

Semestrální práce BDT

JAN PRESERÍN

1. Automatické nahrání dat

Skripty definující strukturu dat jsou uloženy v souboru **OUTPUTS/skripty/data_load.txt**

Skript definující natažení každodenních dat z csv je uložen v souboru **OUTPUTS/skripty/daily_load.txt**

Ve skriptu data_load.txt je popsáno prvotní vytvoření struktury složek na Metacentru, do kterých jsem následně přes FTP nahrál vstupní data sales_train.csv a ostatní číselníky s názvy. Data za celou historii jsou na metacentru ve složce **semestralka/prodeje/all**. Co se týče denních dat, admin je každý den nahraje do složky **semestralka/prodeje/daily**, a já následně každý den spustím skripty ze souboru daily_load.txt, které jednak zkopírují vše ze složky **semestralka/prodeje/daily** na složku se stejným jménem na HDFS, poté zapnu Hive a vytvořím externí tabulku dle zadání, jejíž obsah následně spojím s tabulkou obsahující všechna prodejní data. Soubory na HDFS i na lokálním FS ve složce **semestralka/prodeje/daily** přesunu do složky **semestralka/prodeje/history**. Následně ve Sparku provedu několik operací nad Hive tabulkami, přičemž každodenní report vznikne z dat **s nejnovějším datem** nacházejícím se v souboru se všemi prodejními daty (které v tu chvíli již jsou o nová data aktualizovány). Podle potřeby firmy/chování admina by se to dalo ještě více upravit a parametrizovat, nechal jsem to ale prozatím takto. Skript vygeneruje textový soubor s každodenním reportem a uloží ho na lokální FS. Vzorek souboru se nachází v **OUTPUTS/one_day_report/example_one_day_report.txt**

2. Každodenní report

Jak jsem již psal, vzorek souboru se nachází v **OUTPUTS/one_day_report/example_one_day_report.txt** a report vzniká pomocí Sparku, který bere data z tabulek v Hive. Vše je vidět v souboru **OUTPUTS/skripty/daily_load.txt** a ukázka vygenerovaného souboru se nachází v **OUTPUTS/one_day_report/example_one_day_report.txt**

3. Kumulativní report

Ke každému bodu přikládám všechny požadované obrázky s popisem + náhled tabulek, pokud jsou součástí výstupů. Některé tabulky jsou moc dlouhé, a proto je vše uloženo ještě jako CSV soubor, kde je lze zobrazit celé.

ABC analýza

1. Rozdělení všech produktů do kategorií ABC a XYZ - soubor **OUTPUTS/grafy_tabulky/ABC_abcxyz_fullproducts.csv**

| Produkt | abc | xyz |
|--|-----|-----|
| 1C:Аудиокниги. Английские сказки, часть 1 [PC, Цифровая версия] | C | X |
| 1C:Аудиотеатр. Чехов А.П. Дуэль [Цифровая версия] | C | X |
| 1C:Бухгалтерия 8 (ред.3.0) как на ладони. Изд 4. Гартвич А. | B | Z |
| 1C:Зарплата и Управление Персоналом 8. Базовая версия. | A | Z |
| 1C:Управление небольшой фирмой 8. Самоучитель. Издание 2 [Цифровая версия] | C | Y |
| 50 CENT Curtis (фирм.) | C | X |
| ACCEPT Best Of LP | C | Z |
| AZNAVOUR CHARLES 20 Chansons D'Or | C | Z |
| Assassin's Creed 3. Freedom Edition [PS3, русская версия] | C | Z |
| BEATLES, THE One (фирм.) | B | Z |
| Battlefield 3 Premium. Сборник дополнений (код загрузки, без диска) [PC, русская версия] | A | Z |
| Blades of Time [PS3, русская версия] | C | X |
| CHARLES RAY The Great Ray Charles LP | B | Z |
| CLAPTON ERIC 24 Nights Chronicles 2DVD (фирм.) | C | X |

2. Četnosti jednotlivých kategorií - soubor **OUTPUTS/grafy_tabulky/ABC_abcxyz_count.csv**

| ABC skupina | XYZ skupina | pocet produktu |
|-------------|-------------|----------------|
| A | X | 18 |
| A | Y | 57 |
| A | Z | 1597 |
| B | X | 51 |
| B | Y | 282 |
| B | Z | 3924 |
| C | X | 4387 |
| C | Y | 144 |
| C | Z | 11347 |

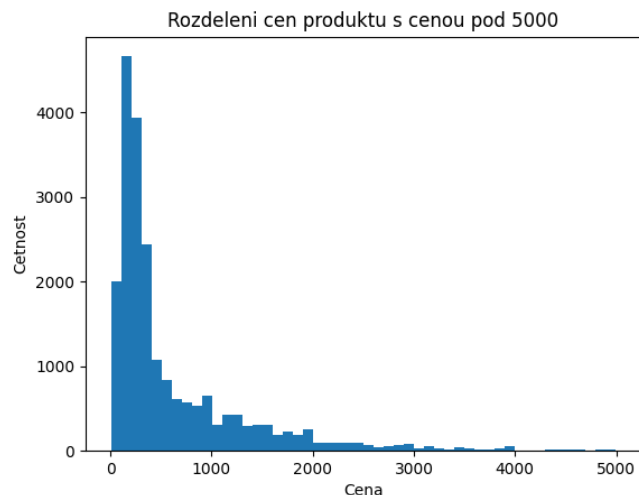
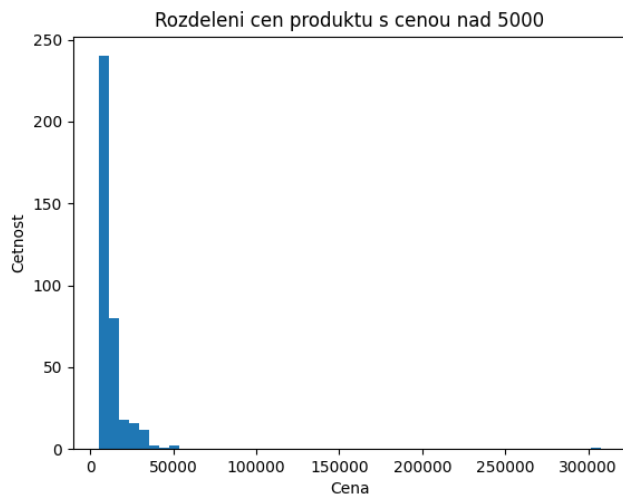
3. Ukázka 3 produktů z každé kategorie ABC/XYZ - soubor **OUTPUTS/grafy_tabulky/ABC_showcase_abcxyz.csv**

| Produkt | abc | xyz |
|--|-----|-----|
| Ведьмак 3: Дикая Охота - Дополнение "Каменные Сердца" (код загрузки, без диска) [PC, русская версия] | A | X |
| Guitar Hero Live (Гитара + игра) [PS4, английская версия] | A | X |
| Radmin 3 - 522 лиц. | A | X |
| Dr.Web Security Space K3 1 ПК/1 год (картонная упаковка) | A | Y |
| PSP E1008: адаптер ARTPLAYS 100-240V | A | Y |
| 1С:Бухгалтерия 8 (ред.3.0) как на ладони. Изд 5. Гартвич А. | A | Y |
| 3D GARDEN (Наш Сад 10) (box) | A | Z |
| Assassin's Creed: Единство. Guillotine Edition [PS4, русская версия] | A | Z |
| FIFA 13 [PS Vita, русская документация] | A | Z |
| РАЗЛОМ САН-АНДРЕАС + коллекционная открытка | B | X |
| Pro Evolution Soccer 2016 [PS4, русские субтитры] | B | X |
| Sony PlayStation 4 (500 Gb) White (CUH-1208A/B01) | B | X |
| СБ. Хиты 80-х в новых версиях | B | Y |

Rozdělení průměrných cen produktů, tabulka 10 produktů s nejvyšší průměrnou cenou

Mám zde 2 rozdělení, jedno pro produkty dražší než 5000, druhé pro levnější produkty. Je to z toho důvodu, že rozptyl cen produktů je dost veliký a v jednom histogramu nebylo nic vidět (long tail), jelikož většina produktů byla v 1. binu a jednotky produktů zase úplně vpravo. Následně přikládám tabulku 10 produktů s nejvyšší pr. cenou. - soubor

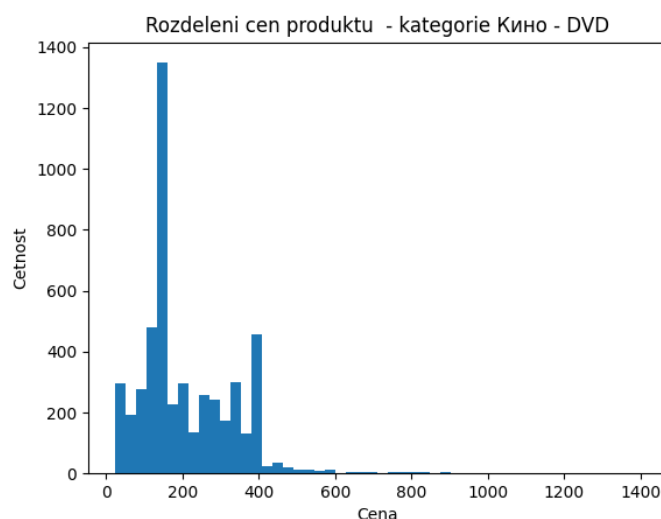
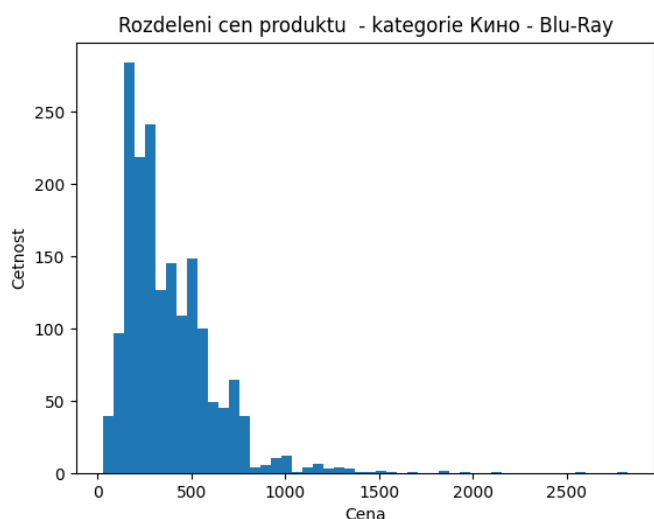
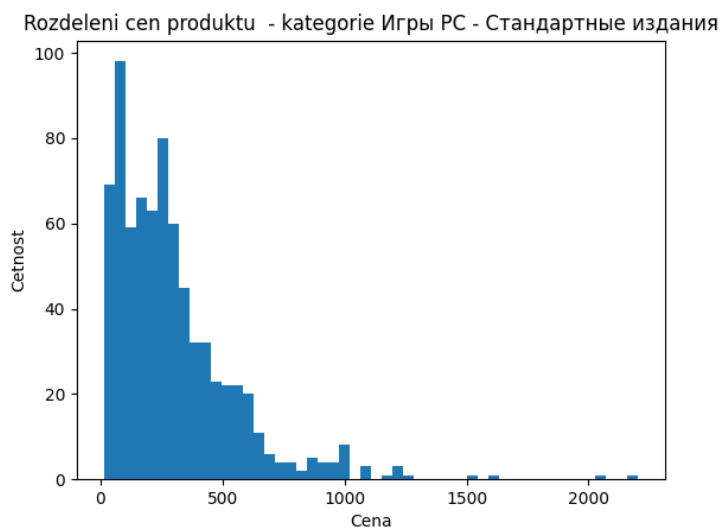
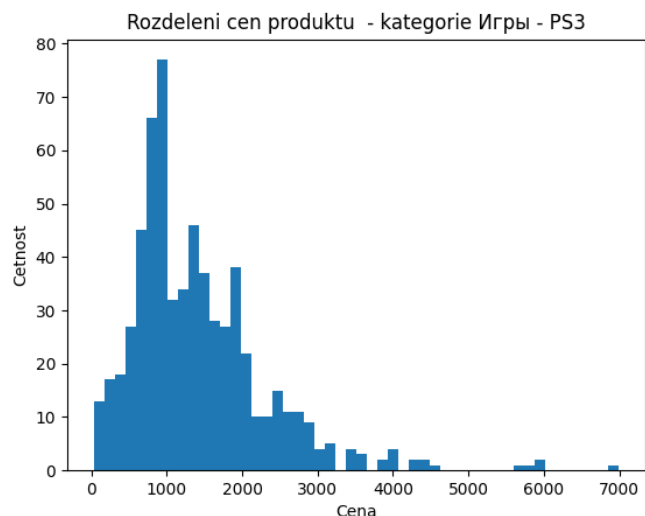
OUTPUTS/grafy_tabulky/GRAF_1_avg_prices_categories.csv.



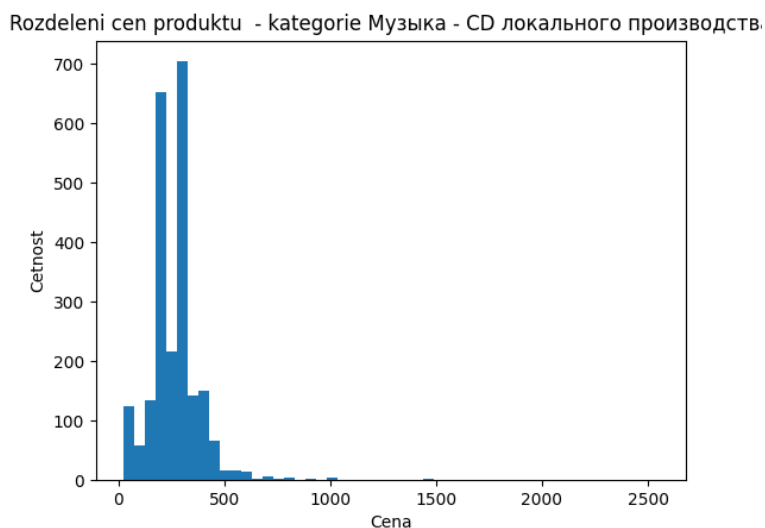
| Produkt | item_id | Prumerna cena |
|--|---------|---------------|
| Radmin 3 - 522 лиц. | 6066 | 307980.0 |
| Коллекционные шахматы (Властелин Колец) | 13199 | 50999.0 |
| UserGate Proxy & Firewall 6.X с модулем фильтрации сайтов Entensys URL Filtering до 200 сессий (Academic/Social) | 7241 | 49782.0 |
| UserGate Proxy & Firewall 6.X до 100 сессий | 7238 | 42000.0 |
| Лицензия 1С-Битрикс: Управление сайтом - Эксперт | 14173 | 40900.0 |
| Комплект "Microsoft Xbox One 1TB Limited Edition Halo 5 + "Halo 5. Limited Edition" [KF6-00012] | 13403 | 39990.5 |
| Photoshop 13 Windows Russian BOX | 5748 | 35132.86 |
| Комплект «Sony PlayStation 4 (1Tb) Black (CUH-1208B)» + игра «Assassin's Creed: Синдикат» + игра «Watch_Dogs» | 13503 | 34990.0 |
| Комплект Sony PlayStation 4 (1Tb) Black (CUH-1208B) + игра Uncharted: Натан Дрейк. Коллекция | 13448 | 34823.33 |
| CorelDRAW Graphics Suite X6 - Small Business Edition Russian BOX | 2410 | 34713.5 |

TOP 10 produktů dle průměrné ceny

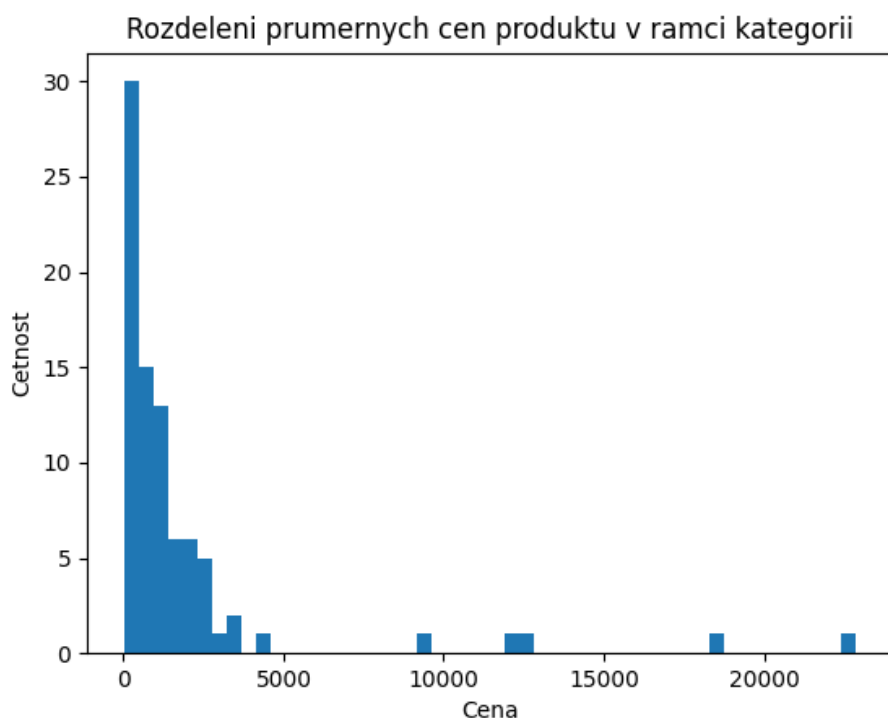
Rozdělení průměrných cen produktů pouze v rámci kategorie - TOP 5 kategorií



Můžeme zde vidět, že například kategorie PS3 obsahuje celkem drahé produkty, naopak ostatní kategorie spíše levnější



Rozdělení průměrných cen za celé kategorie, tabulka 10 kategorií s nejvyšší průměrnou cenou



Ize vidět, že většina kategorií obsahuje spíše produkty s nízkou cenou

| Slovni_nazev_kategorie | kategorie | prumerna_cena |
|-----------------------------|-----------|--------------------|
| Игровые консоли - PS4 | 12 | 22889.299108370662 |
| Игровые консоли - XBOX ONE | 16 | 18682.56100468793 |
| Игровые консоли - PS3 | 11 | 12686.74201778796 |
| Игровые консоли - XBOX 360 | 15 | 12024.564613317336 |
| Игровые консоли - PSVita | 14 | 9263.114423335557 |
| Игровые консоли - PSP | 13 | 4572.157098591541 |
| Программы - MAC (Цифра) | 74 | 3403.5714285714284 |
| Музыка - Подарочные издания | 60 | 3319.4875081967207 |
| Игры - PS4 | 20 | 2791.7798200611965 |
| Аксессуары - XBOX ONE | 7 | 2713.5522953420004 |

soubor **OUTPUTS/grafy_tabulky/**
GRAF_3_avg_prices_categories.csv

AX, BX, CZ - tabulky s 10 TOP produkty

soubory **OUTPUTS/grafy_tabulky/GRAF_4_{AX,BX,CZ}.csv**

| item_name | item_id | abc | xyz | revenue |
|---|---------|-----|-----|------------|
| Uncharted: Натан Дрейк. Коллекция [PS4, русская версия] | 7223 | A | X | 3178372.1 |
| Assassin's Creed: Синдикат. Специальное издание [PS4, русская версия] | 1583 | A | X | 2584789.88 |
| Uncharted: Натан Дрейк. Коллекция. Специальное издание [PS4, русская версия] | 7224 | A | X | 2256936.0 |
| Assassin's Creed: Синдикат. Биг Бен [PS4, русская версия] | 1578 | A | X | 2085006.25 |
| Halo 5: Guardians. Коллекционное издание [Xbox One, русская версия] (CV4-00014) | 3838 | A | X | 1489156.0 |
| Ведьмак 3: Дикая Охота - Дополнение "Каменные Сердца" (код загрузки, без диска) [PS4, русская версия] | 10202 | A | X | 1045725.15 |
| Ведьмак 3: Дикая Охота - Дополнение "Каменные Сердца" (код загрузки, без диска) [PC, русская версия] | 10201 | A | X | 1031492.8 |
| Комплект Sony PlayStation 4 (500 Gb) Black (CUH-1208A/B01) + Дополнительный контроллер черный | 13449 | A | X | 1028598.0 |
| Assassin's Creed: Синдикат. Чаринг-Кросс [PS4, русская версия] | 1586 | A | X | 868142.78 |
| Guitar Hero Live (Гитара + игра) [PS4, английская версия] | 3756 | A | X | 758380.0 |

Kategorie AX

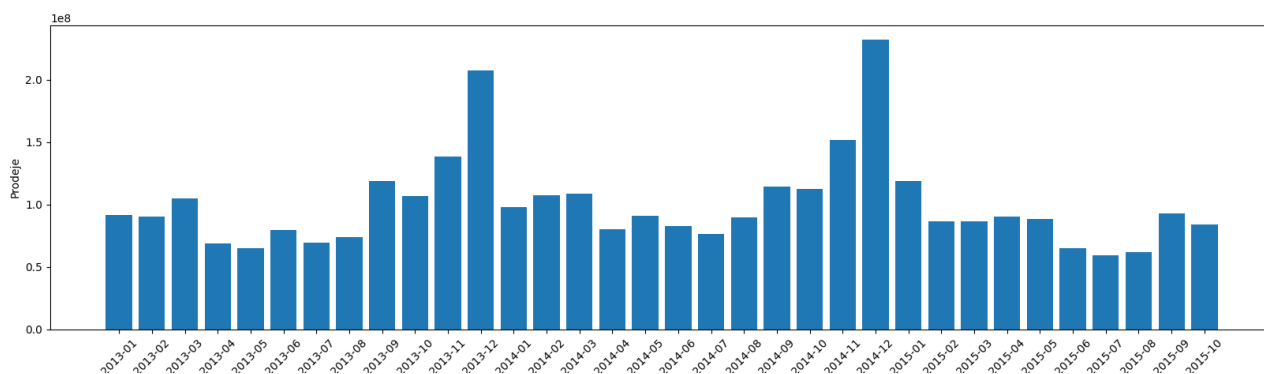
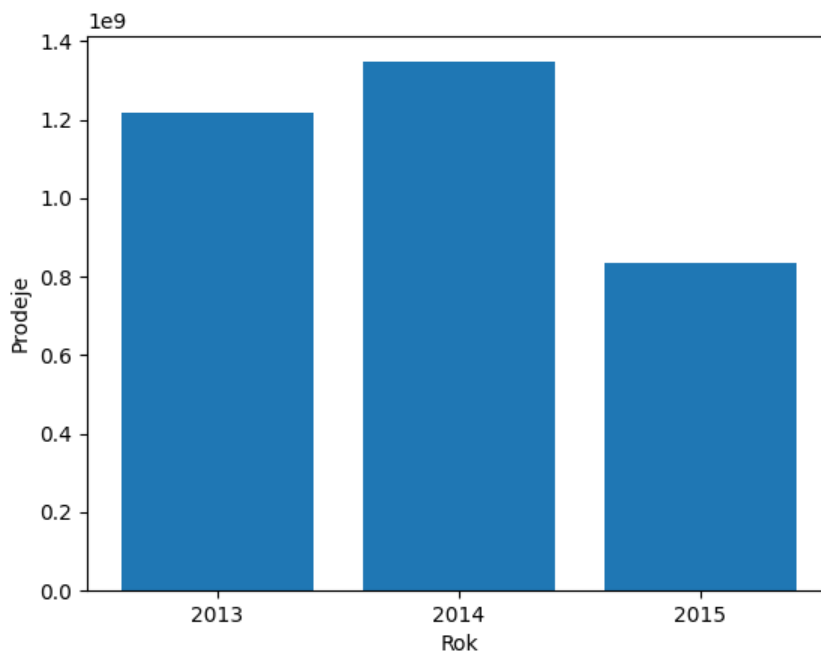
| item_name | item_id | abc | xyz | revenue |
|--|---------|-----|-----|-----------|
| Photoshop 13 Windows Russian BOX | 5748 | B | X | 245930.0 |
| Halo 5: Guardians. Ограниченное издание [Xbox One, русская версия] (CV3-00021) | 3839 | B | X | 244965.0 |
| Assassin's Creed: Синдикат. Специальное издание [Xbox One, русская версия] | 1584 | B | X | 239139.0 |
| Sony PlayStation 4 (500 Gb) White (CUH-1208A/B01) | 6678 | B | X | 228528.0 |
| Assassin's Creed: Синдикат. Биг Бен [Xbox One, русская версия] | 1579 | B | X | 204961.0 |
| Metal Gear Solid V: The Phantom Pain. Коллекционное издание [Xbox One, русские субтитры] | 4844 | B | X | 158043.02 |
| Артбук Мир игры Assassin's Creed Syndicate | 8546 | B | X | 153375.0 |
| Pro Evolution Soccer 2016 [PS4, русские субтитры] | 5907 | B | X | 140563.0 |
| Шапка Adventure Time Jake плюшевая | 21942 | B | X | 137861.0 |
| NHL 16 [Xbox One, русские субтитры] | 5042 | B | X | 121968.0 |

Kategorie BX

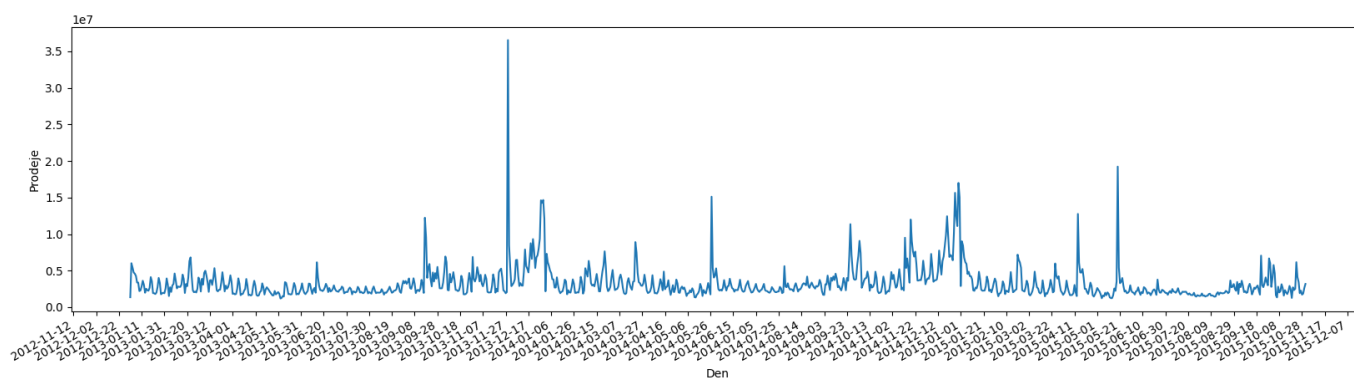
| item_name | item_id | abc | xyz | revenue |
|---|---------|-----|-----|----------|
| Музыкальная шкатулка/LOVE STORY | 15199 | C | Z | 47359.0 |
| ZZ TOP Greatest Hits | 8031 | C | Z | 47341.2 |
| БЕГЛЕЦЫ (регион) | 8701 | C | Z | 47338.3 |
| МАРА Почувствуй разницу Live 2CD+DVD (фирм.) | 14359 | C | Z | 47317.0 |
| STING Ten Summoner`s Tales LP | 6350 | C | Z | 47306.0 |
| Жетоны Wolfenstein The New Order Dogtag Blazkowicz | 11542 | C | Z | 47286.7 |
| ОДНАЖДЫ В ИРЛАНДИИ (BD) | 16515 | C | Z | 47284.41 |
| PS4: Наушники с микрофоном Tritton Kama Stereo Headset - Black PS4/PS Vita | 5676 | C | Z | 47229.5 |
| HELLOWEEN Straight Out Of Hell | 3788 | C | Z | 47225.0 |
| Государственные и муниципальные учреждения: учет в 1С:БГУ 8. Е.А.Кадыш, С.Ю.Рыженкова, К.В.Сергеева | 10697 | C | Z | 47200.0 |

Kategorie CZ

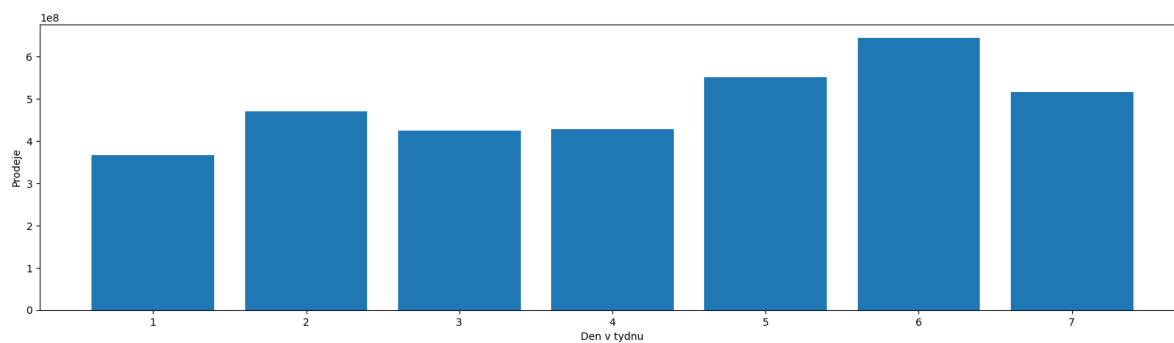
Grafy časového vývoje



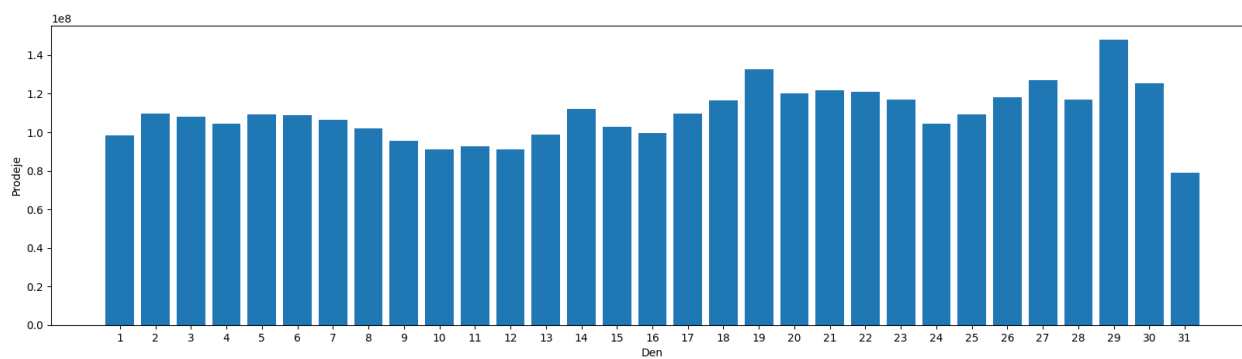
zde lze vypořizovat určitou sezonnost okolo Vánoc



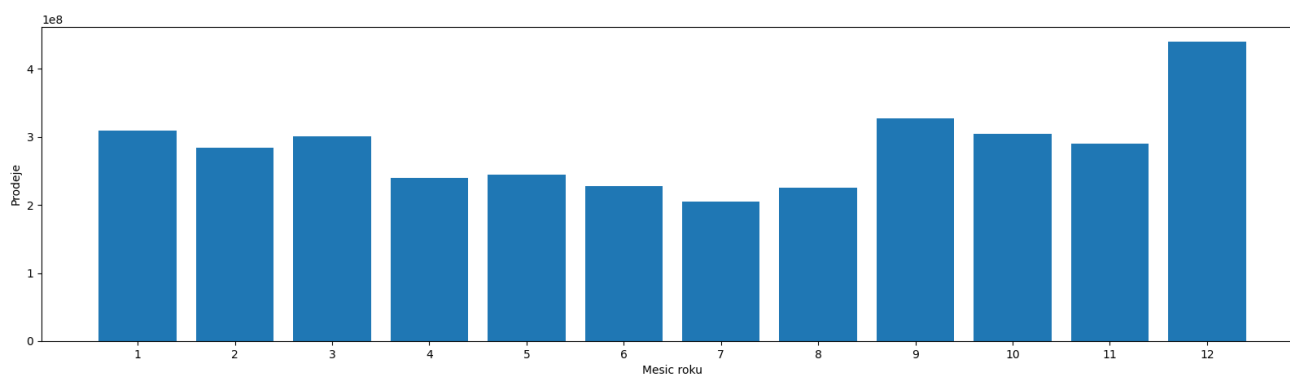
Grafy rozdělení celkové tržby podle času



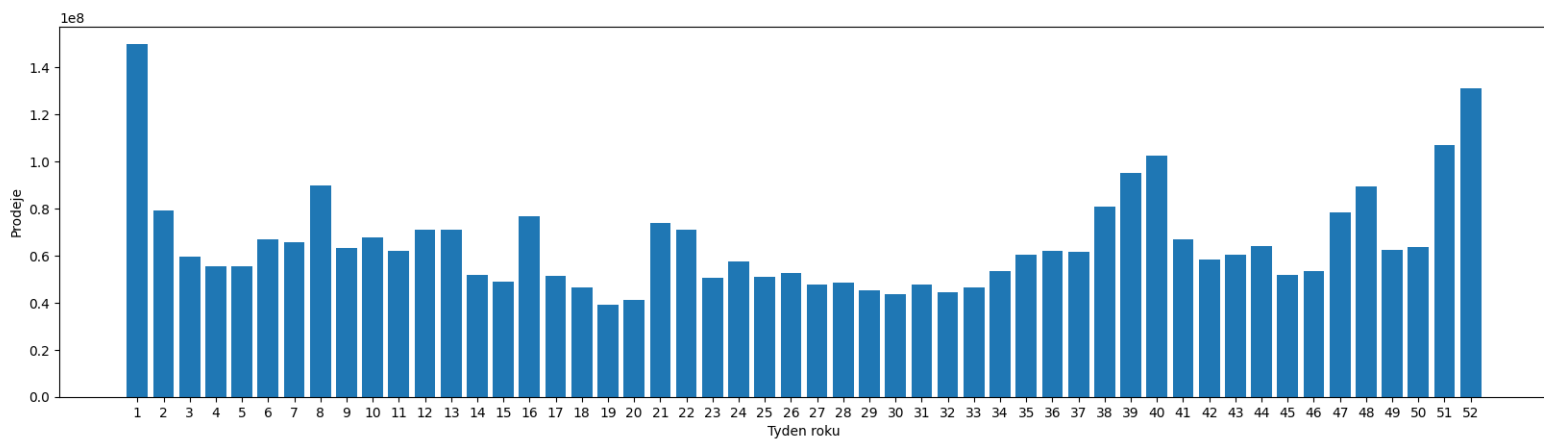
Dny v týdnu - je vidět, že zákazníci nakupují hlavně o víkendu



Dny v měsíci - prodeje v rámci měsíce jsou celkem stabilní, lehčí nárůst na konci

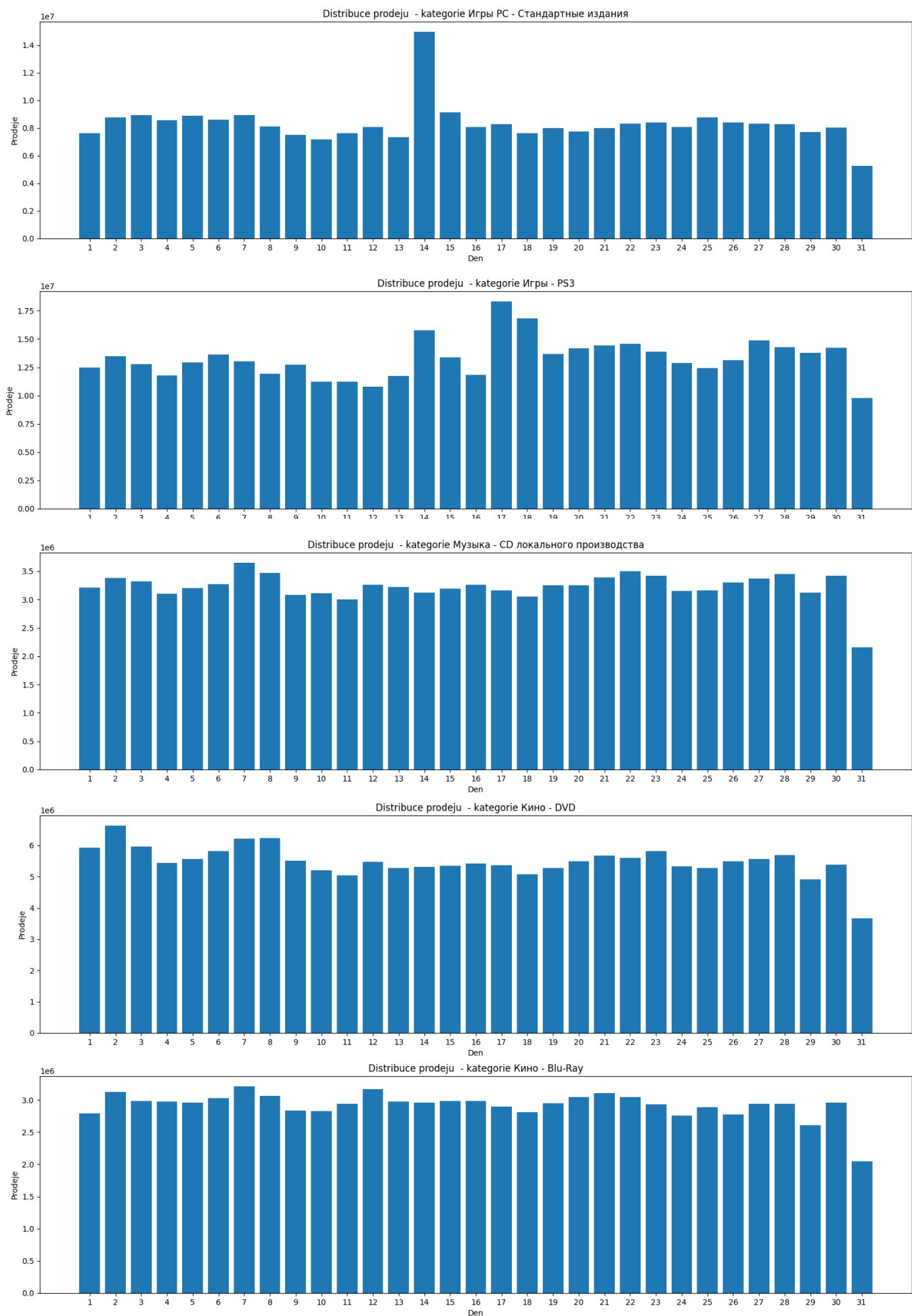


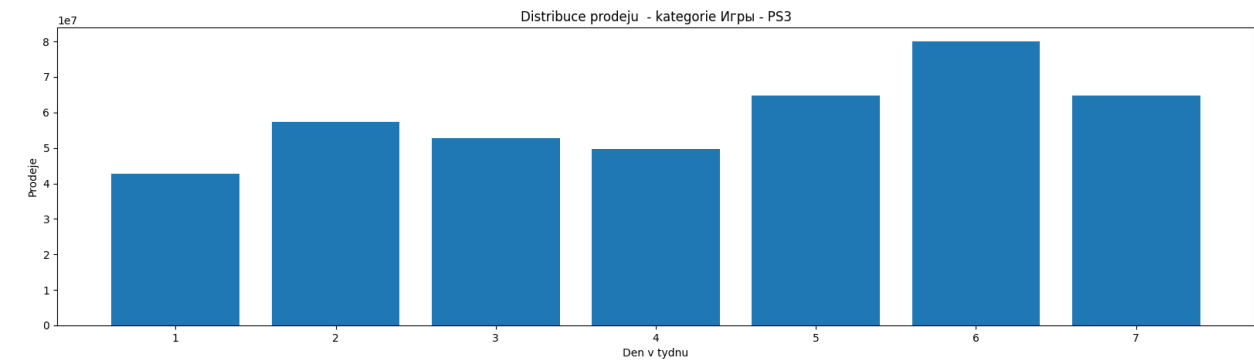
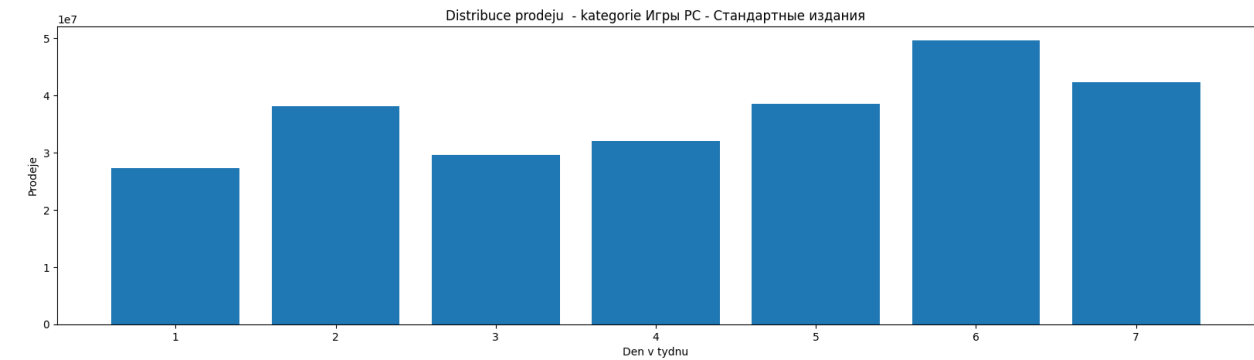
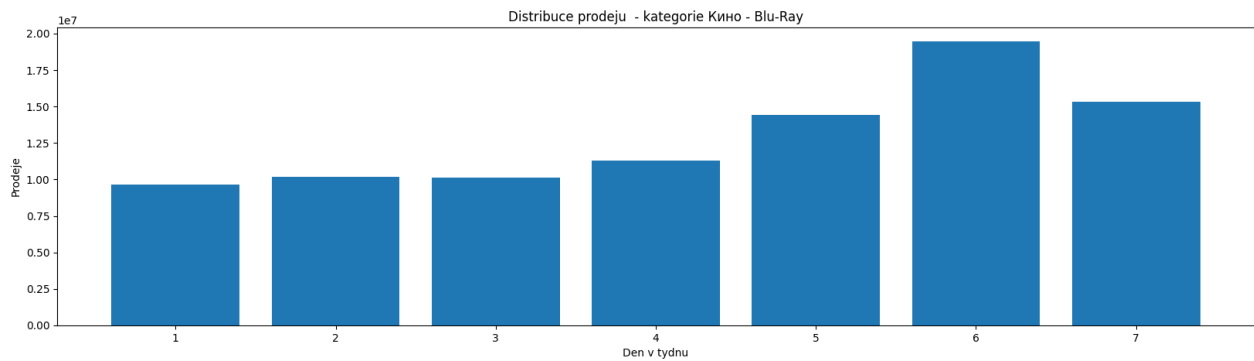
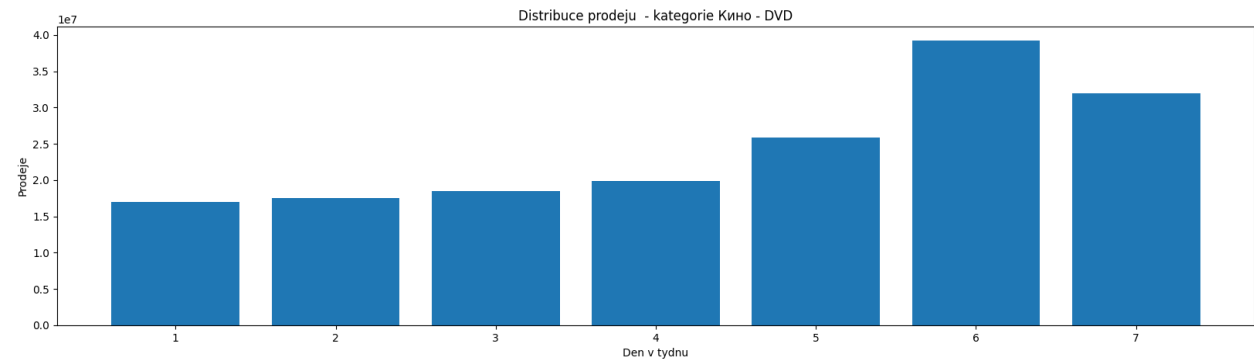
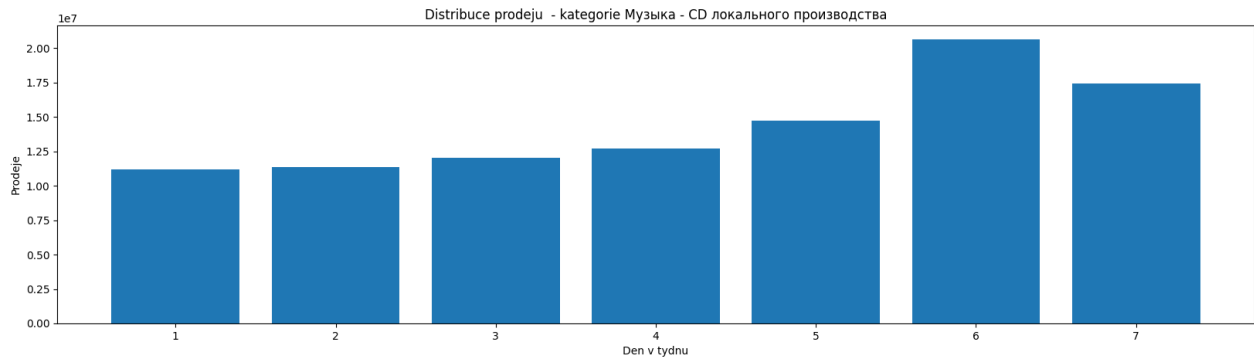
Měsíce roku - nárůst prodeje v prosinci reflektuje vánoční špičku

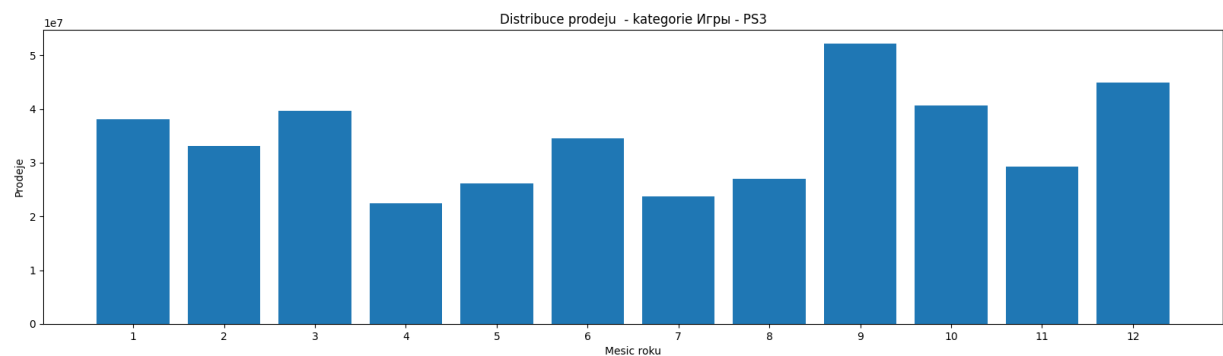
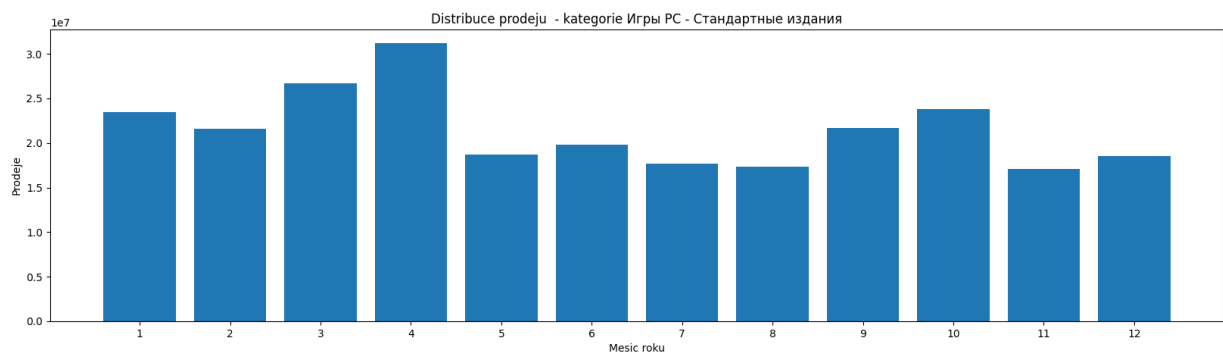
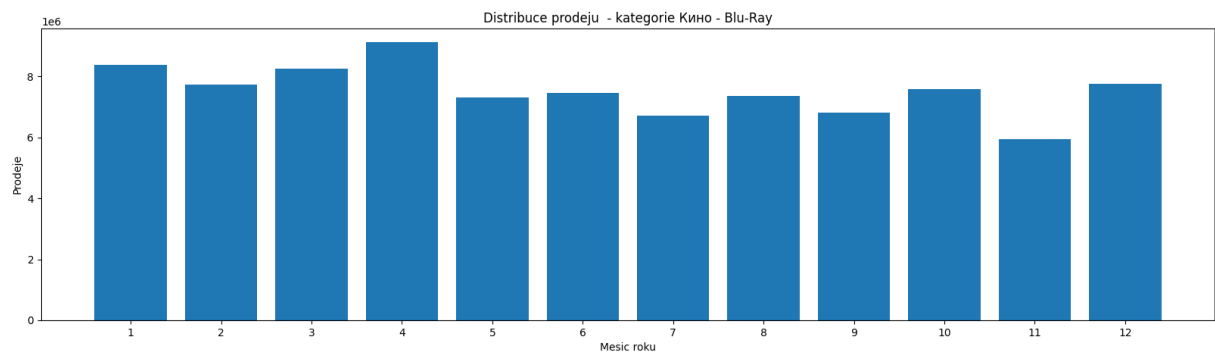
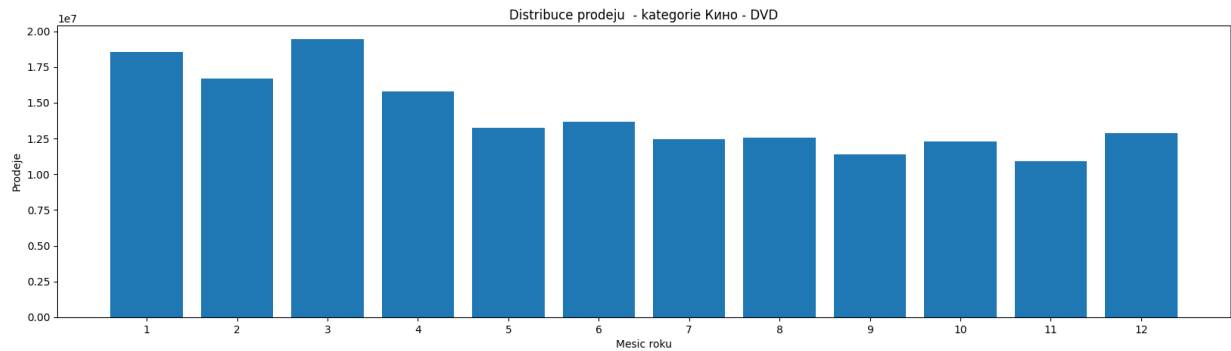
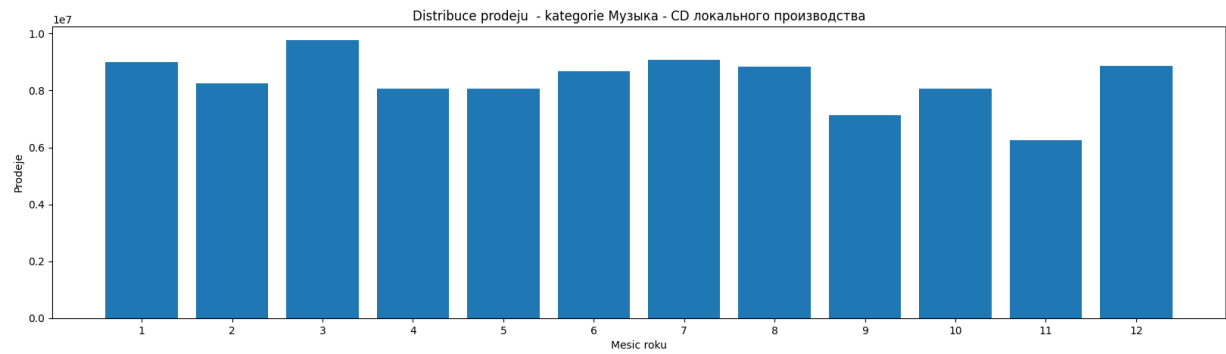


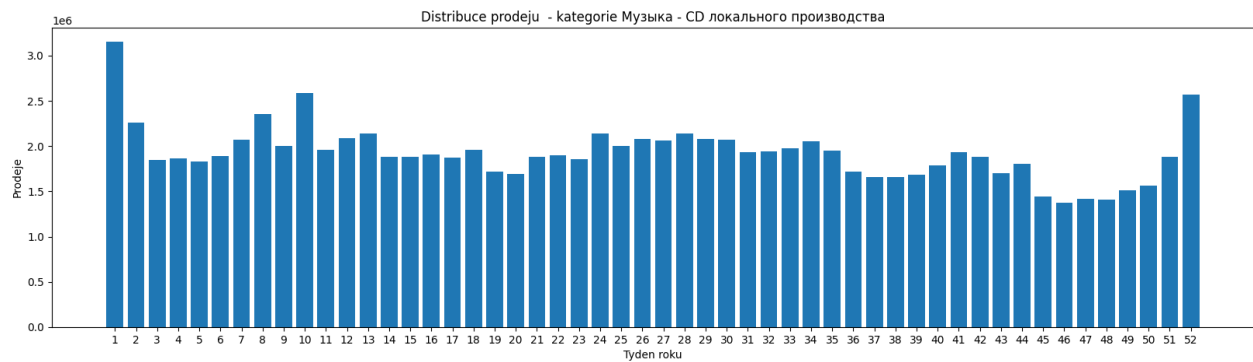
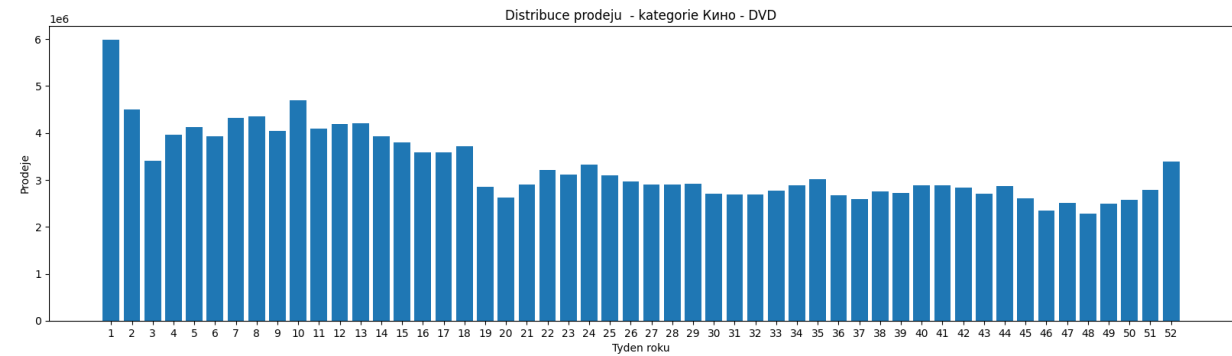
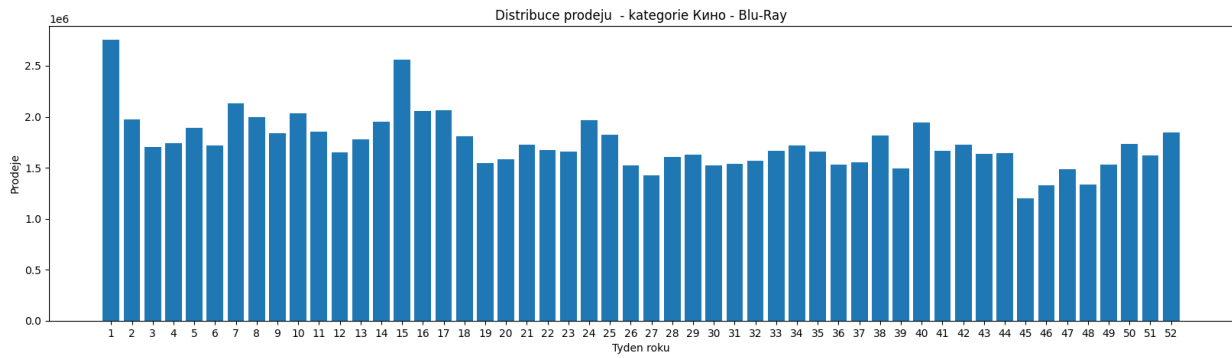
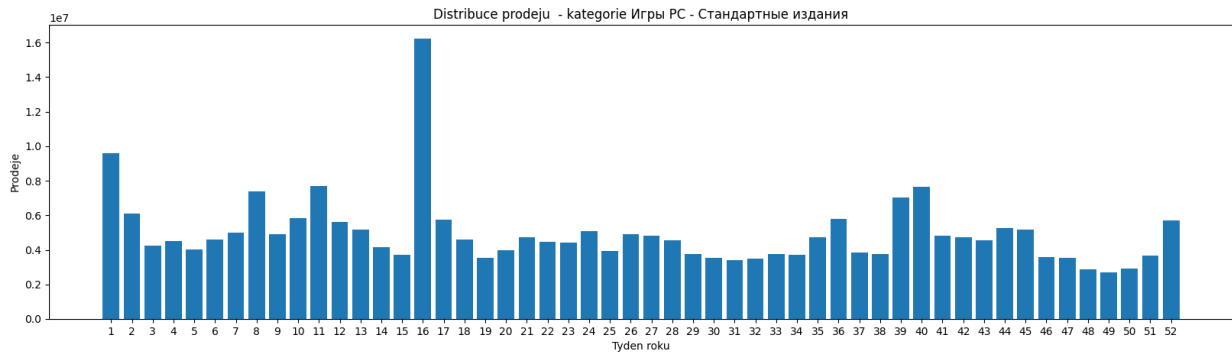
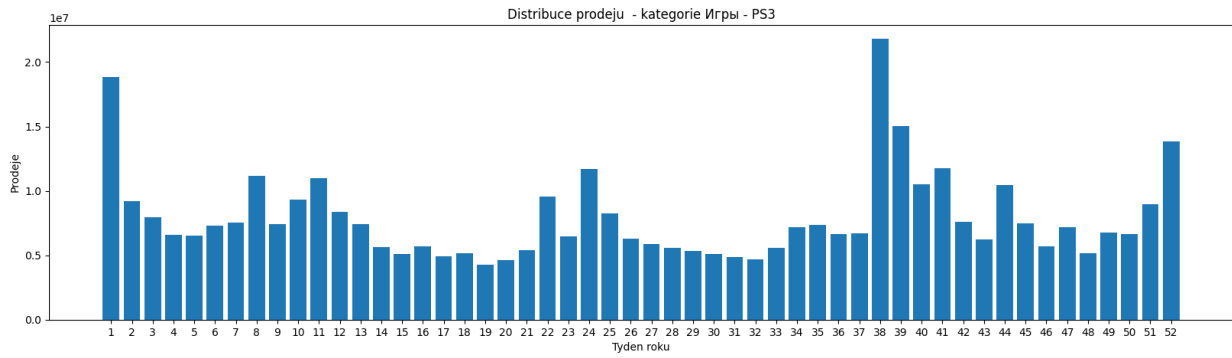
Týdny roku - opět je zřejmá sezonnost okolo Vánoc

Graf rozdělení celkové tržby podle času - 5 TOP kategorií









Shrnutí

Vypozorovali jsme, že existují určité časy/doby, kdy zákazníci nakupují více (vánoční špička, víkendy), proto by bylo vhodné reklamní aktivity zaměřovat na tyto dny. Také by bylo vhodné zaměřit aktivity na produkty z kategorie A(X) a podpořit tím náš hlavní business.