Documentation for Social Media Analysis

Project Overview:

The goal of this project is to integrate datasets collected from different sources to provide a unified view for insightful analysis and decision-making to foster business growth. The project involves data cleaning, data integration, sentiment analysis, trend detection, and data visualization.

Steps and Methodologies

**Data Cleaning:**

Extracting Date Components:

Extract the month and year from the date\_added column for each dataset.

**Delimited Process:**

Use a delimited process on genre, cast, director, and country to separate and normalize values.

**Database Setup:**

Create a database and import the datasets.

Check for duplicate values (none found).

Check and handle null values using imputation and deletion methods where appropriate.

Handling Null Values:

For columns like year\_added, which had null values, SQL limitations prevented imputation directly, so these were addressed later using Python.

**Data Integration:**

**Merging Datasets:**

Use the UNION ALL SQL command to merge datasets into a single unified dataset.

Export this merged dataset for further analysis in Python.

**Data Preprocessing:**

**Data Import:**

Import the merged dataset into Python using Pandas: df = pd.read\_csv('filename.csv').

**Exploratory Data Analysis (EDA):**

**Check for null values and handle them:**

Impute missing year\_added values using df['year\_added'].fillna('value').

Drop rows with missing values in the director column due to the small number of missing values.

**Sentiment Analysis:**

**Using TextBlob Library:**

Perform sentiment analysis on the content descriptions.

Categorize sentiments into positive, negative, and neutral.

Group sentiments by genre to observe trends:

Drama: Highest positive and negative sentiment content.

Action-Adventure: Highest neutral content.

**Trend Detection:**

**Analyze Trends:**

Analyze the release year trends for TV shows and movies across each OTT platform.

Calculate the average count of released content each year.

Determine the percentage distribution of TV shows and movies across platforms.

Identify top content-producing countries.

**Data Visualization:**

**Visualization with Tableau:**

Visualize cleaned and processed data to derive actionable insights.

**Key analyses include:**

* Number of contents on each OTT platform.
* Top 10 genres by the number of contents.
* Top actors with the most content.
* Top directors with the most content.
* Top countries by the number of contents.
* Distribution of contents by type (TV shows vs. movies).
* Top 10 ratings by the number of contents.
* Contents added by month.
* Contents added by year.
* Number of contents released by year.

**Conclusion**

**Unified Data Source:**

Successfully integrated multiple datasets into a unified view, enabling comprehensive analysis.

**Enhanced Data Quality:**

Improved data quality through systematic cleaning and preprocessing, ensuring accuracy in analysis.

**Sentiment Insights:**

Identified sentiment distribution across genres, highlighting trends in content reception.

**Trend Analysis:**

Provided insights into content release patterns over the years, assisting in strategic planning for content production.

**Platform Performance:**

Analyzed performance metrics across OTT platforms, including content volume and genre distribution.

**Actionable Visualizations:**

Developed detailed visualizations for key metrics, facilitating data-driven decision-making and strategy formulation.