Exploratory Data Analysis on YouTube data

Domain: Social media

Context and Content: In a fairly recent move by Youtube, it announced the decision to hide the number of dislikes from users around November 2021. However, the official YouTube Data API allowed you to get information about dislikes until December 13, 2021. Doing an EDA-exercise can help to draw some unseen insights from this dataset.

Learning Outcome:

● Exploratory Data Analysis using Pandas.

Objective:

To do data analysis and explore the youtube dislikes dataset using numpy and pandas libraries and drive meaningful insights by performing Exploratory data analysis.

Data Description:

YouTube Dislikes Dataset:

● This dataset contains information about trending YouTube videos from August 2020 to December 2021 for the USA, Canada, and Great Britain.

● This dataset contains the latest possible information about dislikes,likes,views and more which was collected just before December 13. The information was collected by videos that had been trending in the USA, Canada, and Great Britain for a year prior.

● Dataset link: https://www.kaggle.com/datasets/dmitrynikolaev/youtube-dislikes-dataset

Attribute Information:

SL.No	Column Name	Description
1.	Video ID	Unique video id.
2.	Title	Video title.
3.	Channel ID	Id of the channel.
4.	Channel Title	Title of the channel.
5.	Published at	Video publication date.

6.	View count	Number of views.
7.	Likes	Number of likes.
8.	Dislikes	Number of dislikes.
9.	Comment Count	Number of comments.
10.	Tags	Tags (in one string).
11.	Description	Video description.
12.	Comments	20 Video comments (in one string)

Questions:

1. Import required libraries and read the provided dataset (youtube_dislike_dataset.csv) and retrieve top 5 and bottom 5 records.

2. Check the info of the dataframe and write your inferences on data types and shape of the dataset.

3. Check for the Percentage of the missing values and drop or impute them.

4. Check the statistical summary of both numerical and categorical columns and write your inferences.

5. Convert datatype of column published_at from object to pandas datetime.

6. Create a new column as 'published_month' using the column published_at (display the months only)

7. Replace the numbers in the column published_month as names of the months i,e., 1 as 'Jan', 2 as 'Feb' and so on.....

8. Find the number of videos published each month and arrange the months in a decreasing order based on the video count.

9. Find the count of unique video_id, channel_id and channel_title.

10. Find the top10 channel names having the highest number of videos in the dataset and the bottom10 having lowest number of videos.

11. Find the title of the video which has the maximum number of likes and the title of the video having minimum likes and write your inferences.

12. Find the title of the video which has the maximum number of dislikes and the title of the video having minimum dislikes and write your inferences.

13. Does the number of views have any effect on how many people disliked the video? Support your answer with a metric and a plot.

14. Display all the information about the videos that were published in January, and mention the count of videos that were published in January