Project Overview: Global YouTube Statistics Analysis

Introduction

This project focuses on analyzing the Global YouTube Statistics.csv dataset located in the dataset/ directory. The analysis is performed in the Jupiter notebook Main.ipynb, which covers data cleaning, exploratory data analysis (EDA), and visualization. The goal is to generate insights at multiple levels—channel-level, country-level, geo-spatial, and temporal.

Key Steps

1. Data Cleaning & Pre-processing

    - Handle missing dates and inconsistent numeric formats.
    - Derive a unified Full Date column from year, month, and day fields.
    - Normalize large numerical values (subscribers, views, population).

2. Feature Engineering

    - Compute *subscribers-per-capita* to measure channel penetration relative to country population.
    - Rank top channels by subscribers and video views.
    - Categorize channels by type (Music, Entertainment, Education, etc.).

3. Exploratory Data Analysis

    - Visualize category distributions and subscriber growth trends.
    - Map top channels by latitude/longitude for geo-spatial insights.
    - Compare older vs. newer channels to study longevity and growth patterns.

4. Visualization Outputs

- Charts:

    - Top channels by subscribers/views
    - Subscriber distribution by category
    - Subscriber growth over time

    - Country penetration rates

- Advanced analyses (optional): regression, clustering, time-series forecasting.

Tools & Libraries

- Pandas: Data manipulation, cleaning, and aggregation.
- Numpy: Numerical operations and efficient computation.
- Matplotlib: Custom charts and multi-panel dashboards.
- Sea born: Statistical visualizations and correlation heat maps.

How to Run

1. Open Main.ipynb in Jupiter Notebook or JupyterLab.
2. Run all cells sequentially.
3. Ensure dependencies (pandas, numpy, matplotlib, sea born) are installed in your environment.
4. A virtual environment is available at .venv/ for reproducibility.

Next Steps

- Produce reproducible Python scripts for core ETL and visualization tasks.
- Add a requirements.txt file for dependency management.
- Create a concise README with instructions and sample outputs.

🔎Overview

This project demonstrates how data science workflows can be applied to real-world media datasets. By combining ETL (Extract, Transform, and Load) processes with exploratory analysis and visual storytelling, the notebook provides actionable insights into YouTube's global ecosystem.

Key contributions include:

- A scalable framework for analyzing large datasets with millions of records.
- Geo-spatial mapping of channels to highlight regional dominance.
- Temporal analysis to compare legacy channels with newer entrants.
- Socio-economic correlations (population, education, unemployment) to contextualize digital adoption.

This project is designed to be extendable: future work may include predictive modeling (subscriber growth forecasting), clustering channels by performance, and building interactive dashboards for stakeholders.

📂 GitHub Repository

https://github.com/ajay-8897/python_practice