Project Report

Introduction

This project aims to perform comprehensive data analysis on a dataset involving video content, focusing on identifying top-performing categories based on average views. The goal is to clean, analyze, and transform the data into insights using Python and SQL, and then visualize it using Power BI.

Abstract

The project workflow includes reading a dataset, performing necessary data cleaning, conducting data analysis using Python (Pandas, NumPy), and writing SQL queries to extract actionable insights. Finally, the refined data is exported and visualized in Power BI to uncover patterns, such as the most engaging video categories based on user views.

Tools Used

- Python (Google Collab): For data loading, cleaning, and preliminary analysis.
- Pandas & NumPy: For data manipulation and computation.
- SQL (SQLite): For advanced querying and insights generation.
- Power BI: For building interactive dashboards and visual storytelling.

Steps Involved in Building the Project

1. Import Libraries:

Imported essential Python libraries like pandas, numpy, and sqlite3.

2. Upload and Read Dataset:

Used pandas.read_csv() to load the dataset containing video-related data.

3. Data Cleaning:

- Checked for null values and duplicates.

- Handled missing data.

- Standardized column names for consistency.

4. Data Analysis:

- Explored key statistics like total views, likes, and comments.

- Categorized data based on metrics such as country, tags, and video title.

5. SQL Queries:

- Created a SQLite connection.

- Wrote SQL queries to compute average views per category.

- Extracted top categories based on these averages.

6. Export to Power BI:

- Cleaned and SQL-processed data was exported as .csv.

- Imported into Power BI to design visualizations like bar charts and filters.

Conclusion

The project successfully demonstrated the pipeline from raw data ingestion to polished insights. The SQL-driven analysis highlighted the most engaging content categories. Visualizations in Power BI made these insights accessible and interactive, supporting data-driven decisions in content strategy.