Name: Kyle Salgado-Gouker

Date: April 6, 2025

Class: DSC 680 - Professor Iranitalab

Portfolio Milestone 1 – Project 6

README.md files

Project 6

Global Theme Park Attendance & Ride Analysis

This project integrates and analyzes multiple datasets related to global theme park attendance and ride features. By combining data from Queue-Times, the Roller Coaster Database (RCDB), and Wikipedia, it builds a unified, structured dataset to explore patterns in park popularity, ride characteristics, and visitor behavior.

Summary

The goal was to merge and clean complex, messy data from different sources to answer questions like:

- Which parks consistently draw the most visitors worldwide?
- How do ride attributes (e.g. height, speed, capacity) correlate with popularity?
- Can inconsistent formats across sources be reconciled into a usable SQL-backed dataset?

Data Sources

- **Wikipedia** — Global theme park attendance figures (2006–2022)
- **Queue-Times API** — Real-time and historical ride and park metadata
- **Roller Coaster Database (RCDB)** — Ride features (height, length, speed, duration, etc.)

Methods

- Web scraping and API querying to gather raw data
- Data cleaning:
    - Dropped and renamed inconsistent columns
    - Fixed malformed headers and missing data
    - Removed irrelevant or duplicate records

- Merged datasets by applying fuzzy matching and name normalization
- Built a local SQL database for structured queries and analysis

Key Insights

- Created a comprehensive and queryable dataset of parks and rides
- Identified outliers and high-capacity attractions
- Detected format inconsistencies and documented corrections for future automation

Tools & Dependencies

- Python 3.8+
- JupyterLab
- pandas
- numpy
- sqlite3
- fuzzywuzzy
- requests
- BeautifulSoup
- matplotlib / seaborn (for visualizations)

Limitations & Notes

- Queue-Times API occasionally timed out or delivered incomplete results
- RCDB and Wikipedia data required manual adjustments for structural mismatches
- Some ride and park names had to be standardized manually

Author

Kyle Salgado-Gouker
November 2023

Link

https://github.com/kvsgouker/540FinalProject/tree/main