UNIVERZITA PARDUBICE

FAKULTA EKONOMICKO-SPRÁVNÍ

# DIPLOMOVÁ PRÁCE

2025 Bc. Vojtěch Vejlupek

Univerzita Pardubice

Fakulta ekonomicko-správní

# Analýza účinnosti marketingových kampaní na sociálních sítích

Diplomová práce

2025 Bc. Vojtěch Vejlupek

Univerzita Pardubice
Fakulta ekonomicko-správní
Akademický rok: 2024/2025

# ZADÁNÍ DIPLOMOVÉ PRÁCE

(projektu, uměleckého díla, uměleckého výkonu)

| | |
|---|---|
| Jméno a příjmení: | **Bc. Vojtěch Vejlupek** |
| Osobní číslo: | **E23058** |
| Studijní program: | **N0613A140041 Aplikovaná informatika – Data Science pro business** |
| Téma práce: | **Analýza účinnosti marketingových kampaní na sociálních sítích** |
| Zadávající katedra: | **Ústav systémového inženýrství a informatiky** |

## Zásady pro vypracování

Cílem práce je shrnout současné přístupy k hodnocení účinnosti marketingových kampaní na sociálních sítích, popsat přístupy k datové analýze příspěvků týkající se marketingové kampaně, provést sběr a analýzu dat pro vybranou organizaci a na základě výsledků navrhnout doporučení.
Osnova:

- Marketingové kampaně na sociálních sítích
- Analýza textu na sociálních sítích
- Sběr dat a analýza textu
- Návrh doporučení pro marketingovou kampaň

Rozsah pracovní zprávy: **cca 50 stran**
Rozsah grafických prací:
Forma zpracování diplomové práce: **tištěná/elektronická**

Seznam doporučené literatury:

ALVES, Helena; FERNANDES, Cristina; RAPOSO, Mário. Social media marketing: a literature review and implications. *Psychology & Marketing*, 2016, 33.12: 1029-1038.
APPEL, Gil; GREWAL, Lauren; HADI, Rhonda; STEPHEN, Andrew T. The future of social media in marketing. *Journal of the Academy of Marketing Science*, 2020, 48.1: 79-95.
BARKER, Melissa S.; BARKER, Donald; BORMANN, Nicholas F.; ROBERTS, Mary Lou a ZAHAY, Debra L. *Social media marketing: a strategic approach*. Second edition. Boston: Cengage Learning, 2017. ISBN 978-1305502758.
LIU, Bing. *Sentiment analysis: mining opinions, sentiments, and emotions*. New York: Cambridge University Press, 2015. ISBN 978-110-7017-894.
TUTEN, Tracy L. *Social media marketing*. New York: SAGE Publications Ltd, 2023. 978-1529731989.

Vedoucí diplomové práce: **prof. Ing. Petr Hájek, Ph.D.**
Ústav systémového inženýrství a informatiky

Datum zadání diplomové práce: **1. září 2024**
Termín odevzdání diplomové práce: **30. dubna 2025**

L.S.

**prof. Ing. Jan Stejskal, Ph.D.** v.r.
děkan

**prof. Ing. Petr Hájek, Ph.D.** v.r.
garant studijního programu

V Pardubicích dne 1. září 2024

# PROHLÁŠENÍ AUTORA

Prohlašuji:

Práci s názvem Analýza účinnosti marketingových kampaní na sociálních sítích jsem vypracoval samostatně. Veškeré literární prameny a informace, které jsem v práci využil, jsou uvedeny v seznamu použité literatury. Byl jsem seznámen s tím, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon), ve znění pozdějších předpisů, zejména se skutečností, že Univerzita Pardubice má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle § 60 odst. 1 autorského zákona, a s tím, že pokud dojde k užití této práce mnou nebo bude poskytnuta licence o užití jinému subjektu, je Univerzita Pardubice oprávněna ode mne požadovat přiměřený příspěvek na úhradu nákladů, které na vytvoření díla vynaložila, a to podle okolností až do jejich skutečné výše. Beru na vědomí, že v souladu s § 47b zákona č. 111/1998 Sb., o vysokých školách a o změně a doplnění dalších zákonů (zákon o vysokých školách), ve znění pozdějších předpisů, a směrnicí Univerzity Pardubice č. 7/2019 Pravidla pro odevzdávání, zveřejňování a formální úpravu závěrečných prací, ve znění pozdějších dodatků, bude práce zveřejněna prostřednictvím Digitální knihovny Univerzity Pardubice.

V Pardubicích dne 30. 4. 2025

Bc. Vojtěch Vejlupek v.r.

# PODĚKOVÁNÍ

Rád bych poděkoval panu profesorovi Petru Hájkovi za vedení mé diplomové práce, jeho skvělé rady a také za ochotu a vstřícnost. Dále bych tímto chtěl poděkovat své rodině a přátelům za to, že mě podporují při studiu.

## ANOTACE

Tato diplomová práce se zabývá analýzou účinnosti marketingových kampaní na sociálních sítích pro firmu Lyžebraní. Práce shrnuje současné přístupy k hodnocení účinnosti online marketingových kampaní a popisuje využití procesního rámce CRISP-DM při datové analýze marketingových dat. V rámci tohoto přístupu byly využity analytické metody jako je korelační analýza, analýza hlavních komponent, shluková analýza, vícenásobná lineární regrese, predikce sezónnosti s využitím modelu Prophet a analýza sentimentu textových komentářů. Výstupem práce jsou doporučení pro zlepšení budoucích marketingových aktivit, která vycházejí ze získaných poznatků z provedených analýz.



## KLÍČOVÁ SLOVA

Digitální marketing, sociální sítě, analýza hlavních komponent, shluková analýza, vícenásobná lineární regrese, predikce sezónnosti, analýza sentimentu


## TITLE

Analysis of the effectiveness of social media marketing campaigns


## ANNOTATION

This thesis focuses on evaluating the effectiveness of social media marketing campaigns for the company Lyžebraní. The thesis summarizes current approaches to assessing the effectiveness of online marketing campaigns and describes the use of the CRISP-DM process framework in the data analysis of marketing data. Within this approach, analytical methods such as correlation analysis, principal component analysis, cluster analysis, multiple linear regression, seasonality prediction using the Prophet model, and sentiment analysis of text comments were utilized. The output of the thesis includes recommendations for improving future marketing activities based on the insights gained from the analyses.



## KEYWORDS

Digital marketing, social media, principal component analysis, cluster analysis, multiple linear regression, seasonality prediction, sentiment analysis

# OBSAH

# SEZNAM OBRÁZKŮ

# SEZNAM ZKRATEK A SYMBOLŮ

| | |
|---|---|
| CRISP-DM | Standardní metodika datové analýzy napříč odvětvími |
| CSV | Formát textového souboru |
| CTA | Výzva k akci |
| CTR | Míra prokliku |
| FAQ | Často kladené dotazy |
| ID | Identifikační číslo |
| KPI | Klíčový ukazatel výkonnosti |
| LDA | Metoda tematického modelování textu |
| MAE | Průměrná absolutní chyba |
| MAPE | Průměrná procentuální chyba |
| MSE | Střední kvadratická chyba |
| OLS | Metoda nejmenších čtverců |
| PC1 – PC5 | První až pátá hlavní komponenta |
| PCA | Analýza hlavních komponent |
| $R^2$ | Koeficient determinace |
| ROI | Návratnost investice |
| SSE | Součet čtvercových chyb |
| WCCS | Součet čtverců vzdáleností v rámci shluku |

# ÚVOD

Sociální sítě se v posledních letech staly nedílnou součástí komunikačních a marketingových strategií moderních firem. Platformy jako Facebook a Instagram umožňují nejen cílené oslovování potenciálních zákazníků, ale také měření efektivity jednotlivých marketingových aktivit v reálném čase. Právě schopnost vyhodnotit přínos kampaní na sociálních sítích a jejich vztah k obchodním výsledkům představuje pro firmy důležitý krok směrem k efektivnějšímu rozhodování a optimalizaci nákladů (Barker a kol. 2013).

V návaznosti na rozšiřující se digitální prostředí se hodnocení výkonnosti marketingových aktivit stává zásadním nástrojem pro optimalizaci online strategií. Zvláště v prostředí sociálních sítí, které se vyznačuje rychlou zpětnou vazbou a vysokou mírou interakce, je sledování vhodných metrik klíčem k porozumění tomu, jak jednotlivé kampaně ovlivňují chování spotřebitelů. Pravidelné vyhodnocování klíčových ukazatelů výkonnosti (KPI) přitom napomáhá nejen sledování úspěšnosti kampaní, ale je také předpokladem pro efektivní řízení návratnosti investic (ROI) (Gołąb-Andrzejak, 2023).

Tato diplomová práce se zaměřuje na analýzu účinnosti marketingových kampaní na sociálních sítích Facebook a Instagram, konkrétně v kontextu prodejní akce Lyžebraní.

Cílem této práce je shrnout současné přístupy k hodnocení účinnosti marketingových kampaní na sociálních sítích, aplikovat metodiky datové analýzy na reálná data firmy Lyžebraní a na základě výsledků formulovat doporučení pro optimalizaci budoucích kampaní.

V rámci obsahu práce se první kapitola věnuje teoretickému vymezení sociálních sítí a popisu vybraných platforem Facebook a Instagram. Druhá kapitola je zaměřená na marketingové kampaně, jejich význam a způsoby hodnocení. Ve třetí kapitole je představena analyzovaná firma Lyžebraní. Čtvrtou kapitolu tvoří metodika CRISP-DM, včetně popisu a použití vybraných metod datové analytiky jako jsou: korelační analýza, analýza hlavních komponent, shluková analýza, vícenásobná lineární regrese, analýza sezónnosti a analýza sentimentu. Práce je zakončena shrnutím hlavních zjištění a návrhem doporučení pro optimalizaci budoucích marketingových kampaní.

# 1 SOCIÁLNÍ SÍTĚ

## 1.1 Definice sociálních sítí

Sociální sítě lze z praktického hlediska chápat jako soubor digitálních technologií, které jsou pro uživatele dostupné prostřednictvím aplikací a webových platforem (Appel a kol., 2020). Tyto nástroje jim umožňují fungovat v online prostředích, kde dochází k výměně digitálního obsahu a informací mezi propojenými uživateli. Typickými představiteli těchto platforem jsou například Facebook, Instagram nebo Twitter (od roku 2023 pod novým názvem X), které kombinují technologické možnosti s nástroji pro sociální interakci.

V odborné literatuře se často objevuje i pojem sociální média, pod který bývají sociální sítě zařazovány jako jedna z podskupin. Sociální média lze obecně charakterizovat jako online platformy umožňující uživatelům tvořit, sdílet a vyměňovat si obsah. Podle Karlíčka (2016) mezi tato média patří sociální sítě, blogy, fóra a další komunity. Karlíček zároveň upozorňuje, že hranice mezi jednotlivými formami sociálních médií nejsou vždy zcela jasné, protože se svými funkcemi často překrývají, a právě to přispívá k terminologické nejednotnosti v této oblasti.

Jelikož se digitální prostředí neustále vyvíjí, přibývají nové platformy a mění se způsoby užívání i technologie, skrze které uživatelé vstupují do online světa. Množství různých webových služeb, funkcí a interakcí, které dnes spadají pod pojem sociální sítě je natolik široké, že nelze tento pojem zcela přesně vymezit (Tuten, 2021).

Tuten (2021) upozorňuje, že sociální sítě nejsou pouze technickým nástrojem, ale především platformou umožňující obousměrnou komunikaci, sdílení obsahu a zapojení uživatelů. Sociální sítě podle něj umožňují firmám nejen oslovit publikum, ale také budovat s uživateli vztahy prostřednictvím aktivního zapojení, zpětné vazby a komunitní interakce. Toto zapojení uživatelů odlišuje sociální sítě od tradičních forem médií a činí z nich důležitý nástroj pro moderní marketingové strategie.

Sociální média v sobě spojují různé výhody, které z nich činí atraktivní nástroj zejména pro komerční subjekty. Umožňují podnikům snadno a nízkonákladově navazovat a rozvíjet vztahy se zákazníky, přičemž jejich dostupnost a rychlost šíření informací zvyšují efektivitu komunikace. Vzhledem k tomu, že sociální sítě dnes ovlivňují téměř všechny oblasti života, stávají se nedílnou součástí firemních strategií a marketingových aktivit (Brambilla a kol., 2022).

## 1.2 Vybrané sociální sítě

V rámci této diplomové práce byla pozornost zaměřena výhradně na platformy Facebook a Instagram. Tyto dvě sociální sítě byly zvoleny záměrně, neboť představují hlavní komunikační kanály v rámci on-line marketingové strategie pro firmu Lyžebraní (o firmě více v kapitole 3). Lyžebraní využívá tyto platformy nejen pro propagaci produktů a realizaci reklamních kampaní, ale také pro informování zákazníků o všech aktuálních novinkách, slevách, otevírací době a konkrétních prodejních dnech.

### 1.2.1 Facebook

Facebook, spuštěný v roce 2004, patří mezi nejznámější a nejdéle fungující sociální sítě. Původně vznikl za účelem propojení studentů univerzit, avšak postupem času se vyvinul v globální platformu umožňující komunikaci, sdílení a vytváření online komunit. Jeho hlavním cílem bylo poskytnout uživatelům nástroj pro budování vztahů a spojení s ostatními lidmi napříč světem (Appel a kol., 2020).

Facebook se v průběhu let rozrostl ze sítě pro budování sociálních vazeb do komplexní platformy s rozšířenými možnostmi digitální komunikace a marketingu. Díky nástrojům jako Facebook Live, Marketplace a reklamním kampaním představuje důležitý komunikační kanál, který zasahuje i do oblastí jako e-commerce a mediální distribuce (Tuten, 2021). Jeho vlastnictví dalších služeb jako WhatsApp, Messenger nebo Instagram zároveň zajišťuje silné postavení v ekosystému digitálního marketingu.

Z marketingového pohledu je Facebook vnímán jako jedna z nejdůležitějších platforem pro interakci značek s uživateli na sociálních sítích. Podle Brambilla a kol. (2022) značky využívají Facebook jako prostředek k propojení s vysoce zapojenými uživateli, s cílem vytvořit vztahy a posílit loajalitu ke značce. Díky možnostem jako jsou sdílení obsahu, komentáře a další interakce může Facebook pozitivně ovlivnit celý proces nákupního rozhodování, od počátečního získání informací až po fázi po nákupu, kdy uživatelé sdílí své zkušenosti a hodnocení.

Stejného názoru je i Alves a kol. (2016), jež poukazují na to, že právě Facebook spolu s Twitterem (od roku 2023 pod novým názvem X) patří mezi sociální platformy, které podniky nejčastěji využívají ve své marketingové praxi. Tyto sítě dosahují nejlepších výsledků, pokud jde o formování postojů spotřebitelů ke značce. Vzhledem k tomu, že umožňují obousměrnou komunikaci, vysokou míru uživatelské angažovanosti a snadné šíření obsahu, by firmy měly svou přítomnost na těchto kanálech považovat za klíčovou součást strategického řízení značky.

### 1.2.2 Instagram

Instagram je sociální síť zaměřená na vizuální obsah, která byla spuštěna v roce 2010, jako mobilní aplikace pro sdílení fotografií. Tato aplikace umožňuje uživatelům upravovat a publikovat fotografie či krátká videa a doplňovat je popisky, hashtagy nebo údaji o poloze. Díky těmto funkcím je obsah na platformě snadno dohledatelný a může se šířit i mimo okruh sledujících (Holak a McLaughlin, 2024).

Tato vizuálně zaměřená sociální síť si získala popularitu především díky jednoduchému rozhraní a možnosti snadno upravovat a sdílet obrazový obsah. Právě toto zaměření ji odlišuje od jiných platforem a zároveň z ní dělá účinný nástroj i pro firemní využití. Firmy používají Instagram nejen k propagaci produktů, ale také k navazování a udržování vztahů se zákazníky (Yang, 2021).

Z hlediska celosvětového dosahu patří Instagram mezi čtyři nejvýznamnější sociální sítě na světě a v roce 2022 evidoval více než 1,4 miliardy uživatelů. Platforma podporuje přes 25 milionů obchodních účtů a neustále rozšiřuje své funkce, aby umožnila firmám efektivnější cílení a propagaci nabízených produktů. Kromě původních funkcí, jako jsou příběhy, karusely, živá vysílání a fotografie, Instagram dnes umožňuje přidávat do příspěvků interaktivní odkazy a nákupní prvky, které usnadňují přímý prodej z platformy (Brambilla a kol., 2022).

# 2 MARKETINGOVÉ KAMPANĚ NA SOCIÁLNÍCH SÍTÍCH

## 2.1 Definice marketingové kampaně

Marketingovou kampaň lze chápat jako soubor vzájemně propojených marketingových aktivit, které firmě pomáhají naplňovat její komunikační a obchodní cíle (Valdani a Arbore, 2015). Tyto cíle mohou zahrnovat například propagaci nového produktu, zvýšení povědomí o značce nebo sběr obsahu vytvářeného samotnými uživateli, jenž následně přispívá k posílení důvěryhodnosti značky. Kampaň obvykle využívá kombinaci různých kanálů a formátů, jako jsou sociální média, e-mailový marketing, videoobsah, tištěná inzerce nebo online reklama ve vyhledávačích (Shopify, 2024).

Marketingová kampaň představuje cílenou, časově ohraničenou a strategicky řízenou aktivitu, jejímž cílem je prostřednictvím zvolených komunikačních kanálů ovlivnit vnímání značky, motivovat cílové publikum k určité akci a v konečném důsledku podpořit obchodní výsledky organizace (Tuten, 2021). Tento přístup podtrhuje skutečnost, že moderní marketing již dávno nepředstavuje pouze nástroj prodeje, ale především prostředek k budování dlouhodobého vztahu se zákazníkem, a to včetně prvků jako je zpětná vazba, emoce či aktivní zapojení uživatelů do tvorby obsahu.

## 2.2 Význam marketingové kampaně na sociálních sítích

Marketingové kampaně realizované prostřednictvím sociálních sítí hrají zásadní roli při ovlivňování zákaznického chování v jednotlivých fázích nákupního rozhodovacího procesu. Díky možnosti publikovat cílený obsah na různých platformách mohou značky efektivně budovat povědomí, formovat pozitivní vnímání, posilovat důvěru, vyvolávat zájem o produkt a vést zákazníka k samotné koupi. Sociální sítě tímto způsobem nepředstavují pouze prostor pro šíření sdělení, ale zejména prostředek pro aktivní formování postojů a rozhodnutí spotřebitelů (Tuten, 2021).

Význam marketingových kampaní na sociálních sítích však přesahuje samotnou podporu prodeje. Výzkumy ukazují, že jejich dopad se promítá i do celkové tržní hodnoty firem (Alves a kol., 2016). Klíčem k úspěchu je schopnost značky vyvolat silné emoce, a to zejména prostřednictvím interaktivního, vizuálně přitažlivého a strategicky umístěného obsahu. Pouhá přítomnost značky na sociálních platformách přitom nestačí, jelikož zásadní roli hraje aktivní zapojení uživatelů do kampaní a komunitních aktivit.

Z pohledu strategického řízení značky představují kampaně na sociálních sítích výkonný nástroj, který firmám umožňuje pružně reagovat na aktuální trendy i chování spotřebitelů. Platformy jako Instagram a Facebook jsou dnes využívány nejen pro propagaci produktů a budování značky, ale také jako prostor pro navazování a prohlubování vztahů se zákazníky. Kampaně zde přispívají k efektivní akvizici cílových skupin, zvyšování dosahu sdělení a celkové účinnosti komunikační strategie (Mou, 2020).

## 2.3 Hodnocení marketingových kampaní na sociálních sítí

Marketingová kampaň na sociálních sítích je efektivní pouze tehdy, pokud lze její účinnost objektivně změřit. Vzhledem ke specifickému charakteru těchto platforem však nelze hodnocení kampaní omezit pouze na jednoduché kvantitativní metriky, jako je počet zobrazení či kliknutí. Skutečná efektivita vychází z širšího rámce, který zahrnuje vztah mezi kampaní a chováním uživatelů, úrovní jejich zapojení a dosažením stanovených cílů (Tuten, 2021).

Jedním z nejčastěji používaných přístupů je sledování tzv. engagement metrik, které spadají pod širší skupinu tzv. KPI (Key Performance Indicators – klíčových ukazatelů výkonnosti). Mezi nejběžnější KPI v oblasti sociálních médií patří počet zobrazení (reach), míra prokliku (CTR), počet nových sledujících, sdílení, komentáře a reakce. Tyto ukazatele umožňují zhodnotit úroveň interakce uživatelů s obsahem kampaně a poskytují marketingovým týmům zpětnou vazbu o efektivitě obsahu. Nicméně, jak upozorňuje Tuten (2021), samotné výše uvedené kvantitativní metriky nemusí plně vystihnout skutečný dopad kampaně, jelikož je důležité brát v potaz i kvalitativní faktory, jako je sentiment uživatelských komentářů nebo hloubka zapojení cílové skupiny.

Dalším běžně využívaným přístupem k hodnocení efektivity kampaní na sociálních sítích je měření návratnosti investic, označované jako ROI (Return on Investment). V tomto kontextu představuje ROI nástroj, pomocí kterého firmy posuzují, zda se prostředky investované do marketingových aktivit skutečně promítly do dosažení obchodních cílů. Zásadní roli v procesu měření ROI sehrává propojení dat ze sociálních sítí s reálnými obchodními výsledky. To obvykle zahrnuje integraci výstupů z analytických nástrojů, jako je například Google Analytics, s daty z e-commerce systému či CRM platformy. Pouze tak je možné přesně určit, jaký podíl na dosažených tržbách lze přiřadit konkrétní kampani na sociálních médiích (O'Brien, 2022).

## 2.4 Metody datové analýzy používané v oblasti marketingových kampaní

Datová analýza hraje v marketingových kampaních klíčovou roli, protože umožňuje systematicky vyhodnocovat efektivitu jednotlivých aktivit, porozumět chování zákazníků a optimalizovat marketingové strategie. V praxi se využívá široká škála analytických metod, které pomáhají transformovat nasbíraná data do podoby užitečných informací (Tuten, 2021).

### 2.4.1 Korelační analýza

Korelační analýza je základní statistická metoda sloužící ke zkoumání vztahů mezi proměnnými. V marketingu se běžně využívá ke zjištění, jak silně spolu souvisí různé metriky, například počet zobrazení příspěvku a počet prokliků na webové stránky. Identifikace těchto vztahů napomáhá lépe pochopit, které faktory nejvíce ovlivňují úspěšnost marketingových aktivit (Berman, 2016).

### 2.4.2 Analýza hlavních komponent (PCA)

Analýza hlavních komponent (PCA) slouží ke snížení rozměrnosti dat při zachování co největší části variability. V oblasti marketingu se tato metoda využívá například při zpracování komplexních dat o výkonu příspěvků, kdy umožňuje zjednodušit analýzu tím, že shrne původní velké množství proměnných do menšího počtu hlavních faktorů (Abdi a Williams, 2010). Díky tomu lze efektivněji identifikovat klíčové charakteristiky ovlivňující výsledky kampaní.

### 2.4.3 Shluková analýza

Shluková analýza umožňuje seskupování objektů do skupin (shluků) na základě jejich podobnosti. V marketingové praxi se využívá například ke klasifikaci zákazníků, příspěvků nebo kampaní podle společných rysů (Syakur a kol., 2018). Metody jako K-means nebo DBSCAN tak pomáhají lépe cílit marketingovou komunikaci na jednotlivé segmenty publika.

### 2.4.4 Regresní analýza

Regresní analýza slouží ke zkoumání a modelování vztahů mezi závislou proměnnou a jednou či více nezávislými proměnnými. V marketingu se běžně využívá k predikci výsledků kampaní, například odhadu počtu konverzí na základě investovaného rozpočtu, typu obsahu či cílení reklamy. Výsledky regresní analýzy umožňují kvantifikovat vliv jednotlivých faktorů na úspěšnost kampaně (Lin a kol., 2024).

### 2.4.5 Analýza sentimentu

Analýza sentimentu se zaměřuje na hodnocení emocí a názorů vyjádřených v textových datech, jako jsou komentáře na sociálních sítích. V oblasti marketingu se používá k hodnocení vnímání značky zákazníky a k analýze zpětné vazby na kampaně (Liu, 2012). Vyhodnocení pozitivního, neutrálního nebo negativního sentimentu pomáhá firmám lépe reagovat na potřeby a očekávání cílové skupiny.

### 2.4.6 Analýza časových řad

Predikční modely sezónnosti slouží k odhalování opakujících se vzorců v datech v čase. Model Prophet, vyvinutý společností Facebook, umožňuje efektivně modelovat sezónní trendy a predikovat budoucí vývoj výkonu kampaní. V marketingové praxi se predikce sezónnosti využívá zejména při plánování kampaní v období zvýšené poptávky, například během vánočních svátků nebo sezónních výprodejů (Shakeel a kol. 2023).

# 3 LYŽEBRANÍ

Lyžebraní je specializovaná prodejní akce lyžařského vybavení v České republice, která se každoročně koná v obci Hlavenec ve Středočeském kraji. Tato akce je určena pro široké spektrum zimních sportovců, zejména tedy pro: lyžaře, snowboardisty i běžkaře, bez ohledu na věk či úroveň zkušeností.

Akce probíhá během několika víkendových termínů v období od října do března. Zákazníci mají možnost vybírat z rozsáhlé nabídky více než 70 000 kusů nového i bazarového vybavení, včetně lyží, snowboardů, běžek, lyžařských bot, helem, brýlí, zimního oblečení a dalšího příslušenství. Prodejní plocha o rozloze přes 1 200 m² umožňuje pohodlný výběr a vyzkoušení zboží přímo na místě.

Lyžebraní se vyznačuje širokým sortimentem, cenovou dostupností a odborným poradenstvím. Zkušený personál poskytuje individuální konzultace a pomáhá s výběrem vybavení podle potřeb zákazníků. Důležitou součástí akce je online rezervace nákupního termínu, která zajišťuje plynulý průběh nákupů a minimalizuje čekací doby.

Cílem Lyžebraní je zpřístupnit kvalitní lyžařské vybavení široké veřejnosti za přijatelné ceny a umožnit tak co největšímu počtu rodin i jednotlivců aktivní trávení zimního období na horách. Sortiment zahrnuje produkty pro začínající i zkušené sportovce všech věkových kategorií, přičemž důraz je kladen na kvalitu, funkčnost a finanční dostupnost nabízeného zboží.

## 3.1 Marketingové kampaně firmy Lyžebraní

V oblasti online marketingové komunikace se Lyžebraní opírá o využívání sociálních sítí, zejména platforem Facebook a Instagram. Tyto kanály hrají klíčovou roli při oslovování cílové skupiny, informování o termínech konání akce a budování vztahu se zákazníky. Obsah zde publikovaný je vizuálně přitažlivý, aktuální a zároveň funkčně zaměřený. Konkrétně se jedná o příspěvky upozorňující na slevy, novinky v sortimentu nebo praktické informace k rezervacím.

Firma klade důraz na to, aby komunikace byla jednoduchá, srozumitelná a reflektovala sezónní charakter celého projektu. Významnou roli hraje také rychlá zpětná vazba, jelikož uživatelé mají možnost pokládat dotazy, komentovat příspěvky či sdílet vlastní zkušenosti. Tato interaktivita přispívá nejen k většímu zapojení komunity, ale také k posílení důvěry zákazníků.

Přestože firma aktivně komunikuje jak prostřednictvím Facebooku, tak Instagramu, počet sledujících na jednotlivých platformách se výrazně liší. Facebookový profil má více než 11 tisíc sledujících, zatímco na Instagramu je to přibližně 2 800 uživatelů. Tento poměr do jisté míry odráží složení cílové skupiny, kterou tvoří převážně rodiče nebo lyžaři, kteří hledají vybavení pro sebe i své blízké. Právě Facebook poskytuje pro firmu vhodné prostředí pro komunikaci informací o termínech, výhodách rezervace nebo konkrétním sortimentu.

Analýza obsahu publikovaného na sociálních sítích v období od 1. října 2023 do 16. března 2024 ukazuje, že Lyžebraní využívá na obou platformách různé formáty příspěvků přizpůsobené charakteru daného kanálu. Na Facebooku tvoří základ komunikace textové a obrazové příspěvky, které jsou doplňovány krátkými videi, příspěvky typu reels a odkazy na externí webové stránky. Tento typ obsahu kombinuje textové informace s vizuální složkou a často obsahuje výzvy k akci, například k rezervaci termínu nebo návštěvě webových stránek.

Na Instagramu je komunikace založena především na příbězích (stories) a klasických obrazových příspěvcích. Stories slouží k rychlému sdílení aktuálních informací a navození atmosféry spojené s konáním akce. Obrazové příspěvky v hlavním feedu pak doplňují komunikaci o vizuálně atraktivní prezentaci nabídky a prostředí Lyžebraní. Publikovaný obsah se zaměřuje na udržování pravidelného kontaktu se sledujícími a na podporu povědomí o značce.

# 4 DATOVÁ ANALÝZA MARKETINGOVÉ KAMPANĚ POMOCÍ METODIKY CRISP-DM

CRISP-DM (Cross-Industry Standard Process for Data Mining) představuje robustní a systematicky strukturovaný rámec pro realizaci projektů v oblasti data miningu. Skládá se ze šesti hlavních fází, které poskytují jasně definovaný postup od identifikace problému až po praktické nasazení výsledků (Chumbar, 2023). K hlavním výhodám metodiky CRISP-DM patří její oborová nezávislost, strukturovanost a iterativní charakter, který umožňuje pružně reagovat na zjištěné nedostatky v průběhu projektu. Možnost návratu do předchozích fází podporuje neustálé zlepšování kvality řešení a celkové zvyšování efektivity datově orientovaných projektů. Klíčové fáze CRISP-DM jsou následující (Chumbar, 2023):

První fází je porozumění podnikání (Business Understanding), které se zaměřuje na identifikaci cílů projektu z obchodního hlediska. V rámci této fáze dochází k formulaci zadání, analýze současného stavu a definování cílů data miningu spolu s návrhem projektového plánu.

Druhá fáze je porozumění datům (Data Understanding), během něhož jsou shromažďována data, prováděna jejich deskriptivní analýza, vizualizace a kontrola kvality. Tato fáze slouží jako základ pro posouzení vhodnosti dat pro další zpracování.

Ve třetí fázi s názvem příprava dat (Data Preparation) dochází k čištění, transformaci a formátování dat tak, aby byla připravena pro následné modelování. Součástí této fáze je i konstrukce nových atributů, výběr relevantních proměnných a integrace dat z různých zdrojů.

Čtvrtou fází je modelování (Modeling), kde jsou aplikovány vybrané algoritmy či statistické metody. Tato fáze často vyžaduje zpětnou vazbu a úpravy v přípravě dat, neboť různé modelovací techniky mají specifické požadavky na vstupy.

Pátou část tvoří vyhodnocení (Evaluation), které pak slouží k posouzení výkonnosti vytvořených modelů ve vztahu k původním obchodním cílům. Výsledky jsou analyzovány a ověřovány, přičemž se posuzuje připravenost řešení k nasazení.

Závěrečnou šestou fází je nasazení (Deployment), která představuje přechod od analytického výstupu k jeho praktickému využití. Tato etapa může mít podobu generování reportů, vytvoření automatizovaného systému nebo integrace modelu do provozního prostředí.

## 4.1 Porozumění podnikání

Podle Siobos (2024) je porozumění podnikání úvodní fází metodiky CRISP-DM, ve kterém se zaměřujeme na pochopení podnikatelského kontextu a převedení problému do podoby, která je srozumitelná i z hlediska manažerského rozhodování.

V rámci porozumění podnikatelského kontextu je cílem této fáze ověřit, zda mají marketingové kampaně publikované na sociálních sítích Facebook a Instagram měřitelný vliv na tržby společnosti Lyžebraní. Analýza účinnosti těchto kampaní je pro firmu zásadní, protože prostřednictvím příspěvků na sociálních sítích usiluje o zvýšení povědomí o značce a motivaci potenciálních zákazníků k rezervaci termínu nákupu. Zjištění vztahu mezi kampaněmi a následnou tržbou umožní lépe pochopit, jaký obsah a formáty komunikace jsou z hlediska obchodních cílů nejúčinnější.

## 4.2 Porozumění datům

Siobos (2024) uvádí, že porozumění datům tvoří druhou fází metodiky CRISP-DM, která následuje po vyjasnění podnikatelských cílů. V této fázi je potřeba se zaměřit na sběr dat, jejich kontrolu, prozkoumání a vytvoření základního přehledu o jejich kvalitě, rozsahu a potenciálu pro další analýzu. Součástí této fáze je rovněž identifikace problémů v datech, jako jsou chybějící hodnoty, nevhodné datové typy, duplicity nebo extrémní odlehlé hodnoty. Cílem je vytvořit si dostatečně důkladný přehled o tom, s jakými daty se pracuje.

### 4.2.1 Sběr marketingových dat z Meta Ads Manager

Pro primární sběr dat byla využita platforma Meta Ads Manager, která poskytuje podrobné informace o kampaních a příspěvcích na sociálních sítích Facebook a Instagram. V rámci tohoto rozhraní bylo nejprve nezbytné zvolit vhodné časové období pro analýzu marketingových kampaní. Jako referenční rámec posloužily oficiální termíny prodejní sezóny firmy Lyžebraní v období 2023/24, konkrétně od 19. října 2023 do 16. března 2024. S ohledem na skutečnost, že marketingové aktivity zpravidla začínají ještě před samotným spuštěním sezóny, bylo jako počáteční datum sběru zvoleno 1. října 2023.

Rozhraní Meta Ads Manager však umožňuje export dat pouze za maximálně tříměsíční období (obrázek 1), což vedlo k potřebě rozdělit analyzované období do dvou částí: říjen–prosinec a leden–březen. Pro každý typ obsahu bylo nutné provést samostatný export, což vedlo ke vzniku šesti výstupních souborů: příspěvky na Facebooku, příspěvky na Instagramu a Instagram Stories, vždy ve dvou časových obdobích. Tyto soubory byly následně spojeny a utříděny do tří finálních datových sad podle typu obsahu.


**Obrázek 1:** Export dat z Meta Ads Manageru

Zdroj: Vlastní zpracování

### 4.2.2 Sběr komentářů

Kromě výše uvedených dat o marketingových příspěvcích byly z Meta Ads Manageru exportovány také komentáře, které uživatelé zanechali pod jednotlivými příspěvky. Komentáře byly získány pouze z platformy Facebook, protože na Instagramu nebyly reakce v podobě komentářů dostatečně četné ani relevantní pro plánovanou analýzu sentimentu.

### 4.2.3 Sběr dat o prodejních dnech

Pro potřeby propojení marketingových dat s reálnými výsledky bylo nutné zajistit také data týkající se prodejních dnů. Tato data poskytla firma Lyžebraní s podmínkou anonymizace reálných čísel. Původní hodnoty tržeb byly upraveny tak, aby nebylo možné zpětně určit jejich skutečnou výši, a zároveň zachovávaly relativní rozdíly mezi jednotlivými dny i trendy v čase. Datová sada obsahuje následující atributy: název dne, datum, den v týdnu, počet registrovaných zákazníků a tržbu. Díky těmto údajům bylo možné vytvořit komplexní obraz o souvislosti mezi marketingovými aktivitami a dosaženými obchodními výsledky.

### 4.2.4 Použité nástroje pro práci s daty

Pro zpracování a kontrolu dat byly využity dvě hlavní platformy: Microsoft Excel a webové prostředí Kaggle, které umožňuje práci s daty pomocí programovacího jazyka Python. Excel sloužil především k úvodnímu spojení a filtrování datových souborů, zatímco Kaggle poskytl prostředí pro automatizované skripty a přehlednou vizualizaci klíčových vlastností dat.

### 4.2.5 Načtení dat

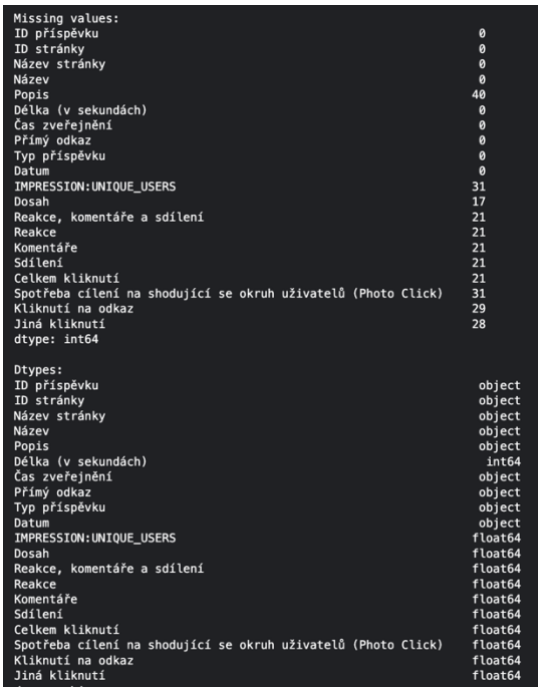
V prostředí Kaggle byly importovány datové sady s informacemi o marketingových kampaních. Na obrázku 2 je zobrazen náhled prvních řádků datového souboru, sloužící k ověření úspěšnosti importu. Zdrojový kód pro načtení dat je uveden v příloze A.

| | ID příspěvku | ID účtu | Uživatelské jméno účtu | Název účtu | Popis | Délka (v sekundách) | Čas zveřejnění | Přímý odkaz | Typ příspěvku |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 1,79222E+16 | 1,78414E+16 | lyzebrani | Lyžebraní | Děkujeme @running_princess_ | 15 | 12/29/2023 06:28 | NaN | Instagram Story |
| 1 | 1,80195E+16 | 1,78414E+16 | lyzebrani | Lyžebraní | Je to tak! přípravy jsou v plném proudu a regi... | 5 | 10/04/2023 12:31 | NaN | Instagram Story |
| 2 | 1,79851E+16 | 1,78414E+16 | lyzebrani | Lyžebraní | Skol! Přišlo nám trapné vás informovat o lyžíc... | 0 | 10/09/2023 09:41 | NaN | Instagram Story |
| 3 | 1,80289E+16 | 1,78414E+16 | lyzebrani | Lyžebraní | Ztráty a nálezy:\n\nNa posledním Lyžebraní jsm... | 0 | 12/22/2023 05:50 | NaN | Instagram Story |
| 4 | 1,79899E+16 | 1,78414E+16 | lyzebrani | Lyžebraní | Děkujeme @mikolasbilek | 15 | 12/29/2023 06:16 | NaN | Instagram Story |

**Obrázek 2:** Náhled datového souboru Instagram Stories

Zdroj: Vlastní zpracování

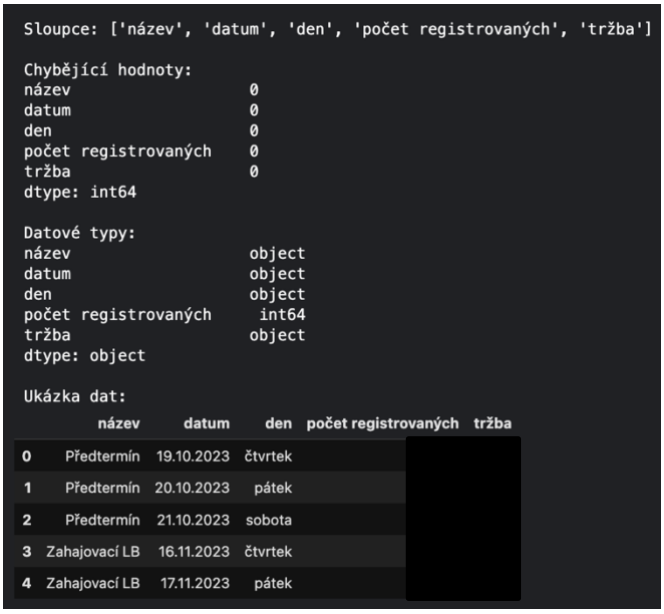
Dále byl na obrázku 3 zobrazen výčet prázdných buněk a seznam datových typů, pro jednotlivé datové sady. Toto zobrazení totiž pomohlo identifikovat možné problémy, které by mohly ovlivnit výsledky analýzy, jako jsou například chybějící hodnoty v důležitých sloupcích nebo nesprávně rozpoznané datové typy (např. čísla jako text, datum jako řetězec apod.).




**Obrázek 3:** Datová sada Facebook příspěvky s nulovými hodnotami a datovými typy

Zdroj: Vlastní zpracování

Stejný postup porozumění dat podstoupil i soubor se statistikami z jednotlivých prodejních dnů, zobrazen níže na obrázku 4.




**Obrázek 4:** Datová sada prodejní dny s nulovými hodnotami a datovými typy

Zdroj: Vlastní zpracování

## 4.3 Příprava dat

Podle Siobos (2024) je fáze přípravy dat třetím krokem metodiky CRISP-DM, který navazuje na porozumění dat. V této fázi se zaměřujeme na úpravu dat do takové podoby, která umožní efektivní použití analytických metod. To zahrnuje výběr relevantních proměnných, odstranění chyb a nekonzistencí, transformaci datových typů, sjednocení informací z více zdrojů a případně tvorbu nových atributů. Cílem této fáze je zajistit, aby vstupní data byla kvalitní, jednotná a připravená na následné modelování, protože právě kvalita vstupních dat výrazně ovlivňuje výsledky celého analytického procesu.

Prvním krokem fáze přípravy dat bylo nahrazení chybějících hodnot ve všech datových souborech obsahujících marketingové příspěvky. Všechny prázdné buňky byly nahrazeny hodnotou 0, a to s využitím skriptu v jazyce Python. Tento krok sloužil ke sjednocení dat a zajištění jejich kompatibility s dalšími analytickými nástroji.

Druhým krokem bylo propojení datových souborů obsahujících marketingové příspěvky s daty o tržbách. Vzhledem k tomu, že bylo potřeba přiřadit každému příspěvku nejbližší následující prodejní den, bylo toto spojení provedeno manuálně v prostředí Microsoft Excel. Tento způsob byl zvolen z důvodu zachování přesnosti při párování dat. Zároveň byl Microsoft Excel využit také pro převod hodnot tržeb z procentního vyjádření na desetinný formát, což usnadnilo následné zpracování v analytických nástrojích. Veškeré provedené úpravy byly následně uloženy do nově pojmenovaných souborů ve formátu CSV.

Po následném načtení nově vytvořených CSV souborů zpět do prostředí Kaggle bylo zapotřebí provést kontrolu, zda byly všechny ručně provedené úpravy úspěšně zachovány. V rámci této kontroly byla ověřena správnost struktury dat, přítomnost nově připojených sloupců, stejně jako správné datové typy jednotlivých sloupců.

V rámci přípravy dat byly komentáře získané z platformy Facebook ručně přeloženy do anglického jazyka, neboť nástroje pro analýzu sentimentu dosahují lepších výsledků při práci s anglickým textem (Kincl a kol., 2019).

## 4.4 Modelování

Modelování je čtvrtou fází metodiky CRISP-DM, v jejímž rámci jsou na připravená data aplikovány vhodně zvolené algoritmy s cílem identifikovat vzory, struktury nebo predikovat chování cílové proměnné (Plotnikova a kol., 2022). Tato fáze zahrnuje i výběr modelovacích technik, nastavení parametrů a jejich vyhodnocení na základě zvolených metrik. Jak upozorňuje Siobos (2024), během tohoto procesu může vzniknout potřeba vrátit se zpět k transformaci dat, pokud se ukáže, že aktuální podoba vstupních proměnných neodpovídá požadavkům vybraného modelu.

### 4.4.1 Analýza korelace

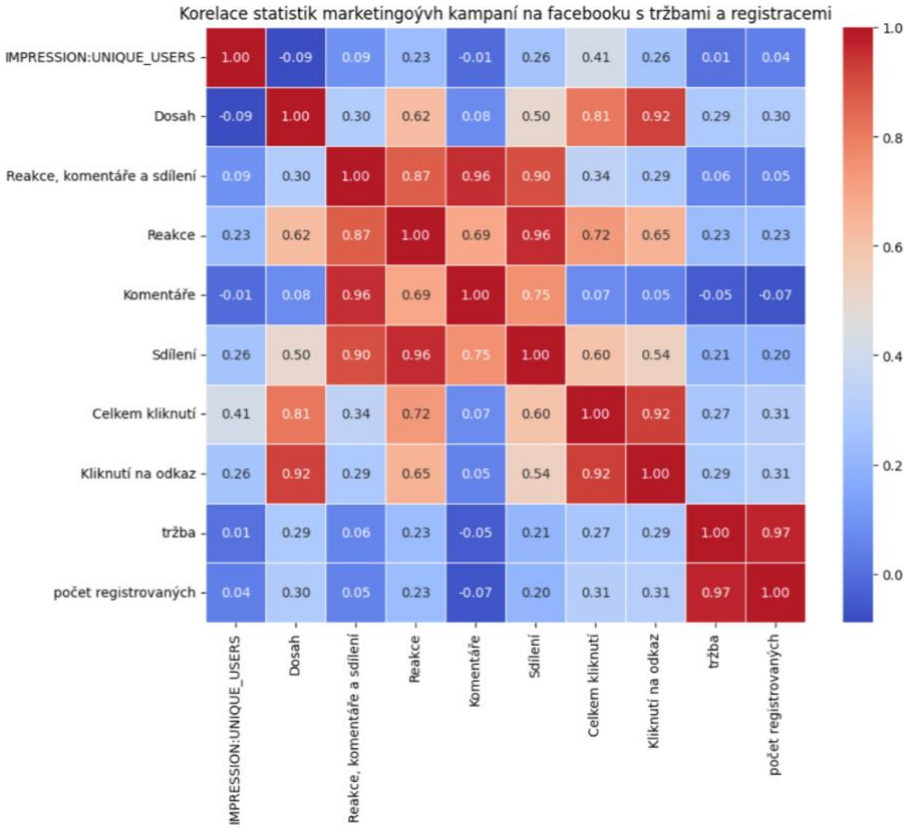
Prvním krokem fáze modelování bylo vytvoření korelačních matic ve formě teplotních map (heatmap), a to pro předem připravené datové sady obsahující marketingové příspěvky. Konkrétně příspěvky publikované na Facebooku, Instagramu a formát Instagram Stories. Cílem této analýzy bylo zjistit, zda mezi jednotlivými metrikami v rámci každé datové sady existují silné korelace, které by mohly negativně ovlivnit výsledky následně aplikovaných modelovacích technik. Vizuální podoba korelačních matic byla vytvořena v prostředí webové platformy Kaggle s využitím programovacího jazyka Python. Kód použitý pro analýzu korelace je uveden v příloze B na příkladu datové sady Instagram příspěvky. Stejný postup byl aplikován i na ostatní datové sady, tedy Facebook příspěvky a Instagram Stories..

Použitá korelační matice vycházela z výpočtu Pearsonova korelačního koeficientu, který je jednou z nejrozšířenějších metod pro vyhodnocování lineárních vztahů mezi dvěma spojitými proměnnými. Tento koeficient kvantifikuje sílu a směr vzájemné závislosti v rozsahu od –1 do +1. Hodnota blízká +1 značí silnou pozitivní korelaci, tedy situaci, kdy růst jedné proměnné je spojen s růstem druhé. Naopak hodnota blízká –1 značí, že s růstem jedné proměnné druhá klesá. Pokud se koeficient pohybuje kolem nuly, mezi proměnnými není prokazatelný lineární vztah. Výpočet je založen na porovnání odchylek jednotlivých hodnot od jejich průměrů a vztahu těchto odchylek mezi analyzovanými proměnnými (Berman, 2016):

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \cdot (y_i - \bar{y})^2}} \quad (1)$$

#### 4.4.1.1 Korelace Facebook příspěvky

Na obrázku 5Obrázek 5 je zobrazena teplotní mapa korelační analýzy vybraných metrik marketingových kampaní na Facebooku ve vztahu k dosažené tržbě a počtu registrovaných zákazníků. Z vizualizace je patrné, že mezi některými proměnnými existují silné korelace. Nejvýraznější pozitivní korelace byla zaznamenána mezi proměnnými „počet registrovaných“ a „tržba“ ($r = 0.97$), což je zcela očekávané vzhledem k přímé vazbě mezi těmito dvěma ukazateli.




**Obrázek 5:** Teplotní mapa korelace datové sady Facebook příspěvky

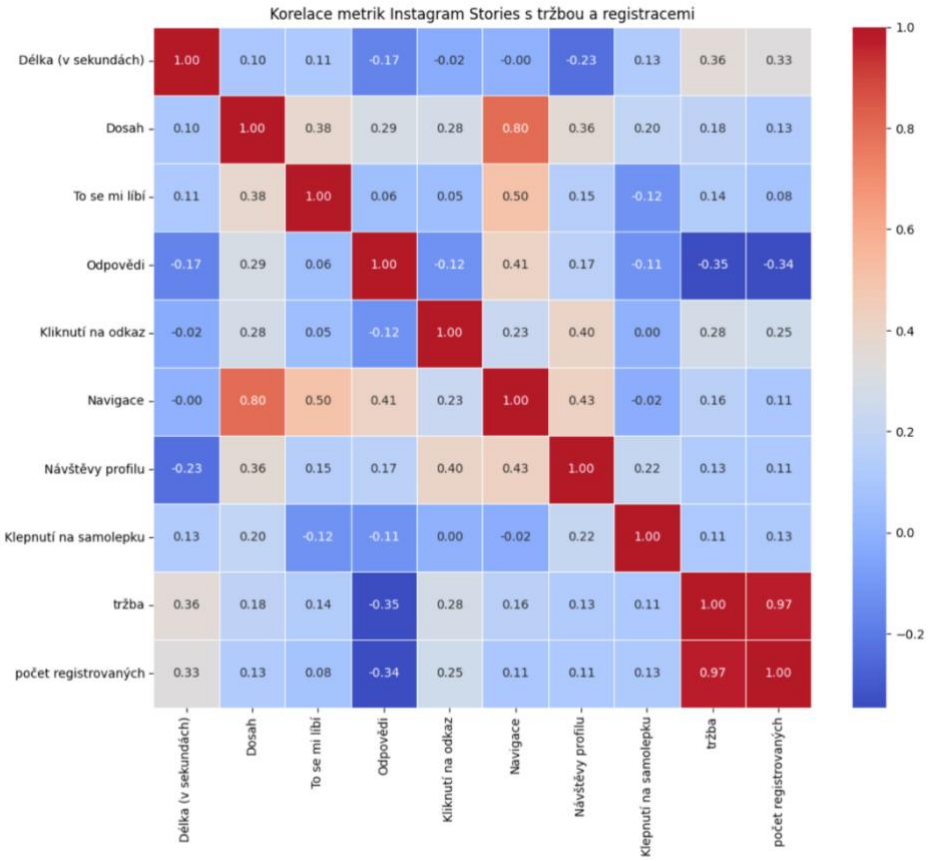
Zdroj: Vlastní zpracování

Silná korelace byla dále identifikována mezi metrikami měřícími interakci uživatelů například „reakce, komentáře a sdílení“ a „reakce“ ($r = 0.87$), případně „komentáře“ a „sdílení“ ($r = 0.75$). Tyto vysoké korelace naznačují, že některé proměnné se překrývají, což by mohlo negativně ovlivnit výkonnost regresních modelů. Tento jev, označovaný jako multikolinearita, je častým důvodem pro aplikaci metod pro redukci dimenze dat, jako je analýza hlavních komponent (PCA).

Na základě těchto zjištění bylo rozhodnuto, že v další fázi modelování bude aplikována metoda PCA, která umožní nahradit vzájemně korelované proměnné novými, vzájemně nezávislými komponentami při zachování většiny informační hodnoty dat.

#### 4.4.1.2 Korelace Instagram stories

Stejný analytický postup byl následně aplikován i na datovou sadu obsahující příspěvky ve formátu Instagram Stories. Výsledky korelační analýzy jsou zobrazeny níže na obrázku 6.



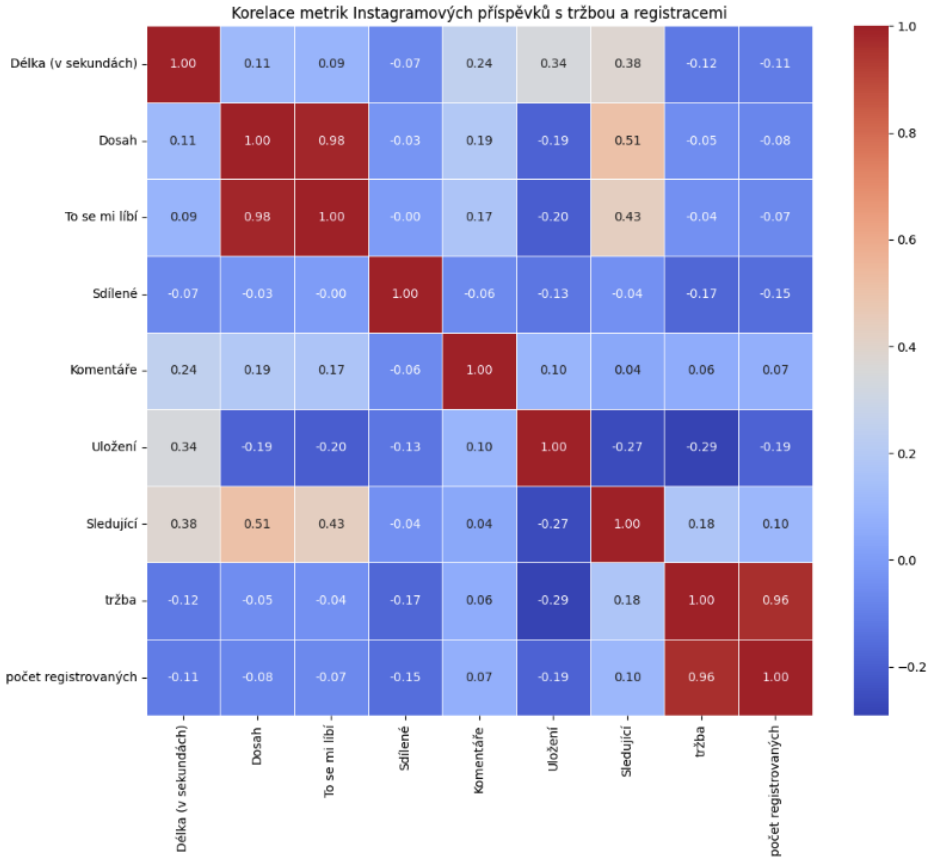

**Obrázek 6:** Teplotní mapa korelace datové sady Instagram Stories

Zdroj: Vlastní zpracování

Také v tomto případě byla potvrzena silná pozitivní korelace mezi proměnnými „tržba“ a „počet registrovaných“ ($r$ = 0.97), což opět odráží přímou vazbu mezi těmito dvěma klíčovými ukazateli výkonnosti kampaně. Významnější korelační vztah se objevil také mezi metrikami „navigace“ a „dosah“ ($r$ = 0.80), které souvisejí s mírou interakce uživatelů s daným formátem obsahu.

#### 4.4.1.3 Korelace Instagram příspěvky

Korelační analýza byla také aplikována na datovou sadu obsahující klasické příspěvky publikované na platformě Instagram. Výsledky jsou vizualizovány níže na obrázku 7.




**Obrázek 7:** Teplotní mapa korelace datové sady Instagram příspěvky

Zdroj: Vlastní zpracování

Také v tomto případě se potvrdila velmi silná pozitivní korelace mezi proměnnými „tržba“ a „počet registrovaných“ ($r$ = 0.96), což opět potvrzuje úzkou vazbu mezi těmito dvěma ukazateli obchodního výkonu kampaně.

Na rozdíl od předchozích datových sad však metriky uživatelského zapojení, jako jsou „To se mi líbí“, „Sdílení“, „Komentáře“ nebo „Uložení“, nevykazují žádný významný lineární vztah k tržbě nebo počtu registrací. Korelační koeficienty se v těchto případech pohybují spíše nízko nebo dokonce mírně záporně. Naopak relativně silnější vztah lze pozorovat mezi některými z těchto metrik navzájem, například „To se mi líbí“ a „Dosah“ ($r$ = 0.98) nebo „Tržba“ a „Počet registrovaných“ ($r$ = 0.96), což značí vnitřní propojenost interakčních prvků v rámci jednoho příspěvku.

### 4.4.2 Analýza hlavních komponent (PCA)

Analýza hlavních komponent PCA je jednou z nejpoužívanějších technik vícerozměrné statistické analýzy, jejímž cílem je redukovat dimenzionalitu dat při zachování co největší míry variability. Principem metody je nahrazení původních korelovaných proměnných novou množinou vzájemně nekorelovaných proměnných, tzv. hlavních komponent. Tyto komponenty jsou lineárními kombinacemi původních proměnných a jsou seřazeny podle toho, kolik vysvětlené variance v datech zachycují (Abdi a Williams, 2010).

Podle Jahangir a kol. (2021) lze metodu PCA rozdělit do následujících kroků:

1. Výpočet průměrného vektoru příznaků:

Nejprve je třeba standardizovat data odečtením průměrné hodnoty každé proměnné. Výsledkem je tzv. centrovaná matice dat, která je základem pro další výpočty:

$$x_{mean} = \frac{1}{n}\sum_{i=1}^{n} x_i. \quad (2)$$

2. Výpočet kovarianční matice

Kovarianční matice *Cov* měří, jak se jednotlivé proměnné mění společně. Je klíčovým prvkem PCA, protože právě na jejím základě se vypočítávají hlavní komponenty:

$$Cov = \frac{1}{n}\sum_{i=1}^{n}(x_i - x_{mean})(x_i - x_{mean})^T. \quad (3)$$

3. Výpočet vlastních čísel a vlastních vektorů kovariační matice

Ze získané kovarianční matice se spočítají tzv. vlastní čísla ($\lambda_k$) a jejich odpovídající vlastní vektory ($e_k$), které definují směry maximální variance v datech:

$$Cov * e_k = \lambda_k * e_k, \quad (4)$$

kde $\lambda_k$ je vlastní číslo a $e_k$ je odpovídající vlastní vektor.

4. Volba komponent s největšími vlastními čísly

Z množiny všech vlastních vektorů se vyberou ty, které odpovídají největším vlastním číslům – tedy ty, které vysvětlují největší část variability:

$$M_p = \{e_1, e_2 \dots e_p\}. \quad (5)$$

5. Projekce původních dat do nového prostoru

Na závěr se původní data $X$ přetransformují pomocí matice hlavních komponent $M_P^T$ do nového prostoru s nižší dimenzionalitou

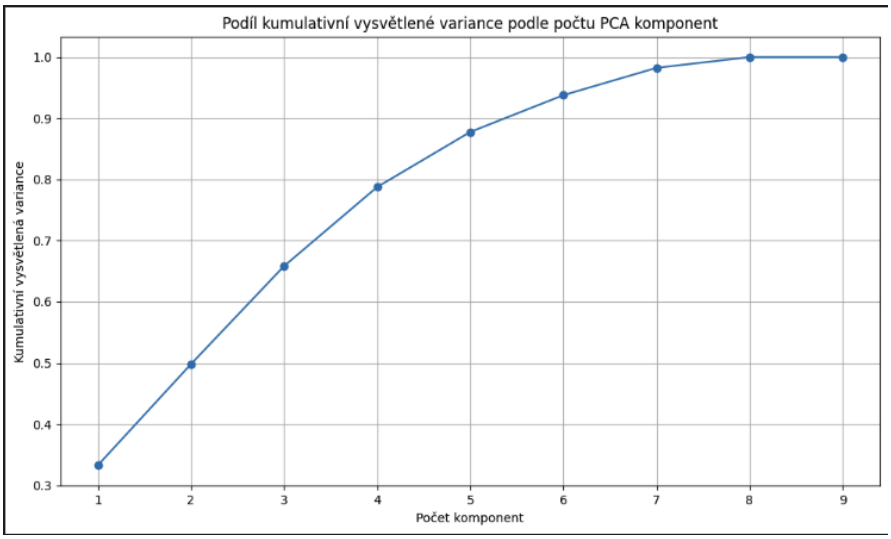
$$N_p = M_P^T * X. \quad (6)$$

Na základě výsledků z Analýza korelace, která potvrdila přítomnost silných korelací mezi jednotlivými metrikami v rámci všech analyzovaných datových sad, byla zvolena aplikace PCA metody. Tato metoda byla provedena samostatně pro každou datovou sadu, s cílem eliminovat problém multikolinearity a zároveň snížit počet vstupních proměnných bez významné ztráty podstatné informační hodnoty.

Ukázkový kód pro aplikaci metody PCA na datovou sadu Instagram příspěvky je uveden v příloze C. Stejným způsobem byla následně tato metoda aplikována také na datové sady Facebook příspěvky a Instagram Stories.

#### 4.4.2.1 Aplikace PCA na vstupní data Instagram Stories

Na základě vývoje kumulativní vysvětlené variance na obrázku 8, bylo rozhodnuto zvolit pět hlavních komponent, které dohromady vysvětlují přibližně 88 % variability v datech. Tato hranice byla zvolena jako dostatečná, neboť další komponenty již přinášely jen minimální nárůst informační hodnoty, zatímco by navyšovaly složitost následného modelování.




**Obrázek 8:** Vývoj kumulativní vysvětlené variance pro datovou sadu Instagram Stories

Zdroj: Vlastní zpracování

Výsledky aplikace PCA metody na datovou sadu jsou ilustrovány v tabulce níže (obrázek 9), která představuje pouze náhled transformovaných dat, konkrétně prvních několik záznamů z celé datové sady. Každý řádek odpovídá jednomu příspěvku a je popsán pomocí pěti hlavních komponent (PC1 až PC5). Tyto komponenty zachycují většinu variability obsažené v datech a zároveň eliminují problém vzájemné korelace mezi jednotlivými vstupními proměnnými. Ke každému příspěvku je dále uvedena hodnota dosažené tržby a odpovídající identifikátor příspěvku (ID), který slouží k zachování propojení mezi transformovanými a původními daty.

| | PC1 | PC2 | PC3 | PC4 | PC5 | tržba | ID příspěvku |
|---|---|---|---|---|---|---|---|
| 0 | -1.198489 | 0.074294 | 1.426538 | -0.287883 | 0.543322 | 1.42 | 17922200000000000 |
| 1 | 6.412796 | 3.697785 | -0.850144 | -1.318696 | 0.893642 | 1.70 | 18019500000000000 |
| 2 | 4.047878 | -0.970967 | 1.192726 | -0.615525 | -1.705305 | 1.70 | 17985100000000000 |
| 3 | 1.062887 | 0.057715 | -0.931782 | 0.164597 | -0.036499 | 0.41 | 18028900000000000 |
| 4 | -1.145496 | 1.792209 | 1.313016 | 1.713282 | 0.431093 | 1.42 | 17989900000000000 |

**Obrázek 9:** Náhled hodnot pro PCA komponenty datové sady Instagram Stories

Zdroj: Vlastní zpracování

Další tabulka zobrazená na obrázku 10 ukazuje tzv. váhy (loadings) jednotlivých marketingových metrik ve vztahu k pěti hlavním komponentám PCA (PC1 až PC5). Tyto váhy ukazují, jak silně každá původní metrika přispívá ke konkrétní komponentě. Hodnoty mohou být kladné i záporné, přičemž vyšší absolutní hodnota značí vyšší míru zastoupení dané metriky v příslušné komponentě. Každá komponenta je tedy lineární kombinací původních proměnných, přičemž právě váhy určují relativní vliv jednotlivých metrik na vytvořenou dimenzi.

Váhy metrik (loadings)

| | PC1 | PC2 | PC3 | PC4 | PC5 |
|---|---|---|---|---|---|
| Délka (v sekundách) | −0.025358 | 0.211870 | 0.736169 | 0.030325 | 0.442980 |
| Dosah | 0.529317 | 0.056488 | 0.182759 | 0.142556 | 0.126680 |
| To se mi líbí | 0.338649 | −0.200356 | 0.387333 | −0.311095 | −0.603540 |
| Sdílené | 0.000000 | 0.000000 | 0.000000 | −0.000000 | −0.000000 |
| Odpovědi | 0.266086 | −0.511969 | −0.206953 | 0.420470 | 0.396548 |
| Kliknutí na odkaz | 0.252947 | 0.443009 | −0.221350 | −0.540678 | 0.415404 |
| Navigace | 0.559884 | −0.154237 | 0.101636 | 0.013264 | 0.040007 |
| Návštěvy profilu | 0.389444 | 0.293204 | −0.409862 | 0.011979 | −0.173546 |
| Klepnutí na samolepku | 0.067655 | 0.586217 | 0.066979 | 0.642286 | −0.248860 |

**Obrázek 10:** Váhy metrik pro PCA komponenty v datové sadě Instagram Stories

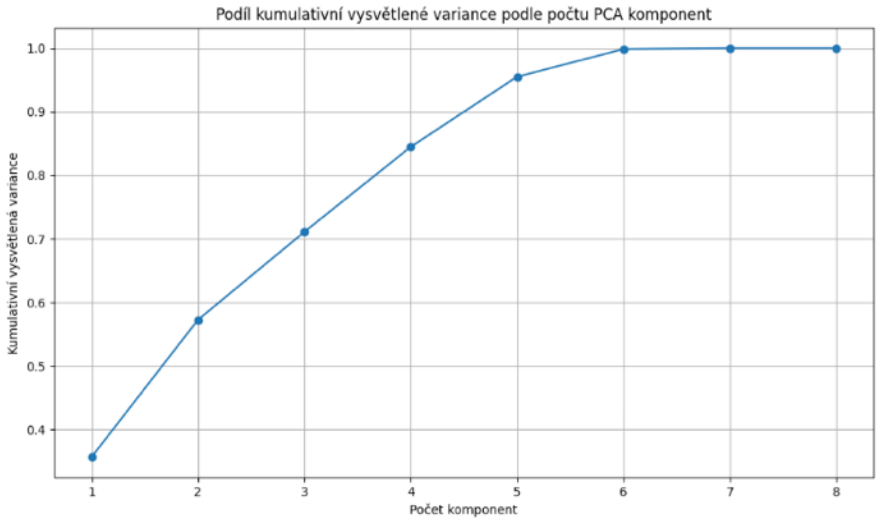
Zdroj: Vlastní zpracování

Přehled nejvýznamnějších metrik pro každou komponentu:

- PC1: Nejvíce ovlivněna metrikou „Navigace“ s váhou 0.559884,
- PC2: Nejvíce ovlivněna metrikou „Klepnutí na samolepku“ s váhou 0.586217,
- PC3: Nejvíce ovlivněna metrikou „Délka (v sekundách)” s váhou 0.736169,
- PC4: Nejvíce ovlivněna metrikou „Klepnutí na samolepku” s váhou 0.642286,
- PC5: Nejvíce ovlivněna metrikou „To se mi líbí“ s váhou –0.603540.

#### 4.4.2.2 Aplikace PCA na vstupní data Instagramové příspěvky

Na základě předchozího postupu byla metoda hlavních komponent aplikována také na datovou sadu obsahující klasické příspěvky publikované na platformě Instagram. Pomocí analýzy kumulativní vysvětlené variance (obrázek 11), byl stanoven optimální počet hlavních komponent. Z obrázku je patrné že pět komponent postačovalo k zachycení více než 90 % celkové variability dat.




**Obrázek 11:** Vývoj kumulativní vysvětlené variace pro datovou sadu Instagramové příspěvky

Zdroj: Vlastní zpracování

Aplikace PCA metody poskytla dva hlavní výstupy. Prvním výstupem je transformovaná datová matice (obrázek 12), ve které jsou jednotlivé příspěvky vyjádřeny pomocí nových komponent (PC1–PC5) namísto původních korelovaných metrik.

| | PC1 | PC2 | PC3 | PC4 | PC5 | tržba | ID příspěvku |
|---|---|---|---|---|---|---|---|
| 0 | 4.675398 | 1.128473 | 1.948977 | 4.080791 | -1.512289 | 1.42 | 18220023712270152 |
| 1 | -0.386831 | -0.964016 | -0.304097 | 0.219396 | -0.327477 | 0.41 | 18015026009060112 |
| 2 | -0.362971 | -0.968233 | -0.308956 | 0.211418 | -0.314369 | 0.71 | 18032296156735200 |
| 3 | 0.266562 | -0.305508 | 0.229915 | 0.927680 | -0.568173 | 0.71 | 18015893125860888 |
| 4 | -0.370003 | -0.969991 | -0.307309 | 0.214692 | -0.321987 | 1.01 | 18149115490306808 |

**Obrázek 12:** Náhled hodnot PCA komponent pro datovou sadu Instagramové příspěvky

Zdroj: Vlastní zpracování

Druhým výstupem je tzv. matice váhových koeficientů (loadings) zobrazená na obrázku 13, která určuje, jak výrazně každá původní metrika ovlivňuje příslušnou komponentu.

Váhy metrik (loadings)

| | PC1 | PC2 | PC3 | PC4 | PC5 |
|---|---|---|---|---|---|
| Délka (v sekundách) | 0.185375 | 0.623947 | 0.373576 | 0.298038 | -0.059887 |
| Komentář s daty | 0.000000 | -0.000000 | 0.000000 | -0.000000 | -0.000000 |
| Dosah | 0.597016 | -0.073710 | -0.116649 | -0.152217 | 0.282904 |
| To se mi líbí | 0.580970 | -0.103153 | -0.118372 | -0.194952 | 0.319815 |
| Sdílené | -0.024488 | -0.280724 | 0.887562 | -0.320143 | 0.153692 |
| Komentáře | 0.175827 | 0.387643 | -0.031460 | -0.677146 | -0.576810 |
| Uložení | -0.180067 | 0.602713 | -0.004261 | -0.115713 | 0.616061 |
| Sledující | 0.455788 | 0.047757 | 0.209876 | 0.524966 | -0.279703 |

**Obrázek 13:** Váhy metrik pro PCA komponenty v datové sadě Instagramové příspěvky

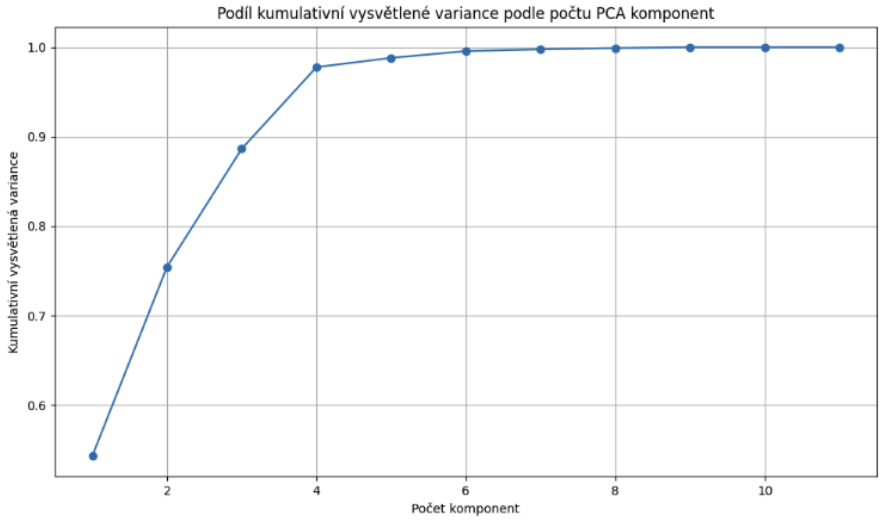
Zdroj: Vlastní zpracování

Přehled nejvýznamnějších metrik pro každou komponentu**:**

- PC1: Nejvíce ovlivněna metrikou „Dosah“ s váhou 0.597016,
- PC2: Nejvíce ovlivněna metrikou „Délka (v sekundách)“ s váhou 0.623947,
- PC3: Nejvíce ovlivněna metrikou „Sdílené“ s váhou 0.887562,
- PC4: Nejvíce ovlivněna metrikou „Sledující“ s váhou 0.524966,
- PC5: Nejvíce ovlivněna metrikou „Uložení“ s váhou 0.610661.

#### 4.4.2.3 Aplikace PCA na Facebook příspěvky

PCA metoda byla aplikována i na třetí datovou sadu obsahující marketingové příspěvky zveřejněné na platformě Facebook. Opět byl vytvořen graf kumulativní vysvětlené variance (obrázek 14), na jehož základě byly vybrány čtyři hlavní komponenty. Tyto komponenty společně přes téměř 95% variability v datech.




**Obrázek 14:** Vývoj kumulativní vysvětlené variance pro datovou sadu Facebook příspěvky

Zdroj: Vlastní zpracování

V tabulce níže (obrázek 15) jsou zobrazeny hodnoty transformované datové matice, ve které jsou jednotlivé příspěvky vyjádřeny pomocí nových komponent (PC1–PC4) namísto původních korelovaných metrik.

| | PC1 | PC2 | PC3 | PC4 | tržba | ID příspěvku |
|---|---|---|---|---|---|---|
| 0 | -1.139474 | -0.068011 | 0.112952 | -0.361244 | 1.42 | 1820370000000000 |
| 1 | -1.139518 | -0.067983 | 0.113010 | -0.361232 | 1.42 | 1423950000000000 |
| 2 | -0.913638 | -0.150952 | -0.619455 | 1.667300 | 0.87 | 828233000000000 |
| 3 | -0.541649 | 0.395813 | 0.365005 | -0.124893 | 0.87 | 828234000000000 |
| 4 | -1.139489 | -0.068002 | 0.112972 | -0.361240 | 0.87 | 289148000000000 |

**Obrázek 15:** Náhled hodnot PCA komponent pro datovou sadu Facebook příspěvky

Zdroj: Vlastní zpracování

Níže přiložená tabulka (obrázek 16), znázorňuje tzv. váhy (loadings) jednotlivých metrik přiřazené ke čtyřem hlavním komponentám. Tyto váhy vyjadřují relativní míru vlivu jednotlivých metrik na každou z komponent.

Váhy metrik (loadings)

| | PC1 | PC2 | PC3 | PC4 |
|---|---|---|---|---|
| Délka (v sekundách) | 0.007005 | -0.087506 | -0.407191 | 0.853676 |
| IMPRESSION:UNIQUE_USERS | 0.144288 | -0.138544 | 0.670703 | 0.398057 |
| Dosah | 0.320096 | -0.207517 | -0.422723 | -0.089077 |
| Reakce, komentáře a sdílení | 0.276430 | 0.480652 | -0.006942 | 0.032009 |
| Reakce | 0.380971 | 0.219496 | -0.009668 | -0.017614 |
| Komentáře | 0.182688 | 0.574824 | -0.006307 | 0.051529 |
| Sdílení | 0.350167 | 0.297847 | 0.022404 | 0.093339 |
| Celkem kliknutí | 0.375019 | -0.251517 | 0.044392 | -0.074850 |
| Spotřeba cílení na shodující se okruh uživatelů... | 0.323426 | -0.225557 | 0.397054 | 0.065853 |
| Kliknutí na odkaz | 0.352340 | -0.268042 | -0.178305 | 0.076970 |
| Jiná kliknutí | 0.360193 | -0.213919 | -0.116443 | -0.276200 |

**Obrázek 16:** Váhy metrik PCA komponent pro datová sada Facebook příspěvky

Zdroj: Vlastní zpracování

Přehled nejvýznamnějších metrik pro každou komponentu:

- PC1: Nejvíce ovlivněna metrikou „Celkem kliknutí“ s váhou 0.375019,
- PC2: Nejvíce ovlivněna metrikou „Komentáře“ s váhou 0.578424,
- PC3: Nejvíce ovlivněna metrikou „IMPRESSION:UNIQUE_USERS“ s váhou 0.670703,
- PC4: Nejvíce ovlivněna metrikou „Délka (v sekundách)“ s váhou 0.853676.

### 4.4.3 Shluková analýza K-means

Shluková analýza patří mezi základní nástroje využívané v oblasti dolování dat. Jejím účelem je rozdělit zkoumaná data do skupin (tzv. shluků) tak, aby si prvky v rámci jedné skupiny byly co nejvíce podobné, zatímco rozdíly mezi jednotlivými skupinami byly co nejvýraznější. Tento přístup umožňuje odhalit skryté struktury a vztahy v datech, které by jinak mohly zůstat nepovšimnuty (Syakur a kol., 2018).

Jednou z nejpoužívanějších metod pro shlukování je algoritmus K-means, který patří mezi metody rozdělující data do pevně stanoveného počtu shluků *K*. Algoritmus funguje tak, že se postupně přiřazují data k nejbližšímu centroidu (středu shluku) a iterativně se přepočítávají nové pozice těchto středů až do dosažení konvergence. Výsledkem je rozdělení dat do K shluků, přičemž každý shluk má svůj centroid, který minimalizuje vzdálenost ke všem prvkům daného shluku (Syakur a kol., 2018).

Výpočet centroidu je dle Syakur a kol. (2018) dán vzorcem:

$$c_j = \frac{1}{n_j} \sum_{i=1}^{n_j} x_i, \tag{7}$$

kde $c_j$ je centroid *j*-tého shluku, $n_j$ je počet pozorování v *j*-tém shluku a $x_i$ jsou jednotlivá pozorování v tomto shluku.

Pro měření vzdálenosti mezi jednotlivými body a centroidy se typicky používá Eukleidovská vzdálenost:

$$d(x, c) = \sqrt{\sum_{i=1}^{p} (x_i - c_i)^2}. \tag{8}$$

Cílem algoritmu je minimalizovat tzv. sumu čtvercových chyb (SSE – Sum of Squared Error), která je definována jako:

$$SSE = \sum_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - c_k\|^2, \tag{9}$$

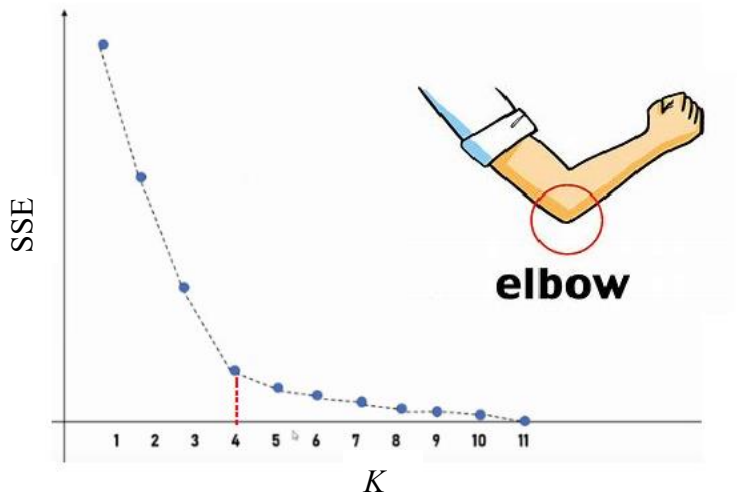
kde $C_k$ je množina pozorování přiřazených ke *k*-tému shluku a $c_k$ je jeho centroid.

#### 4.4.3.1 Metoda lokte

Jelikož algoritmus K-means vyžaduje jako vstupní parametr předem definovaný počet shluků *K*, je nezbytné před samotnou aplikací tohoto algoritmu určit vhodný počet těchto skupin. Nesprávně zvolená hodnota může totiž vést buď k příliš hrubému rozdělení, nebo naopak ke zbytečně složitému modelu, který bude obtížně interpretovatelný (Syakur a kol., 2018).

Jedním z nejčastěji využívaných přístupů pro výběr optimálního počtu shluků je metoda lokte (Elbow Method). Princip této metody spočívá ve výpočtu sumy čtvercových chyb (SSE) pro různé hodnoty *K*, obvykle v rozmezí od 1 do 10. Výsledné hodnoty SSE se následně zobrazí v grafu, kde osa X znázorňuje počet shluků a osa Y odpovídá velikosti chyby (Syakur a kol., 2018).

V grafu níže (Obrázek 17) se pak hledá tzv. „loket“, neboli bod, kde se rychlost poklesu chyby začne výrazně zpomalovat. Tento zlomový bod naznačuje, že přidávání dalších shluků již nepřináší zásadní zlepšení, a představuje tak rovnováhu mezi kvalitou rozdělení dat a jednoduchostí výsledného modelu. Metoda lokte tak nabízí praktický způsob, jak určit optimální počet shluků, aniž by docházelo k jejich zbytečnému nárůstu, který by mohl vést ke ztrátě přehlednosti (Syakur a kol., 2018).



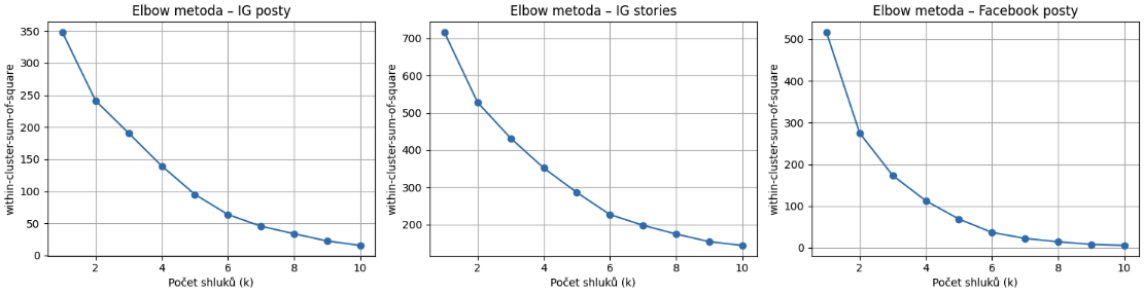

**Obrázek 17:** Metoda lokte

Zdroj: Převzato z Zalarushi (2023)

#### 4.4.3.1.1 Aplikace metody lokte na analyzovaná data

Na základě výše uvedených teoretických principů byla metoda lokte aplikována na všechny tři připravené datové sady, konkrétně na příspěvky publikované na Instagramu a Facebooku, společně i s Instagram Stories. Je důležité zdůraznit, že tyto datové sady nepředstavují původní korelovaná data, ale datové matice vzniklé transformací pomocí metody PCA. Do analýzy byly zahrnuty pouze ty komponenty, které dohromady vysvětlovaly alespoň 88% variability původních dat, čímž se zajistilo zachování informační hodnoty při současné eliminaci multikolinearity mezi proměnnými. Kód, který zahrnuje celý postup shlukové analýzy K-means, včetně aplikace metody lokte, je obsažen v příloze D.

Metoda lokte byla využita ke stanovení optimálního počtu shluků *K*, které následně sloužily jako vstupní parametr pro algoritmus K-means. Výsledky této analýzy jsou znázorněny v následujících grafech (obrázek 18). Na ose Y je uvedena hodnota WCSS (Within-Cluster Sum of Squares), která odpovídá sumě čtvercových chyb (SSE) v rámci shluku. Jedná se o součet kvadrátů vzdáleností všech bodů od jejich příslušných centroidů. V grafech je možné sledovat, jak se hodnota WCSS mění v závislosti na počtu zvolených shluků *K*.




**Obrázek 18:** Metoda lokte pro datové sady (vlevo) Instagramové příspěvky, (uprostřed) Instagramové Stories, (vpravo) Facebookové příspěvky

Zdroj: Vlastní zpracování

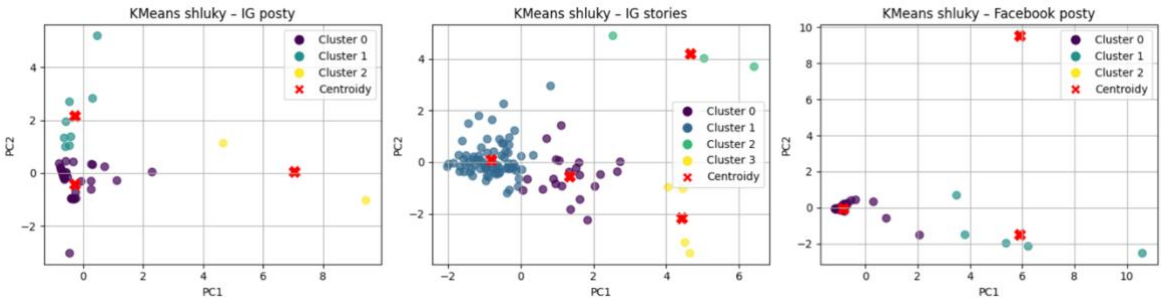
V levém grafu (Instagram příspěvky) lze inflexní bod pozorovat při hodnotě $K = 3$. Po tomto bodu již klesá hodnota WCSS jen velmi mírně, což znamená, že přidávání dalších shluků by vedlo k minimálnímu zlepšení kvality shlukování.

Střední graf ukazuje vývoj WCSS pro datovou sadu Instagram Stories. Inflexní bod je zde patrný při hodnotě $K = 4$, kde dochází k výraznému zlomu ve vývoji křivky.

V pravém grafu lze obdobně identifikovat inflexní bod pro Facebookové příspěvky kolem hodnoty $K = 3$. I v tomto případě je pokles WCSS po této hodnotě pozvolný, což potvrzuje, že tři shluky poskytují dostačující rozlišení v rámci této datové sady.

#### 4.4.3.2 Výsledky K-means shlukování

Na základě výsledků metody lokte byl pro každou ze tří transformovaných datových sad (Instagram příspěvky, Instagram Stories a Facebook příspěvky) stanoven optimální počet shluků, který byl následně použit jako vstupní parametr pro K-means algoritmus. Výsledkem je rozdělení dat do jednotlivých shluků, které jsou znázorněny v následující vizualizaci (obrázek 19).




**Obrázek 19:** Vizualizace shluků K-means pro komponenty PC1 a PC2 z transformovaných datových sad

Zdroj: Vlastní zpracování

Grafy zobrazují dvourozměrnou projekci příspěvků do prostoru prvních dvou hlavních komponent (PC1 a PC2), které byly získány aplikací PCA. Každý bod představuje jeden příspěvek, barva znázorňuje příslušnost ke konkrétnímu shluku a červené křížky označují centroidy jednotlivých skupin.

Instagramové příspěvky (vlevo) byly rozděleny do tří shluků. Dva z těchto shluků jsou si relativně blízké a soustředěné v levé části grafu, zatímco třetí shluk (žlutý) se nachází výrazně vpravo, což může naznačovat jeho odlišnost z hlediska souhrnné metriky PC1. Takové rozdělení může poukazovat na existenci specifického typu příspěvků s odlišným chováním či výkonem.

U datové sady Instagram Stories (uprostřed) došlo k vytvoření čtyř shluků. Rozložení je rovnoměrnější a ukazuje na rozmanitost v typu obsahu Stories. Centroidy jsou dobře oddělené a shluky mají relativně kompaktní podobu, což naznačuje kvalitní segmentaci. Tyto rozdíly mohou být způsobeny například variací v počtu interakcí, délkou sledování nebo prokliky.

Facebookové příspěvky (vpravo) vykazují identifikované shluky vykazují poměrně specifické rozložení. Jeden z klastrů (žlutý) se skládá z několika odlehlých bodů, které se od ostatních výrazně liší v hodnotě PC1. Dva další shluky se nacházejí blízko středu a naznačují standardní typy příspěvků s podobnými vlastnostmi.

#### 4.4.3.3 Analýza výkonnosti shluků dle tržby

V rámci navazující fáze analýzy byla provedena interpretace výkonnosti jednotlivých shluků na základě dosažené tržby. Pro každý vytvořený shluk byla vypočtena průměrná hodnota tržby, přičemž zvláštní pozornost byla věnována nejvýkonnějšímu shluku, tedy takovému, který vykazoval nejvyšší průměrnou tržbu napříč celou datovou sadou. Zároveň byl v každém z těchto shluků identifikován také nejlepší příspěvek, tj. ten, který dosáhl nejvyšší individuální tržby.

Pro Instagramové příspěvky (obrázek 20Obrázek 20) byl jako nejvýkonnější identifikován shluk s označením Cluster 2. Tento shluk dosáhl nejvyšší průměrné tržby, konkrétně 0.96. Nejvýkonnější jednotlivý příspěvek v rámci tohoto shluku nesl ID 18220023712270152 a jeho tržba činila 1.42.

```
ANALÝZA – IG posty
Nejvýkonnější cluster: Cluster 2
Průměrná tržba podle clusterů:
cluster_label
2    0.960000
0    0.794286
1    0.550000
Name: tržba, dtype: float64

Nejlepší příspěvek (ID a tržba) z nejvýkonnějšího clustru:
     ID příspěvku  tržba
18220023712270152   1.42
```

**Obrázek 20:** Výkonnost shlukování pro datovou sadu Instagramové příspěvky

Zdroj: Vlastní zpracování

V případě Instagram Stories (obrázek 21) byl jako nejvýnosnější vyhodnocen shluk Cluster 2, který vykázal průměrnou tržbu 1.61, tedy nejvyšší hodnotu napříč všemi analyzovanými skupinami. Nejvýkonnější příspěvek v rámci tohoto shluku, nesoucí ID 18019500000000000, dosáhl dokonce maximální sledované tržby 1.70, čímž potvrdil výjimečnou efektivitu obsahu zařazeného do této skupiny.

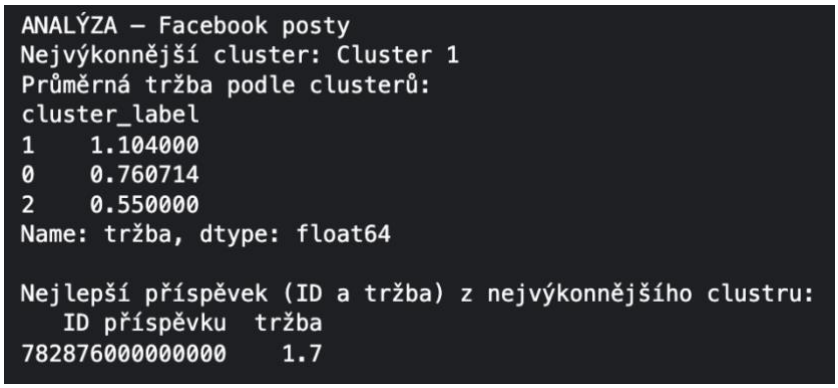
```
ANALÝZA – IG stories
Nejvýkonnější cluster: Cluster 2
Průměrná tržba podle clusterů:
cluster_label
2    1.606667
3    1.012500
1    0.917027
0    0.898095
Name: tržba, dtype: float64

Nejlepší příspěvek (ID a tržba) z nejvýkonnějšího clustru:
     ID příspěvku  tržba
18019500000000000    1.7
```

**Obrázek 21:** Výkonnost shlukování pro datovou sadu Instagram Stories

Zdroj: Vlastní zpracování

U Facebookových příspěvků (obrázek 22) byl za nejúspěšnější považován Cluster 1, kde průměrná tržba dosahovala 1.10. Také zde byl identifikován nejlepší jednotlivý příspěvek s ID 782876000000000, který se rovněž vyznačoval tržbou 1.70, stejně jako nejlepší Story na Instagramu. Tato shoda v maximální hodnotě tržby u příspěvků z rozdílných platforem naznačuje, že podobně efektivní typy obsahu mohou být přítomny napříč více kanály.




**Obrázek 22:** Výkonnost shlukování pro datovou sadu Facebook příspěvky

Zdroj: Vlastní zpracování

Tato interpretace slouží jako důležitý základ pro následné hledání vzorců chování a návrh doporučení pro optimalizaci obsahu. Identifikace příspěvků s nejvyšší výkonností v rámci dominantních shluků otevírá prostor pro další zkoumání charakteristik těchto příspěvků, (například na základě jejich metrik zapojení nebo vizuálního formátu) a může být využita k predikci nebo cílené tvorbě marketingových sdělení.

### 4.4.4 Regresní analýza

Regresní analýza je statistická technika, která slouží k prozkoumání a popsání vztahů mezi proměnnými. Umožňuje zkoumat, jak změny jedné nebo více nezávislých proměnných ovlivňují jinou, závislou proměnnou a tím napomáhá k jejímu lepšímu pochopení nebo předpovědi (Freund a kol., 2010)

#### 4.4.4.1 Vícenásobná lineární regrese

Vícenásobná lineární regrese rozšiřuje základní principy jednoduché regrese na situace, kdy na hodnotu závislé proměnné současně působí více nezávislých faktorů. Cílem tohoto přístupu je nalézt takový matematický model, který co nejlépe vystihuje vztah mezi zkoumanými proměnnými. Vícenásobná lineární regrese umožňuje nejen odhadnout hodnotu cílové proměnné na základě kombinovaného vlivu několika vstupních veličin, ale také posoudit, které z těchto veličin mají na výsledek největší dopad (Lin a kol., 2024).

Základním principem této metody je nalezení nejvhodnějších parametrů modelu tak, aby se minimalizoval rozdíl mezi skutečnými a predikovanými hodnotami. Tento rozdíl je vyjádřen pomocí tzv. sumy čtvercových chyb (SSE), přičemž nejčastěji používanou metodou výpočtu je metoda nejmenších čtverců (Ordinary Least Squares – OLS). Výsledný regresní model je následně možné využít pro predikci hodnot závislé proměnné (Lin a kol., 2024)

Obecný tvar modelu vícenásobné lineární regrese lze podle Lin a kol. (2024) zapsat následovně:

$$y = a + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \varepsilon, \qquad (10)$$

kde $y$ je závislá proměnná, $x_1, x_2, \ldots + x_n$ jsou nezávislé proměnné, $\beta_1, \beta_2, \ldots \beta_n$ jsou regresní koeficienty, $a$ je konstanta (intercept) a $\varepsilon$ je chyba modelu.

#### 4.4.4.2 Aplikace vícenásobné lineární regrese na datové sady

Na základě výsledků PCA analýzy byla provedena vícenásobná lineární regrese, jejímž cílem bylo zjistit, zda a v jaké míře dokáží jednotlivé hlavní komponenty vysvětlit variabilitu v hodnotách tržeb. Tato analýza byla aplikována zvlášť pro každou ze tří datových sad, které pocházely z výstupů metody PCA. Kompletní kód pro vícenásobnou lineární regresi je obsažený v příloze E.

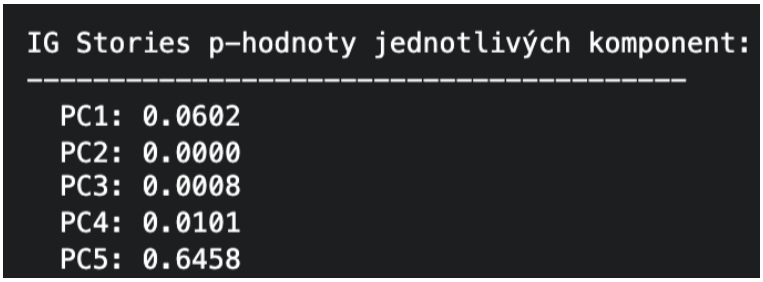
Výsledky aplikace vícenásobné lineární regrese (obrázek 23) na datovou sadu PCA Instagram Stories ukazují, že model dosáhl koeficientu determinace $R^2 = 0.308$, což znamená, že přibližně 30.8 % variability ve výši tržeb lze vysvětlit na základě pěti hlavních komponent získaných pomocí PCA. Hodnota MSE (Mean Squared Error) dosáhla 0.177, což vyjadřuje průměrnou kvadratickou odchylku predikovaných hodnot od skutečných tržeb.

```
IG Stories
----------
R² score: 0.308
MSE: 0.177
Koeficienty:
  PC1: 0.0501
  PC2: 0.1675
  PC3: 0.1309
  PC4: -0.1109
  PC5: 0.0233
```

**Obrázek 23:** Výsledky vícenásobné lineární regrese pro datovou sadu Instagram Stories

Zdroj: Vlastní zpracování

Při interpretaci výsledků vícenásobné lineární regrese je důležité zohlednit nejen velikost regresních koeficientů, ale také jejich statistickou významnost, která je posuzována pomocí *p*-hodnot. Ty jsou pro jednotlivé komponenty zobrazeny na obrázku 24 níže.



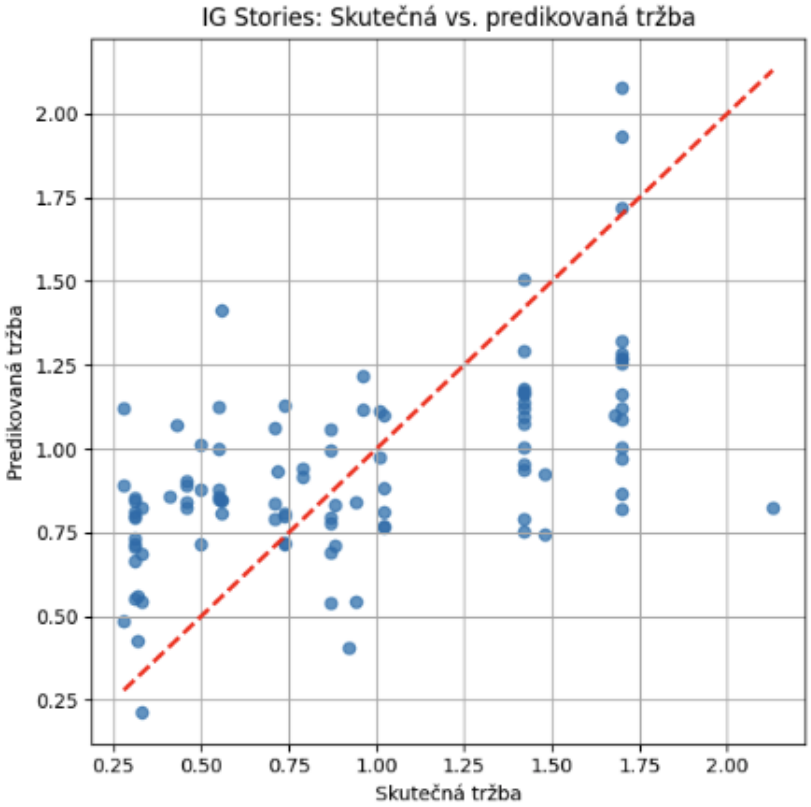

**Obrázek 24:** *p*-hodnoty pro datovou sadu Instagram Stories

Zdroj: Vlastní zpracování

Z hlediska jednotlivých hlavních komponent:

- PC1 vykazuje pozitivní vliv na tržby (koeficient: +0.0501), avšak tento vztah není statisticky významný na běžné hladině významnosti ($\alpha = 0.05$). Komponenta je tvořena zejména metrikami jako navigace, návštěvy profilu a dosah.
- PC2 představuje nejsilnější pozitivní vztah s tržbou (koeficient: +0.1675) a zároveň je statisticky vysoce významná ($p < 0.001$). Největší váhu zde mají metriky klepnutí na samolepku, kliknutí na odkaz a odpovědi.
- PC3 rovněž přispívá pozitivně (koeficient: +0.1309) a její vliv je statisticky významný ($p = 0.0008$). Tato komponenta je ovlivněna zejména metrikami délka sledování, to se mi líbí a dosah.
- PC4 má záporný vliv na tržby (koeficient: -0.1109) a tento vztah je statisticky významný ($p = 0.0101$). Nejvýrazněji ji utvářejí metriky odpovědi, kliknutí na odkaz a klepnutí na samolepku.
- PC5 vykazuje velmi slabý pozitivní vztah (koeficient: +0.0233), který však není statisticky významný ($p = 0.6458$). Tvoří ji především metriky délka sledování, kliknutí na odkaz a navigace.

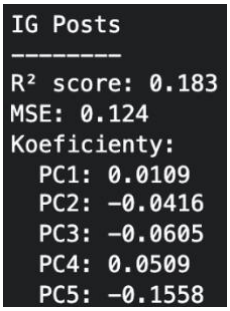
Grafické znázornění modelu pro PCA Instagram Stories (obrázek 25) ukazuje srovnání mezi skutečnými a predikovanými hodnotami tržeb. Přestože model nevysvětluje veškerou variabilitu dat, patrné je určité soustředění bodů okolo diagonální čáry, což potvrzuje existenci lineárního vztahu mezi komponentami a tržbou, a tedy i částečnou predikční schopnost tohoto modelu.




**Obrázek 25:** Porovnání skutečné a predikované tržby pro datovou sadu Instagram Stories

Zdroj: Vlastní zpracování

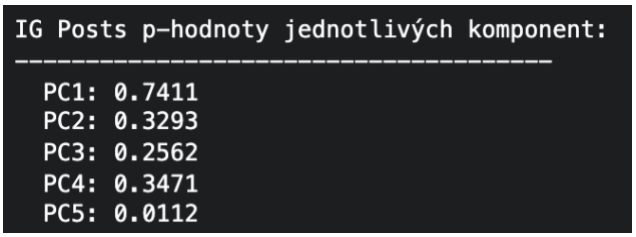
Vícenásobné lineární regrese byla také aplikována na datovou sadu PCA Instagramových příspěvků. Výsledky na obrázku 26Obrázek 26 ukazují nižší hodnotu koeficientu determinace než v případě Stories, konkrétně $R^2 = 0.183$. To znamená, že pouze 18.3 % variability tržeb lze vysvětlit pomocí hlavních komponent získaných metodou PCA. Hodnota MSE (průměrná kvadratická odchylka predikovaných hodnot od skutečných) dosáhla hodnoty 0.124. Tato hodnota je ovlivněna nižší variabilitou tržeb u této datové sady.




**Obrázek 26:** Výsledky vícenásobné lineární regrese pro datovou sadu Instagramové příspěvky

Zdroj: Vlastní zpracování

*P*-hodnoty pro datovou sadu PCA Instagramových příspěvků jsou zobrazeny níže na obrázku 27.




**Obrázek 27:** *p*-hodnoty pro datovou sadu Instagram příspěvky

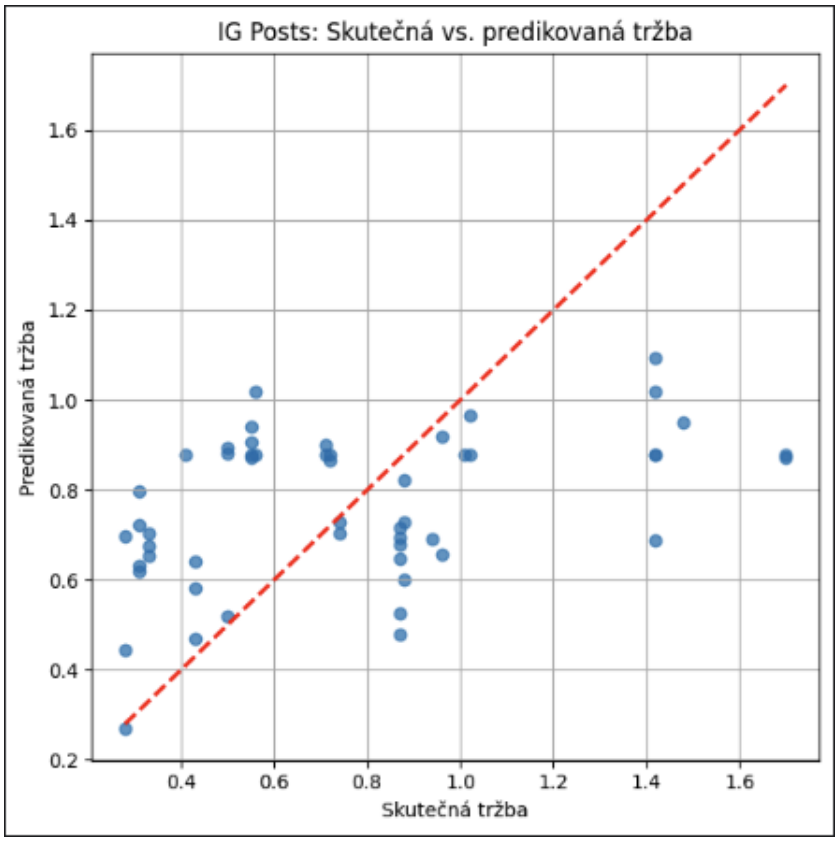
Zdroj: Vlastní zpracování

Z hlediska jednotlivých hlavních komponent:

- PC1 vykazuje velmi slabý pozitivní vliv na tržby (koeficient: +0.0109), ale tento vztah není statisticky významný ($p$ = 0,7411). Komponenta je tvořena zejména metrikami dosah, to se mi líbí a sledující.
- PC2 má záporný vztah k tržbám (koeficient: -0.0416), avšak statisticky není významná ($p$ = 0.3293). Největší váhu zde mají metriky délka videa, uložení a komentáře.
- PC3 rovněž vykazuje negativní vliv (koeficient: -0.0605), přičemž ani tento vztah není statisticky významný ($p$ = 0.2562). Tato komponenta je ovlivněna metrikami sdílení, délka a zásah.

- PC4 přispívá mírně pozitivně (koeficient: +0.0509), ale statisticky nevýznamně ($p$ = 0.3471). Dominují jí metriky sledující, délka videa a komentáře.
- PC5 je nejsilnějším negativním faktorem (koeficient: -0,1558) a zároveň jediná komponenta, která je statisticky významná ($p$ = 0,0112). Formují ji zejména metriky uložení, to se mi líbí a dosah.

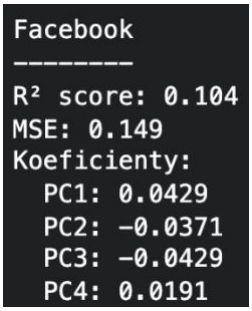
Grafické znázornění modelu vícenásobní lineární regrese pro datovou sadu PCA Instagramové příspěvky skutečná vs. predikovaná tržba (obrázek 28) ukazuje větší rozptyl bodů okolo diagonály, přičemž některé predikce výrazně podhodnocují či nadhodnocují skutečné hodnoty. Model tedy zachycuje obecný trend, avšak predikční přesnost je nižší.




**Obrázek 28:** Porovnání skutečné a predikované tržby pro datovou sadu Instagram příspěvky

Zdroj: Vlastní zpracování

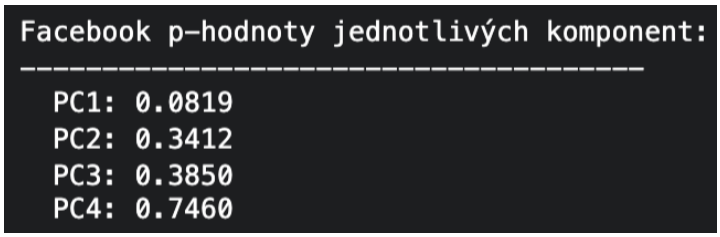
Vícenásobná lineární regrese byla také aplikována na datovou sadu PCA Facebookové příspěvky. Výsledky (obrázek 29) vykazují nejnižší hodnotu koeficientu determinace ze všech tří zkoumaných případů, konkrétně $R^2 = 0.104$. To znamená, že pouhých 10.4 % variability v hodnotách tržeb lze vysvětlit na základě hlavních komponent získaných metodou PCA. Hodnota MSE (průměrná kvadratická chyba predikovaných hodnot od skutečných) dosáhla hodnoty 0.149, což je opět dáno nižším rozptylem v původních datech.




**Obrázek 29:** Výsledky vícenásobné lineární regrese pro datovou sadu Facebook příspěvky

Zdroj: Vlastní zpracování

P-hodnoty pro datovou sadu PCA Facebook příspěvky jsou zobrazeny níže (obrázek 30).



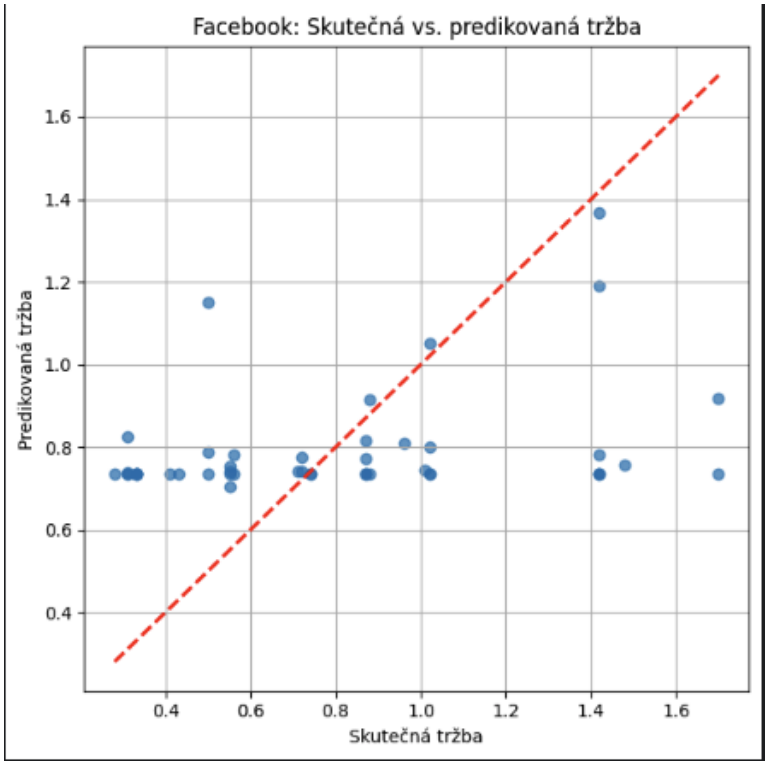

**Obrázek 30:** *p*-hodnoty pro datovou sadu Facebook příspěvky

Zdroj: Vlastní zpracování

Z hlediska jednotlivých hlavních komponent:

- PC1 vykazuje pozitivní vliv na tržby (koeficient: +0.0429), přičemž tento vztah je statisticky významný pouze na hladině $\alpha = 0.10$. Komponenta je tvořena především metrikami reakce, celkem kliknutí a jiná kliknutí.
- PC2 má záporný vztah k tržbám (koeficient: -0.0371), přičemž statistická významnost nebyla potvrzena ($p = 0.3412$). Největší váhu zde mají metriky komentáře, uložení a délka sledování.
- PC3 rovněž vykazuje negativní vliv (koeficient: -0.0429), který není statisticky významný ($p = 0.3850$). Tato komponenta je formována především metrikou sdílení.
- PC4 vykazuje slabý pozitivní vliv (koeficient: +0.0191), ale bez statistické významnosti ($p = 0.7460$). Nejvíce ji ovlivňují metriky délka videa a zásah (reach).

Z grafického znázornění modelu vícenásobné lineární regrese (obrázek 31) pro datovou množinu facebookové příspěvky skutečných a predikovaných hodnot je patrné, že body jsou značně rozptýlené podél diagonály a mnohé predikce nedokážou přesně vystihnout reálné hodnoty.




**Obrázek 31**: Porovnání skutečné a predikované tržby pro datovou sadu Facebook příspěvky

Zdroj: Vlastní zpracování

Je třeba podotknout, že v rámci této fáze nebyla data rozdělena na trénovací a testovací množinu. Cílem zde nebylo vytvořit prediktivní model pro odhad nových hodnot, ale spíše ověřit, zda existuje významná lineární souvislost mezi hlavními komponentami a tržbou. Výsledky tedy slouží především k orientačnímu vyhodnocení síly vztahu, nikoliv k přesné predikci.

K vyhodnocení vztahu mezi hlavními komponentami a výstupní proměnnou byly využity tři základní metriky. Koeficient determinace $R^2$ sloužil k posouzení, jak velkou část variability v hodnotách tržeb dokáže regresní model vysvětlit. Střední kvadratická chyba (MSE) poskytla informaci o průměrné odchylce mezi skutečnými a predikovanými hodnotami, čímž napomohla zhodnotit přesnost modelu v rámci použitého datového rozsahu. P-hodnoty regresních koeficientů pak umožnily identifikovat, které z hlavních komponent měly statisticky významný vliv na výstupní proměnnou. Vzájemná kombinace těchto ukazatelů tak umožnila komplexní zhodnocení síly i relevance lineárních vztahů v rámci jednotlivých modelů.

### 4.4.5 Analýza časových řad pomocí modelu Prophet

Model Prophet byl vyvinut výzkumným týmem Core Data Science společnosti Facebook jako nástroj pro predikci časových řad, který kombinuje vysokou přesnost, interpretovatelnost a uživatelskou přívětivost. Jeho hlavní výhodou je schopnost vytvářet kvalitní predikční modely i bez hlubších znalostí statistiky či programování, bez nutnosti složitého ladění parametrů (Shakeel a kol., 2023). Model je zároveň navržen s důrazem na flexibilitu, neboť si dokáže poradit s výraznými trendovými změnami, sezónními výkyvy i výskytem odlehlých hodnot (Aditya Satrio a kol., 2021).

Dle Shakeel a kol. (2023) základní koncept modelu vychází z aditivního rozkladu časové řady, který předpokládá, že každou hodnotu v čase lze vyjádřit jako součet několika dílčích složek:

$$y(t) = g(t) + s(t) + h(t) + \varepsilon, \tag{11}$$

kde $g(t)$ představuje dlouhodobý trend vývoje hodnoty v čase, $s(t)$ zachycuje sezónní variace, tedy opakující se cykly, $h(t)$ vyjadřuje vliv mimořádných událostí či specifických svátků, které mohou mít výrazný dopad na výsledné hodnoty a $\varepsilon$ je reziduální složka představující náhodnou chybu nebo nepozorované vlivy.

Model Prophet je založen na nelineárním modelu saturačního růstu a po částech (piecewise) lineární modelové funkci. Logistický růstový model v jeho základní formě je používán pro modelování nelineárního trendu a je vyjádřen rovnicí:

$$g(t) = \frac{C}{1+\exp\left(-k(t-m)\right)}, \tag{12}$$

kde $C$ představuje kapacitní limit, $k$ je míra růstu a $m$ je parametr, který určuje bod zlomu.

Když se proměnná míra růstu $k$ mění v čase, je třeba zároveň upravit i proměnnou $m$, aby bylo možné spojit různé fáze růstu do jedné spojité křivky. Modifikovaná rovnice logistického růstu pak vypadá následovně (Shakeel a kol., 2023):

$$g(t) = \frac{C_t}{1+\exp\left(-[k+a(t)^T\delta)-(t-(m+(t)^T\gamma))]\right)}, \tag{13}$$

kde $a(t)$ je vektor, který reprezentuje změny růstu v čase, $\delta$ značí velikost těchto změn (tzv. rate change) a $\gamma$ je korekční vektor pro úpravu pozice zlomu.

Míra růstu se tedy může měnit v předem definovaných bodech (tzv. changepoints), které odpovídají například sezónním výkyvům nebo vlivu kampaní. Trendová složka modelu se pak obecně zapisuje jako:

$$g(t) = [(k + a(t)^T\delta)(t - \left(m + a(t)^T\gamma\right)\big)]. \quad (14)$$

Sezónní složka *s(t)*, která zachycuje pravidelně se opakující efekty, jako jsou dny v týdnu nebo měsíce v roce, je modelována pomocí Fourierovy řady. Ta se vyjadřuje následovně (Shakeel a kol., 2023):

$$s(t) = \sum_{n=1}^{N}(a_n \cos(\frac{2n\pi t}{P}) + b_n \sin(\frac{2n\pi t}{P})), \quad (15)$$

kde $N$ je počet Fourierových párů (čím vyšší, tím detailnější sezónní efekt),$P$ je perioda sezónnosti (např. 7 pro týden, 365.25 pro rok) a $a_n, b_n$ jsou váhové koeficienty pro jednotlivé harmonické složky.

Velkou předností modelu Prophet je jeho robustnost vůči výpadkům v datech a odolnost vůči výkyvům, díky čemuž je vhodný pro aplikaci na reálné datové sady, které často obsahují šum, chybějící hodnoty či nepravidelnosti. Prophet navíc automaticky detekuje a optimalizuje tzv. body zlomu (changepoints), což umožňuje lepší přizpůsobení trendu náhlým změnám ve vývoji časové řady (Aditya Satrio a kol., 2021).

Jeho použití je založeno na jednoduché struktuře vstupních dat, kdy je třeba připravit datový rámec obsahující sloupce *ds* (datum) a *y* (hodnota, kterou chceme predikovat). Po inicializaci modelu a jeho natrénování na historických datech lze snadno vygenerovat predikce pro zvolené časové období. Výstupní hodnoty jsou obsaženy v proměnné $\hat{y}$, která reprezentuje predikované hodnoty pro dané časové značky (*ds*). Tyto výstupy lze dále vizualizovat a analyzovat například z hlediska trendů nebo sezónních vzorců (Shakeel a kol., 2023).

#### 4.4.5.1 Aplikace modelu Prophet

Na základě teoretických poznatků o modelu Prophet, které byly uvedeny v předchozí části, byl tento model následně aplikován i na reálná data firmy Lyžebraní. Cílem však nebyla predikce budoucího vývoje tržeb, ale především snaha popsat stávající situaci a identifikovat případné sezónní vzorce chování zákazníků na základě historických dat. Kompletní kód pro aplikaci modelu Prophet je uveden v příloze F na příkladu datové sady Instagram příspěvky. Stejným způsobem byla analýza sezónnosti provedena také na datech z Facebook příspěvků a Instagram Stories.

Do modelu vstupovala vždy dvojice proměnných: datum zveřejnění marketingového příspěvku (sloupec *ds*) a odpovídající hodnota tržby (sloupec *y*). Ukázka vstupu proměnných je zobrazena níže na obrázku 32.

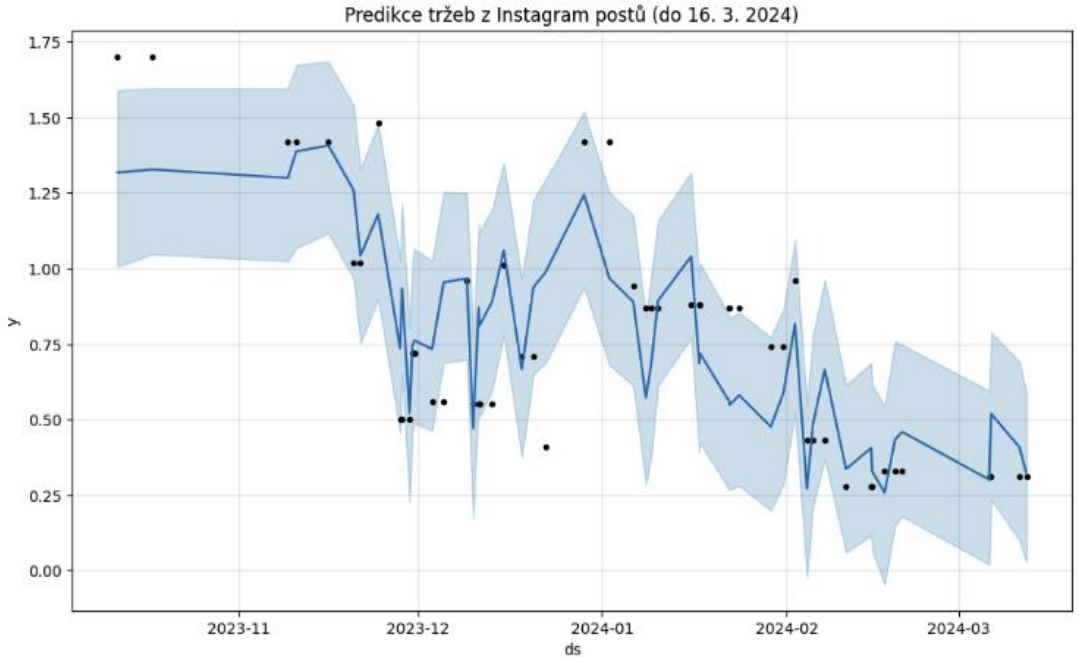
```
df_prophet = df[['Čas zveřejnění', 'tržba']].copy()
df_prophet.columns = ['ds', 'y']
```

**Obrázek 32:** Vstupní proměnné pro model Prophet

Zdroj: Vlastní zpracování

Hodnota tržby byla připojena ke každému příspěvku na základě nejbližšího následujícího prodejního dne, tedy dne, kdy měl daný příspěvek potenciálně největší dopad na skutečný nákupní výkon zákazníků. Tento přístup umožnil vytvořit časovou řadu, ve které bylo možné sledovat vývoj tržeb v kontextu jednotlivých marketingových aktivit.

Model Prophet byl aplikován na datovou sadu obsahující Instagramové příspěvky, přičemž výsledkem je graf časového vývoje tržeb (obrázek 33). Modrá křivka představuje odhadovanou hodnotu tržby v čase, zatímco černé body znázorňují skutečně dosažené tržby přiřazené k jednotlivým příspěvkům. Modře stínovaná oblast pak označuje interval spolehlivosti modelu, ve kterém se predikovaná hodnota s vysokou pravděpodobností nachází.

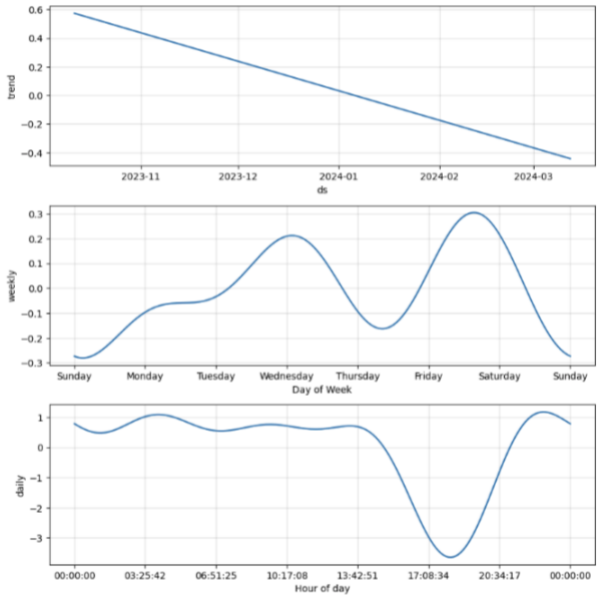



**Obrázek 33:** Časový vývoj tržeb z datové sady Instagram příspěvky

Zdroj: Vlastní zpracování

Z celkového průběhu je patrný postupný pokles tržeb směrem ke konci analyzovaného období, což potvrzuje sezónní charakter poptávky po zimním sortimentu.

Na základě výstupů modelu Prophet lze detailně analyzovat jednotlivé složky ovlivňující vývoj tržeb v čase. Jednotlivé složky jsou zobrazeny na obrázek 34 níže.




**Obrázek 34:** Trend, týdenní sezonní složka, denní sezónnost pro datovou sadu Instagram příspěvky

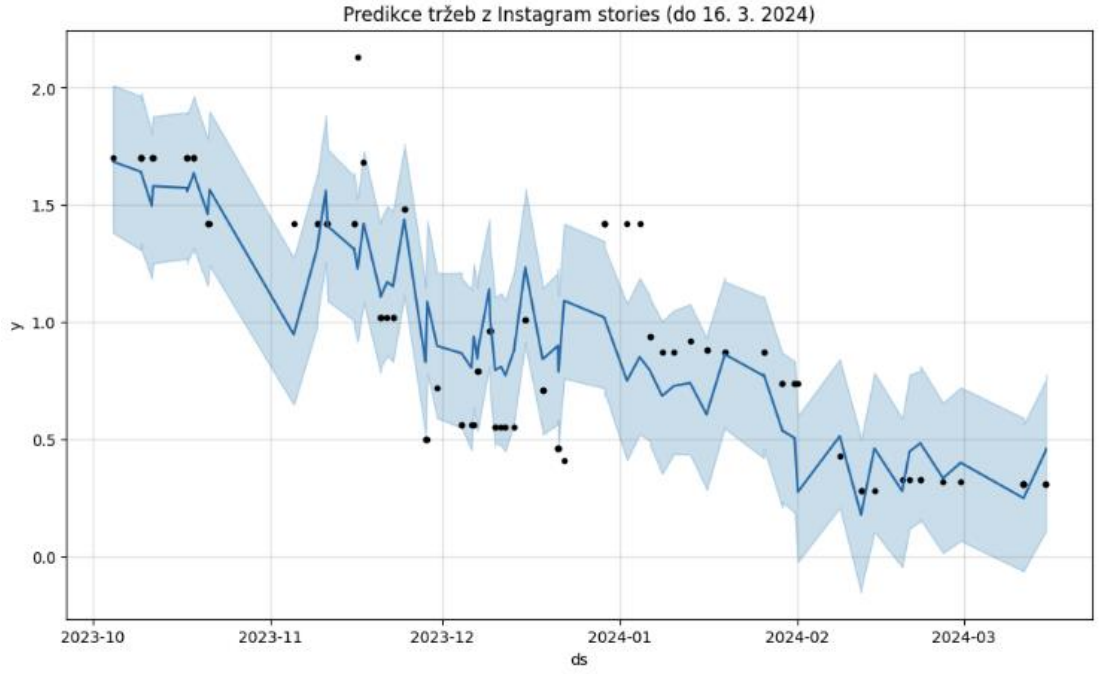
Zdroj: Vlastní zpracování

V prvním grafu je zobrazen celkový trend, který ukazuje na postupný pokles tržeb v průběhu celé sledované sezóny. Tento vývoj odpovídá reálnému chování zákazníků, jelikož nejvyšší zájem tradičně přichází na začátku sezóny (říjen a listopad), zatímco s příchodem března poptávka slábne.

Druhý graf představuje týdenní sezónní složku, která zachycuje systematické rozdíly v tržbách podle dnů v týdnu. Nejvyšších hodnot sezónní složky je dosaženo v pátek a sobotu, což odpovídá skutečným prodejním dnům firmy Lyžebraní, které podle dat nejčastěji připadaly právě na tyto dny. Také neděle vykazuje zvýšené hodnoty, což koresponduje s její rolí jako častého prodejního dne. Naopak pondělí až středa vykazují výrazně nižší hodnoty, což je v souladu s tím, že v těchto dnech se prodej prakticky nerealizoval.

Třetí graf zobrazuje denní sezónnost, tedy vzorce kolísání v rámci jednotlivých hodin dne. Nejvýraznější poklesy jsou patrné v odpoledních a večerních hodinách kolem 17. až 20. hodiny, což může reflektovat pokles aktivity nebo zájmu uživatelů v této části dne. Naopak dopolední a noční hodiny vykazují mírně zvýšenou složku, což může naznačovat, kdy uživatelé na příspěvky nejčastěji reagovali.

Na základě aplikace modelu Prophet na původní datovou sadu obsahující Instagram Stories bylo možné detailně analyzovat časový vývoj tržeb a identifikovat sezónní vzorce chování zákazníků. Časový vývoj tržeb je zobrazen na obrázek 35 níže.

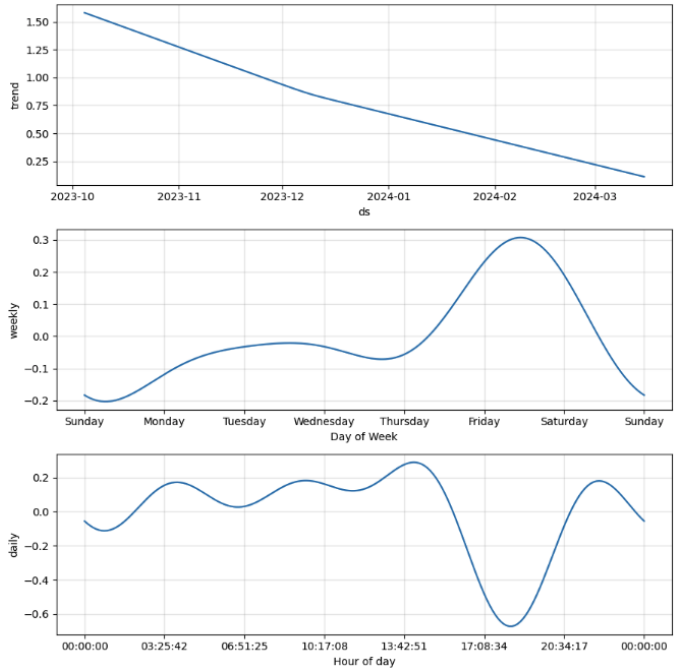



**Obrázek 35:** Časový vývoj tržeb z datové sady Instagram Stories

Zdroj: Vlastní zpracování

Z grafu (obrázek 35) je patrné, že tržby měly na začátku sezóny (říjen a listopad 2023) vysoké hodnoty, přičemž některé tržby přesahovaly i hodnotu 1.5. Postupem času však docházelo k pozvolnému poklesu, přičemž od ledna 2024 je patrný setrvalý trend snižující se efektivity tohoto formátu. Nejnižší hodnoty tržeb byly zaznamenány v únoru a březnu 2024, což může souviset s přirozeným útlumem poptávky po skončení hlavní zimní sezóny.

Z výstupů modelu Prophet lze opět detailně analyzovat jednotlivé složky datové sady Instagram Stories, které ovlivňovaly vývoj tržeb v čase. Tento vývoje je zobrazen na obrázku 36 níže.




**Obrázek 36:** Trend, týdenní sezonní složka, denní sezónnost pro datovou sadu Instagram Stories

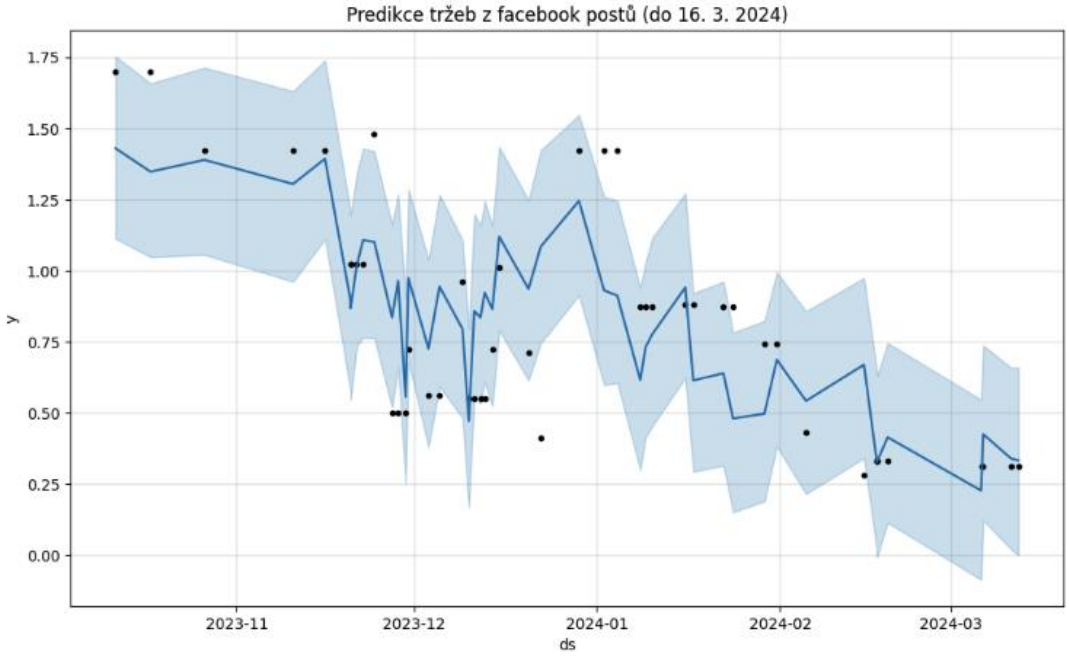
Zdroj: Vlastní zpracování

První graf zachycuje celkový trend tržeb, který má jednoznačně klesající tendenci. Nejvyšší hodnoty tržeb byly dosaženy na začátku sledovaného období, tj. v říjnu a listopadu 2023, což odpovídá zahájení prodejní sezóny. Postupně však dochází k výraznému poklesu, který trvá až do konce sezóny v březnu 2024. Tento vývoj odpovídá realitě firmy Lyžebraní, kdy poptávka zpravidla slábne s přibývajícími měsíci sezóny.

Druhý graf vizualizuje týdenní sezónní složku a ukazuje, že nejvyšší sezónní vliv tržeb připadá na pátky a soboty, tedy hlavní prodejní dny firmy Lyžebraní, což bylo potvrzeno i ve vstupní datové sadě obsahujícím statistiku reálných prodejních dnů. Zvýšené hodnoty jsou patrné také u nedělí, zatímco pondělí až středa vykazují slabé nebo negativní vlivy, což odpovídá faktu, že v těchto dnech se prodej nerealizoval.

Třetí graf zobrazuje denní sezónnost, tedy systematické výkyvy v rámci jednotlivých hodin dne. Nejnižší hodnota je zaznamenána mezi 17. a 20. hodinou, což může souviset s poklesem uživatelské aktivity nebo nižší efektivitou sdíleného obsahu v tomto čase. Naopak mírně zvýšená aktivita je patrná v ranních, dopoledních a pozdně večerních hodinách, což může naznačovat, kdy uživatelé Stories nejčastěji sledují a kdy může být sdílený obsah nejefektivnější.

Model Prophet byl aplikován i na původní datovou sadu obsahující marketingové příspěvky publikované na Facebooku. Z výsledků modelu (obrázek 37), bylo možné zhodnotit vývoj tržeb v čase a analyzovat přítomnost sezónních vzorců chování zákazníků.

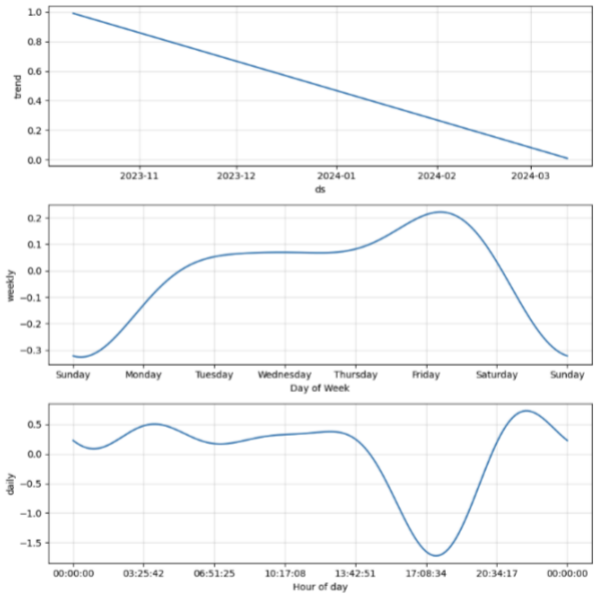



**Obrázek 37:** Časový vývoj tržeb z datové sady Facebook příspěvky

Zdroj: Vlastní zpracování

Z grafu je patrné, že tržby dosahovaly nejvyšších hodnot v říjnu a listopadu 2023, tedy v období zahájení prodejní sezóny. Některé příspěvky vykázaly tržby převyšující hodnotu 1.5, což potvrzuje jejich vysokou efektivitu na začátku kampaně. V prosinci již docházelo k mírnému útlumu, který se postupně prohluboval, přičemž od ledna 2024 lze sledovat setrvalý trend poklesu. Nejnižších hodnot bylo dosaženo v únoru a březnu 2024, kdy většina tržba generovala nižší hodnoty než 0.5. Tento vývoj odpovídá přirozenému sezónnímu cyklu a naznačuje, že efektivita Facebookového obsahu ve vztahu k tržbám klesala s přibývajícím časem sezóny.

Z výstupů modelu Prophet lze opět detailně analyzovat jednotlivé složky, které ovlivňovaly vývoj tržeb v čase. Jednotlivé složky jsou zobrazeny na obrázku 38 níže.




**Obrázek 38:** Trend, týdenní sezonní složka, denní sezónnost pro datovou sadu Facebook příspěvky

Zdroj: Vlastní zpracování

První graf zachycuje celkový trend vývoje tržeb, které mají jednoznačně klesající charakter, s nejvyššími hodnotami v říjnu a listopadu 2023 a s trvalým poklesem až do března 2024. Tento vývoj věrně odráží realitu sezónního podnikání firmy Lyžebraní, kdy poptávka po zimním sortimentu s postupem času výrazně slábne.

Druhý graf vizualizuje týdenní sezónní složku. Nejvyšší hodnoty této složky připadly na pátky a soboty, tedy dny, které byly nejčastěji využívány jako hlavní prodejní termíny, jak vyplývá ze vstupních dat. Naopak nejnižší hodnoty byly zaznamenány na začátku týdne, kdy se žádné prodeje neuskutečňovaly.

Třetí graf zachycuje denní sezónnost. Zde je patrný znatelný pokles tržeb v odpoledních a podvečerních hodinách, zejména mezi 17. a 20. hodinou. Tento časový úsek může být méně vhodný pro zveřejňování příspěvků z důvodu snížené uživatelské aktivity nebo nízké ochoty reagovat na obsah. Naopak zvýšené hodnoty sezónní složky jsou patrné v brzkých ranních a pozdních večerních hodinách, což naznačuje, kdy může být publikace obsahu nejefektivnější z hlediska oslovení publika a potenciálního vlivu na tržby.

V návaznosti na aplikaci modelu Prophet pro jednotlivé datové sady (Facebook, Instagramové příspěvky a Instagram Stories) byla kromě vizuální interpretace jednotlivých složek modelu (trend, sezónnost) provedena také kvantitativní analýza přesnosti modelu pomocí standardních regresních metrik. Cílem této doplňující analýzy bylo zjistit, do jaké míry model Prophet dokáže zpětně popsat vývoj tržeb na základě známých historických dat a zda vykazuje vyšší vysvětlovací schopnost ve srovnání s vícenásobnou lineární regresí aplikovanou v předchozí části práce.

Pro výpočet regresních metrik byl využit kód v programovacím jazyce Python, který vyčíslil čtyři klíčové ukazatele výkonnosti: $R^2$ (koeficient determinace), RMSE (Root Mean Squared Error), MAE (Mean Absolute Error) a MAPE (Mean Absolute Percentage Error). Výsledky těchto výpočtů jsou shrnuty v tabulce níže (obrázek 39).

📈 Regresní metriky pro Prophet predikce:

| | Zdroj | R² | RMSE | MAE | MAPE (%) |
|---|---|---|---|---|---|
| 0 | Facebook | 0.622 | 0.25 | 0.19 | 29.36 |
| 1 | Instagram Posty | 0.669 | 0.22 | 0.18 | 27.75 |
| 2 | Instagram Stories | 0.758 | 0.25 | 0.19 | 27.80 |

**Obrázek 39:** Regresní metriky pro model Prophet

Zdroj: Vlastní zpracování)

Na základě dosažených výsledků lze konstatovat, že model Prophet vykázal výrazně vyšší schopnost vysvětlit variabilitu tržeb oproti dříve aplikované vícenásobné lineární regresi. Nejvyšší hodnotu koeficientu determinace ($R^2$) model dosáhl u datové sady Instagram Stories ($R^2$ = 0.758), následované Instagramovými příspěvky ($R^2$ = 0.669) a příspěvky z Facebooku ($R^2$ = 0.622). Tyto hodnoty naznačují, že Prophet dokázal poměrně přesně vystihnout vývoj tržeb na základě historických dat a identifikovaných sezónních vzorců. Také hodnoty chybových metrik RMSE a MAE byly ve všech případech nízké a vzájemně srovnatelné, což potvrzuje relativně malý rozptyl mezi predikovanými a skutečnými hodnotami tržeb. Míra průměrné absolutní procentuální chyby (MAPE) se pohybovala v rozmezí 27.75 % až 29.36 %, což sice svědčí o nižší predikční přesnosti, avšak vzhledem k charakteru marketingových dat zůstává tato hodnota v akceptovatelném rozmezí.

### 4.4.6 Analýza sentimentu

Analýza sentimentu, někdy označovaná také jako opinion mining, je metoda, která se zabývá automatickým rozpoznáváním a klasifikací názorů, postojů či emocí obsažených v přirozeném jazyce. Jejím cílem je zjistit, zda text vyjadřuje pozitivní, negativní nebo neutrální postoj, případně určit další související prvky (Liu, 2012).

V současnosti nabývá tato metoda na významu především v prostředí sociálních sítí, kde uživatelé spontánně a ve velkém množství publikují své názory, zkušenosti a doporučení. Textový obsah generovaný uživateli jako jsou komentáře, příspěvky či recenze se stal cenným zdrojem informací pro firmy i výzkumníky. Analýza sentimentu zde slouží jako nástroj pro zachycení celkové nálady veřejnosti vůči určité značce či produktu a zároveň umožňuje identifikovat hlubší postoje a emoce v digitální komunikaci (Agüero-Torales a kol., 2019).

Sentiment může být vyhodnocován na různých úrovních od celkového hodnocení celého dokumentu, přes analýzu jednotlivých vět, až po detailní rozbor postojů k jednotlivým aspektům určité entity (Lv a kol., 2021).

Jádrem analýzy sentimentu jsou tzv. sentimentová slova, které nesou určité hodnotící zabarvení (např. výborný, zklamání, nepříjemný). Tato slova bývají obvykle součástí tzv. lexikonů sentimentu, tedy seznamů slov, které umožňují přiřadit danému textu odpovídající hodnotu polarity. Samotná přítomnost těchto slov však pro správnou interpretaci často nestačí, jelikož analýza musí zohlednit i jazykový kontext, například výskyt negací, ironie nebo sarkasmu (Liu, 2012).

Metody založené na lexikonech představují přístup, kdy jsou jednotlivým slovům přiřazeny hodnoty polarity a subjektivity na základě jejich jazykového významu. Tento princip využívá také knihovna TextBlob, která je založena na pravidlech a předem sestaveném slovníku. Každému textu přiřazuje skóre polarity v rozsahu od –1 (silně negativní) po +1 (silně pozitivní) a zároveň určuje míru subjektivity vyjádřeného sdělení (Aljedaani a kol., 2022).

Lexikonový přístup se dle Liu (2012) skládá z následujících čtyř kroků:

1. Označení sentimentových slov a frází: Každé kladné slovo získá skóre +1, každé záporné –1.
2. Aplikace sentimentových posunovačů (shifters): Např. „not“ změní význam na záporný. Věta se tak změní na „not good [-1]“.
3. Zpracování větných spojek typu „but“: Věty obsahující „ale“ obvykle vyjadřují kontrastní sentiment. Význam před a po spojce je často opačný.
4. Agregace sentimentu k aspektům: Výsledné skóre se spočítá pomocí této rovnice:

$$score(a, s) = \sum_{ow_j \in S} \frac{sw_j.so}{dist(sw_j, a_i)}, \quad (16)$$

kde $a_i$ je analyzovaný aspekt ve větě $s$, $ow_j$ je názorové (sentimentové) slovo ve větě $s$, $w_j.so$ je hodnota skóre sentimentu slova $sw_j$ a $dist(sw_j, a_i)$ je vzdálenost mezi slovem $sw_j$ a aspektem $a_i$ve větě $s$.

Čím je sentimentové slovo dál od aspektu, tím menší váhu ve výpočtu má. Pokud je výsledné skóre kladné, je sentiment k aspektu pozitivní, pokud záporné, pak negativní, jinak neutrální.

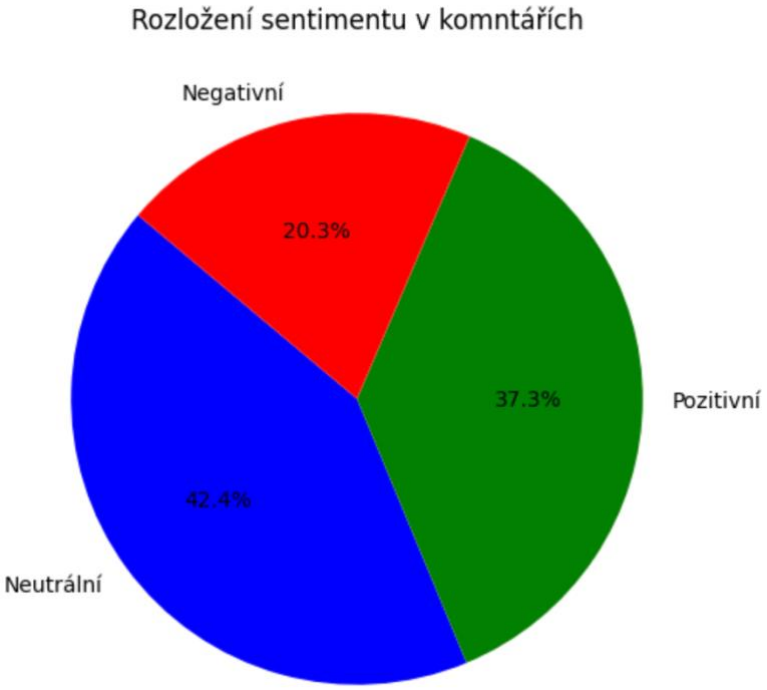
Vedle základní analýzy polarity lze z textových dat dále získat i hlubší vhled do toho, jaká témata se v komentářích objevují. Tematická analýza pomocí metody Latent Dirichlet Allocation (LDA) umožňuje identifikovat latentní témata na základě shlukování slov, která se v textech často vyskytují společně. LDA tak zároveň plní funkci objevování klíčových pojmů i jejich automatického seskupování do tematických okruhů (Liu, 2012).

#### 4.4.6.1 Aplikace analýzy sentimentu

Na základě výše uvedeného teoretického rámce byla provedena analýza sentimentu komentářů k příspěvkům pro firmu Lyžebraní. Cílem bylo zjistit, jaké emocionální ladění jednotlivé komentáře nesou a zda převažuje spíše pozitivní, negativní či neutrální hodnocení.

Pro zpracování dat byla využita knihovna TextBlob v jazyce Python, která implementuje lexikonový přístup k analýze sentimentu. Každému komentáři bylo přiřazeno skóre sentimentu (polarity score) v rozmezí od –1 do +1, přičemž záporné hodnoty indikují negativní sentiment, kladné hodnoty značí pozitivní ladění, hodnoty blízké nule odpovídají neutrálnímu tónu sdělení. Kompletní kód pro analýzu sentimentu je obsažen v příloze G.

Rozložení sentimentu v komentářích je zobrazeno na obrázku 40 níže. Z koláčového grafu je patrné, že největší podíl tvoří neutrální komentáře (42.4 %), následované pozitivními (37.3 %) a negativními (20.3 %).




**Obrázek 40:** Rozložení sentimentu v komentářích

Zdroj: Vlastní zpracování

Dalším krokem analýzy textu bylo určení nejčastěji se vyskytujících klíčových slov v komentářích. Texty komentářů byly nejprve vyčištěny od běžných anglických stop-slov (např. „the“, „is“, „and“) a také od často se opakujících výrazů bez informační hodnoty (např. „hello“, „thank“, „also“). Poté byla vypočtena frekvence slov, přičemž byla identifikována pětice nejčastějších výrazů, která jsou uvedena na obrázku 41 níže.

```
5 nejčastějších klíčových slov v komentářích:
 - skis (16x)
 - secondhand (8x)
 - shoes (6x)
 - boots (6x)
 - please (5x)
```

**Obrázek 41:** Nejčastější slova v komentářích

Zdroj: Vlastní zpracování

Z přehledu klíčových slov vyplývá, že nejčastěji se v komentářích objevovalo slovo „skis“ (lyže), následované výrazy „secondhand“ (bazarové zboží) a „shoes“ (lyžáky). Z toho lze usoudit, že hlavními tématy dotazů a reakcí uživatelů byly lyže, bazarové vybavení a lyžařská obuv. Výskyt slova „please“ (prosím) dále naznačuje, že řada komentářů měla formu dotazu nebo žádosti, a tedy že se uživatelé aktivně zajímali o konkrétní nabídku zboží.

Pro hlubší porozumění obsahu komentářů byla provedena tematická analýza pomocí metody LDA (obrázek 42). Tato metoda umožňuje na základě spolu výskytu slov identifikovat skrytá témata, která se v textech opakují. Model byl nastaven tak, aby identifikoval tři hlavní témata, přičemž pro každé z nich byla uvedena trojice klíčových slov, která dané téma nejvíce charakterizují.

```
3 hlavní témata komentářů (po 3 klíčových slovech):
 - Topic 1: price, cm, secondhand
 - Topic 2: skis, thanks, cm
 - Topic 3: skis, shoes, ski
```

**Obrázek 42:** Hlavní témata komentářů

Zdroj: Vlastní zpracování

První téma se vztahuje k ceně a velikosti zboží (price, cm, secondhand), což může odkazovat na dotazy ohledně velikostí bazarového vybavení a jeho cenové dostupnosti. Druhé téma (skis, thanks, cm) naznačuje zájem o konkrétní lyže a potvrzuje častou interakci formou dotazů a poděkování. Třetí téma (skis, shoes, ski) se zaměřuje na samotné lyže a lyžařskou obuv, což potvrzuje hlavní zaměření komentářů na konkrétní typy zboží.

## 4.5 Hodnocení

Fáze hodnocení představuje klíčový moment v procesu CRISP-DM, kdy se výsledky modelování hodnotí ve vztahu k původnímu zadání a očekáváním. V této fázi nejde pouze o technickou analýzu přesnosti modelů, ale především o posouzení toho, zda výstupy skutečně odpovídají potřebám podniku. Hodnocení tedy zajišťuje propojení mezi analytickými zjištěními a reálnými obchodními cíli (Chumbar, 2023).

Hlavním obchodním cílem bylo analyzovat účinnost marketingových kampaní realizovaných na sociálních sítích Facebook a Instagram ve prospěch firmy Lyžebraní. Konkrétně šlo o vyhodnocení toho, jak jednotlivé příspěvky, formáty a metriky zapojení uživatelů přispívají k obchodním výsledkům, zejména k výši dosažené tržby.

#### 4.5.1.1 Korelační analýza

Korelační analýza byla aplikována na všechny tři datové sady, tedy na datové sady z Facebooku, Instagramu a Instagram Stories. Jejím cílem bylo identifikovat vztahy mezi metrikami uživatelského zapojení a obchodními výsledky, zejména tržbou. Ve všech případech se potvrdila velmi silná korelace mezi počtem registrovaných zákazníků a tržbou, což je z hlediska výkladu účinnosti očekávané. Analýza zároveň odhalila vysoké korelace mezi některými metrikami zapojení (např. komentáře, sdílení, reakce), což poukázalo na výskyt multikolinearity. Tento poznatek se stal podnětem pro další fázi, kde došlo k redukci dimenze dat pomocí metody PCA. Celkově korelační analýza poskytla první orientaci v datech a pomohla odhalit vzorce související s výkonností příspěvků.

#### 4.5.1.2 Metoda hlavních komponent (PCA)

Metoda PCA byla použita pro všechny tři datové sady s cílem redukovat dimenzi a odstranit problém multikolinearity mezi jednotlivými metrikami. Pro každou datovou sadu bylo na základě vysvětlené variance vybráno 4–5 hlavních komponent, které zachycovaly převážnou většinu informací. Výsledné komponenty umožnily transformovat původní metriky zapojení do nových, vzájemně nezávislých veličin, čímž se vytvořil vhodný základ pro další analýzy. Zároveň bylo možné z PCA získat přehled o tom, které kombinace metrik nejvíce ovlivňují výkonnost příspěvků. Metoda tak přispěla k lepšímu porozumění strukturálním vzorcům v datech a podpořila další fáze hodnocení účinnosti kampaní, jako byla shluková a regresní analýza.

#### 4.5.1.3 Shluková analýza K-means

Shluková analýza pomocí algoritmu K-means byla aplikována na všechny transformované datové sady z PCA s cílem rozdělit příspěvky do skupin na základě jejich podobnosti. Pro každou datovou sadu byl pomocí metody lokte stanoven optimální počet shluků, což umožnilo identifikovat typické vzorce chování v rámci kampaní. Výsledky ukázaly, že jednotlivé shluky se výrazně lišily z hlediska dosažené tržby, což pomohlo odhalit nejvýkonnější typy příspěvků.

Nejúspěšnějším Instagramovým Stories bylo vyhodnoceno krátké video z přípravy prodejní haly před prvním prodejním dnem sezóny. Důležitým faktorem zde bylo především správné načasování publikace, které umožnilo včas informovat zákazníky o blížící se akci a vytvořit očekávání.

Nejúčinnější příspěvek na Instagramu představovalo Reels video, které bylo vytvořeno profesionální reklamní agenturou a výrazně podpořeno placenou propagací. Video s využitím humoru a jednoduchého příběhu zobrazovalo, jak závodník v kompletní výbavě z Lyžebraní běží přes prodejní sklad a je symbolicky „sražen nízkými cenami“. Vizuální zpracování bylo doplněno o popisek ve formě strukturovaného seznamu s emoji, který zdůrazňoval výhody nákupu: výrazné slevy, nové zásoby, široký sortiment a odborné poradenství. Tento příspěvek efektivně spojil kreativní storytelling, produktovou prezentaci a přehledné sdělení klíčových benefitů.

Nejúspěšnější příspěvek na Facebooku kombinoval vizuálně poutavou koláž produktů s výrazným textovým sdělením. Důraz byl kladen na časovou omezenost nabídky, výzvu k akci a přímý odkaz na rezervační systém. Díky kombinaci dobře zvoleného textu a atraktivních fotografií produktů (lyže, helmy, oblečení) dokázal příspěvek přitáhnout pozornost uživatelů a motivovat je ke kliknutí na přiložený odkaz. Tento příspěvek zaznamenal 13 komentářů, což výrazně přispělo k vyšší organické viditelnosti a potvrzuje, že příspěvek rezonoval s cílovou skupinou.

#### 4.5.1.4 Regresní analýza

Na základě dat transformovaných metodou PCA byla provedena vícenásobná lineární regrese. Cílem této analýzy nebylo predikovat budoucí tržby, ale analyzovat současný stav a zjistit, jak velkou část variability dosažené tržby lze vysvětlit pomocí hlavních komponent. Hlavní důraz byl kladen na hodnotu koeficientu determinace $R^2$, která udává míru vysvětlené variability cílové proměnné. Nejvyšší hodnotu koeficientu determinace vykazoval model vytvořený pro datovou sadu Instagram Stories ($R^2 = 0.308$), což naznačuje, že v tomto případě lze pomocí hlavních komponent vysvětlit pouze 30 % variability tržby. O něco nižší hodnoty byly zaznamenány u Instagramových příspěvků ($R^2 = 0.224$), zatímco nejslabší lineární vztah se projevil u Facebookových příspěvků ($R^2 = 0.104$).

Jelikož regresní analýza založená na hlavních komponentách vysvětlila pouze omezenou část variability tržeb (maximálně 30 %), vyvstala otázka, zda některé další faktory jako je například sezónnost, nemohou mít ještě vyšší vysvětlující sílu. Tato úvaha vycházela ze samotné povahy podnikání firmy Lyžebraní, která se zaměřuje na sezónní prodej lyžařského a sportovního vybavení na zimu. Prodejní aktivity probíhají v časově omezeném období (zpravidla od října do března) a soustřeďují se především na víkendové dny, kdy jsou organizovány hlavní akce pro zákazníky. Z tohoto důvodu bylo dále zapotřebí provést analýzu sezónnosti.

#### 4.5.1.5 Analýza sezónnosti pomocí modelu Prophet

Vzhledem k tomu, že regresní analýza založená na PCA komponentách vysvětlila pouze omezenou část variability tržeb, bylo nutné ověřit, zda analýza sezónnosti neposkytne silnější vysvětlení obchodních výsledků.

Za tímto účelem byla provedena analýza časových řad tržeb pomocí modelu Prophet. Model byl aplikován zvlášť na datové sady s příspěvky z Facebooku, Instagramu a Instagram Stories. Pro přímé porovnání jeho výkonnosti s předchozí regresní analýzou na PCA datech, byly na výstupy modelu Prophet aplikovány také regresní metriky ($R^2$, RMSE, MAE, MAPE).

Výsledky jednoznačně ukázaly, že sezónní efekt má silnější vysvětlující sílu než jednotlivé marketingové metriky. Ve všech případech dosáhl model Prophet výrazně vyšší hodnoty koeficientu determinace:

- Instagram Stories: Prophet $R^2 = 0.758$, oproti PCA regresi $R^2 = 0.308$.
- Instagramové příspěvky: Prophet $R^2 = 0.669$, oproti PCA regresi $R^2 = 0.224$.
- Facebookové příspěvky: Prophet $R^2 = 0.622$, oproti PCA regresi $R^2 = 0.104$.

Z výsledků vyplývá, že sezónní rytmus a časové načasování kampaní mají zásadní dopad na tržby, a model Prophet tak poskytl robustnější rámec pro pochopení výkonnosti kampaní v kontextu sezónního podnikání než analýzy založené výhradně na marketingových metrikách.

#### 4.5.1.6 Analýza sentimentu v komentářích

V rámci hodnocení účinnosti marketingových kampaní byla provedena také analýza sentimentu komentářů uživatelů na sociální síti Facebook. Cílem bylo zjistit, jaké postoje zákazníci v diskusích vyjadřují, zda převažuje pozitivní či negativní naladění.

Výsledky ukázaly, že největší podíl tvořily komentáře neutrálního rázu (42.4 %), následovaly pozitivní komentáře (37.3 %) a negativní komentáře tvořily 20.3 % všech analyzovaných případů. Celkově lze konstatovat, že uživatelská zpětná vazba byla převážně pozitivní nebo věcná, což svědčí o dobrém vnímání značky a produktu ze strany zákazníků.

Analýza klíčových slov odhalila, že nejčastěji se v komentářích objevovala slova jako „lyže“, „bazarové“, „lyžáky“, nebo „prosím“, což potvrzuje zaměření dotazů na konkrétní zboží a typicky zákaznické dotazy před nákupem. Tematická analýza pomocí LDA identifikovala tři hlavní témata konverzací. První téma se týkalo cen a původu, druhé se vztahovalo ke konkrétním produktům a třetí se soustředilo opět na sortiment.

Analýza sentimentu poskytla cenný vhled do toho, co zákazníky skutečně zajímá a na co se nejčastěji ptají. Informace o tom, jaké produkty vzbuzují největší zájem, jaká témata se opakují a jak zákazníci formulují své dotazy, mohou sloužit jako důležitý podklad pro tvorbu relevantnějších marketingových sdělení.

## 4.6 Nasazení

Fáze nasazení představuje poslední krok v procesu CRISP-DM, kdy jsou výsledky datové analýzy připraveny k praktickému využití. Nejde jen o předání datových modelů nebo statistik, ale o smysluplné začlenění zjištění do podnikových procesů. Výstupy z analýzy je třeba interpretovat tak, aby byly srozumitelné a použitelné pro rozhodování v marketingu, plánování kampaní nebo optimalizaci komunikace se zákazníky. Efektivní nasazení výsledků zahrnuje jejich přehledné prezentování, případně návrh dalších kroků, které firmě pomohou využít poznatky pro dosažení obchodních cílů (Chumbar, 2023).

Na základě výsledků ze všech aplikovaných analýz lze formulovat následující doporučení pro budoucí marketingové kampaně:

**Zaměřit se na formát Instagram Stories**

Z výsledků vícenásobné lineární regrese i shlukové analýzy K-means vyplývá, že příspěvky publikované ve formátu Instagram Stories mají nejvyšší prediktivní sílu ve vztahu k dosažené tržbě. Tento formát je totiž efektivní komunikaci firmy s cílovou skupinou, zejména v kontextu sezónní kampaně, kdy je potřeba rychle a vizuálně přitažlivě sdělit klíčové informace.

Instagram Stories jsou efektivní, jelikož umožňují využití široké škály interaktivních prvků (tzv. CTA), které nejen podporují zapojení uživatelů, ale zároveň přivádějí sledující k dalším krokům, například přechodu na web, rezervaci termínu nebo interakci s nabídkou. Mezi nejefektivnější CTA prvky patří:

- Ankety a hlasování, které podporují rychlou zpětnou vazbu a zároveň zvyšují míru zapojení sledujících.
- Odkazy typu “Swipe-up” nebo “Odkaz v příběhu”, které umožňují okamžitý přechod na rezervační formulář nebo konkrétní produkt.
- Interaktivní samolepky (např. otázky, odpočítávání), které mohou zvyšovat očekávání nebo upozorňovat na časově omezené akce.

V rámci Instagram Stories je důležité pravidelně publikovat v období hlavní sezóny, ideálně v návaznosti na konkrétní prodejní dny, aby byl obsah časově relevantní. Dále je vhodné testovat různé kombinace CTA prvků a sledovat jejich vliv na míru prokliků a následné rezervace.

**Optimalizovat obsah na Facebooku**

Platforma Facebook představuje pro firmu Lyžebraní zásadní komunikační kanál, neboť zde disponuje nejvyšším počtem sledujících a zároveň oslovuje značnou část své cílové skupiny. Vzhledem k těmto okolnostem je vhodné této platformě věnovat zvýšenou pozornost při plánování a tvorbě marketingového obsahu.

Analýza sentimentu ukázala, že většina komentářů pod příspěvky je neutrálního charakteru, přičemž dominují dotazy související se zbožím, konkrétně s dostupností, velikostí či podmínkami nákupu. Tento poznatek naznačuje, že uživatelé Facebooku využívají komentáře primárně jako komunikační kanál pro získání konkrétních informací, nikoliv pro emocionální reakce na značku.

Na základě těchto zjištění lze doporučit následující opatření pro optimalizování obsahu na platformě Facebook:

- Zvýšit informační hodnotu příspěvků, například doplněním častých dotazů přímo do textu nebo grafiky (FAQ, popisky zboží, odkazy na velikostní tabulky).
- Zajistit včasnou a aktivní správu komentářů, a to jak formou odpovědí, tak případným připnutím důležitých informací.
- Vytvářet příspěvky, které podporují interakci, například prostřednictvím otázek na uživatele, soutěží nebo uživatelských recenzí, čímž se zvýší míra zapojení (engagement).

**Důraz na vizuální atraktivitu a strukturované sdělení**

Nejvýkonnější příspěvky (např. Reels na Instagramu) kombinovaly vizuálně silné prvky, humor, jasně strukturované výhody nákupu a příběh (tzv. storytelling). Kreativní formát v kombinaci se srozumitelným sdělením benefitů vede k vyššímu zapojení uživatelů i pozitivnímu vnímání značky.

**Zavedení monitoringu sezónnosti**

Prediktivní model Prophet potvrdil sezónní vzorec zájmu zákazníků, přičemž vrchol nastával na přelomu listopadu a prosince a následný pokles nastával po Vánocích. S ohledem na tuto skutečnost se doporučuje plánovat intenzitu kampaní tak, aby podpořily hlavní prodejní vlnu, a současně hledat cesty, jak prodloužit zájem i do klidnějších období sezóny.

# 5 ZÁVĚR

Tato diplomová práce se zabývala problematikou hodnocení účinnosti marketingových kampaní na sociálních sítích Facebook a Instagram, a to konkrétně v prostředí sezónní prodejní akce Lyžebraní. Cílem práce bylo shrnout současné přístupy k měření efektivity marketingu na těchto platformách, aplikovat metodiky datové analýzy na reálná firemní data a na základě výsledků formulovat konkrétní doporučení pro zlepšení budoucích kampaní.

V rámci práce byly využity různé přístupy a metody datové analýzy, a to v rámci procesu CRISP-DM. Pro hodnocení vztahu mezi marketingovými aktivitami a dosaženými tržbami byly aplikovány metody korelační analýzy, metoda PCA, shluková analýza K-means, vícenásobná lineární regrese, predikce sezónnosti pomocí modelu Prophet a analýza sentimentu textových komentářů.

Výsledky analýz ukázaly, že formát Instagram Stories vykazoval nejvyšší prediktivní sílu vůči tržbám. Facebooková platforma byla identifikována jako důležitý informační kanál, přičemž interakce zde mají charakter dotazů a žádostí o konkrétní informace. Na základě těchto zjištění byla navržena sada doporučení, zahrnující mimo jiné zvýšení frekvence a informační hodnoty příspěvků v hlavní sezóně, širší využití interaktivních prvků (CTA) v Instagram Stories, optimalizaci odpovědí na Facebooku nebo strategické načasování kampaní v souladu se sezónními trendy.

Práce tak prokázala, že propojením metod datové analýzy s praktickými potřebami firmy lze získat cenné informace pro zefektivnění marketingového rozhodování. Výstupem práce jsou doporučení, která mohou být využita ke zlepšení výkonu budoucích kampaní firmy Lyžebraní a přispět tak k efektivnějšímu využití marketingového rozpočtu i lepšímu zacílení komunikace na zákazníky.

Přestože tato práce přinesla konkrétní poznatky využitelné v marketingové praxi firmy Lyžebraní, je vhodné zmínit i její určitá omezení. Analýza byla založena pouze na datech z jednoho sezónního období. Sběr dat z více sezón by přinesl vyšší variabilitu a umožnil vytvoření robustnějších modelů s vyšší vypovídací schopností. Do analýzy rovněž nebyly zahrnuty informace o výši investic do jednotlivých kampaní, které by umožnily detailnější vyhodnocení efektivity, například pomocí metriky návratnosti investic (ROI). V rámci použitých analytických přístupů by bylo možné práci dále rozšířit například o A/B testování, které umožňuje přímé porovnání výkonnosti různých variant příspěvků v reálném čase.

# POUŽITÁ LITERATURA

ABDI, Hervé a WILLIAMS, Lynne J., 2010. Principal component analysis. Online. *WIREs Computational Statistics*. Roč. 2, č. 4, s. 433-459. ISSN 1939-5108. Dostupné z: https://doi.org/10.1002/wics.101. [cit. 2025-04-14].

ADITYA SATRIO, Christophorus Beneditto; DARMAWAN, William; NADIA, Bellatasya Unrica a HANAFIAH, Novita, 2021. Time series analysis and forecasting of coronavirus disease in Indonesia using ARIMA model and PROPHET. Online. *Procedia Computer Science*. Roč. 179, s. 524-532. ISSN 18770509. Dostupné z: https://doi.org/10.1016/j.procs.2021.01.036. [cit. 2025-04-15].

AGÜERO-TORALES, M.M.; COBO, M.J.; HERRERA-VIEDMA, E. a LÓPEZ-HERRERA, A.G., 2019. A cloud-based tool for sentiment analysis in reviews about restaurants on TripAdvisor. Online. *Procedia Computer Science*. Roč. 162, s. 392-399. ISSN 18770509. Dostupné z: https://doi.org/10.1016/j.procs.2019.12.002. [cit. 2025-04-25].

ALJEDAANI, Wajdi; RUSTAM, Furqan; MKAOUER, Mohamed Wiem; GHALLAB, Abdullatif; RUPAPARA, Vaibhav et al., 2022. Sentiment analysis on Twitter data integrating TextBlob and deep learning models: The case of US airline industry. Online. *Knowledge-Based Systems*. Roč. 255, s. 109780. ISSN 09507051. Dostupné z: https://doi.org/10.1016/j.knosys.2022.109780. [cit. 2025-04-18].

ALVES, Helena; FERNANDES, Cristina a RAPOSO, Mário, 2016. Social media marketing: A literature review and implications. Online. *Psychology & Marketing*. Roč. 33, č. 12, s. 1029-1038. ISSN 0742-6046. Dostupné z: https://doi.org/10.1002/mar.20936. [cit. 2025-04-06].

APPEL, Gil; GREWAL, Lauren; HADI, Rhonda a STEPHEN, Andrew T., 2020. The future of social media in marketing. Online. *Journal of the Academy of Marketing Science*. Roč. 48, č. 1, s. 79-95. ISSN 0092-0703. Dostupné z: https://doi.org/10.1007/s11747-019-00695-1. [cit. 2025-04-02].

BARKER, Melissa S., BARKER, D., BORMANN, N. F., NEHER, K. E., ZAHAY, D. *Social media marketing: A strategic approach*. Mason, OH: South-Western Cengage Learning, 2013.

BERMAN, Jules J. *Data Simplification* [online]. 1st ed. Oxford: Elsevier, 2016. ISBN 978-0-12-803781-2. Dostupné z: https://www.sciencedirect.com/topics/computer-science/pearson-correlation. [cit. 2025-04-17].

BRAMBILLA, Matteo, Hossein BADRIZADEH, Niloofar MALEK MOHAMMADI a Ali JAVADIAN SABET. 2022. Analyzing brand awareness strategies on social media in the luxury market: The case of Italian fashion [online]. *Digital*, roč. 3, č. 1, s 1-17. Dostupné z: https://www.mdpi.com/2673-6470/3/1/1. [cit. 2025-04-06].

FREUND, Rudolf J.; WILSON, William J. a MOHR, Donna L., 2010. Linear Regression. Online. In: *Statistical Methods*. Elsevier, s. 321-374. ISBN 9780123749703. Dostupné z: https://doi.org/10.1016/B978-0-12-374970-3.00007-X. [cit. 2025-04-15].

GOŁĄB-ANDRZEJAK, Edyta, 2023. Measuring the effectiveness of digital communication – social media performance: an example of the role played by AI-assisted tools at a university. Online. *Procedia Computer Science*. Roč. 225, s. 3332-3341. ISSN 18770509. Dostupné z: https://doi.org/10.1016/j.procs.2023.10.327. [cit. 2025-04-26].

HOLAK, Brian a McLAUGHLIN, Emily. *Instagram* [online]. TechTarget, 2024. Dostupné z: https://www.techtarget.com/searchcio/definition/Instagram [cit. 2025-04-08]

CHUMBAR, Shawn. *The CRISP-DM Process: A Comprehensive Guide* [online]. Medium, 24. září 2023, Dostupné z: https://medium.com/@shawn.chumbar/the-crisp-dm-process-a-comprehensive-guide-4d893aecb151. [cit. 2025-04-19].

JAHANGIR, Rashid; TEH, Ying Wah; NWEKE, Henry Friday; MUJTABA, Ghulam; AL-GARADI, Mohammed Ali et al., 2021. Speaker identification through artificial intelligence techniques: A comprehensive review and research challenges. Online. *Expert Systems with Applications*. Roč. 171, s. 114591. ISSN 09574174. Dostupné z: https://doi.org/10.1016/j.eswa.2021.114591. [cit. 2025-04-14].

KARLÍČEK, Miroslav, 2016. *Marketingová komunikace: Jak komunikovat na našem trhu - 2., aktualizované a doplněné vydání*. Grada. ISBN 978-80-271-9064-5. Dostupné také z: https://www.bookport.cz/kniha/marketingova-komunikace-2536/. [cit. 2025-04-03].

KINCL, Tomáš; NOVÁK, Michal a PŘIBIL, Jiří, 2019. Improving sentiment analysis performance on morphologically rich languages: Language and domain independent approach. Online. *Computer Speech & Language*. Roč. 56, s. 36-51. ISSN 08852308. Dostupné z: https://doi.org/10.1016/j.csl.2019.01.001. [cit. 2025-04-25].

LIN, Liuyan; JIANG, Wei; CHEN, Biao; YU, Jing a ZHENG, Chenhong, 2024. Construction and Application of Cost Prediction Model Based on Multiple Linear Regression Analysis. Online. *Procedia Computer Science*. Roč. 247, s. 617-623. ISSN 18770509. Dostupné z: https://doi.org/10.1016/j.procs.2024.10.074. [cit. 2025-04-15].

LIU, Bing, 2012. *Sentiment analysis and opinion mining*. Synthesis lectures on human language technologies. San Rafael: Morgan & Claypool. ISBN 978-1-60845-884-4. [cit. 2025-04-15].

LV, Yanxia; WEI, Fangna; CAO, Lihong; PENG, Sancheng; NIU, Jianwei et al., 2021. Aspect-level sentiment analysis using context and aspect memory network. Online. *Neurocomputing*. Roč. 428, s. 195-205. ISSN 09252312. Dostupné z: https://doi.org/10.1016/j.neucom.2020.11.049. [cit. 2025-04-25].

MOU, Jessie Boxin. *Study on Social Media Marketing Campaign Strategy – TikTok and Instagram* [online]. Cambridge, MA: Massachusetts Institute of Technology, Sloan School of Management, 2020. Dostupné z: https://dspace.mit.edu/handle/1721.1/127010. [cit. 2025-04-17].

O'BRIEN, Clodagh. *How to Measure Social Media ROI* [online]. Digital Marketing Institute, 20.5.2022, Dostupné z: https://digitalmarketinginstitute.com/blog/how-to-measure-social-media-roi.[cit. 2025-04-20].

PLOTNIKOVA, Veronika; DUMAS, Marlon a MILANI, Fredrik P., 2022. Applying the CRISP-DM data mining process in the financial services industry: Elicitation of adaptation requirements. Online. *Data & Knowledge Engineering*. Roč. 139, s. 102013. ISSN 0169023X. Dostupné z: https://doi.org/10.1016/j.datak.2022.102013. [cit. 2025-04-11].

SHAKEEL, Asim; CHONG, Daotong a WANG, Jinshi, 2023. Load forecasting of district heating system based on improved FB-Prophet model. Online. *Energy*. Roč. 278, s. 127637. ISSN 03605442. Dostupné z: https://doi.org/10.1016/j.energy.2023.127637. [cit. 2025-04-15].

SHOPIFY. What is a marketing campaign? [online]. 2024. Dostupné z: https://www.shopify.com/blog/marketing-campaign. [cit. 2025-04-07].

SIOBOS, Aneta. *Metodologie data miningu – CRISP-DM proces a používané techniky* [online]. Praha: ITnetwork.cz, 2024 [cit. 2025-04-10]. Dostupné z: https://www.itnetwork.cz/python/data-mining/metodologie-data-mining-procesu-a-pouzivane-techniky

SYAKUR, M. A., B. K. KHOTIMAH, E. M. S. ROCHMAN a B. D. SATOTO, 2018. Integration K-means clustering method and elbow method for identification of the best customer profile cluster. In: *IOP Conference Series: Materials Science and Engineering* [online]. 336(1), s. 012017. DOI: 10.1088/1757-899X/336/1/012017 [cit. 2025-04-14]

TUTEN, Tracy L., [2021]. *Social media marketing*. 4th edition. Los Angeles: SAGE. ISBN 978-1-5297-3199-6. [cit. 2025-04-03].

VALDANI, Enrico a ARBORE, Alessandro, 2015. Marketing Strategies. Online. In: *International Encyclopedia of the Social & Behavioral Sciences*. Elsevier, s. 555-558. ISBN 9780080970875. Dostupné z: https://doi.org/10.1016/B978-0-08-097086-8.73026-1. [cit. 2025-04-25].

YANG, Chen, 2021. Research in the Instagram context: Approaches and methods. Online. *The Journal of Social Sciences Research*. 2021-1-13, č. 71, s. 15-21. ISSN 24136670. Dostupné z: https://doi.org/10.32861/jssr.71.15.21. [cit. 2025-04-07].

ZALARUSHI RAJSINH. *The Elbow Method – Finding the optimal number of clusters* [online]. Medium, 2023. Dostupné z: https://medium.com/@zalarushirajsinh07/the-elbow-method-finding-the-optimal-number-of-clusters-d297f5aeb189. [cit. 2025-04-14].

# SEZNAM PŘÍLOH

Příloha A – Kód pro načtení dat

```python
import pandas as pd

# Načtení souborů s korektním oddělovačem a kódováním
fb = pd.read_csv('/kaggle/input/propojeno/facebook_1.10.23-16.3.24.csv', sep=';', encoding='utf-8')
ig_posts = pd.read_csv('/kaggle/input/propojeno/ig_post_1.10.23-16.3.24.csv', sep=';', encoding='utf-8')
ig_stories = pd.read_csv('/kaggle/input/propojeno/ig_stories_1.10.23-16.3.24.csv', sep=';', encoding='utf-8')

# Funkce pro základní přehled o datech
def data_overview(df, name):
    print(f"--- {name} ---")
    print("Shape:", df.shape)
    print("Columns:", df.columns.tolist())
    print("\nMissing values:")
    print(df.isnull().sum())
    print("\nDtypes:")
    print(df.dtypes)
    print("\nSample data:")
    display(df.head())
    print("\n" + "-"*50 + "\n")

# Přehled pro každý dataset
data_overview(fb, "Facebook")
data_overview(ig_posts, "Instagram Posts")
data_overview(ig_stories, "Instagram Stories")
```

```python
# Nahrazení všech prázdných hodnot nulou
fb.fillna(0, inplace=True)
ig_posts.fillna(0, inplace=True)
ig_stories.fillna(0, inplace=True)

# Funkce pro kontrolu počtu chybějících hodnot
def check_missing(df, name):
    print(f"--- {name} ---")
    missing = df.isnull().sum()
    if missing.sum() == 0:
        print("✅ Žádné chybějící hodnoty.")
    else:
        print("⚠️ Chybějící hodnoty:")
        print(missing[missing > 0])
    print("-" * 50 + "\n")

# Kontrola po doplnění
check_missing(fb, "Facebook")
check_missing(ig_posts, "Instagram Posts")
check_missing(ig_stories, "Instagram Stories")
```

```python
import pandas as pd

# Načtení dat o tržbách
trzby = pd.read_csv('/kaggle/input/trzby2/trzby.csv', sep=';', encoding='utf-8')

# Přehled o datech
print("--- Přehled dat o tržbách ---")
print("Shape:", trzby.shape)
print("\nSloupce:", trzby.columns.tolist())
print("\nChybějící hodnoty:")
print(trzby.isnull().sum())
print("\nDatové typy:")
print(trzby.dtypes)

# Ukázka dat
print("\nUkázka dat:")
display(trzby.head())
```

```python
import pandas as pd

# Načtení nových propojených dat
fb_new = pd.read_csv('/kaggle/input/propojena-data/propojena_data_facebook_trzby.csv', sep=';', encoding='utf-8')
ig_posts_new = pd.read_csv('/kaggle/input/propojena-data/propojena_data_instagram_trzby.csv', sep=';', encoding='utf-8')
ig_stories_new = pd.read_csv('/kaggle/input/propojena-data/propojena_data_instagram_stories_trzby.csv', sep=';', encoding='utf-8')

# Funkce pro základní přehled o datech
def data_overview(df, name):
    print(f"--- {name} ---")
    print("\nMissing values:")
    print(df.isnull().sum())
    print("\nDtypes:")
    print(df.dtypes)
    print("\nSample data:")
    display(df.head())
    print("\n" + "-"*50 + "\n")

# Přehled pro každý nový dataset
data_overview(fb_new, "Facebook (propojeno)")
data_overview(ig_posts_new, "Instagram Posts (propojeno)")
data_overview(ig_stories_new, "Instagram Stories (propojeno)")
```

Příloha B – Kód pro analýzu korelace

```python
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

# Načtení dat
file_path = "/kaggle/input/ig-story/propojena_data_instagram_trzby.csv"
df = pd.read_csv(file_path, delimiter=";")
print(df.dtypes)
```

```python
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

# Načtení dat
file_path = ("/kaggle/input/ig-story/propojena_data_instagram_trzby.csv")
df = pd.read_csv(file_path, delimiter=";")

# Převod sloupců na numerický formát
df["tržba"] = pd.to_numeric(df["tržba"], errors="coerce")
df["počet registrovaných"] = pd.to_numeric(df["počet registrovaných"], errors="coerce")

# Výběr marketingových metrik z datového souboru pro Instagramové příspěvky
marketing_columns = [
    "Délka (v sekundách)",
    "Dosah",
    "To se mi líbí",
    "Sdílené",
    "Komentáře",
    "Uložení",
    "Sledující",
    "tržba",
    "počet registrovaných"
]

# Výběr pouze těchto sloupců a konverze na číselné hodnoty + odstranění NaN
numeric_df = df[marketing_columns].apply(pd.to_numeric, errors="coerce").dropna()

# Výpočet korelační matice
correlation_matrix = numeric_df.corr()

# Zobrazení korelační heatmapy
plt.figure(figsize=(12, 10))
sns.heatmap(correlation_matrix, annot=True, cmap="coolwarm", fmt=".2f", linewidths=0.5)
plt.title("Korelace metrik Instagramových příspěvků s tržbou a registracemi")
plt.show()

# Výpis top 4 metrik s nejvyšší pozitivní korelací s tržbou
top_correlations = correlation_matrix["tržba"].drop("tržba").sort_values(ascending=False).head(4)
print("\n📈 Top 4 metriky s nejvyšší pozitivní korelací s tržbou:")
for metric, corr_value in top_correlations.items():
    print(f"- {metric}: korelace {corr_value:.2f}")
```

Příloha C – Kód pro analýzu hlavních komponent

```python
import pandas as pd

# Cesta k souboru s daty z IG postů
file_path = "/kaggle/input/ig-post/propojena_data_instagram_trzby.csv"

# Načtení dat do DataFrame
df = pd.read_csv(file_path, delimiter=";")

# Zobrazení informací o datových typech a počtu hodnot
df.info()
```

```python
# Výběr číselných vstupních metrik vhodných pro PCA
pca_input_columns = [
    "Délka (v sekundách)",
    "Komentář s daty",
    "Dosah",
    "To se mi líbí",
    "Sdílené",
    "Komentáře",
    "Uložení",
    "Sledující"
]
```

+ Code + Markdown

```python
# Převod vstupních metrik a tržby na číselný formát
pca_inputs = df[pca_input_columns].apply(pd.to_numeric, errors="coerce")
trzba = df["tržba"].apply(pd.to_numeric, errors="coerce")

# Přidání ID
id_prispevku = df["ID příspěvku"]

# Spojení do jednoho dataframe včetně ID
combined = pd.concat([pca_inputs, trzba, id_prispevku], axis=1).dropna()
```

+ Code + Markdown

```python
from sklearn.preprocessing import StandardScaler

# Oddělení vstupních metrik a tržby po očištění
clean_inputs = combined[pca_input_columns]
clean_trzba = combined["tržba"]

# Standardizace vstupních metrik
scaler = StandardScaler()
scaled_data = scaler.fit_transform(clean_inputs)
```

```python
from sklearn.decomposition import PCA
import numpy as np
import matplotlib.pyplot as plt

# PCA bez omezení počtu komponent
pca = PCA()
pca_result = pca.fit_transform(scaled_data)

# Kumulativní vysvětlená variance
cumulative_variance = np.cumsum(pca.explained_variance_ratio_)

# Vykreslení grafu
plt.figure(figsize=(10, 6))
plt.plot(range(1, len(cumulative_variance) + 1), cumulative_variance, marker='o')
plt.xlabel("Počet komponent")
plt.ylabel("Kumulativní vysvětlená variance")
plt.title("Podíl kumulativní vysvětlené variance podle počtu PCA komponent")
plt.grid(True)
plt.tight_layout()
plt.show()
```

```python
# PCA s 4 komponentami
pca_final = PCA(n_components=5)
pca_result = pca_final.fit_transform(scaled_data)
```

```python
# Výsledný DataFrame s komponentami
pca_df = pd.DataFrame(
    pca_result,
    columns=[f"PC{i+1}" for i in range(5)],
    index=clean_inputs.index
)

# Připojení tržby a ID příspěvku
pca_df["tržba"] = clean_trzba.values
pca_df["ID příspěvku"] = combined.loc[clean_inputs.index, "ID příspěvku"].values
pca_df.head()
```

```python
# Vytvoření DataFrame s váhami metrik pro každou komponentu
loadings = pd.DataFrame(
    pca_final.components_.T,
    index=pca_input_columns,
    columns=[f"PC{i+1}" for i in range(5)]
)


# Výpis loadings tabulky
print("\n Váhy metrik (loadings)")
print(loadings.round(6))
```

```python
# Uložení PCA výsledků do souboru
output_path = "/kaggle/working/pca_ig_posty_5komponent.csv"
pca_df.to_csv(output_path, index=False)

print(f" Výstupní soubor byl uložen do: {output_path}")
```

Příloha D – Kód pro shlukovou analýzu

```python
import pandas as pd

# Načtení souborů
stories_df = pd.read_csv('/kaggle/input/pca-pro-shlukovku/pca_IG_stories_5komponent_ID.csv')
facebook_df = pd.read_csv('/kaggle/input/pca-pro-shlukovku/pca_facebook_4komponenty_ID.csv')
ig_posty_df = pd.read_csv('/kaggle/input/pca-pro-shlukovku/pca_ig_posty_5komponent_ID.csv')

# Náhled dat a typy sloupců
print(" IG STORIES:")
print(stories_df.head())
print(stories_df.dtypes)
print("\n" + "-"*50 + "\n")

print(" FACEBOOK:")
print(facebook_df.head())
print(facebook_df.dtypes)
print("\n" + "-"*50 + "\n")

print(" IG POSTY:")
print(ig_posty_df.head())
print(ig_posty_df.dtypes)
```

```python
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans

# Dataset a sloupce s PCA komponentami
datasets = {
    'IG posty': ig_posty_df[['PC1', 'PC2', 'PC3', 'PC4', 'PC5']],
    'IG stories': stories_df[['PC1', 'PC2', 'PC3', 'PC4', 'PC5']],
    'Facebook posty': facebook_df[['PC1', 'PC2', 'PC3', 'PC4']],
}

# Nastavení grafu
plt.figure(figsize=(15, 4))

for i, (name, X) in enumerate(datasets.items(), start=1):
    inertias = []
    k_range = range(1, 11)

    for k in k_range:
        kmeans = KMeans(n_clusters=k, random_state=42, n_init=10)
        kmeans.fit(X)
        inertias.append(kmeans.inertia_)

    # Vykreslení každého elbow grafu do samostatného subplotu
    plt.subplot(1, 3, i)
    plt.plot(k_range, inertias, marker='o')
    plt.title(f'Elbow metoda – {name}')
    plt.xlabel('Počet shluků (k)')
    plt.ylabel('within-cluster-sum-of-square')
    plt.grid(True)

plt.tight_layout()
plt.show()
```

```python
import matplotlib.pyplot as plt
from matplotlib.lines import Line2D
from sklearn.cluster import KMeans

# Základní datasety a jejich počet shluků
datasets = {
    'IG posty': (ig_posty_df.copy(), 3),
    'IG stories': (stories_df.copy(), 4),
    'Facebook posty': (facebook_df.copy(), 3),
}

# Colormap
cmap = plt.cm.viridis

# Grafy
plt.figure(figsize=(15, 4))

for i, (name, (df, n_clusters)) in enumerate(datasets.items(), start=1):
    X = df[['PC1', 'PC2']]
    kmeans = KMeans(n_clusters=n_clusters, random_state=42, n_init=10)
    clusters = kmeans.fit_predict(X)

    # Přidáme výsledný cluster do původního dataframe
    df['cluster_label'] = clusters
    datasets[name] = (df, n_clusters)  # uložíme zpět s aktualizovanými daty

    # Vykreslení
    plt.subplot(1, 3, i)
    plt.scatter(X['PC1'], X['PC2'], c=clusters, cmap=cmap, s=50, alpha=0.7)
    plt.scatter(kmeans.cluster_centers_[:, 0], kmeans.cluster_centers_[:, 1],
                c='red', s=100, marker='X', label='Centroidy')

    handles = [Line2D([0], [0], marker='o', color='w',
                      markerfacecolor=cmap(k / (n_clusters - 1)), markersize=10,
                      label=f'Cluster {k}') for k in range(n_clusters)]
    handles.append(Line2D([0], [0], marker='X', color='w',
                          markerfacecolor='red', markersize=10, label='Centroidy'))

    plt.legend(handles=handles, loc='best')
    plt.title(f'KMeans shluky – {name}')
    plt.xlabel('PC1')
    plt.ylabel('PC2')
    plt.grid(True)

plt.tight_layout()
plt.show()
```

```python
# TEXTOVÁ ANALÝZA
# --------------------
for name, (df, _) in datasets.items():
    print(f"\nANALÝZA – {name}")

    # Výpočet průměrné tržby podle clusteru
    mean_revenue = df.groupby('cluster_label')['tržba'].mean().sort_values(ascending=False)
    best_cluster = mean_revenue.idxmax()

    print(f"Nejvýkonnější cluster: Cluster {best_cluster}")
    print("Průměrná tržba podle clusterů:")
    print(mean_revenue)

    # Nejlepší 1 příspěvek z nejlepšího clustru
    top_post = df[df['cluster_label'] == best_cluster][['ID příspěvku', 'tržba']].sort_values(by='tržba', ascending=False).head(1)
    print("\nNejlepší příspěvek (ID a tržba) z nejvýkonnějšího clustru:")
    print(top_post.to_string(index=False))
```

Příloha E – Kód pro vícenásobnou lineární regresi

```python
import pandas as pd
from sklearn.linear_model import LinearRegression
from sklearn.metrics import r2_score, mean_squared_error
import matplotlib.pyplot as plt

# Načtení souborů
ig_stories = pd.read_csv('/kaggle/input/pca-pro-shlukovku/pca_IG_stories_5komponent_ID.csv', encoding='utf-8', delimiter=',')
facebook = pd.read_csv('/kaggle/input/pca-pro-shlukovku/pca_facebook_4komponenty_ID.csv', encoding='utf-8', delimiter=',')
ig_posts = pd.read_csv('/kaggle/input/pca-pro-shlukovku/pca_ig_posty_5komponent_ID.csv', encoding='utf-8', delimiter=',')

# Funkce pro lineární regresi + vrácení predikcí
def run_regression_and_plot(df, name, target_col='tržba', id_col='ID příspěvku'):
    X = df.drop(columns=[target_col, id_col])
    y = df[target_col]

    model = LinearRegression()
    model.fit(X, y)
    y_pred = model.predict(X)

    # Výpis metrik
    r2 = r2_score(y, y_pred)
    mse = mean_squared_error(y, y_pred)
    print(f"\n{name}")
    print("-" * len(name))
    print(f"R² score: {r2:.3f}")
    print(f"MSE: {mse:.3f}")
    print("Koeficienty:")
    for feature, coef in zip(X.columns, model.coef_):
        print(f"  {feature}: {coef:.4f}")

    # Vizualizace
    plt.figure(figsize=(6, 6))
    plt.scatter(y, y_pred, alpha=0.7)
    plt.plot([y.min(), y.max()], [y.min(), y.max()], 'r--', linewidth=2)
    plt.xlabel("Skutečná tržba")
    plt.ylabel("Predikovaná tržba")
    plt.title(f"{name}: Skutečná vs. predikovaná tržba")
    plt.grid(True)
    plt.tight_layout()
    plt.show()

# Spuštění pro všechny datasety
run_regression_and_plot(ig_stories, "IG Stories")
run_regression_and_plot(ig_posts, "IG Posts")
run_regression_and_plot(facebook, "Facebook")
```

```python
import statsmodels.api as sm

def print_p_values(df, name, target_col='tržba', id_col='ID příspěvku'):
    X = df.drop(columns=[target_col, id_col])
    y = df[target_col]

    X_with_const = sm.add_constant(X)
    model = sm.OLS(y, X_with_const).fit()

    print(f"\n{name} p-hodnoty jednotlivých komponent:")
    print("-" * (len(name) + 30))
    for feature, pval in model.pvalues.items():
        if feature != 'const':
            print(f"  {feature}: {pval:.4f}")
```

```python
print_p_values(ig_stories, "IG Stories")
print_p_values(ig_posts, "IG Posts")
print_p_values(facebook, "Facebook")
```

Příloha F – Kód pro model Prophet

```
# Načtení CSV (Instagram posty)
df_instagram = pd.read_csv("/kaggle/input/og-data/propojena_data_instagram_trzby.csv", sep=";", encoding="utf-8")

# Převod 'tržba' na číselný formát
df_instagram['tržba'] = pd.to_numeric(df_instagram['tržba'], errors='coerce')

# Převod času zveřejnění na datetime
df_instagram['Čas zveřejnění'] = pd.to_datetime(df_instagram['Čas zveřejnění'], errors='coerce')

# Odstranění řádků s chybějícími hodnotami
df_instagram = df_instagram.dropna(subset=['Čas zveřejnění', 'tržba'])

# Příprava dat pro Prophet
df_prophet_ig = df_instagram[['Čas zveřejnění', 'tržba']].copy()
df_prophet_ig.columns = ['ds', 'y']

# Trénink modelu
model_ig = Prophet()
model_ig.fit(df_prophet_ig)

# Budoucí data (do 2024-03-16)
last_date = '2024-03-16'
future_ig = model_ig.make_future_dataframe(periods=0)
future_ig = future_ig[future_ig['ds'] <= last_date]

# Predikce
forecast_ig = model_ig.predict(future_ig)

# Graf predikce
fig1_ig = model_ig.plot(forecast_ig)
plt.title("Predikce tržeb z Instagram postů (do 16. 3. 2024)")
plt.show()

# Sezónní komponenty
fig2_ig = model_ig.plot_components(forecast_ig)
plt.show()
```

```python
from sklearn.metrics import r2_score, mean_squared_error, mean_absolute_error
import pandas as pd
from prophet import Prophet
import matplotlib.pyplot as plt
import numpy as np

# Funkce pro trénink Prophet modelu, výstup forecastu + metriky
def forecast_trzby(file_path, label, last_date='2024-03-16'):
    df = pd.read_csv(file_path, sep=";", encoding="utf-8")

    # Zpracování dat
    df['tržba'] = pd.to_numeric(df['tržba'], errors='coerce')
    df['Čas zveřejnění'] = pd.to_datetime(df['Čas zveřejnění'], errors='coerce')
    df = df.dropna(subset=['Čas zveřejnění', 'tržba'])

    df_prophet = df[['Čas zveřejnění', 'tržba']].copy()
    df_prophet.columns = ['ds', 'y']

    # Trénink modelu
    model = Prophet()
    model.fit(df_prophet)

    future = model.make_future_dataframe(periods=0)
    future = future[future['ds'] <= last_date]

    forecast = model.predict(future)

    # Spojení forecastu a skutečných hodnot
    merged = pd.merge(df_prophet, forecast[['ds', 'yhat']], on='ds')
    y_true = merged['y']
    y_pred = merged['yhat']

    # Výpočet metrik
    r2 = r2_score(y_true, y_pred)
    rmse = mean_squared_error(y_true, y_pred, squared=False)
    mae = mean_absolute_error(y_true, y_pred)
    mape = np.mean(np.abs((y_true - y_pred) / y_true)) * 100

    # Uložení do slovníku
    metrics = {
        'Zdroj': label,
        'R²': round(r2, 3),
        'RMSE': round(rmse, 2),
        'MAE': round(mae, 2),
        'MAPE (%)': round(mape, 2)
    }

    return model, forecast, label, metrics
```

```python
# Cesty k souborům
paths = {
    "Facebook": "/kaggle/input/og-data/propojena_data_facebook_trzby.csv",
    "Instagram Posty": "/kaggle/input/og-data/propojena_data_instagram_trzby.csv",
    "Instagram Stories": "/kaggle/input/og-data/propojena_data_instagram_stories_trzby.csv"
}

# Spuštění predikcí a výpočtu metrik
results = []
vystupy_metrik = []

for label, path in paths.items():
    model, forecast, name, metrics = forecast_trzby(path, label)
    results.append((model, forecast, name))
    vystupy_metrik.append(metrics)

# Vykreslení všech tří predikcí vedle sebe
fig, axes = plt.subplots(1, 3, figsize=(18, 5))

for i, (model, forecast, name) in enumerate(results):
    model.plot(forecast, ax=axes[i])
    axes[i].set_title(f"Predikce tržeb z {name} (do 16. 3. 2024)")
    axes[i].set_xlabel("Datum")
    axes[i].set_ylabel("Tržba")

plt.tight_layout()
plt.show()

# Výpis metrik
print("\n📈 Regresní metriky pro Prophet predikce:")
df_metriky = pd.DataFrame(vystupy_metrik)
print(df_metriky)
```

Příloha G – Kód pro analýzu sentimentu

```python
import pandas as pd
import matplotlib.pyplot as plt
from textblob import TextBlob
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.decomposition import LatentDirichletAllocation
import re
from collections import Counter
import nltk
from nltk.corpus import stopwords

# === 0. Stažení stop slov z nltk ===
nltk.download('stopwords')
stop_words = set(stopwords.words('english'))  # používáme set pro rychlost při filtrování
stop_words_list = list(stop_words)  # ale CountVectorizer potřebuje list

# === 1. Načtení dat ===
file_path = "/kaggle/input/analyza-sentimentu-nove/analyza_sentimentu_1.10.23-16.3.24.csv"
df = pd.read_csv(file_path, encoding="utf-8", sep=";")

# === 2. Čištění textu ===
def clean_text(text):
    text = str(text).lower()
    text = re.sub(r"http\S+|www\S+|https\S+", '', text)  # odstranění URL
    text = re.sub(r"[^a-zA-Z\s]", '', text)  # odstranění čísel a interpunkce
    return text

df["clean_text"] = df["text"].apply(clean_text)

# === 3. Analýza sentimentu ===
def analyze_sentiment(text):
    analysis = TextBlob(str(text))
    return analysis.sentiment.polarity

df["Sentiment Score"] = df["clean_text"].apply(analyze_sentiment)
df["Sentiment Label"] = df["Sentiment Score"].apply(
    lambda x: "Pozitivní" if x > 0 else "Negativní" if x < 0 else "Neutrální"
)
```

```python
# === 4. Vizualizace sentimentu ===
sentiment_counts = df["Sentiment Label"].value_counts()
colors = {"Neutrální": "blue", "Pozitivní": "green", "Negativní": "red"}

# Koláčový graf
plt.figure(figsize=(8, 6))
plt.pie(sentiment_counts, labels=sentiment_counts.index, autopct='%1.1f%%',
        startangle=140, colors=[colors[label] for label in sentiment_counts.index])
plt.title("Rozložení sentimentu v komntářích")
plt.show()

# Sloupcový graf
plt.figure(figsize=(8, 6))
plt.bar(sentiment_counts.index, sentiment_counts.values,
        color=[colors[label] for label in sentiment_counts.index])
plt.xlabel("Sentiment Label")
plt.ylabel("Number of Comments")
plt.title("Number of Comments per Sentiment Label")
plt.show()
```

```python
# === 5. Klíčová slova bez stop slov + přidání "hello" a "thank" a "also"===
custom_stopwords = stop_words.union({"hello", "thank", "also"})
stop_words_list = list(custom_stopwords)  # pro CountVectorizer

def extract_keywords(text):
    words = re.findall(r'\b\w{4,}\b', text.lower())
    return [word for word in words if word not in custom_stopwords]

all_words = []
for comment in df["clean_text"].dropna():
    all_words.extend(extract_keywords(comment))

word_counts = Counter(all_words)
common_words = word_counts.most_common(5)

print("5 nejčastějších klíčových slov v komentářích:")
for word, count in common_words:
    print(f" - {word} ({count}x)")

# === 6. LDA – tématická analýza (3 témata × 3 slova) ===
vectorizer = CountVectorizer(max_features=1000, stop_words=stop_words_list)
X = vectorizer.fit_transform(df["clean_text"].dropna())

lda = LatentDirichletAllocation(n_components=3, random_state=42)
lda.fit(X)

words = vectorizer.get_feature_names_out()
topics = []
for topic_idx, topic in enumerate(lda.components_):
    top_words = [words[i] for i in topic.argsort()[:-4:-1]]  # 3 slova na téma
    topics.append(f"Topic {topic_idx + 1}: {', '.join(top_words)}")

print("\n3 hlavní témata komentářů (po 3 klíčových slovech):")
for topic in topics:
    print(f" - {topic}")
```