Univerzita Jana Evangelisty Purkyně v Ústí nad Labem

Přírodovědecká fakulta

Katedra informatiky

OLAP a DuckDB

Seminární práce

Rok: 2025

Vypracoval: Valdemar Pospíšil

Obsah

1 Instalace DuckDB

Pro projekt jsem si zvolil databázový systém **DuckDB**, který je vhodný pro OLAP analýzy a snadno se instaluje pomocí Python knihovny:

```
pip install duckdb
```

Výhodou DuckDB je jeho jednoduché použití přímo z Pythonu bez nutnosti serveru.

2 Dataset

Pro projekt jsem si stáhl dataset **UFO Sightings** z Kaggle (`https://www.kaggle.com/datasets/sahityasetu/ufo-sightings`). Dataset obsahuje informace o pozorování UFO včetně času, místa, popisu a tvaru objektu.

3 Vytvoření datového skladu

Vytvořil jsem datovou strukturu ve tvaru **hvězdy (star schema)** s jednou faktovou tabulkou a třemi dimenzními tabulkami.

3.1 Dimenzní tabulky

- **dim_ufo**: obsahuje různé tvary UFO.
- **dim_time**: obsahuje časové údaje (rok, měsíc, den, hodina).
- **dim_location**: obsahuje údaje o zemi, regionu a lokalitě.

3.2 Faktová tabulka

- **fact_sightings**: obsahuje jednotlivá pozorování, která jsou propojena na dimenze přes cizí klíče.

4 Načítání dat a vytvoření tabulek

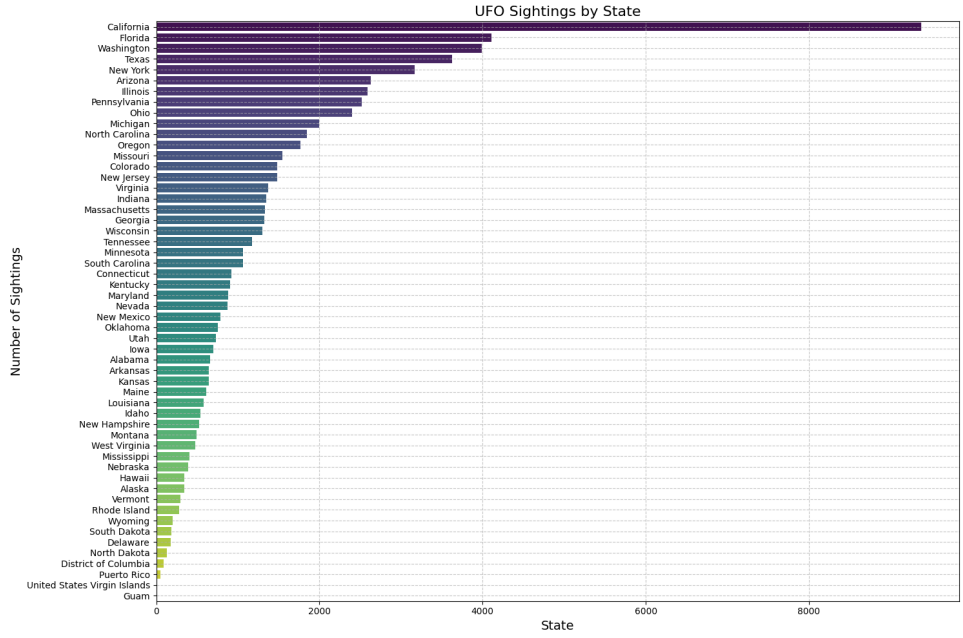
Celý proces byl realizován v Python skriptu pomocí knihovny DuckDB. Byla načtena původní data a vytvořeny potřebné tabulky.

5 Analytické dotazy

Bylo vytvořeno několik analytických dotazů nad datovou strukturou. Výstupy byly vizualizovány pomocí knihoven jako **Matplotlib**, **Seaborn**, **Tabulate** a **Folium**.

5.1 Rozložení pozorování UFO podle států

- Výstup: mapa počtu pozorování v jednotlivých státech USA.



Obrázek 1: Počet pozorování UFO podle států

5.2 Distribuce délky pozorování

- Výstup: histogram délky pozorování UFO v sekundách.

5.3 Pozorování v průběhu dne

- Výstup: graf ukazující počet pozorování v různých hodinách dne.

6 Data mining

Pro analýzu dat jsem vyzkoušel několik metod dolování znalostí:

6.1 Shlukování (Clustering)

Použil jsem metodu K-Means clusteringu k identifikaci oblastí s častým výskytem pozorování UFO.

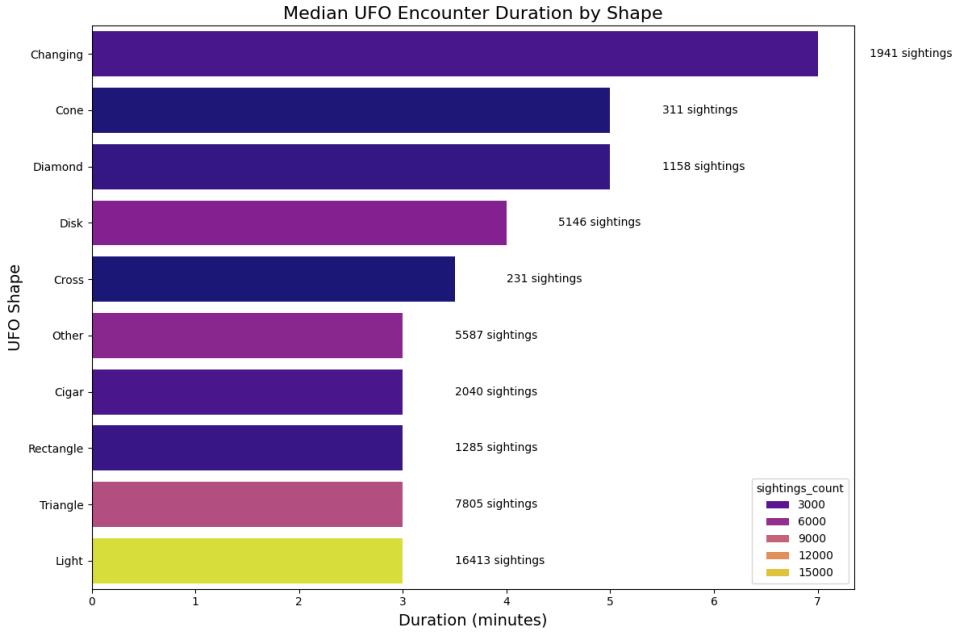
6.2 Asociační pravidla

Analyzoval jsem textová data z popisů pozorování a hledal časté kombinace slov.

7 Závěr

V projektu jsem úspěšně:

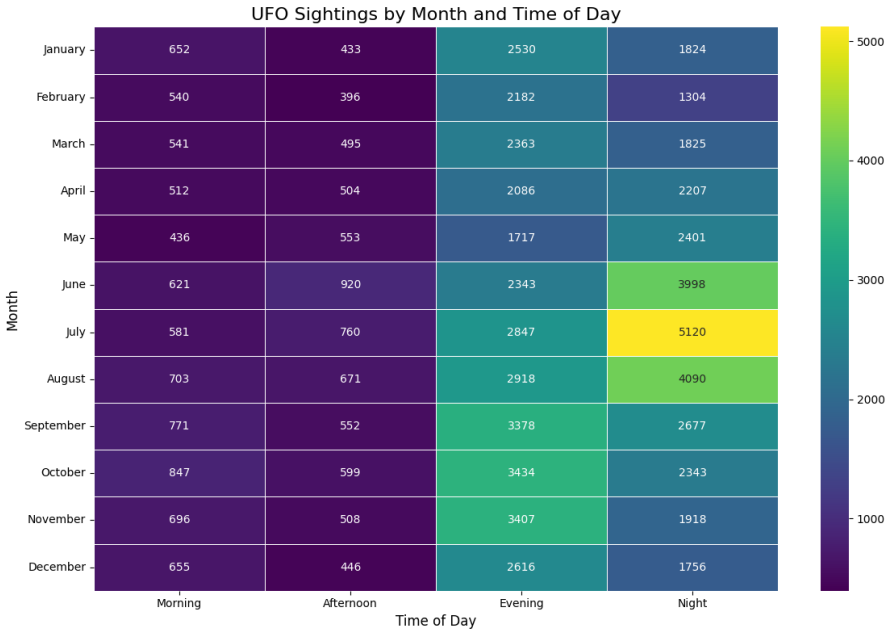
- nainstaloval a využil DuckDB pro OLAP analýzy,
- vytvořil datový sklad ve struktuře hvězdy,



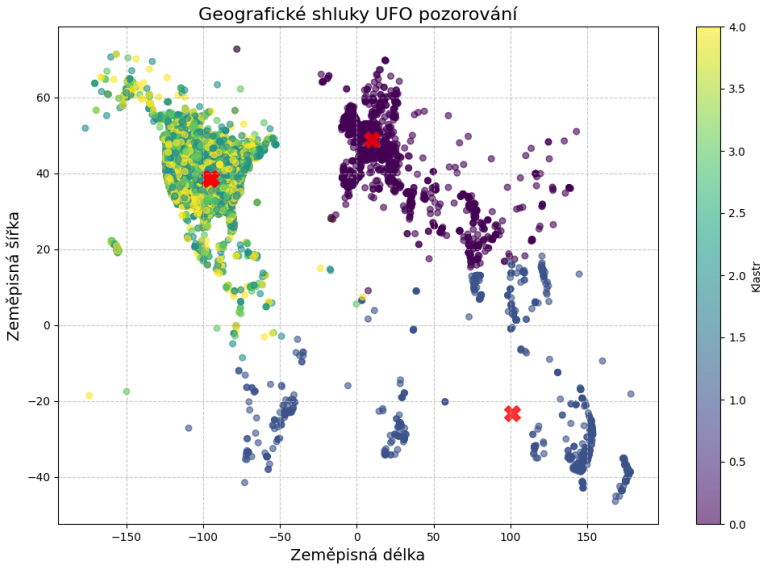
Obrázek 2: Délka pozorování UFO

- provedl analýzy nad daty pomocí SQL dotazů a Pythonu,
- aplikoval metody dolování dat (shlukování, asociační pravidla).

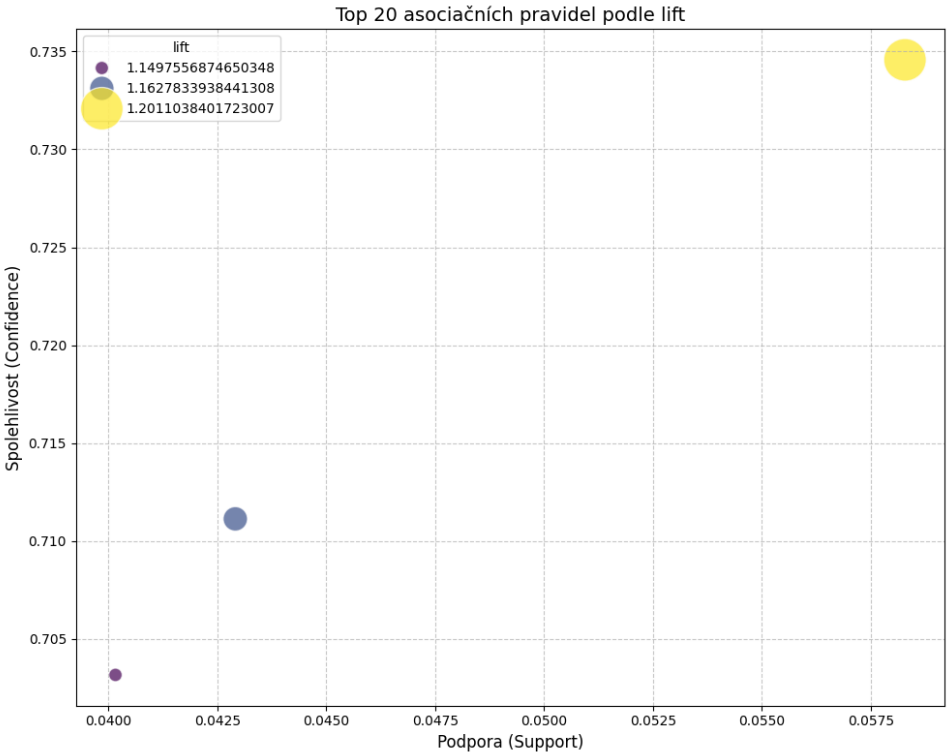
Projekt je kompletně připraven v repozitáři a doplněn o vizualizace výsledků.



Obrázek 3: Distribuce pozorování podle hodin



Obrázek 4: Mapa shluků pozorování UFO



Obrázek 5: Asociační pravidla z popisů pozorování