Exploratory Data Analysis (EDA)

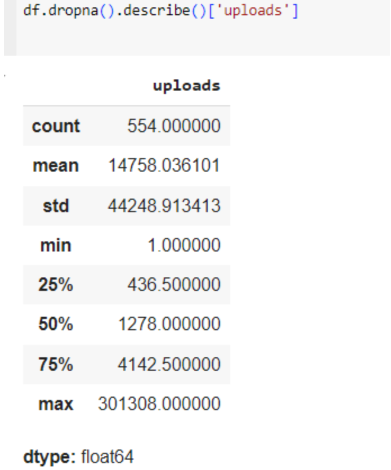
Brief Overview of the Dataset

The dataset contains 995 entries and 28 columns. It primarily includes information about YouTubers, such as:

- rank: Position of the YouTuber in terms of popularity.
- Youtuber: Name of the YouTube channel.
- subscribers: Number of subscribers the channel has.
- video views: Total views the channel has received.
- category: Content category (e.g., Music, Education).
- uploads: Number of videos uploaded by the channel.
- Country: Country of origin of the channel.
- created_year and created_date: Date the channel was created.
- earnings: Estimates of monthly and yearly earnings.
- demographic: Includes population, unemployment rate, and urban population for the country, etc.

The null values were dropped from the dataset before performing the requirements asked.

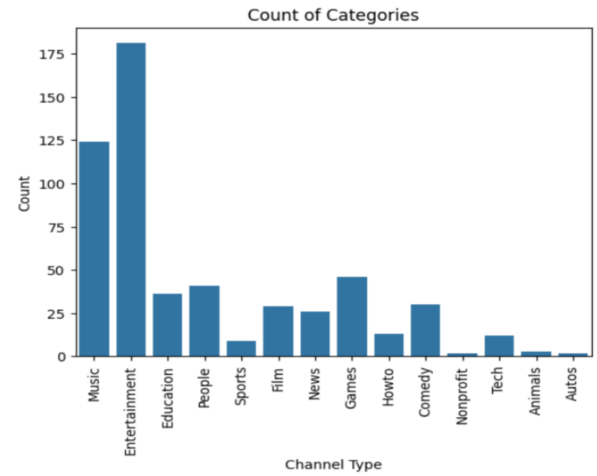
1. Descriptive summary for the “uploads” feature/attribute.

Types	Descriptive Summary of the Data			
	Method	Data Description		
		Technique		
			Purpose	To get a general descriptive summary of the data and features in the dataset
			Python Syntax	*Modules*: import pandas as pd *Syntax*: df.dropna().describe()['uploads']
			Example Plot	 df.dropna().describe()['uploads'] uploads count 554.000000 mean 14758.036101 std 44248.913413 min 1.000000 25% 436.500000 50% 1278.000000 75% 4142.500000 max 301308.000000 dtype: float64
Interpretation	The summary statistical analysis of the dataset's "uploads" column, computed using the describe () function in Python's Pandas, reveals a highly skewed distribution. The mean value is approximately 14,758, representing the average number of uploads in the dataset. The standard deviation is 44,248, indicating a high standard deviation, possibly indicating extreme values or outliers. The minimum number of uploads is 1, providing insight into the lower bound of the dataset. The 25th percentile is 436.5 uploads, indicating 25% of the data lies below 436.5 uploads. The median is 1,278, indicating 50% of data points have uploads equal to or less than this value. The 75th percentile is 4,142.5 uploads, indicating 75% of the data has uploads equal to or fewer than 4,142.5. The maximum number of uploads is 301,308, likely an outlier, significantly impacting the standard deviation and mean.			

2. One Univariate analysis (one plot) for any categorical feature/attribute of your choice.

Types	Univariate analysis			
	Method	Feature distribution		
		Techniques	Bar chart (Bar plot)	
			Purpose	A bar chart is a visual representation of categorical data, simplifying complex information, highlighting differences, and identifying trends, making it essential for data presentation in various fields, including business performance metrics.

			Python syntax	*Modules:* import seaborn as sns
				import matplotlib.pyplot as plt
				sns.countplot(data=df_cleaned, x='channel_type')
				plt.xlabel('Channel Type')
				plt.ylabel('Count')
				plt.title('Count of Categories')
				plt.xticks(rotation=90);

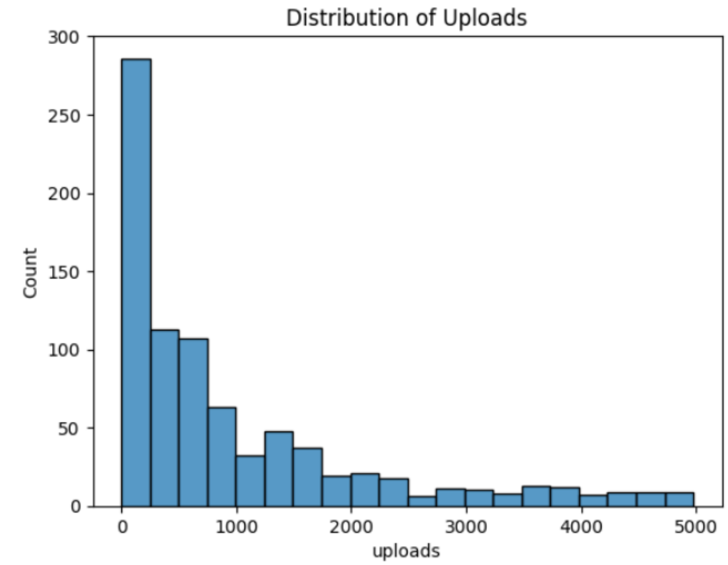
			Example Plot	

Interpretation	This chart shows the number of occurrences for each category, "Channel Type," with each bar representing a distinct channel type. The X-axis labels represent various channels, such as music, entertainment, education, people, sports, film, news, games, comedy, nonprofit, tech, animals, and autos. The Y-axis represents the count of channels in each category, with higher bars indicating more channels. The largest bars in the dataset are categorized as "Music" and "Entertainment", with "Music" having less than 125 counts and "Entertainment" exceeding 175, indicating their dominance in online content, attracting a wide audience. "Education," "Games," and "Howto" are popular categories, with channels focusing on instructional content, tutorials, and resources, while "Games" may involve video gaming, streaming, or reviews. The dataset shows moderate representation of categories like Film, News, People, Sports, and Comedy, but lower counts compared to leading categories like Music and Entertainment. The rare categories like "Nonprofit," "Tech," "Animals," and "Autos" with low counts, suggesting their low frequency may be due to specialized audiences or limited scope.

3. One Univariate analysis (One plot) for any continuous features/attribute of your choice.

Type	Univariate analysis			
	Method	Feature distribution and outliers		
		Technique	Histogram	
			Purpose	It shows the distribution of continuous numerical variables by dividing data into intervals and displaying frequency. It helps identify patterns like normal distribution and outliers, making them crucial in data analysis and interpretation.

			Python syntax	sns.histplot(data=df[df['uploads']<=5000], x='uploads')
				plt.title('Distribution of Uploads')

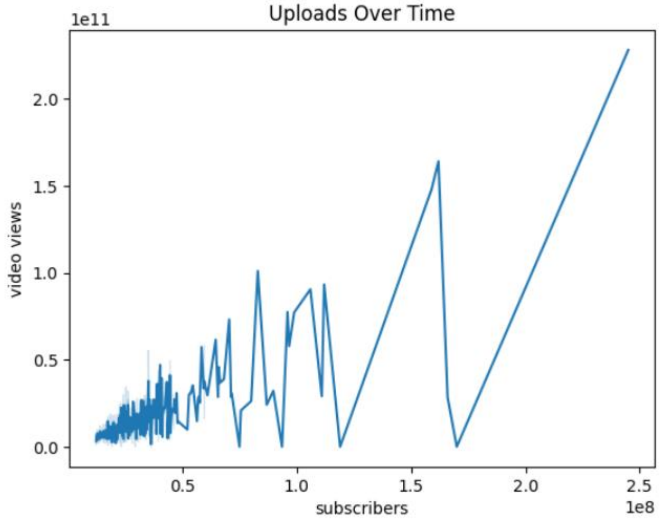
			Example Plot	
Interpretation	The displayed histogram is a visual representation of the distribution of uploads across different ranges. It shows a heavily skewed distribution, with most uploads falling into lower ranges. The largest bar, representing uploads between 0 and 500, indicates that most users have relatively low upload numbers. After the first bin (0-500 uploads), there is a noticeable and steady decline in frequency as the number of uploads increases. The counts gradually decrease as the X-axis progresses, indicating fewer entities with higher upload counts. Towards the right			

	side of the histogram (higher ranges of uploads), very few entities have large upload numbers. This suggests that only a small proportion of the entities in the dataset are responsible for very high upload numbers. The long-tail distribution in this histogram could indicate several things about the dataset, such as the concentration of activity, engagement trends, potential for segmentation, impact on infrastructure, and opportunities for growth. A grounded understanding of this distribution can lead to targeted strategies for engagement or monetization, as well as inform infrastructure scaling. To simply put, this histogram effectively visualizes the distribution of uploads within the dataset, demonstrating a strong right-skew with most entities falling into the lower upload ranges. By understanding the distribution of uploads, platforms can better cater to both low- and high-volume users, ensuring a balanced ecosystem that supports diverse levels of activity.

4. Two Bivariate analysis (One plot) for any two features of your choice

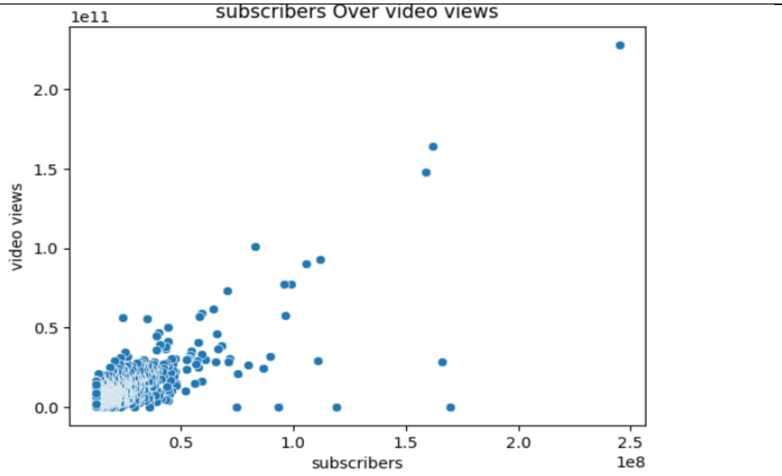
Types	Bivariate analysis			
	Method	Show and explains the relationship between two or more variables		
		Technique	Line plot	
			Purpose	A line plot displays data points connected by straight lines, illustrating trends, changes, or patterns over time. It's useful for tracking performance, progress, and comparing changes in multiple datasets. Line plots are simple and clear, especially for time series data.

			Python syntax	sns.lineplot(data=df, x='subscribers', y='video views')
				plt.title('subscribers Over video views')

			Example Plot	
Interpretation	The graph shows a positive correlation between subscribers and video views over time for a content platform or channel. It shows a gradual increase in views as subscribers grow, with frequent small spikes and dips. As subscriber count increases, larger swings in view counts may indicate viral hits and less popular content. A major spike in views occurs when subscribers reach 1.5 x 10^8 (150 million), suggesting a viral video or series. A sharp decline follows, possibly indicating a return to typical viewing patterns. The graph shows a complex relationship between audience size and content engagement over time.			

Types	Bivariate analysis			
	Method	Relationship between two or more variables		
		Technique	Scatter plot	
			Purpose	Scatter plots visually represent the relationship between two continuous variables by plotting data points on a two-dimensional graph. They help identify patterns, trends, and potential outliers, and assess linear, non-linear, or no correlation between variables, making them essential for exploratory data analysis and regression modeling.

			Python syntax	sns.lineplot(data=df, x='subscribers', y='video views')
				plt.title('subscribers Over video views')

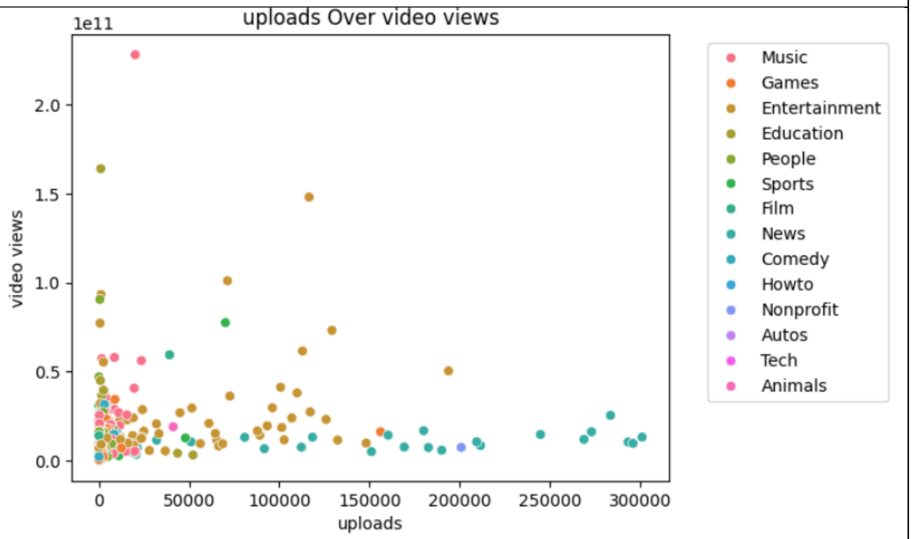
			Example Plot	
Interpretation	This scattered plot of subscriber and video views shows a positive correlation, suggesting that a larger subscriber base leads to more views. However, there is significant variability in the data, suggesting that subscriber count alone does not determine view count. A dense cluster of data points in the lower left corner represents channels			

	with fewer subscribers and lower view counts. Outliers in the upper right quadrant indicate viral content or highly shareable videos. The relationship between subscribers and views is non-linear, with a wider range of possible views counting as subscriber counts increase. An upper limit to the number of views appears, possibly indicating a saturation point or the platform's maximum reach. Gaps in the data suggest that channels tend to remain small or grow to large sizes.

5. One Multivariate analysis (One plot) for any three features of your choice

Types	Multivariate analysis			
	Method	Reveals the relationship between two or more variables		
		Technique	Scatter plot	
			Purpose	To compare the degree of linear relationship between two continuous variables and the third variable.

			Python syntax	sns.scatterplot(data=df, x='uploads', y='video views', hue= 'channel_type')
				plt.legend(loc='upper left', bbox_to_anchor= (1.05, 1))
				plt.title('uploads Over video views')

			Example Plot	
Interpretation	This scattered plot of video uploads and views across 14 different content categories on a video-sharing platform reveals a variety of content, inconsistent correlations, high performers, cluster patterns, and spread in upload numbers. Notable outliers in categories like Music, Entertainment, and Education are seen, while categories like Animals and Tech have lower overall view counts. This graph also shows upload frequency variation, with News having points spread out to 300,000 uploads. The lack of a strong correlation between uploads and views emphasizes the importance of content quality over quantity. Different categories show distinct patterns, suggesting			

	that success strategies may need to be tailored to specific genres. High-view outliers across multiple categories indicate the potential for content to go viral. The graph basically reveals the complex ecosystem of the platform, were factors beyond upload frequency influence viewership. However, niche opportunities may be found in less populated areas.