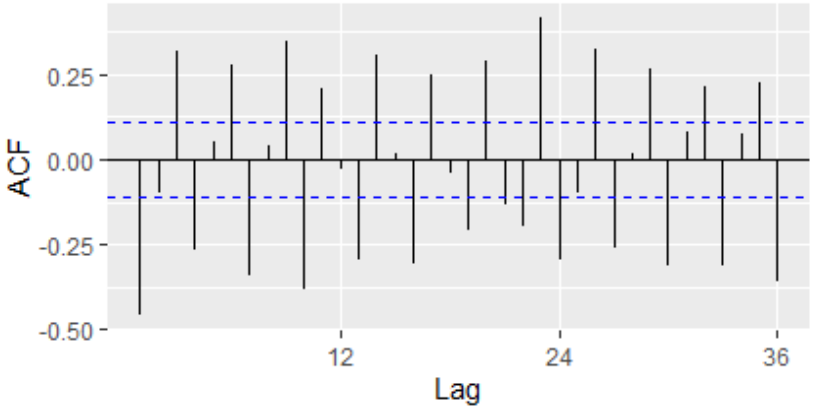
Obrázok č. 10: Výstup zo sezónneho naivného modelu

```
Forecast method: Seasonal naive method

Model Information:
Call: snaive(y = DY)

Residual sd: 304.1848
```

Zdroj: Vlastné spracovanie v programovacom jazyku R

Veľkosť reziduálnej odchýlky je 304,1848 miliónov USD. To znamená, že model sa mýli v primere o 304,1848 milióna USD. Túto hodnotu môžeme považovať za štandard – benchmark. Tento typ modelov slúži na porovnanie presnosti predpovede s ostatnými prognostickými modelmi.

Obrázok č. 11: Korelogram sezónneho naivného modelu.



Zdroj: Vlastné spracovanie v programovacom jazyku R

Na obrázku č. 11 môžeme vidieť, že nielen prvá, ale aj viaceré ďalšie hodnoty prekračujú hranice modrej prerušovanej čiary. Cieľom je aby všetky čierne čiary ostali pod úrovňou modrej prerušovbanej čiary. Na základe týchto informácii môžeme tvrdiť, že existuje model, ktorý vysvetľuje vačšie percento hodnôt ako tento sezónny naivný model.

Exponenciálne vyrovnávanie

Ďalšou metóda ktorá by mohla byť vhodná na analýzu údajov je metóda exponenciálneho vyrovnávania. Existujú rôzne typy týchto modelov, niektoré zohľadňujú trend a naopak niektoré trend nezohľadňujú. Funkcia ets() vyskúša všetky možné modely exponenciálneho vyrovnávania a vyberie ten najvhodnejší.

Obrázok č. 12: Funkcia ets() pre exponenciálny model

```
model2 <- ets(Y)
summary(model2)
checkresiduals(model2)
```

Zdroj: Vlastné spracovanie v programovacom jazyku R

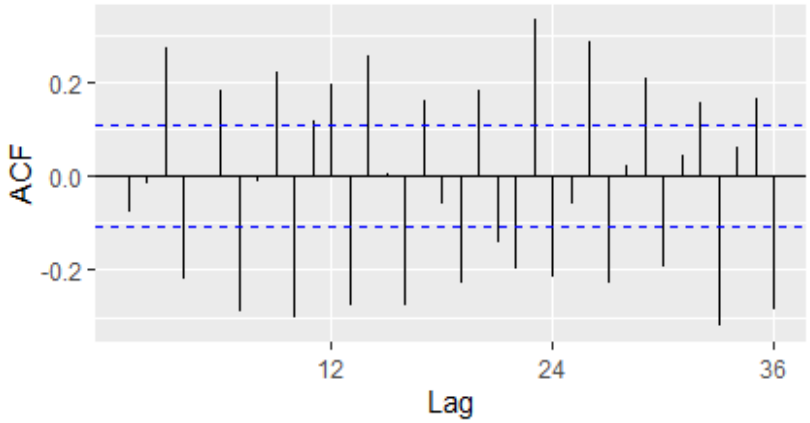
Obrázok č. 13: Výstup z modelu exponenciálneho vyrovnávania

```
ETS(A,N,A)

Call:
 ets(y = DY)

  Smoothing parameters:
    alpha = 3e-04
    gamma = 1e-04

  Initial states:
    l = 17.1995
    s = -3429.852 1651.299 646.2562 -36.3164 -469.8328 276.1686
           -477.9216 60.0648 276.2533 146.2995 401.4845 956.0963

  sigma:  282.1828
```

Zdroj: Vlastné spracovanie v programovacom jazyku R

Veľkosť odchýlky je 282,1828 miliónov USD čo je nižšie ako v predošlom modely. To znamená, že tento model je lepší ako sezónny naivný model.

Obrázok č. 14: Korelogram exponenciálneho modelu



Zdroj: Vlastné spracovanie v programovacom jazyku R

Avšak, keď sa pozrieme na korelogram, väčšina hodnôt presahuje interval spoľahlivosti, to znamená, že stále existujú informácie, ktoré tento model nedokáže efektívne využiť. Inými slovami, stále existuje model, ktorý bude lepší na prognózovanie.

ARIMA model

Pomocou funkcie auto.arima() jazyk R vyskúša ARIMA modely s rôznimi parametrami a vyberie model, ktorý sa najlepšie hodí. Ak chceme pracovať s modelom ARIMA, musíme použiť stacionárne dáta. Keďže cieľom práce je prognóza priemerných denných maloobchodných tržieb v USA, musíme premennú zapísať, takým spôsobom, aby sme ju očistili o trend a sezónnosť.

Obrázok č. 15: Funkcia auto.arima() pre ARIMA modely

```
model3 <- auto.arima(Y,d=1,D=1, stepwise = FALSE, approximation = FALSE)
print(summary(model3))
checkresiduals(model3)
```

Zdroj: Vlastné spracovanie v programovacom jazyku R

Táto metóda má rôzne nastavenia. Namiesto toho aby sme skúšali všetky možné kombinácie ARIMA modelov, nastavíme parameter stepwise na FALSE, to zapríčíni, že každý ďalší model bude trošku pozmenený, aby sa našiel ten najefektívnejší.

Vzhľadom na to, že táto metóda bola vyvynutá na to aby sedela na rôzne časové rady, auto.arima() v základnom nastavení robí analýzu čo najrýchlejšie. Niektoré premenné iba odhaduje, napríklad odhaduje AIC (Akaikeho informačné kritérium) namiesto toho aby túto hodnotu presne určil. Touto metódou sa hlavne šetrí čas, ale keďže pracujeme iba s jedným

časovým radom, nepotrebujeme šetriť čas a preto sme pramtetru approximation priradili hodnotu FALSE.

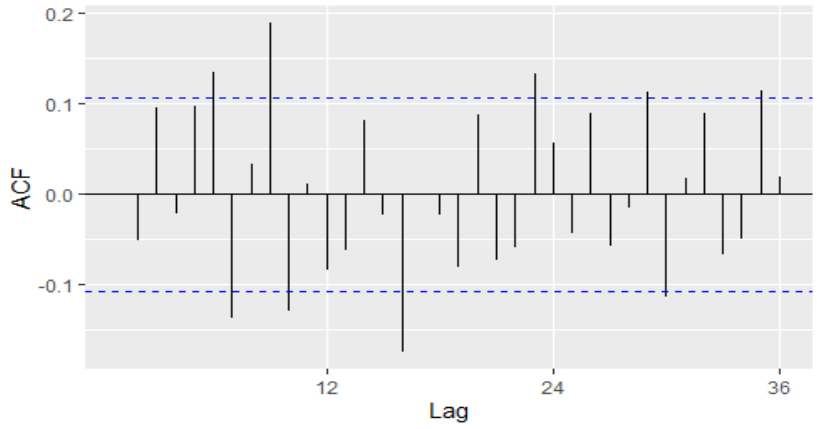
Obrázok č. 16: Výstup z ARIMA modelu

```
Series: DY
ARIMA(0,0,1)(2,1,2)[12]

Coefficients:
          ma1     sar1     sar2     sma1    sma2
      -0.4828  0.7872  -0.6642  -1.1991  0.5242
s.e.   0.0479  0.0646   0.0736   0.0700  0.0830

sigma^2 estimated as 43102:  log likelihood=-2187.43
AIC=4386.86   AICc=4387.12   BIC=4409.52
```

Zdroj: Vlastné spracovanie v programovacom jazyku R

Na obrázku č. 19 vyššie môžeme vidieť aký typ modelu sedel najlepšie k našim dátam. Vo výstupe môžeme vidieť rozptyl sigma^2 = 43102. Ak chceme získať odchýlku, musíme túto hodnotu odmocniť a tak získame hodnota štandardnej odchýlky 207,6102.

Obrázok č. 17: Korelogram druhých mocnín reziduí ARIMA modelu



Zdroj: Vlastné spracovanie v programovacom jazyku R

Na korelograme taktiež vidieť zlepšenie, pretože sa väčšina hodnôt nachádza v rozmedzí modrej prerušovanej čiary. Stále sa tu nachádza autokorelácia, ale už v menšej miere ako v predošlích modeloch.

S istotou môžeme povedať, že spomedzi modelov, ktoré sme testovali, je ARIMA model ten najlepší, ale nie perfektný, kvôli autokrelácii. Tým pádom určite existuje model, ktorý je efektívnejši, avšak tento model nepoznáme a pravdepodobne bude aj o mnoho komplikovanejší ako modely, ktoré sme testovali.

4.4 Prognóza pomocou ARIMA modelu a porovnanie s reálnym stavom

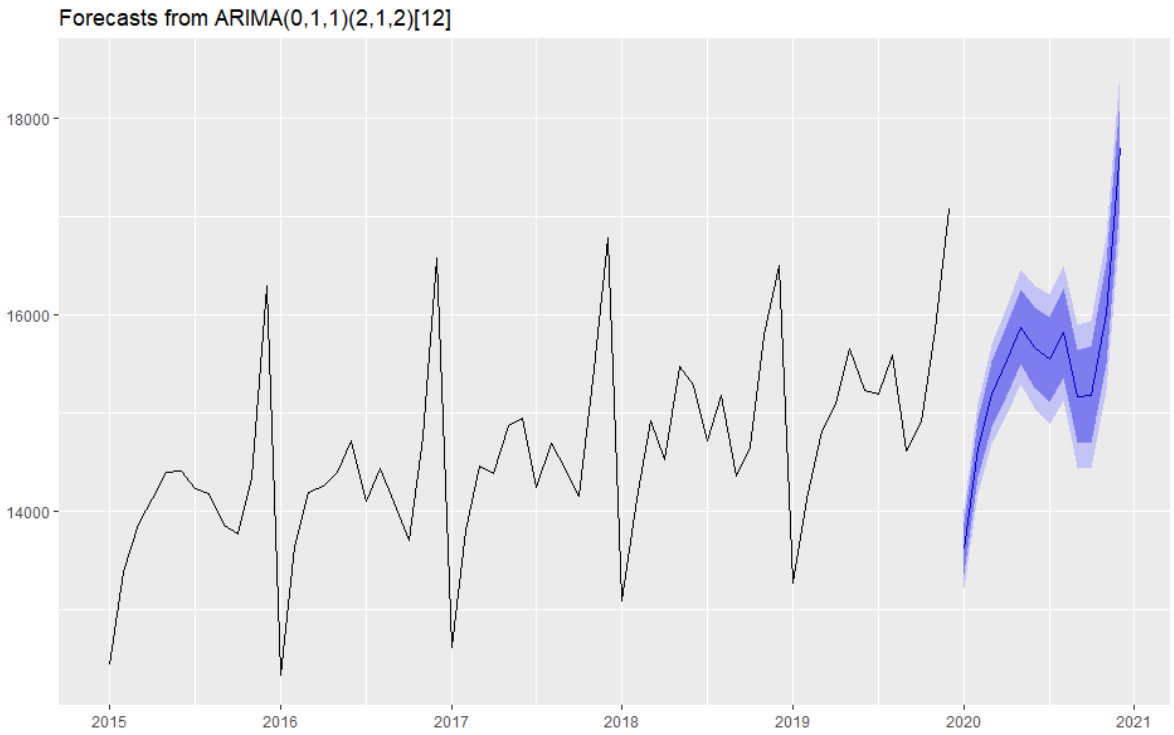
V jazyku R prognózujeme pomocou funkcie forecast(), kde ako premennú zapíšeme ARIMA model. Prameter h určuje na aký dlhý časový úsek chceme robiť našu predikciu, v tomto prípade na 1 rok dopredu, ale vyhľadom, že pracujeme s mesačnými dátami, hodnota prametra bude 12. Použijeme funkciu autoplot() na zobrazenie grafu, využijeme aj parameter include na zobrazenie posledných 5 rokov – 60 mesiacov pre lepšie zobrazenie a summary() pre zobrazenie konkrétnych výsledkov.

Obrázok č. 18: Funkcia forecast() pre prognózu časového radu

```
ARIMAprognoza <- forecast(model3, h=12)
autoplot(ARIMAprognoza,include=60)
summary(ARIMAprognoza)
```

Zdroj: Vlastné spracovanie v programovacom jazyku R

Graf. č. 6: Prognóza vývoja priemerných maloobchodných tržieb pre rok 2020



Zdroj: Vlastné spracovanie v programovacom jazyku R

Obrázok č. 19: Zhrnutie predikcie ARIMA modelu

Forecasts:

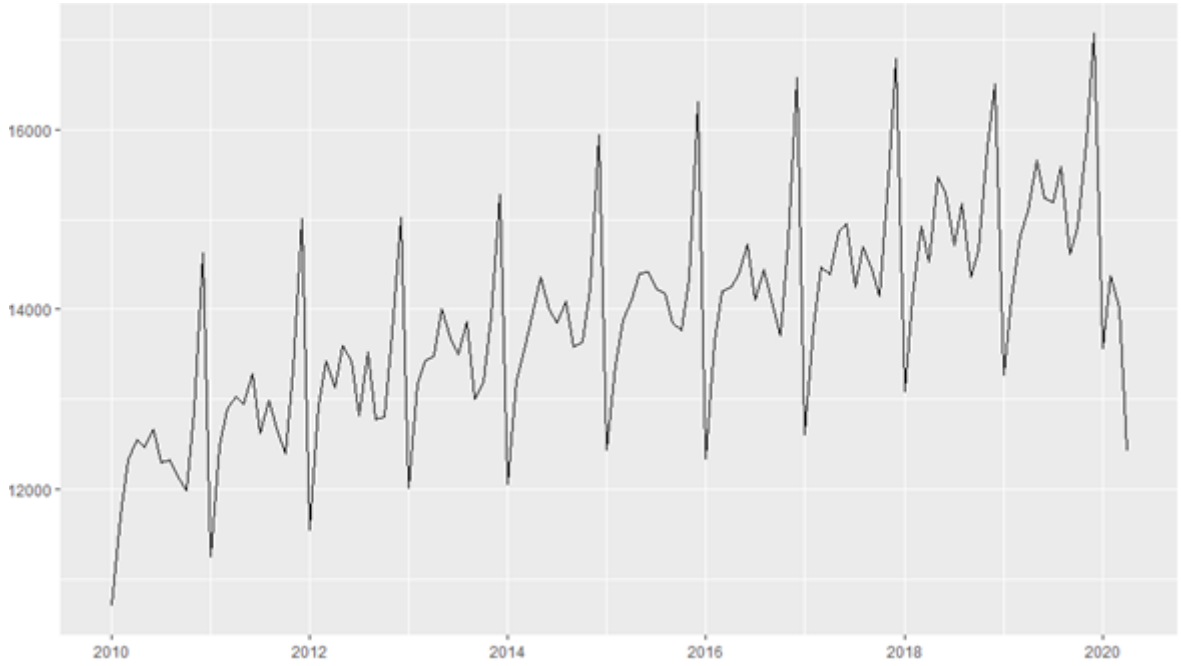
	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
Jan 2020	13622.34	13356.26	13888.41	13215.41	14029.26
Feb 2020	14644.43	14344.87	14943.99	14186.30	15102.57
Mar 2020	15194.24	14864.58	15523.90	14690.07	15698.41
Apr 2020	15539.37	15182.13	15896.60	14993.03	16085.71
May 2020	15879.10	15496.28	16261.93	15293.62	16464.59
Jun 2020	15671.15	15264.33	16077.96	15048.98	16293.31
Jul 2020	15547.97	15118.51	15977.43	14891.17	16204.78
Aug 2020	15817.77	15366.80	16268.74	15128.06	16507.47
Sep 2020	15174.18	14702.68	15645.69	14453.08	15895.29
Oct 2020	15186.35	14695.17	15677.53	14435.15	15937.54
Nov 2020	16017.18	15507.08	16527.28	15237.06	16797.31
Dec 2020	17694.06	17165.72	18222.39	16886.04	18502.08

Zdroj: Vlastné spracoanie v programovacom jazyku R

Na grafe č. 6 vyššie je porgnóza vyznačená modrou čiarou. Základom grafu je centrálna predikcia – najpravdepodobnejší vývoj priemerných denných maloobchodných tržieb v USA pre dané obdobie. Predikcia vyzerá celkom realisticky, keďže zachytáva rastúci trend a sezónnosť, o ktorej sme hovorili. Ďalej môžeme vidieť, že modrá čiara vyjadruje exaktné hodnoty v danom čase – tzv. bodová predikcia, ale graf obsahuje aj tmavo modrú a svetlo modrú zónu, ktorá vyjadruje interval v akom sa môžu hodnoty pohybovať. Tmavo modrá zobrazuje 80% interval spoľahlivosti, čo znamená, že je 80% šanca, že sa hodnoty budú pohybovať v tomto intervale, svetlo modrá vyjadruje 95% interval spoľahlivosti. Dôležité je aby sme sa neopierali o konkrétne body, ale počítali aj s učitou odchýlkou.

V zhrnutí na obrázku č. 22 vyššie sa nachádzajú konkrétne hodnoty bodovej predikcie a taktiež aj hodnoty horných a dolných hraníc intervalov spoľahlivosti.

Graf č. 6: Reálny vývoj priemerných denných maloobchodných tržieb



Zdroj: Vlastné spracoanie v programovacom jazyku R

Na grafe č. 6 je zobrazený priebeh maloobchodných tržieb od začiatku roku 2010 po apríl roku 2020. Na prvý pohľad je zrejmé, že priebeh hodnôt začiatkom roka 2020 je odlišný od prognózovaného pohybu. Januárová a feruárová predikcia vyšla podľa očakávaní, ale v marci a apríli sme predpokladali nárast priemerných maloobchodných tržieb ako za uplnulých pár rokov. Realitou je, že priemerné denné maloobchodné tržby sú v marci a apríli hlboko negatívne. Takýto hlboký prepad naznačuje príchod ekonomickej recesie. Tento aprílový prepad bol zatiaľ jeden z najvyšších prepadov v celej historií meriania tohto makroekonomického ukazovateľa.

Tabuľka č. 3: Porovnanie skutočných hodnôt a prognózy

Obdobie	Skutočnosť	Prognóza	% chyba
Jan 2020	13563,91156	13622,34	0,430
Feb 2020	14372,40301	14644,43	1,893
Mar 2020	14027,1977	15194,24	8,320
Apr 2020	12421,94628	15539,37	25,096

Zdroj: Vlastné spracovanie v MS Excel

Na základe tabuľky č. 3 môžeme tvrdiť, že v prvom období (január) sa model mýlil iba o 0,430 % a v druhom o 1,893 % čo je v rozmedzí 95 % intervalu spoľahlivosti. V marci

môžeme pozorovať vysokú odchýlku reálneho stavu od prognózy 8,32 % a v apríli je rozdiel ešte vyšší a to 25,096 %. Príčinou týchto vysokých rozdielov medzi reálnou a prognózovanou hodnotou, nie je zlé prognózovanie ale neočakávaná situácia ohľadom pandémie COVID-19. Pandémia síce vypukla už koncom roku 2019, ale dopady na americkú ekonomiku má až v týchto mesiacoch (marec, apríl). Spojené štáty americké sú v čase písania práce lídrom v počte nakazených. Vzhľadom na túto situáciu, jednotlivé štáty prijali opatrenia a nariadili tzv. "lockdown" kedy dočasne zatvorili firmy a obchody, s cieľom obmedziť kontakt medzi obyvateľmi, aby sa vírus nešíril ďalej. Jedným z dôvodom prečo maloobchodné tržby klesali je jednak strach spotrebiteľov z nákazy, čiže radšej ostanú doma a nevytvárajú tržby a druhým hlavným dôvodom je, že aj keby spotrebitelia chceli míňať svoje peniaze, väčšina firiem a obchodov je zatvorených a teda majú obmedzené možnosti pri kúpe tovarov a služieb.

Ďalším neželaným následkom pandémie nie je len pokles maloobchodných tržieb ale aj prudký nárast nezamestnanosti, ktorá sa v apríli tohoto roku zvýšila zo 4,4 % na 14,7 % čo predstavuje nárast zhruba o 20 miliónov nezamestnaných. Mnoho firiem muselo prepustit zamestnatncov z jednoduchého dôvodu, keďže nemajú tržby nemôžu si dovoliť vyplácať mzdy. Rastú preto obavy, že americký trh práce zažije podobný šok ako počas veľkej hospodárskej krízy v 30. rokoch minulého storočia, kedy ekonomika niekoľko rokov po sebe klesala a nezamestnanosť stúpla až k 25 %. Prudký nárast nezamestnatnosť je teda ďalším faktorom, ktorý ovplyvnil vývoj tržieb, vzhľadom na to, že obyvatelia prišli o príjem a nemínajú toľko, koľko sa očakávalo.

Americká národná maloobchodná federácia vo februári tohto roku vydala ročnú predpoveď kde očakávala nárast maloobchodných tržieb medzi 3,5 % až 4,1 % čo by predstavovalo celkovo 3,9 biliónov USD v roku 2020. Tento odhad ale nepočítal s tým, že sa z koronavírusu stane globálna pandémia. Analytici očakávajú, že po skončení epidémie nastane výrazné oživenie, čiže sa očakáva, že ekonomika bude zase rásť už na jeseň tohto roku, avšak na úroveň spred krízi sa vráti najskôr budúci rok.

Vačšina podnikov je kvôli tejto situácii na pokraji bankrotu a bez zásahu štátu neprežijú nasledujúce mesiace.

ZÁVER

Cieľom bakalárskej práce bolo podanie základných informácií z oblasti problematiky spracovania veľkých objemov údajov – Big Data. Dôležitosť ich získavania, triedenia, spracovania pomocou rôznych metód a následnej analýzy demonštrujeme na prognóze maloobchodných tržieb v USA. Tieto tržby považujeme za jeden z hlavných makroekonomických ukazovateľov smerovania svetovej ekonomiky.

V prvých troch kapitolách sme charakterizovali Big Data a základné kroky, ktoré sú potrebné pre ich spracovanie. Uviedli sme niekoľko praktických príkladov, v akých odvetviach sa Big Data plne využívajú a prispievajú k zlepšeniu konkurencieschopnosti podnikov, ale aj kvality života občanov a vytýčili sme hlavný cieľ a sekundárne ciele, ktoré boli potrebné pre úspešné splnenie hlavného cieľa. Následne sme popísali metódy, bez ktorých by sme tieto ciele nedokázali splniť.

Posledná štvrtá časť sa týkala praktickej aplikácie popísaných metód v programovacom jazyku R, kde sme na základe metodológie postupovali pri analýze maloobchodných tržieb v USA. Najprv sme získané dáta pretransformovali do podoby denných priemerných maloobchodných tržieb a upravili sme ich o infláciu. Potom sme ich integrovali do jazyka R. Vzhľadom na to, že ekonomické časové rady sú vo väčšine prípadov nestacionárne, analyzovali sme základné vlastnosti ako sú trend a sezónnosť. Grafickým zobrazením sme zistili, že dáta obsahujú trendovú ako aj sezónnu zložku. Na základe týchto informácii sme vybrali najvhodnejšie metódy, ktorými by sa dal tento časový rad najlepšie opísať. Dáta sme transformovali pomocou 1. diferencie na stacionárne, pretože stacionárne dáta boli podmienkou ďalšieho testovania. Na testovanie sme použili vybrané metódy a to sezónny naivný model, exponenciálne vyrovnávanie a ARIMA model. Na základe štandardnej odchýlky a korelogramu sme určili, že ARIMA model je najvhodnejší na prognózovanie nasledujúceho roka. Výsledkom našej práce je prognóza vývoja priemerných denných maloobchodných tržieb na prognózovaný mesiac. Predikciu sme následne porovnali so skutočným stavom za prvé štyri mesiace v roku 2020.

Na záver by som chcel podotknúť, že aj napriek vysokej kvalite, množstva pozorovaní a správneho postupu pri výbere najlepšej metódy, sa prognózované hodnoty v 3. a 4. období významne odlišujú od skutočných hodnôt. Je to dôsledok neočakávanej vírusovej pandémie. Vzhľadom na túto skutočnosť všetky predikcie makroekonomických ukazovateľov, ktoré dodatočne nezohľadnia tento fakt, nebudú relevantné.

ZOZNAM POUŽITEJ LITERATÚRY

[1] LUNGU, Ion, 2012. Database Systems Journal vol. III, no.4/2012. Database Systems Journal. III(4), 3-4

[2] PERRY, J Steven. What is Big Data? More than volume, velocity and variety. In: IBM Developer May 22, 2017. Dostupné na: https://developer.ibm.com/dwblog/2017/what-is-big-data-insight/

[3] MARR, B. Why only one of the 5 Vs of big data really matters. IBM Big Data & Analytics Hub, 2015 Dostupné na: https://www.ibmbigdatahub.com/blog/why-only-one-5-vs-big-data-really-matters

[4] LEVÁRSKY, Stanislav, Big Data sa už dnes dajú využiť v každom segmente, 2014. Dostupné na: https://zive.aktuality.sk/clanok/95598/big-data-sa-uz-dnes-daju-vyuzit-v-kazdom-segmente/

[5] MAYER-SCHÖNBERGER, Viktor a CUKIER Kenneth, Big Data: Revoluce, která zmení způsob, jak žijeme, pracujeme a myslíme, Brno: Computer Press, 2014. ISBN 978-80-251-4119-9.

[6] OUSSOUS, Ahmed, BENJELLOUN, Fatima-Zahra, LAHCEN, Ayoub Ait a BELFKIH, Samir. Big Data technologies: A survey. Journal of King Saud University - Computer and Information Sciences. 2018, vol. 30, no. 4, ISSN 13191578 DOI: 10.1016/j.jksuci.2017.06.001. Dostupné na: http://www.sciencedirect.com/science/article /pii/S1319157817300034

[7] GANDOMI, Amir a HAIDER, Murtaza. Beyond the hype: Big data concepts, methods, and analytics, International Journal of Information Management. 2015, vol. 35, no. 2, ISSN 0268-4012. DOI: 10.1016/j.ijinfomgt.2014.10.007. Dostupné na: http://www.sciencedirect.com/science/article/pii/S0268401214001066

[8] IDC, 2014, Executive Summary: Data Growth, Bussiness Opportunities, and the IT Imperatives. The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things. Dostupné na:

https://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm

[9] AGRAWAL, D, BERNSTEIN, Philip, BERTINO, Elisa,... WIDOM, Jennifer. Challenges and Opportunities with Big Data: A white paper prepared for the Computing Community Consortioum comitee of the Computing Research Association. 2012. Dostupné na: https://cra.org/ccc/wp-content/uploads/sites/2/2015/05/bigdatawhitepaper.pdf

[10] KHAN, Nawsher, ALSAGER, Mohammed, SHAH, Habib, GRAN, Badsha, ABBASI, Aftab a SALEHIAN, Solmaz. The 10Vs, issues and challenges of big data. Proceedings of the 2018 International Conference on Big Data and Education (ICBDE '18). 2018. DOI: 10.1145/3206157.3206166. Dostupné na: https://dl.acm.org/citation.cfm?id=3206166

[11] KHAN, Nawsher, YAGOOB, Ibrar, HASHEM, Ibrahim, INAYAT, Zakira, KAMALELDIN, Waleed, ALAM, Muhammad, SHIRAZ, Muhammad a GANI, Abdullah. Big data: Survey, technologies, opportunities, and challenges. The Scientific World Journal 2014, s. 1-18. DOI: 10.1155/2014/712826. Dostupné na: https://www.researchgate.net/publication/264159615_Big_Data_Survey_Technologies_ Opportunities_and_Challenges

AGGARWAL, A. Managing Big Data Integration in the Public Sector. Piscataway, NJ: IGI Global, 2016.

[12] HYNDMAN, J Rob, ATHANASOPOULOS, George, Forecasting: principles and practice, 2nd edition. OTexts, 2018. Dostupné na: https://otexts.com/fpp2/

[13] MCKINSEY GLOBAL INSTITUTE, Why Big Data is the new competitive advantage. 2012, Dostupné na : https://www.mckinsey.com/media/McKinsey/Business%20Functions/McKinsey%20Digital/Our%20Insights/Big%20data%20The%20next%20frontier%20for%20innovation/MGI_big_data_full_report.ashx

[14] http://www.fhi.sk/files/katedry/kove/veda-vyskum/prace/2006/Lukacik-Pekar2006.pdf

[15] http://fsi.uniza.sk/kkm/files/publikacie/pp/pp_kap_8.pdf

[16] https://medium.com/@peerxp/the-6-stages-of-data-processing-cycle-3c2927c466ff

[17] https://www.oracle.com/big-data/guide/big-data-use-cases.html

[18] https://www.talend.com/resources/what-is-data-processing

[19] https://towardsdatascience.com/5-industries-becoming-defined-by-big-data-and-analytics-e3e8cc0c0cf

[20] https://publications.iadb.org/publications/english/document/Big-Data-in-the-Public-Sector-Selected-Applications-and-Lessons-Learned.pdf

[21] https://touchit.sk/big-data-fenomen-ktoreho-sa-mame-bat-alebo-ho-vyuzit/84736

[22] https://archives.erepublic.com/GT/GT_Mag_Oct_2018.pdf

[23] https://statetechmagazine.com/article/2020/01/how-states-overcome-big-data-analytics-challenges

[24] https://imaginenext.ingrammicro.com/data-center/six-big-data-use-cases-for-the-public-sector