**Data Analysis Report**

Table of Contents

[**Project Overview** 2](#_Toc173267930)

[**Dataset Information** 2](#_Toc173267931)

[**Data Loading** 2](#_Toc173267932)

[**Data Information** 2](#_Toc173267933)

[**Data Cleaning** 3](#_Toc173267934)

[**Data Analysis and Visualizations** 4](#_Toc173267935)

[**Conclusion** 7](#_Toc173267936)

# **Project Overview**

This project aims to analyze YouTube statistics using a dataset obtained from Kaggle. The analysis includes cleaning the data, exploring relationships between various features, and generating visualizations to provide insights.

## **Dataset Information**

* **Source**: Kaggle
* **File**: Global YouTube Statistics.csv

## **Data Loading**

The dataset was loaded using the pandas library. The first few rows of the dataset were inspected to understand its structure and content.

import pandas as pd

import Loading\_file as ld

data = ld.load\_file()

print(data.head())

## **Data Information**

**Data Types**

The data types of each column were reviewed to ensure proper handling of data.

print(infoFile.data\_types(data))

**Data Description**

Descriptive statistics of the dataset were obtained to understand the distribution and summary of each column.

print(infoFile.describe\_data(data))

## **Data Cleaning**

**Removing Duplicate Values**

Duplicate values in the rank column were removed to ensure data integrity.

cleaned\_data = cf.clean\_duplicates\_col(data, "rank")

**Dropping Zero Values**

Rows where the video views column had zero values were removed as they are considered useless.

Cleaned\_data = cf.clean\_col\_null(data, 'video views')

**Removing Null Values**

Rows with null or empty values in the category column were removed.

cleaned\_data = data[~data['category'].isna() & (data['category'].str.strip() != '')]

**Dropping Unnecessary Columns**

Columns that are not needed for the analysis were dropped to simplify the dataset.

cleaned\_data = cf.drop\_column(data, ["Longitude", "Latitude", "Urban\_population", "Population",

"Gross tertiary education enrollment (%)", "created\_date", "created\_month",

"lowest\_monthly\_earnings", "highest\_monthly\_earnings",

"video\_views\_for\_the\_last\_30\_days", "channel\_type\_rank",

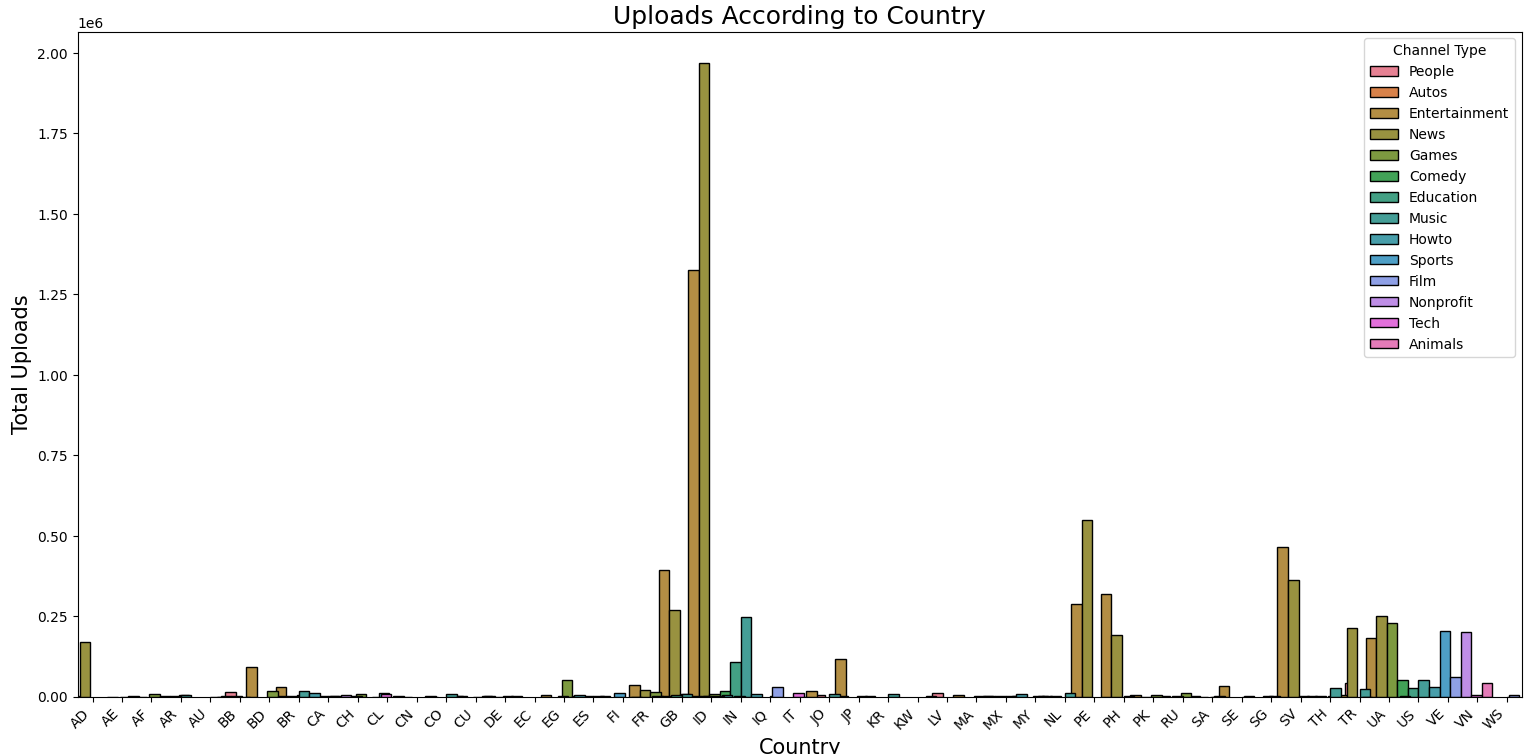
"Unemployment rate", "subscribers\_for\_last\_30\_days"])

## **Data Analysis and Visualizations**

Several visualizations were created to explore the relationships in the data.

**Bar Plot**

A bar plot was created to show the total uploads by country, separated by channel type.



**Scatter Plot**

A scatter plot was created to explore the relationship between different numeric features.

A graph of a number of dots

Description automatically generated

**Count Plot**

A count plot was created to show the distribution of a categorical feature.A diagram of a graph

Description automatically generated with medium confidence

**Category Views Bar Plot**

A bar plot was created to show the total views for each category.

A graph of different colored columns

Description automatically generated

**Scatter Plot with Regression**

A scatter plot with a regression line was created to show the relationship between the created year and rank.

A graph of a number of channels

Description automatically generated with medium confidence

**Saving the Cleaned Data**

The cleaned dataset was saved to a new Excel file for future use.

cleaned\_data.to\_excel("Data\_sets/cleaned\_data.xlsx", index=False, engine="openpyxl")

## **Conclusion**

This project involved cleaning and analyzing a YouTube statistics dataset to uncover insights and patterns. The visualizations provided a clear understanding of how different features are related. The cleaned dataset is now ready for further analysis or machine learning tasks.