**Analysis on “Global YouTube Statistics” dataset**

Prepared by: Sushma Sree Laskar

Student Number: 23032632

GitHub Repository:

<https://github.com/Sushma897sree/Global_YouTube_Statistics.git>

In this report, we will conduct an exploratory data analysis on the "Global YouTube Statistics 2023" dataset. This dataset contains information about YouTube videos worldwide, including features such as video title, channel title, category, views, likes, dislikes, and comment count. We will explore various relationships within the dataset using descriptive statistics and visualization techniques.

Dataset Description

The "Global YouTube Statistics 2023" dataset consists of 5 rows and 28 columns. Below are the features included in the dataset:

Video id: Unique identifier for each video.

Trending\_date: Date when the video was trending.

title: Title of the video.

Channel\_title: Title of the YouTube channel that uploaded the video.

Catageory\_id: ID representing the category of the video.

views: Number of views the video received.

likes: Number of likes the video received.

These features provide a comprehensive overview of the performance and engagement metrics associated with each video in the dataset, allowing us to analyse various aspects of YouTube content and viewer interactions.

**Data Preparation**

First, we will load the dataset and before conducting the exploratory analysis, we perform essential data cleaning and preprocessing steps. This includes handling missing values, converting data types, and ensuring data consistency to ensure the reliability and accuracy of our analysis results.

**Exploratory Data Analysis**

1. Bar Chart

We begin the analysis by visualizing the distribution of video categories using a bar chart. This allows us to understand the prevalence of different content categories among the YouTube videos in our dataset.

A graph with blue bars and text

Description automatically generated

The bar chart illustrates the distribution of videos across various content categories on YouTube. Each bar represents a specific category, with the height of the bar indicating the number of videos within that category. The graph reveals notable variability in category popularity, with certain categories attracting a larger number of videos compared to others. This insight provides valuable guidance for content creators and marketers, enabling them to tailor their strategies to capitalize on popular trends and audience preferences. By optimizing content strategies based on category performance, stakeholders can enhance engagement and maximize impact on the platform. We can observe that people are keenly interested in Entertainment related videos and viewership is higher in such domain.

2.Scatter Plot: Exploring the Relationship between Video Views and Subscribers

The scatter plot below visualizes the relationship between the number of video views and the number of subscribers for YouTube channels in the dataset.

A screen shot of a graph

Description automatically generated

The scatter plot above visualizes the relationship between the number of video views and the number of subscribers for YouTube channels in the dataset. Each point on the plot represents a YouTube channel, with its position determined by the values of video views and subscribers. The x-axis represents the number of video views, while the y-axis represents the number of subscribers. By examining the distribution of points on the plot, we can gain insights into the correlation or relationship between these two metrics. This analysis can help content creators and marketers understand how video views correlate with channel subscribers, providing valuable insights into audience engagement and channel growth strategies. Further analysis, such as calculating correlation coefficients, can provide quantitative measures of the relationship between these variables, enhancing our understanding of audience behavior on YouTube.

This explanation provides a clear understanding of the scatter plot and its significance in analysing the relationship between video views and subscribers on YouTube.

3.Correlation Matrix: Analysing Relationships between Numeric Variables

A close-up of a graph

Description automatically generatedA screen shot of a computer code

Description automatically generatedThe correlation matrix heatmap illustrates the strength and direction of relationships between numeric variables in the dataset.