VYSOKÁ ŠKOLA EKONOMICKÁ V PRAZE

Semestrální práce z předmětu Úvod do datové analytiky – 4IZ260

Téma: Analýza sdílení jízdních kol v Londýně – 2. část

ZS 2024/2025

Autoři: Filip Morschl, Polina Stonaieva

Cvičící: Ing. David Chudán, Ph.D.

Datum vytvoření: 8.12.2024

Obsah

Úvod

Tato semestrální práce navazuje na svou první část, ve které je vysvětlena problematika, popsány atributy datasetu a sepsána obecná specifikace zadání. Následující část se zaměřuje na analýzu dat o sdílení jízdních kol v Londýně v letech 2015 a 2017.

Cílem je pomocí analytických metod získat podrobnější vhled do faktorů, které ovlivňují poptávku po těchto službách. Tato část práce zahrnuje vizualizaci dat, tvorbu agregovaných ukazatelů a aplikaci metod CF-Miner a 4ft-Miner.

Výsledky této analýzy mohou být přínosné pro správce systémů sdílených kol při optimalizaci služeb a mohou také přispět k širšímu porozumění vlivu environmentálních a sezónních faktorů na udržitelnou městskou mobilitu.

1. Vizualizace klíčových atributů

1.1. Diskretizace numerických atributů

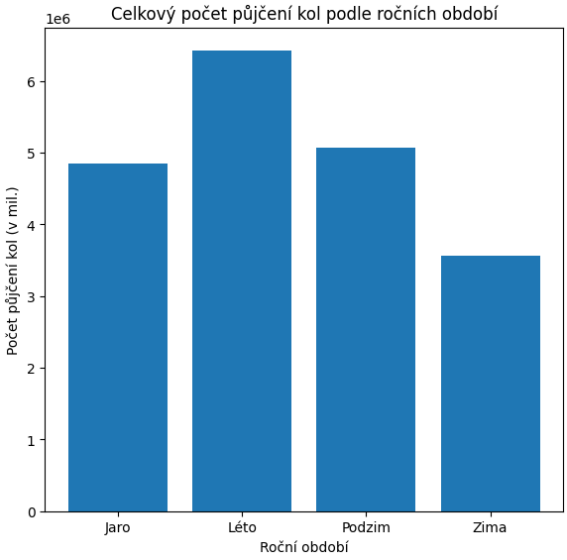
Byla provedena diskretizace atributů, které vyjadřují teplotu, pocitovou teplotu a rychlost větru.

```
df['t1_intervals'] = pd.cut(df['t1'],
                            bins = [-2, 0,  6, 12, 18, 24, 34],
                            labels = ['(-1,5 ; 0]', '(0 ; 6]', '(6 ; 12]','(12 ; 18]','(18 ; 24]','(24 ; 34]' ])
df['t2_intervals'] = pd.cut(df['t2'],
                            bins = [-7, 0,  6, 12, 18, 24, 34],
                            labels = ['(-1,5 ; 0]', '(0 ; 6]', '(6 ; 12]','(12 ; 18]','(18 ; 24]','(24 ; 34]' ])
df['wind_intervals'] = pd.cut(df['wind_speed'],
                              bins=[-1, 10, 20, 30, 40, 50, 60],
                              labels=['(0 ; 10]', '(10 ; 20]', '(20 ; 30]', '(30 ; 40]', '(40 ; 50]', '(50 ; 60]'])
```

Obrázek 1: Diskretizace numerických atributů

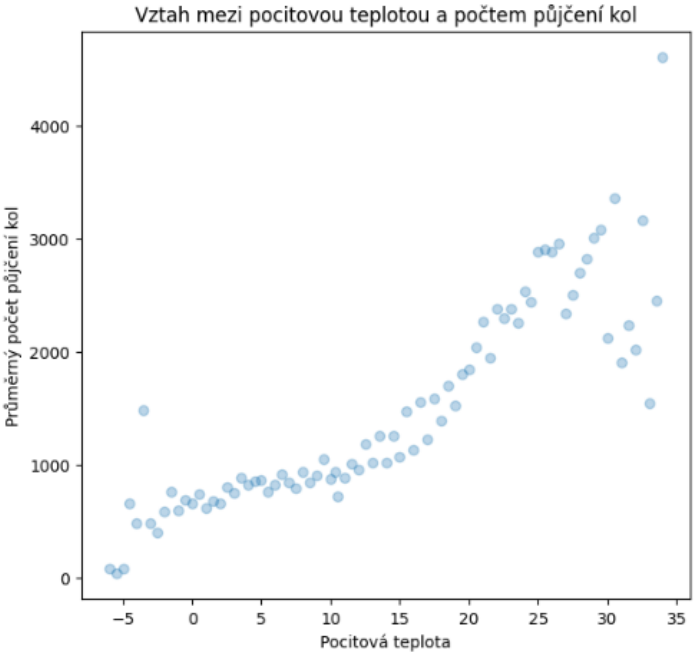
1.2. Vizualizace atributů

V prvním sloupcovém grafu lze vidět, že roční období mají vliv na počet půjčovaných kol, kdy během teplejších ročních období dochází po větší poptávce sdílených kol.

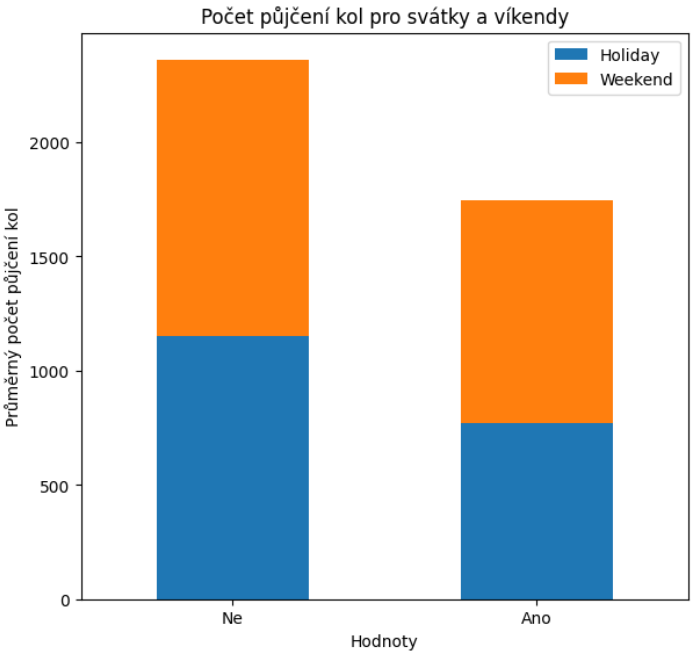


Obrázek 2: Graf – roční období

Bodový graf níže ukazuje pozitivní korelaci (0,39) mezi pocitovou teplotou a počtem půjčených kol.

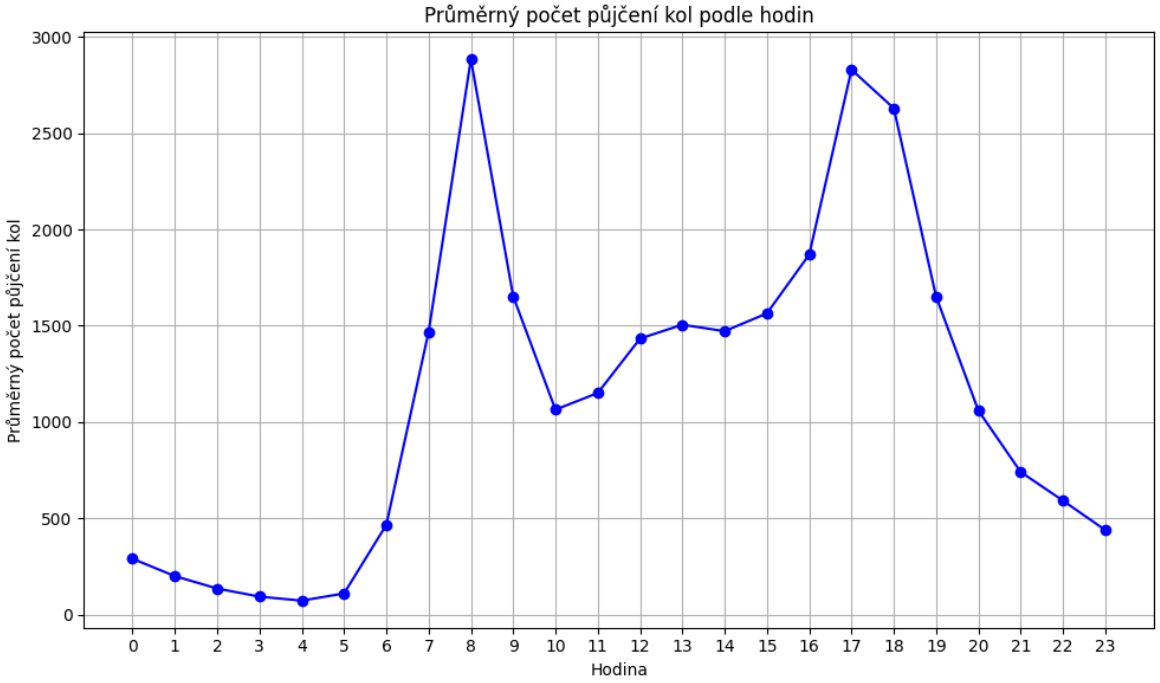


Obrázek 3: Graf – pocitová teplota



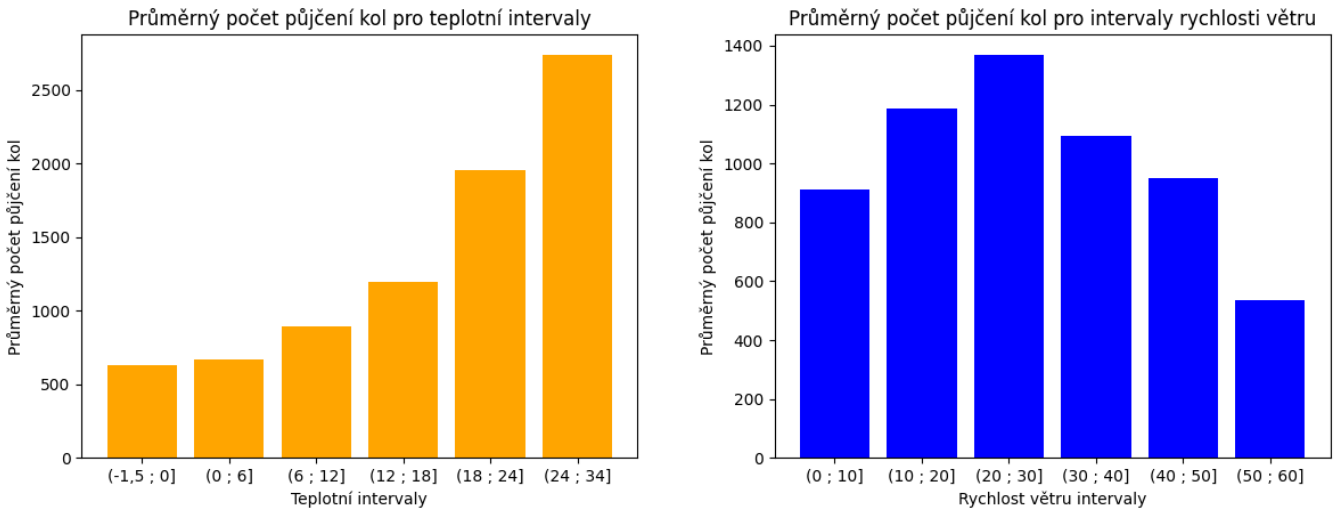
Obrázek 4: Graf – svátky a víkendy

Z následujících dvou grafů je patrné, že služby sdílených kol jsou často využívány jako dopravní prostředek do zaměstnání. Průměrný počet půjčených kol denně je vyšší v pracovní dny přibližně o 500-600 kol, viz obrázek 4 (záznamy v datasetu obsahují pozitivní hodnoty pouze v jednom ze sloupců *is_weekend* nebo *is_holiday*, nikdy v obou zároveň). Navíc proběhne nejvíce půjčení v čase mezi 7:00 – 9:00 a následně v 16:00 – 19:00.

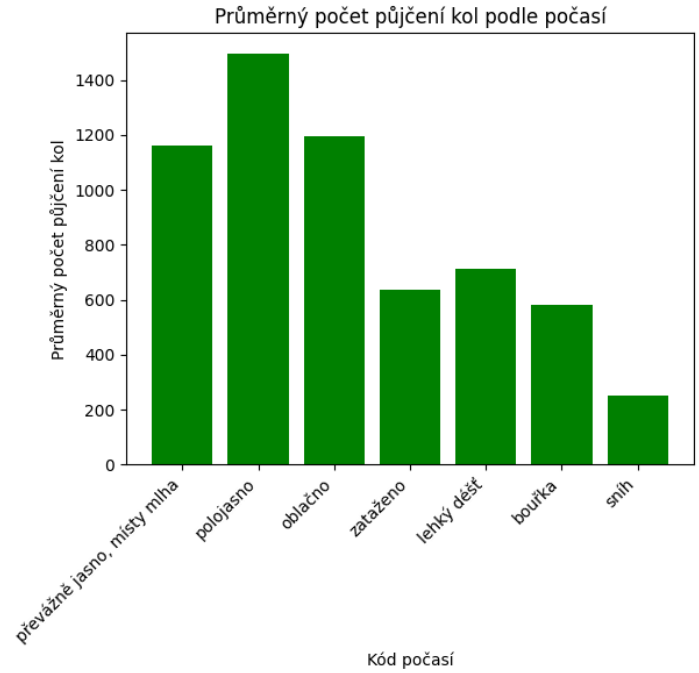


Obrázek 4: Graf – čas

Následující tři sloupcové grafy ukazují závislost mezi různými faktory a průměrným počtem půjčených kol. Z prvního grafu je patrné, že s rostoucí teplotou roste i průměrný počet půjčených kol. Naopak, rychlost větru neprokazuje výrazný vliv na využívání těchto služeb, přičemž větší pokles počtu půjčených kol je pozorován až při rychlostech větru mezi 50 a 60 km/h. Poslední graf zobrazuje, jak různá počasí ovlivňují průměrný počet půjčených kol.



Obrázek 5: 2 grafy - teplota, rychlost větru



Obrázek 6: Graf - počasí

2. Tvorba agregovaných údajů

V Python notebooku byly vytvořeny různé agregované údaje, které jsou znázorněny v následujících tabulkách.

Tabulka 1: Půjčky kol celkem

Rok	Počet půjčených kol
2015	9 738 746
2016	10 129 546
2017	37 680
Celkem	19 905 972

Tabulka 2: Průměrný počet půjčených kol pro běžné dny a svátky

Svátek	Prům. počet půjč. kol	Počet záznamů
Ne	1151,53	17 030
Ano	769,53	384

Tabulka 2 ukazuje, že svátečních dnů je v datasetu velmi málo, což naznačuje, že z pohledu byznysu není efektivní se na tyto dny zaměřovat.

Tabulka 3: Průměrný počet půjčených kol pro pracovní dny a víkend

Víkend	Prům. počet půjč. kol	Počet záznamů
Ne	1 209,27	12 444
Ano	977,42	4 970

Z tabulky 3 je patrné, že během pracovních dnů jsou služby sdílených kol využívanější než během víkendů, což potvrzuje, že lidé sdílená kola využívají ve velké míře jako prostředek dopravy do zaměstnání.

Tabulka 4: Procentuální podíl různých počasí v datasetu

Počasí	Procentuální podíl v datasetu [%]
převážně jasno, místy mlha	35,32
polojasno	23,17
oblačno	20,39
zataženo	8,41
lehký déšť	12,29
bouřka	0,08
sníh	0,34

Z tabulky 4 je zřejmé, že většina dní je pro jízdu na kole příznivá, protože převládají podmínky jako převážně jasno, polojasno nebo oblačno, které dohromady tvoří téměř 80 % všech zaznamenaných případů. To značí dobré podmínky pro úspěch byznysu.

3. Analytická otázka 1

Analytické otázky byly formulovány tak, že vycházejí z obecné specifikace zadání z první části práce. Byl přidán sloupec *is_workday*, který je pozitivní, pokud *is_holiday* a *is_weekday* jsou negativní. Poptávka po sdílených kolech byla rozdělena do pěti kategorií na základě hodnot ve sloupci *cnt*, který vyjadřuje počet půjčených kol. Rozdělení bylo provedeno pomocí hranic stanovených podle percentilů. Hodnoty do 20. percentilu byly označeny jako nízká poptávka, mezi 20. a 40. percentilem jako středně nízká poptávka atd.

Znění:

Existuje kombinace údajů o ročním období, počasí, teplotou a zdali je pracovní den, která s vysokou pravděpodobností předpovídá zvýšenou poptávku po sdílených kolech v Londýně?

3.1. Formální zápis

4ft: Ƀ(season) ∧ Ƀ(weather_code) ∧ Ƀ(t1_intervals) ∧ Ƀ(is_workday) $\Rightarrow_{0.9,\ 400}$ demand_category(*)

3.2. Nalezené odpovědi na analytickou otázku

```
List of rules:
RULEID BASE  CONF  AAD    Rule
     1   421 0.923 +1.309 season(0 1) & weather_code(2) & t1_intervals((18 ; 24]) & is_workday(1) => demand_category(Středně vysoká poptávka Vysoká poptávka) | ---
     2   444 0.927 +1.319 season(0 1) & weather_code(2) & t1_intervals((18 ; 24] (24 ; 34]) & is_workday(1) => demand_category(Středně vysoká poptávka Vysoká poptávka) | ---
     3   418 0.923 +1.308 season(1) & weather_code(2) & t1_intervals((18 ; 24] (24 ; 34]) & is_workday(1) => demand_category(Středně vysoká poptávka Vysoká poptávka) | ---
     4   464 0.926 +1.317 season(1 2) & weather_code(2) & t1_intervals((18 ; 24]) & is_workday(1) => demand_category(Středně vysoká poptávka Vysoká poptávka) | ---
     5   495 0.930 +1.327 season(1 2) & weather_code(2) & t1_intervals((18 ; 24] (24 ; 34]) & is_workday(1) => demand_category(Středně vysoká poptávka Vysoká poptávka) | ---
```

Obrázek 7: Výstup 4ft-Mineru

3.3. Interpretace vybraného pravidla

Pravidlo č. 2:

Podmínky:

- roční období: jaro a léto,
- počasí: polojasno,
- teplota: 18-34 °C,
- pracovní den: ano.

Výsledek:

- Středně vysoká až vysoká poptávka.

Toto pravidlo naznačuje, že během jarního a letního období, kdy je polojasno, teplota je mezi 18 a 34 °C a je pracovní den, tak se poptávka v 92,7 % případů pohybuje mezi středně vysokou a vysokou.

AAD +1,308 znamená, že tato kombinace faktorů také zvyšuje pravděpodobnost vysoké poptávky po kolech oproti průměru.

Byla provedena interpretace pouze jednoho pravidla, protože všechna nalezená pravidla jsou si velmi podobná a výsledné interpretace by se příliš nezměnily. Nejvýznamnějšími faktory pro zvýšenou poptávku jsou polojasné počasí, teplota v rozmezí 18-34 °C a zdali je pracovní den. Roční období naopak nemá výrazný vliv, protože pravidla pokrývají různé sezóny, ale výsledky zůstávají podobné.

4. Analytická otázka 2

Znění:

Pro kterou podmnožinu danou kombinací ročního období, počasí, teploty a rychlosti větru je rostoucí četnost pro kategorie poptávky?

4.1. Formální zápis

CF: $\approx_{\text{STEPS-UP=4}}$ demand_category / season, weather_code, t1_intervals, wind_intervals

4.2. Nalezené odpovědi na analytickou otázku

```
List of rules:
RULEID BASE  S_UP  S_DOWN Condition
     1  1396     4     0 season(0 1) & weather_code(3 4 7) & t1_intervals((12 ; 18] (18 ; 24]) & wind_intervals((10 ; 20] (20 ; 30])
     2  1301     4     0 season(1) & weather_code(2 3 4) & t1_intervals((12 ; 18] (18 ; 24]) & wind_intervals((0 ; 10] (10 ; 20])
     3  1043     4     0 season(1) & weather_code(3 4 7) & t1_intervals((12 ; 18] (18 ; 24]) & wind_intervals((10 ; 20] (20 ; 30])
```

Obrázek 8: Výstup CFmineru

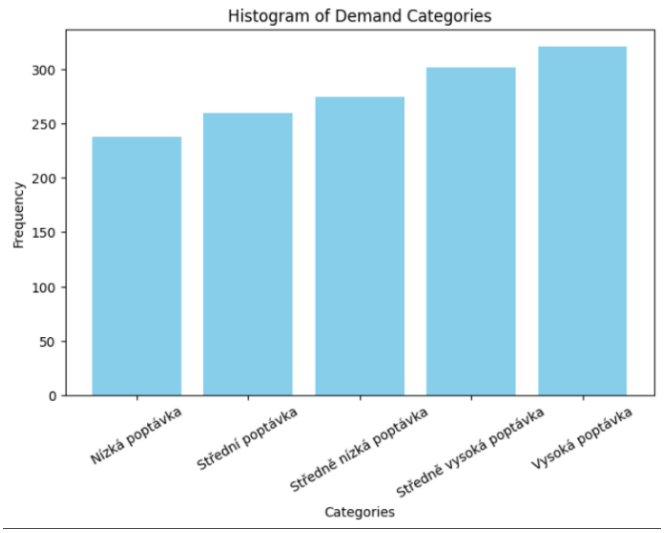
4.3. Interpretace vybraného pravidla

Pravidlo č. 1:

Podmínky:

- roční období: jaro a léto,
- počasí: oblačno, zataženo, lehký déšť nebo přeháňka,
- teplota: mezi 12–24 °C,
- rychlost větru: mezi 10–20 20–30 km/h.

Během jara a léta, když je oblačno, zataženo nebo lehký déšť a teplota je mezi 12–24 °C, s rychlostí větru mezi 10–30 km/h, je vidět rostoucí četnost pro kategorie poptávky.



Obrázek 9: Graf - CFminer

5. Analytická otázka 3

Znění:

Existuje taková podmnožina, která je definovaná kombinací ročního období, počasí, teploty a rychlosti větru, ve které převažuje pouze jedna kategorie poptávky?

5.1. Formální zápis

CF: $\approx_{RelMax=0,6}$ demand_category / season, weather_code, t1_intervals, wind_intervals

5.2. Nalezené odpovědi na analytickou otázku

```
List of rules:
RULEID BASE  S_UP  S_DOWN Condition
     1   222     2      1 season(0 1) & weather_code(1) & t1_intervals((18 ; 24] (24 ; 34]) & wind_intervals((20 ; 30])
     2   204     2      1 season(0 1) & weather_code(1 2) & t1_intervals((24 ; 34]) & wind_intervals((10 ; 20] (20 ; 30])
     3   209     2      1 season(0 1) & weather_code(1 2 3) & t1_intervals((24 ; 34]) & wind_intervals((10 ; 20] (20 ; 30])
     4   204     2      1 season(1) & weather_code(1 2 3) & t1_intervals((24 ; 34]) & wind_intervals((10 ; 20] (20 ; 30])
     5   221     2      1 season(1 2) & weather_code(1) & t1_intervals((18 ; 24] (24 ; 34]) & wind_intervals((20 ; 30])
     6   204     2      1 season(1 2) & weather_code(1 2) & t1_intervals((24 ; 34]) & wind_intervals((0 ; 10] (10 ; 20])
     7   236     2      1 season(1 2) & weather_code(1 2) & t1_intervals((24 ; 34]) & wind_intervals((10 ; 20] (20 ; 30])
     8   207     2      1 season(1 2) & weather_code(1 2 3) & t1_intervals((24 ; 34]) & wind_intervals((0 ; 10] (10 ; 20])
     9   241     2      1 season(1 2) & weather_code(1 2 3) & t1_intervals((24 ; 34]) & wind_intervals((10 ; 20] (20 ; 30])
```

Obrázek 10: Výstup - CFminer 2

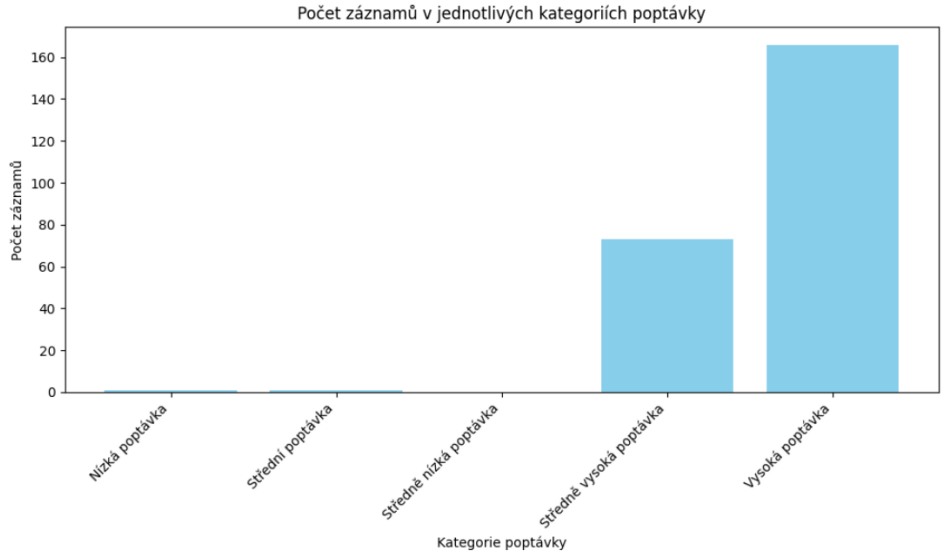
5.3. Interpretace vybraného pravidla

Pravidlo č. 9:

Podmínky:

- roční období: léto a podzim,
- počasí: jasno, polojasno nebo oblačno,
- teplota: mezi 24–34 °C,
- rychlost větru: mezi 10–30 km/h.

Během léta a podzimu, za jasného, polojasného nebo oblačného počasí, s teplotou mezi 24–34 °C a rychlostí větru mezi 10–30 km/h, je vidět výrazná převaha pro kategorii „vysoká poptávka“.



Obrázek 11: Graf – CFminer 2

Závěr

Tato semestrální práce se zaměřila na analýzu dat o sdílení jízdních kol v Londýně v letech 2015 až 2017 a prozkoumání faktorů ovlivňujících poptávku po této službě. Práce se věnovala úvodní analýze a následně aplikovala různé analytické metody, včetně CF-Miner a 4ft-Miner, k získání hlubších poznatků.

Pomocí vizualizací a tvorby agregovaných údajů byly zjištěny klíčové body:

- teplejší dny zvyšují poptávku po sdílených kolech,
- rychlost větru nemá významný vliv na počet půjčených kol,
- nejvyšší poptávka se objevuje během pracovních dní, zejména v dopravních špičkách ráno a odpoledne.

Díky aplikaci metod CF-Miner a 4ft-Miner byla identifikována pravidla, která potvrdila, že klíčovými faktory pro vysokou poptávku je teplota, počasí a pracovní dny.

Tato zjištění mohou být přínosná pro správce systému sdílených kol při optimalizaci služeb, zejména v období vysoké poptávky.

Seznam obrázků

Seznam tabulek