Project Title: Youtube Data Analysis EDA

Introduction:

This project involved performing exploratory data analysis (EDA) on a Youtube dataset to uncover insights. The key steps taken were:

1. Data Loading and Cleaning: The raw Youtube data with over 500,000 rows was loaded into a Pandas DataFrame. The data was checked for null values and these were handled appropriately.

2. Sentiment Analysis: The textblob package was used to perform sentiment analysis on the video comment text. This generated a polarity score between -1 and 1 for each comment, indicating negative, neutral or positive sentiment.

3. Word Cloud Analysis: Word clouds were generated to visualize the most frequent positive and negative words in the comments.

4. Emoji Analysis: The top 10 most used emojis in the comments were analyzed.

5. Video Category Analysis: Analytics were performed to find insights related to video categories, including: most viewed categories, audience engagement, correlation between views/likes/dislikes, top channels by trending videos, most liked channels, most viewed channels.

6. Title and Tag Analysis: An analysis was done to uncover any relationships between punctuations used in titles & tags and the video views, likes, dislikes and comments.

Data Loading and Cleaning:

The raw Youtube dataset was loaded into a Pandas DataFrame in Python. Initial exploratory data analysis revealed that the dataset had over 500,000 rows of data.

The data was checked for any missing or null values, which can often skew results if not handled properly. A few columns were found to have some null values the amount was smaller so null values were dropped.

Handling null values ensured cleaner data and more accurate analysis during the later stages of the project.

Sentiment Analysis:

Sentiment analysis was performed on the 'comment_text' column to determine if each comment had a positive, negative or neutral sentiment.

The textblob Python package was used for this sentiment analysis, as it provides an easy way to generate a polarity score for text in Python.

First, the textblob library was imported. Then, the 'comment_text' column was iterated through and each comment was passed into a textblob object to generate the polarity score

This generated a polarity score between -1 and 1 for each row, where:

- Score > 0 indicates positive sentiment
- Score = 0 indicates neutral sentiment
- Score < 0 indicates negative sentiment

The new 'polarity' column was added to the DataFrame containing sentiment scores for each comment.

Word Cloud Analysis:

Word cloud visualizations were generated to identify the most frequent words in the positive and negative video comments.

The wordcloud Python library was used to generate these word clouds. The steps were:

- Filter the data to only include positive or negative comments based on the 'polarity' column.
- Join all these filtered comments into one large text string.
- Pass this text into the WordCloud() object and generate the word cloud image.

This revealed words like 'love', 'like', 'awesome', 'best' were most common in positive remarks.

A similar process was followed for negative comments, with words like 'boring', 'worst', 'hate' emerging.

These word clouds provided quick visualization of important words in positive and negative sentiment comments.

Conclusion and Summary:

In summary, this project provided valuable exploratory data analysis on a Youtube dataset. Key skills demonstrated include:

- Loading and cleaning large datasets in Pandas
- Applying text analysis techniques like sentiment analysis using textblob
- Visualizing text data as word clouds using wordcloud library
- Uncovering insights about Youtube videos, channels and comments through category/column analysis

The analysis uncovers interesting trends about audience engagement, popular channels & categories, sentiment of comments, and impact of titles on video performance.

Next steps for this project could be building a machine learning model to predict video performance based on attributes like title and tags. The analysis done provides a solid starting point for further predictive modeling.