Dinosauří svět

Jste hlavní datový analytik ve světově nejrozšířenějším Dinosauřím parku. Tento park, rozprostírající se na tisících hektarech, je domovem pro desítky druhů dinosaurů, od malých a obratných Dromeosaurů po mohutné a impozantní Brachiosaury. Jakožto datový analytik je vaším úkolem sledovat a analyzovat velké množství dat shromážděných parkem - od informací o jednotlivých druzích dinosaurů po data o jejich populaci v jednotlivých zónách a oblastech parku.

Vaše data pocházejí z řady zdrojů a jsou v různých formátech, což vyžaduje komplexní předzpracování dat a čištění dat. Výsledky vaší analýzy a vizualizace dat jsou klíčové pro strategické rozhodování parku a jeho řízení, včetně rozhodnutí o rozpočtu, plánování a přidělování zdrojů. Při výkonu vašeho úkolu se budete muset vypořádat s řadou úkolů, včetně čištění dat, agregace dat, tvorby přehledových tabulek a vizualizace dat. Vaše výsledky budou mít rozhodující dopad na to, jak je park řízen a jak se budou vyvíjet populace dinosaurů. Připravte se na výzvu a nezapomeňte, že na vaší práci závisí životy našich prehistorických přátel!

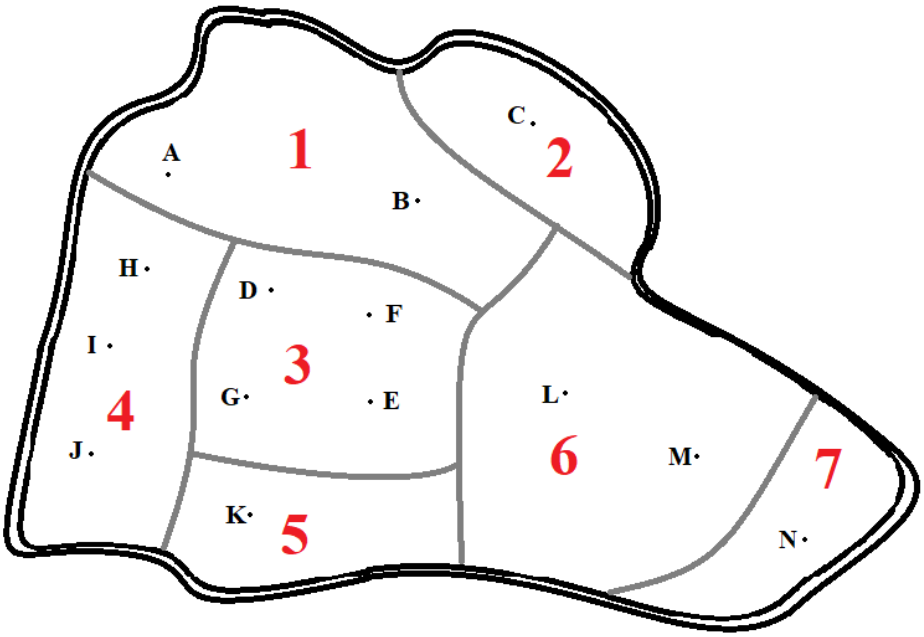
Budete odevzdávat:

- **html (nebo pdf) export z jupyter-notebooku** (kde budou všechny buňky evaluované - vyhodnocené). Nadpisem (markdown buňka začínající #) zde pak zdůrazněte výsledky k jednotlivým cílům – budou zde vidět grafy.
- **Samotný jupyter-notebook**.
- **Výsledné soubory** (DinoList.csv, agregace.csv, zona_1.csv…zona_8.csv) – viz v textu.

Rozdělení parku

Dinosauří park je rozdělen do 7 zón (1-7) ve kterých jsou čidla které monitorují určité oblasti (A-N). Tak jak ukazuje následující graf:

Obr. 1. Mapa Dinosauřího Parku



Zdroj: Vlastní zpracování

Poznámka:

Vyjeďte z předpokladu, že snímače čidel se nikde nepřekrývají a dohromady vždy pokrývají 100% dané zóny. Tedy můžete provést součet hodnot z čidel a dostat agregátní hodnotu pro celou zónu.

Data o dinosaurech

Použijte přiložený datový soubor jménem **„dinolist.txt“.** Není nutné stahovat z webu žádná další data. Jedná se o základní encyklopedická data používaná v parku.

Data jsou ve stavu, v jakém vám jej vyexportoval místní IT pracovník. :)

1 Cíl:
Vytvořte soubor DinoClean.csv který bude obsahovat vyčištěný datový soubor. Prohlédněte si jednotlivé záznamy a vyřešte problematické záznamy. Tedy udělejte minimálně:

- **Sjednoťte v souboru značení chybějících proměnných tak aby odpovídala pythonu**
- **Odstraňte proměnnou Teeth, která neobsahuje skoro žádné záznamy**
- **Hodnoty proměnných pro detekci výskytu v jednotlivých kontinentech převeďte na True/False hodnoty (budou z nich teda booleans)**
- **Z proměnných milion_years, feet a případně year odstraňte text – zanechte zde jen čísla. Proměnné pak převeďte pak na číselné hodnoty – numerické.**
- **Chybějící hodnoty i numerických proměnných nahraďte prostým průměrem pro danou proměnnou – zaokrouhlete na celé číslo podle pravidel zaokrouhlování.**

Poznámka:

Na extrakci čísel z stringu lze dělat mnoha způsoby Ale asi nejjednodušší pro vás je použít z knihovny re následující postup:

```
import re

re.findall("[0-9]+", "test 1234")
```

V tomto případě ze stringu „test 1234“ extrahujeme všechna čísla. První argument, který slouží jako pattern, zůstane i u vás stejný a není třeba ho měnit. Samozřejmě, toto lze provést i mnoha jinými způsoby.

Data o oblastech:

Tyto data jsou očištěná. Obsahují jen jména dinosaurů a jejich počty v dané oblasti v jednotlivých letech.

Vytvořte následně:

Cíl 2: Cílem je zjistit počty dinosaurů v celém parku dohromady

- **Vytvořte souhrnnou tabulku za celý park. V řádkách budou jednotliví dinosauři a v sloupcích jejich počty – v celém parku v jednotlivých letech**

- **Tedy v řádkách budou pro jednotlivé dinosaury záznamy a ve sloupečkách jednotlivé roky.**
- **Tabulku setřiďte podle jmen (abecedně), zobrazte její hlavičku v jupyter notebooku**
- **Uložte a odevzdejte jako soubor „agregace.csv“**

Tabulky může vypadat například nějak takto (jen ukázkový příklad prvních pár sloupců a řádků):

	rok_1986	rok_1987	…
Aurecil	10	15	…
Pterodaktyl	11	19	…
…	…	…	..

Cíl 3: Přehledová tabulka pro každou zónu
Pro jednotlivé zóny (1-7) vytvořte přehledové tabulky, kde bude:

- **Průměrná váha dinosaura v dané zóně na základě encyklopedických dat.**
- **Průměrná výška dinosaura v dané zóně na základě encyklopedických dat.**
- **Výpočet těchto statistik bude vycházet z počtu dinosaurů v dané zóně a dat po očištění ze souboru DinoList.csv – vypočtěte vážný průměr.**
- **Tedy pro každou zónu vznikne jedna tabulka, která bude mít dvě řádky (pro váhu a výšku) a 10 sloupců (pro jednotlivé roky)**
- **Tabulky uložte jako „zona_1.csv“…. až „zona_7.csv“**

Poznámka: Zpracování dat, a i ukládání udělejte v cyklu. Pokud to tak nebude budu to hodnotit drobnou penalizací.

Příklad tabulky pro jednu zónu:

	Ausunecosaurus	Betasaurus	…
vaha	100	15	…
vyska	50	4	…

Cíl4: Vyberte náhodně (zcela náhodně vygenerujte – pomocí numpy random) jméno pro jednoho dinosaura a pro toho dinosaura vytvořte grafy:

- **Koláčový graf zobrazující jeho počty v jednotlivých zónách (1-7) pro rok 1990**
- **Koláčový graf zobrazující jeho počty v jednotlivých oblastech (A-N) pro rok 1990**
- **Linioví graf zobrazující vývoj počtu kusů dinosaura v čase (osa x – roky, osa y – počty). Chci, aby zde byla linie pro každou zónu kde se vyskytuje (tedy sedm čar)**
- **Chci, aby kód fungoval i v situaci kdy ho pustím znova a náhodně si tak vyberu jiného dinosaura!!!**