Analysis of Top 5000 YouTube Channels

Intro

This project analyses top-5000-youtube-channels dataset from Kaggle.

The data is pre-processed and analyzed using Python Pandas library and visualized using Seaborn library.

With a clear objective in mind, the next step is to load and prepare the dataset for analysis. This ensures that all subsequent operations, including data exploration and visualization, are performed on clean and well-structured data.

Loading Data

The data is in CSV format and is loaded to pandas dataframe for analysis.

Once the dataset is successfully loaded, the next step involves a preliminary review to understand its structure and contents. This includes summarizing key statistics and identifying potential data quality issues.

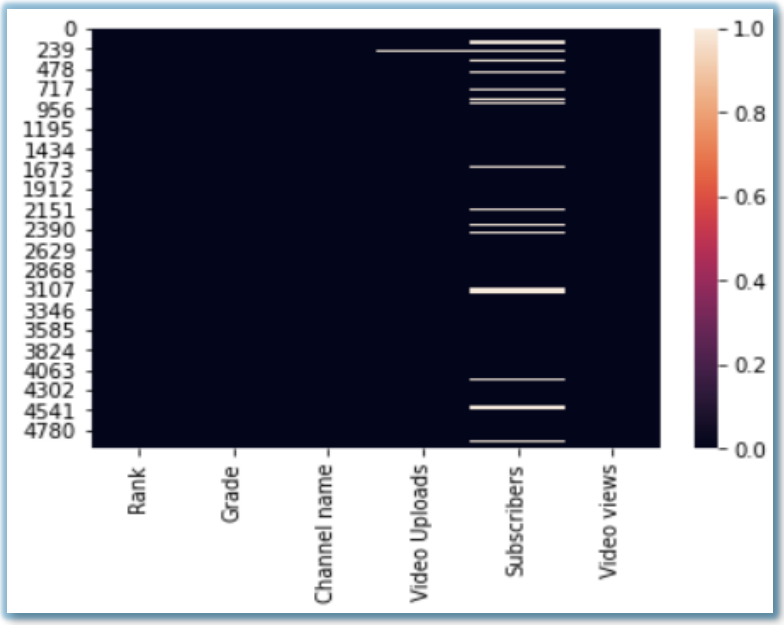
Reviewing Data

The preliminary review of the data is done using info() and describe() functions to see summary statistics of the dataset. The dataset contains 5000 rows and 6 columns, totaling 30000 data points. The column names are Rank, Grade, Channel name, Video uploads, Subscribers, Video views. Only the "Video views" column is numerical, while the other columns are of object type, which can include a mix of strings (words), numbers, or missing values.

After understanding the dataset's structure, the focus shifts to data cleaning. This critical step ensures the dataset is free from inconsistencies, missing values, and formatting issues, making it suitable for analysis and visualization.

Data Cleaning

To prepare data for analysis, visualization and potential use in machine learning model, the following steps are performed: remove missing values, clean columns and convert them to numeric data type, replace grade values with numerical values, standardize column names.



1. **Handle missing values**

At first glance there appears to be no missing values. However, upon further review, the missing values are marked as "--". There are 7.74% of missing values in column Subscribers and 0.12% of missing values in column Video uploads (See the heatmap of missing values.) The "--" symbols are replaced with NaN values, and rows containing these missing values are dropped.

2. **Clean “Rank” column**

Values in the “Rank” column are formatted as strings (e.g. “1,000th”). To clean up the values remove the comma. Remove the “st”, “nd”, “th” endings. Convert column to integer data type.

3. **Clean “Grade” column**

Map the grade from letters to numbers. Since the grades are ordered alphabetically, they are mapped to numerical values (1, 2, 3, 4, 5 etc.) to enable analysis and convert the column to numeric type.

4. **Clean “Video uploads” and “Subscribers” columns**

Both column values are only numerical so convert both columns to numeric type.

Now only the column “Channel name” remains string type column, and the rest of the columns are numerical.

5. **Standardize column names**

The column names are inconsistent; some start with uppercase letters while others start with lowercase. Standardized naming convention for columns prevents errors in future use and simplifies column retrieval.

With a cleaned and standardized dataset, additional features can now be created to enhance the analysis. These engineered features provide deeper insights and enable more nuanced exploration.

Feature engineering

A new column, “Average views”, is created by dividing the “Video views” column by the “Video uploads” column for each channel.

The engineered features are now integrated into the dataset, allowing for comprehensive exploratory data analysis (EDA). This phase uncovers patterns, trends, and relationships between variables, providing answers to key questions.

Exploratory data analysis (EDA)

EDA involves reviewing data to analyze key metrics such as average views, top channels by “Video uploads”, and creating correlation matrix to understand the relationships between numeric features.

The analysis aims to answer the following questions:

1. What are the top ten channels with the maximum number of video uploads?
2. Which grade has the maximum number of video uploads?
3. Which grade has the highest average views?
4. Which grade has the highest number of subscribers?
5. Which grade has the highest video views?

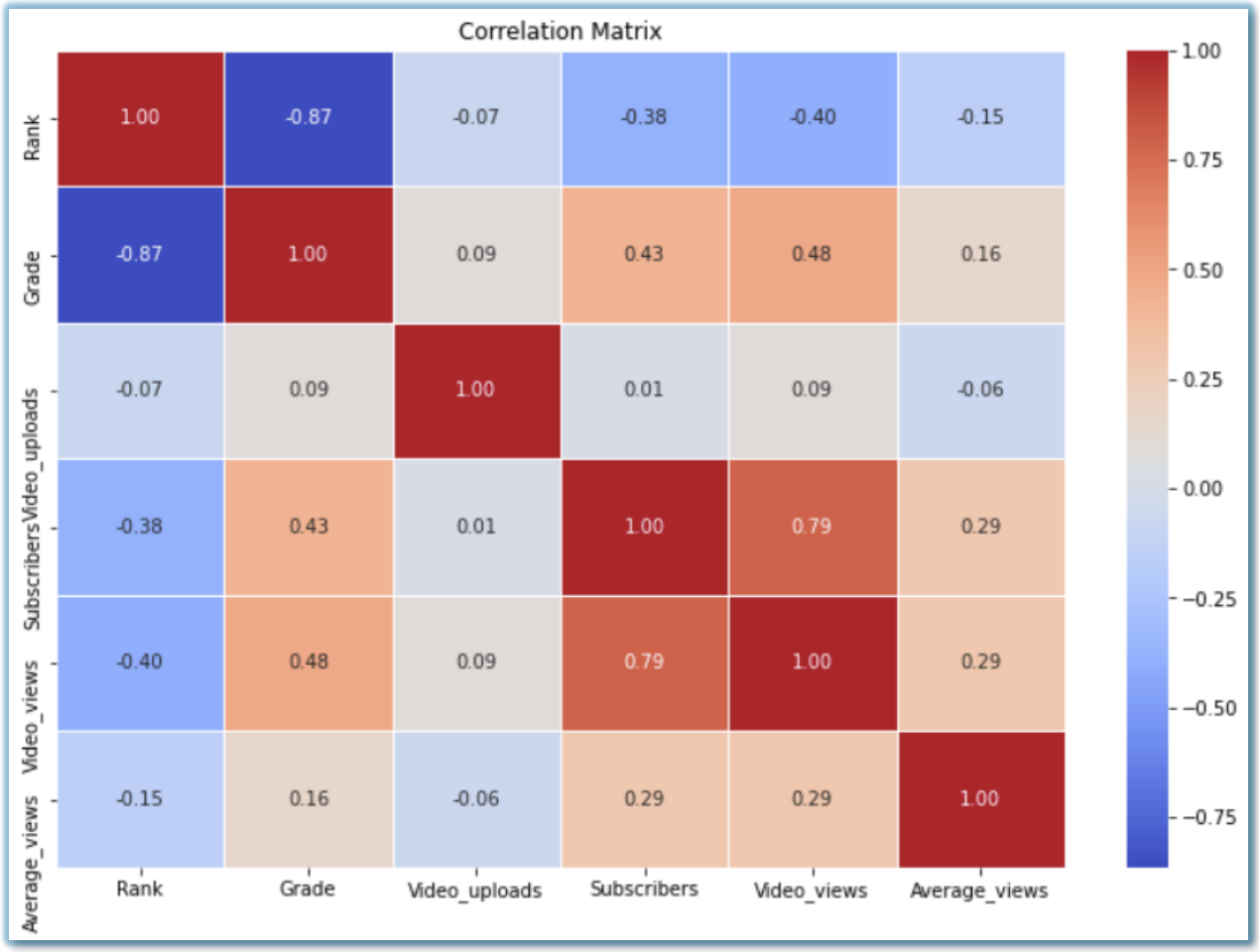
This information is obtained by using “group by” function and computing the averages and is presented in the table below. Also, it can be visualized using the graphs shown in the following pages.

Grade	Rank	Video_uploads	Subscribers	Video_views	Average_views
1	3520.54	3136.16	1535207.95	555183839.09	3280380.88
2	1533.99	4382.58	2798520.38	1102450027.69	5254804.04
3	534.29	5709.86	5107136.29	2497972949.11	10540908.45
4	31.32	16960.30	11726947.47	6168741772.73	11577080.32
5	5.50	37450.70	22281762.50	21199091192.80	5688267.96

In addition to visualizing patterns through graphs, examining the relationships between variables numerically adds depth to the analysis. A correlation matrix offers insights into the strength and direction of these relationships.

Correlation matrix

The correlation matrix visually represents the strength and direction of relationships between variables.



The Pearson correlation coefficient measures the linear association between two variables. It has a value between -1 and 1 where:

- -1 indicates a perfectly negative linear correlation between two variables
- 0 indicates no linear correlation between two variables
- 1 indicates a perfectly positive linear correlation between two variables

A coefficient closer to -1 or 1 indicates a stronger linear relationship, while a value near 0 indicates no linear correlation.

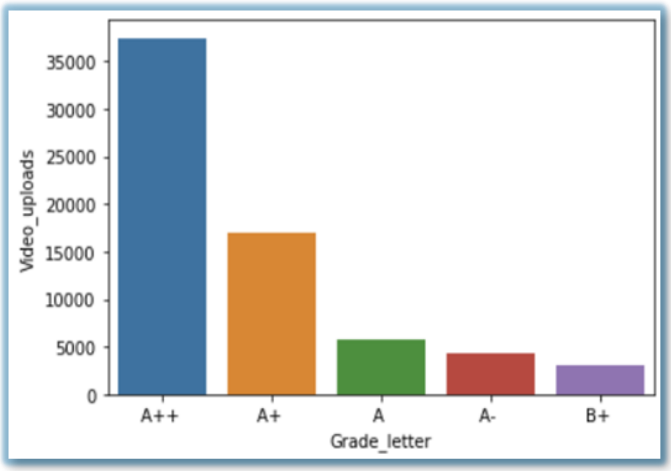
With a clear understanding of relationships between variables, the analysis focuses on answering specific questions derived from the dataset. These insights will highlight trends and characteristics of the top YouTube channels.

Answer questions with data:

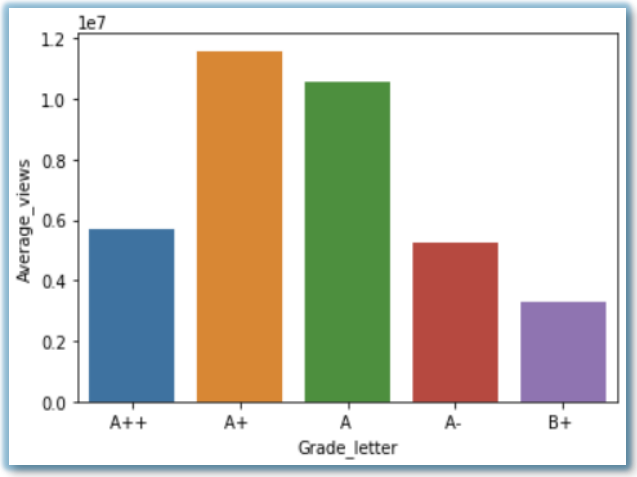
1. Which five channels have the highest number of video uploads?

Rank	Grade	Channel_name	Video_uploads	Subscribers	Video_views	Grade_letter	Average_views
3454	1	AP Archive	422326	746325	548619569	B+	1299.04
1150	2	YTN NEWS	355996	820108	1640347646	A-	4607.77
2224	1	SBS Drama	335521	1418619	1565758044	B+	4666.65
324	3	GMA News	269065	2599175	2786949164	A	10357.90
2957	1	MLB	267649	1434206	1329206392	B+	4966.23

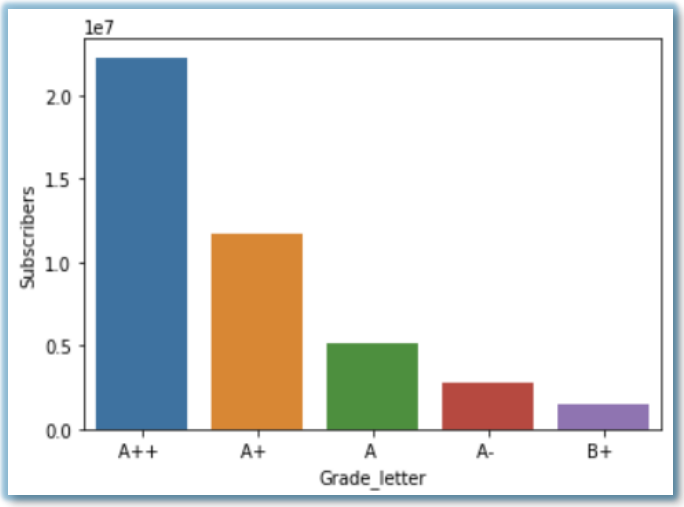
2. Which grade has the maximum number of video uploads?



3. Which grade has the highest average views?



4. Which grade has the highest number of subscribers?



5. Which grade has the highest video views?