FUTURE VISION™
COMPUTER INSTITUTE

PROBLEM STATEMENT

5000 Youtube Channel Data Analysis

PROBLEM STATEMENT

The objective of this analysis is to clean and analyze the dataset of the Top 5000 YouTube Channels, which includes key metrics such as rank, grade, channel name, video uploads, subscribers, and video views.

The analysis aims to derive meaningful insights from this dataset, identify correlations between key metrics, and perform necessary data cleaning and transformation tasks.

By conducting this analysis, the goal is to understand how the grade, video uploads, and subscriber counts correlate with each other and to identify patterns that can be useful for content creators, marketers, or analysts interested in YouTube trends.

This analysis also helps in cleaning and transforming the dataset to make it suitable for deeper insights or predictive modeling tasks.

BUSINESS PROBLEM OVERVIEW

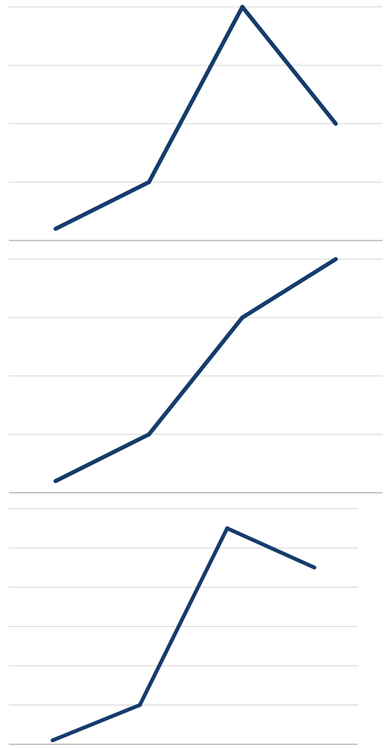
In the digital age, YouTube has become one of the largest platforms for content creation and consumption, hosting millions of channels across various genres. For businesses, marketers, and content creators, understanding the performance and characteristics of top YouTube channels is essential for gaining insights into market trends, audience preferences, and potential partnership opportunities.

The given dataset, which provides data on the Top 5000 YouTube Channels, offers a valuable opportunity to analyze key metrics such as video uploads, subscribers, and video views, helping businesses make informed decisions.

The primary business problem revolves around understanding the factors that drive YouTube channel success and how various attributes, such as channel grade, video uploads, and subscribers, are interrelated.

This analysis is not just a data exploration exercise; it has direct applications in strategic planning, content creation, influencer marketing, and audience engagement. By understanding the underlying patterns and correlations in the dataset, businesses can make data-driven decisions to enhance their presence on YouTube, tailor their content strategies, and optimize their marketing efforts.

The Top 5000 YouTube Channels dataset is a comprehensive collection of data about the most popular YouTube channels, providing valuable insights into channel performance and audience engagement. This dataset includes key attributes such as the channel's rank, grade, video uploads, subscribers, and total video views, all of which help assess the success and reach of the channels.



Each channel is assigned a rank, which indicates its relative position in terms of popularity, often based on factors like subscriber count, video views, and content uploads. This rank is initially represented with an ordinal suffix (e.g., "1st", "2nd"), but is cleaned to allow for numerical analysis.

This dataset serves multiple purposes, offering a valuable resource for businesses, marketers, and content creators looking to understand the dynamics of top YouTube channels. It can help analyze performance correlations, such as how video uploads, subscriber numbers, and views are related, and can provide insights into the most effective content strategies.

Furthermore, the data is highly useful for influencer marketing, as it allows companies to identify successful influencers based on performance metrics, leading to more targeted marketing campaigns. Overall, the Top 5000 YouTube Channels dataset provides a deeper understanding of what drives success on YouTube, and can be leveraged for strategic decision-making in digital marketing, content creation, and audience engagement.

UNDERSTANDING & DEFINING DATASET

PROJECT PIPELINE

The project pipeline can be briefly summarized in the following steps:

- Data Understanding: Here, we need to load the data and understand the features present in it. This would help us choose the features that we will need for your final model.

- Exploratory data analytics (EDA): Normally, in this step, we need to perform univariate and bivariate analyses of the data, followed by feature transformations, if necessary. For the current data set, because Gaussian variables are used, we do not need to perform Z-scaling. However, you can check if there is any skewness in the data and try to mitigate it, as it might cause problems during the model-building phase.

FUTURE VISION™
COMPUTER INSTITUTE

THANK YOU