Analysis of YouTube Data

Finlatics Project

Praneeth Devanabanda

PROBLEM STATEMENT AND OBJECTIVES

- As part of a comprehensive analysis project of Finlatics, I delved into YouTube data. This dataset encompassed the over 1000 top YouTubers ranked by subscriber count, including details such as channel name, uploads, views, earnings, country of origin, and more.
- My role as a data analyst involved several key objectives: preprocessing the dataset, conducting exploratory data analysis (EDA), and uncovering nuanced patterns within the data. Through EDA, I aimed to extract valuable insights, ranging from basic to advanced level questions.
- YouTube stands as the world's foremost video streaming platform, originating in the mid-2005 and maintaining its prominence since. For aspiring creators, delving into datasets like this one offers a wealth of insights. By conducting extensive analyses, creators can discern successful strategies, identify audience preferences, and gain valuable insights into video content dynamics.

DATA DESCRIPTION

→ The dataset comprised fundamental features such as YouTuber name, country of origin, channel creation date, number of views, number of subscribers, and rank based on subscriber count.

→ Additionally, slightly more intricate features included the channel's rank within its country, channel type, channel category, as well as the lowest and highest monthly and yearly incomes.

→ For country-specific data, the dataset contained the following attributes:

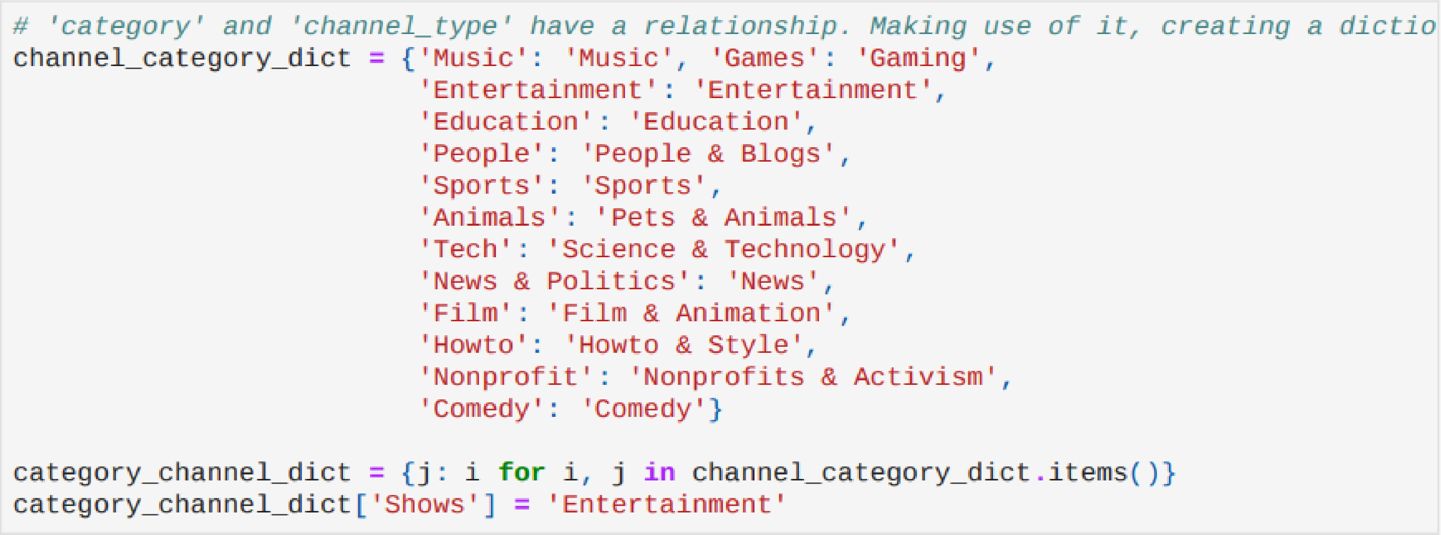
- Gross tertiary education enrollment (%): The percentage of the population enrolled in tertiary education in the country.
- Population: The total population of the country.
- Unemployment rate: The unemployment rate in the country.
- Urban population: The percentage of the population residing in urban areas.
- Latitude: The latitude coordinate of the country's location.
- Longitude: The longitude coordinate of the country's location.

PREPROCESSING OF DATASET

→ I utilized Jupyter Notebook, a specialized platform for Python users, to streamline my workflow by incorporating multiple code snippets within a single notebook.

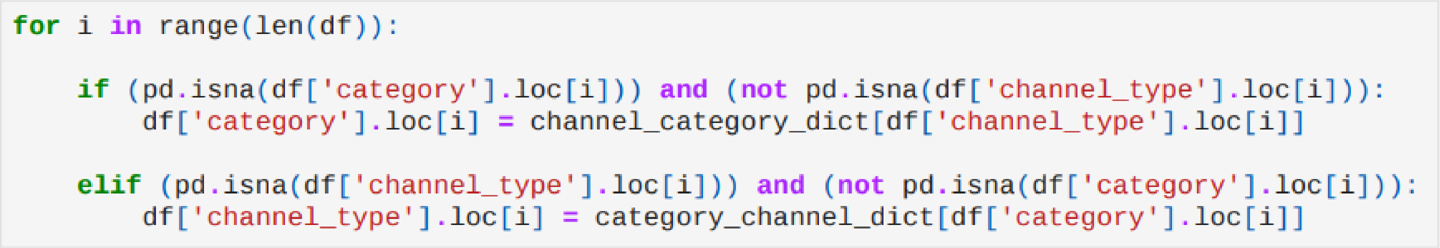
→ The data cleaning phase consumed a significant portion of the project timeline. My preprocessing steps included:

- Initial inspection of the CSV file to identify anomalies such as incorrect YouTuber names, entirely null records, and duplicate rows. These anomalies were promptly removed to ensure data integrity.
- Approximately 150 records were missing country information along with associated country-specific attributes such as population, unemployment rate, longitude, and latitude. To address this, I created multiple dictionaries with country names as keys and corresponding features as values. Leveraging a custom imputer function, I provided the index and country name as inputs, enabling the imputation of all missing country-related features.
- The 'Channel Type' and 'Category' columns held pivotal significance for subsequent analyses. Recognizing their interrelation, I established a mapping dictionary between them. This facilitated the imputation of missing values in one column by referencing the mapped value from the other column. Code snippet of this process is as follows:



```
# 'category' and 'channel_type' have a relationship. Making use of it, creating a dictio
channel_category_dict = {'Music': 'Music', 'Games': 'Gaming',
                         'Entertainment': 'Entertainment',
                         'Education': 'Education',
                         'People': 'People & Blogs',
                         'Sports': 'Sports',
                         'Animals': 'Pets & Animals',
                         'Tech': 'Science & Technology',
                         'News & Politics': 'News',
                         'Film': 'Film & Animation',
                         'Howto': 'Howto & Style',
                         'Nonprofit': 'Nonprofits & Activism',
                         'Comedy': 'Comedy'}

category_channel_dict = {j: i for i, j in channel_category_dict.items()}
category_channel_dict['Shows'] = 'Entertainment'
```



```
for i in range(len(df)):

    if (pd.isna(df['category'].loc[i])) and (not pd.isna(df['channel_type'].loc[i])):
        df['category'].loc[i] = channel_category_dict[df['channel_type'].loc[i]]

    elif (pd.isna(df['channel_type'].loc[i])) and (not pd.isna(df['category'].loc[i])):
        df['channel_type'].loc[i] = category_channel_dict[df['category'].loc[i]]
```

- The 'Creation Date' column presented encoding issues and contained missing values. To address this, I manually assigned values to some records and subsequently converted the entire column to the datetime format.
- Following the completion of necessary imputations and final datatype assignments, the dataset comprised a total of 973 records. I then transformed it into a 'Cleaned Dataset.csv' file, paving the way for further analysis and answering of pertinent questions.

PROBLEMS AND SOLUTIONS

- I had to address 20 questions using Python's pandas, matplotlib, and seaborn libraries as part of the case study. While explaining detailed code here isn't practical, I'll walk through each question, detailing my approach and findings. To fully understand the analysis, I suggest referring to the PDF of my Jupyter Notebook along with this presentation.

1. What are the top 10 YouTube channels based on the number of subscribers?

→ Since the Subscriber Rank column is already available and sorted in the data frame, I simply extracted the first 10 records and created a horizontal bar chart. The top YouTuber by subscribers is the renowned Indian music and cinema production company 'T-Series', followed by the emerging American sensation MrBeast.

2. Which category has the highest average number of subscribers?

→ I utilized pandas' '.groupby()' function to group the number of subscribers for each category, using the 'mean' aggregation function. After sorting the values, I plotted a horizontal bar chart. It's evident that the 'Shows' category leads with an average of 46.6 million subscribers, followed by the 'Trailers' category. The 'Travel & Events' category has the fewest subscribers on average.

3. How many videos, on average, are uploaded by YouTube channels in each category?

→ Similarly, I employed pandas' '.groupby()' function, this time grouping 'uploads' along with the 'category' feature, using the 'mean' aggregation function. A horizontal bar chart illustrates that 'News & Politics' leads with an average of around 112,000 videos across all YouTube channels. This trend is understandable, as YouTube has become a platform for news networks to promptly upload relevant news following any incident. Conversely, the 'Travel & Events' category has the fewest uploads on average.

4. What are the top 5 countries with the highest number of YouTube channels?

→ A straightforward '.value_counts()' and top 5 analysis would suffice, but I also went for an alternative approach using the '.groupby()' function on the 'Youtubers' column along with 'Country', using the 'size' aggregation function. Both methods yield identical results. A simple horizontal bar chart highlights the dominance of the United States, followed closely by India.

5. What is the distribution of channel types across different categories?

→ Both 'Channel Types' and 'Categories' are categorical features. To explore their relationship, I generated a cross-tabulation and created a heatmap chart. This visualization makes it easy to discern which channel types have more or fewer YouTube channels in specific categories. Notably, the 'Entertainment' channel type exhibits a high concentration of YouTube channels in the 'Entertainment' category. This observation aligns with my findings during the preprocessing stage, indicating that these two features are closely related.

Additionally, I plotted a stacked bar chart with categories on the x-axis and channel types as stacks. However, this chart was less effective for visualization purposes compared to the heatmap.

6. Is there a correlation between the number of subscribers and total video views for YouTube channels?

→ Utilizing the '.corr()' function, I computed the correlation between the number of subscribers and total views, yielding a value close to 0.82. This high correlation suggests a strong positive relationship, which I confirmed by plotting a scatter plot. Indeed, the scatter plot revealed a clear upward positive trend between these two features.

7. How do the monthly earnings vary throughout different categories?

→ I conducted separate 'groupby()' operations on the 'low_monthly_earnings' and 'high_monthly_earnings' columns, grouping them by category with the mean aggregation function. Subsequently, I created two distinct horizontal bar charts. Interestingly, the 'Show' category emerged with the highest earnings in both features. Furthermore, the order of categories remains consistent across both charts, indicating a stable pattern. Conversely, 'Travel & Events' consistently exhibited the lowest earnings in both cases.

8. What is the overall trend in subscribers gained in the last 30 days across all channels?

→ I grouped the column specifying the number of subscribers gained in the past 30 days with the channel type column using the sum aggregation function. The resulting horizontal bar chart illustrates that the 'Entertainment' channel type leads with a subscriber gain of around 91 million, followed by the 'People' channel type with 31 million subscriber gains. Conversely, the 'Autos' and 'Non-Profit' channels exhibited the least subscriber gains.

9. Are there any outliers in terms of yearly earnings from YouTube channels?

→ In relation to yearly earnings, the dataset includes two features named 'lowest_yearly_earnings' and 'highest_yearly_earnings'. To analyze these features, I opted for the simplest approach of plotting boxplots for each column. Upon examination, I observed numerous data points exceeding the upper whisker of the boxplot, indicating values greater than (Q3 + 1.5xIQR). However, upon closer inspection, I found that these outliers were not merely random anomalies or errors that required removal from the dataset.

Interestingly, there was one record belonging to a YouTuber named 'KIMPRO', a South Korean channel, which boasted the highest yearly earnings. To validate this finding, I verified the channel and confirmed its authenticity.

10. What is the distribution of channel creation dates? Is there any trend over time?

→ A straightforward KDE (Kernel Density Estimation) distribution plot is an effective method for this analysis. However, I encountered an anomaly where the official YouTube channel had a creation date listed as 1970, which is actually accurate according to many sources indicating YouTube's creation date.

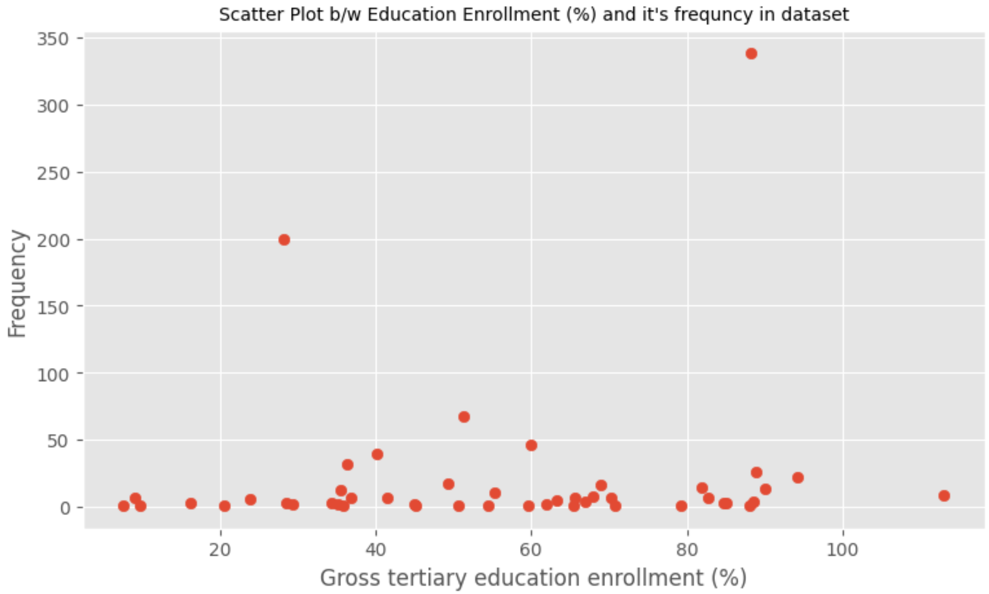
Temporarily excluding this anomaly, I proceeded to plot the KDE plot. The visualization revealed two distinct peaks: one around 2005-2006, coinciding with YouTube's inception, and another peak in 2014-2015. This latter peak corresponds to the time when YouTube started gaining widespread popularity due to the introduction of smartphones in the market. This facilitated increased access to YouTube and prompted individuals to view it as a legitimate source of income.

11. Is there a relationship between gross tertiary education enrollment and the number of YouTube channels in a country?

→ The 'Gross tertiary education enrollment' column is country-specific, with the same value across countries in the dataset. To analyze this feature, I used the '.value_counts()' function to determine the frequency of each 'GTEE' value in the dataset. I then plotted a scatter plot between 'GTEE' and the frequency/number of YouTube channels.

The scatter plot revealed no discernible visual trend between these two features. Despite this, it's worth noting that the country with the most YouTube channels was the United States, with a tertiary education enrollment rate of 88.2%. This relationship is evident in the scatter plot. Conversely, India ranked second with a tertiary education enrollment rate of 28.1%.

Interestingly, irrespective of whether the tertiary education enrollment percentage is high or low, the number of YouTube channels does not seem to depend on it. For instance, Australia boasts the highest gross tertiary education enrollment of 113.1%, yet it has a relatively low number of YouTube channels. This observation suggests that the number of YouTube channels is not directly correlated with the level of tertiary education enrollment.



12. How does the unemployment rate vary among the top 10 countries with the highest number of YouTube channels?

→ To identify the top 10 countries with the highest number of YouTube channels, I extracted the first 10 indices of the '.value_counts()' function. Since the unemployment rate is specific to each country and remains fixed across the dataset, I created a dictionary with countries as keys and their corresponding unemployment rates as values. Subsequently, I created a simple bar chart to visualize this data.

The bar chart revealed that the United States has the highest unemployment rate, followed by Spain and Brazil. Conversely, Thailand boasts the lowest unemployment rate among the top 10 countries.

13. What is the average urban population percentage in countries with YouTube channels?

→ The dataset included both a population column and an urban population column. To calculate the average urban population percentage in each country, I divided the urban population by the total population and multiplied the result by 100. I then plotted a horizontal bar chart to visualize this data.

The bar chart revealed that Kuwait and Singapore showed 100% urban population, indicating no rural population or YouTubers according to the data. Conversely, Samoa exhibited the lowest urban population percentage. Notably, India had a 34.47% urban population percentage.

14. Are there any patterns in the distribution of YouTube channels based on latitude and longitude coordinates?

→ Initially, I plotted KDE distributions of longitudes and latitudes, but these visualizations did not reveal significant patterns. Then I created a Choropleth map with color grading based on the number of YouTube channels in each country. The Choropleth map can be viewed in the PDF.

In the Choropleth map, countries with high numbers of YouTube channels, such as India and the United States, appear greenish, while countries like Mexico and Brazil, which follow closely behind, appear yellowish. Countries with lower numbers of YouTube channels appear more reddish.

Overall, there doesn't appear to be a clear pattern based on longitudes and latitudes. The number of YouTube channels does not seem to depend on geographic coordinate system.

15. What is the correlation between the number of subscribers and the population of a country?

→ The population of each country is a specific feature within the 'Country' column, meaning that the population value remains constant across the entire dataset for each country. To analyze the relationship between population and the number of subscribers, I used the '.groupby()' function to aggregate the number of subscribers across each population, using the sum aggregation function.

Upon calculating the correlation between these two newly created features, I found a correlation coefficient of 0.46. When plotted on a scatterplot, a slight positive trend between population and total subscribers can be observed. From this, we can conclude that the total subscribers and the population of a country are slightly correlated with each other.

16. How do the top 10 countries with the highest number of YouTube channels compare in terms of their total population?

→ For this analysis, I began by extracting the top 10 countries with the most YouTube channels using the '.value_counts()' function on the 'Country' column and selecting the top 10 indices. Since population is a country-specific feature, I created a dictionary with countries as keys and their corresponding populations as values. Finally, I created a simple bar chart to visualize this data.

The bar chart highlights that India significantly leads in terms of population, followed by the United States. Conversely, Spain and Canada have lower populations among the top 10 countries with the most YouTube channels.

17. Is there a correlation between the number of subscribers gained in the last 30 days and the unemployment rate in a country?

→ Similar to the 15th solution, unemployment rate is a country-specific column. Using the groupby function and calculating the correlation, the correlation coefficient comes out as 0.19. Upon plotting a scatterplot, no particular trend between the total subscribers gained in the past 30 days and unemployment rate is evident. From this, we can conclude that unemployment rate and subscribers gained in the past 30 days are not correlated.

18. How does the distribution of video views for the last 30 days vary across different channel types?

→ Utilizing the column specifying video views for the past 30 days for each YouTube channel, I employed the groupby function across channel_type with the sum aggregation function. Subsequently, I plotted a horizontal bar chart to visualize the results.

The bar chart revealed that the 'Entertainment' channel type had the highest total views in the past 30 days, with 64.26 billion views, followed by the 'Music' channel type with 38 billion views. Conversely, the 'Non-Profit' and 'Autos' channel types had the least views in the past 30 days.

19. Are there any seasonal trends in the number of videos uploaded by YouTube channels?

→ This question cannot be directly answered without data on date-wise uploads of each YouTube channel. In the absence of such data, I examined the 'Created Dates' and quarters to gain insights.

Firstly, I created a new column named 'Year-Quarter' based on the 'Created Dates' and plotted a horizontal time series graph to visualize the frequency of video uploads in YouTube channels for each quarter. Interestingly, YouTube channels created in the 3rd quarter of 2006 had the highest number of video uploads till now. Additionally, channels created more recently, within the past 6-7 years, tend to have fewer video uploads, which is understandable.

Secondly, I focused on the total uploads in each quarter of YouTube creation dates. Channels created in the 3rd quarter seem to have a higher number of uploads in total, as indicated by the second plot.

20. What is the average number of subscribers gained per month since the creation of YouTube channels till now?

→ To calculate the average number of subscribers gained from the creation date till the present, I subtracted each creation date from April 8, 2024, using the '.days()' function to obtain the total number of days elapsed between the two dates. Dividing this number by 30 and rounding off gives the number of months elapsed between the two dates. Then, I divided the total number of subscribers by the number of months elapsed to determine the average number of subs gained per month.

Finally, I plotted a horizontal bar chart to identify the top 10 YouTube channels with the highest average subs gain from the creation date till now. According to the chart, the '5-Minute Crafts' channel leads with 1.78 million subs gained each month, followed by 'Vlad and Niki' and 'MrBeast'.

TAKEAWAYS AND CONCLUSIONS

- Through this project, I've enhanced my proficiency with both basic and advanced functionalities of NumPy and Pandas. From fundamental data manipulation techniques to more complex group-by operations and plotting libraries, I've gained a deeper understanding of data analysis tools.
- The preprocessing phase proved to be particularly time-consuming, requiring the application of numerous new functions and techniques. Despite the challenges, navigating through various preprocessing steps provided invaluable learning experiences.
- The skills acquired during this project will undoubtedly be instrumental in my professional endeavors. As data analysis continues to play a crucial role across various industries, the knowledge gained from this project will empower me to tackle real-world challenges with confidence.

THE END

THANK YOU