CS732: Data Visualisation Assignment 1 Report

Sarthak Harne
IMT2020032
Sarthak.Harne@iiitb.ac.in

Sougandh Krishna K S
IMT2020120
Sougandh.Krishna@iiitb.ac.in

Monjoy Narayan Choudhury
IMT2020502
Monjoy.Choudhury@iiitb.ac.in

I. Dataset

In this assignment, we work with the Global YouTube Statistics 2023 dataset. This comprises of the top 995 creators on YouTube, based on their subscriber counts, with comprehensive details on subscriber counts, video views, upload counts, country of origin, earnings, and more. The data fields present in the dataset are:

1) rank: Position of the YouTube channel based on the number of subscribers
2) Youtuber: Name of the YouTube channel
3) subscribers: Number of subscribers to the channel
4) video views: Total views across all videos on the channel
5) category: Category or niche of the channel
6) Title: Title of the YouTube channel
7) uploads: Total number of videos uploaded on the channel
8) Country: Country where the YouTube channel originates
9) Abbreviation: Abbreviation of the country
10) channel_type: Type of the YouTube channel (e.g. individual, brand)
11) video_views_rank: Ranking of the channel based on total video views
12) country_rank: Ranking of the channel based on the number of subscribers within its country
13) channel_type_rank: Ranking of the channel based on its type (individual or brand)
14) video_views_for_the_last_30_days: Total video views in the last 30 days
15) lowest_monthly_earnings: Lowest estimated monthly earnings from the channel
16) highest_monthly_earnings: Highest estimated monthly earnings from the channel
17) lowest_yearly_earnings: Lowest estimated yearly earnings from the channel
18) highest_yearly_earnings: Highest estimated yearly earnings from the channel
19) subscribers_for_last_30_days: Number of new subscribers gained in the last 30 days
20) created_year: Year when the YouTube channel was created
21) created_month: Month when the YouTube channel was created
22) created_date: Exact date of the YouTube channel's creation
23) Gross tertiary education enrollment (%): Percentage of the population enrolled in tertiary education in the country
24) Population: Total population of the country
25) Unemployment rate: Unemployment rate in the country
26) Urban_population: Percentage of the population living in urban areas
27) Latitude: Latitude coordinate of the country's location
28) Longitude: Longitude coordinate of the country's location

Apart from this, we have made columns of our own based on the available data. These columns are:

1) Grouped Categories: Grouped into Big, Medium, Small, and Other Categories based on the number of channels
2) Days Since Created = (2022-12-31) - Date of Creation (inferred)
3) Avg Daily Subscribers = Subscribers / Days Since Created
4) Avg Monthly Subscribers = Subscribers / Months Since Created
5) Avg Daily Views = Views / Days Since Created
6) Avg Monthly Views = Views / Months Since Created

Task

Through visual exploratory analysis, we target to gain the following insights and expect the one to reproduce the following tasks:

1) T1: View, Subscribers and Uploads Based
2) T2: Category and Channel Revenue Based
3) T3: Unemployment Rate and Education Based

Assumption/Data Filtration

Since the data points were very large in number, a lot of visualization used won't make much clear sense. Due to this reason, we applied some sort of data filtration which mostly included the following constraints.

1) Allowing contribution of data entries that have a certain (more or less) number of Youtubers/ YouTube channels.
2) Using only Top/Bottom n based on a field.
3) The 'Date Since Created' attribute which is made considers the final date to be December 2022 as the dataset comprised of records till December 2022 only.

Data Stories

A. View, Subscribers and Uploads Based

Hypothesis 1: Subscribers are more correlated to some factors like Views and categories and not to factors like

Uploads.

The idea behind this hypothesis is that if more people watch a particular channel, more people will subscribe to that channel. Similarly, categories like Music and Entertainment are more popular with people, so the category of the channel should also play a role in determining its number of subscribers.

On the other hand, someone could upload many low-quality videos, without many people watching them. Moreover, some YouTube channels might upload occasional, yet high-quality and well-received videos, while others might upload videos more frequently, which are all moderately received. An example for the same is the case of SET India and Mr. Beast. SET India has 159,000,000 subscribers and Monthly Uploads of 597, while Mr Beast has 166,000,000 subscribers and Monthly Uploads of 5.

This hypothesis is verified visually, by using scatter plots between the Number of Subscribers and the Views 1 and Uploads 2 fields. This is also easily verified by the correlation between these two (Corr(Uploads, Subscribers) = 0.077, Corr(Views, Subscribers) = 0.752)

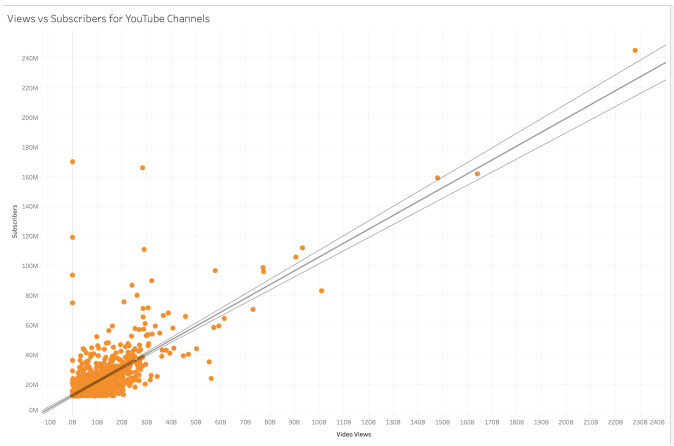


Fig. 1. Scatter Plot between Views and Subscribers

For testing out the correlation between Categories and Country, a Density Plot 3 and Cartograph 4 are plotted, respectively.

With these plots, we can conclude that the Number of Subscribers is correlated to Views, Categories, and Country, while not very correlated to Uploads.

Some considerations:

- The line in the scatter plot for Views and Subscribers 1 denotes the trend as the line fits to the data, along with its confidence interval.
- The colours for the Views 1 and Uploads 2 plot are chosen to look aesthetically pleasing.
- The colour map in the density plot 3 is such that red signifies a higher density of points. (The legend for the same is not available in Tableau).

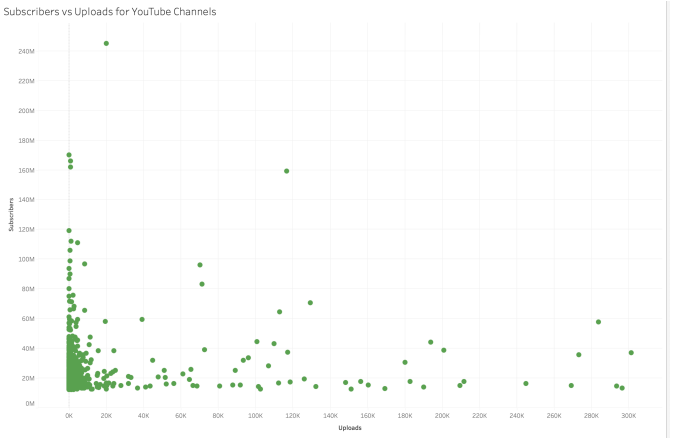


Fig. 2. Scatter Plot between Uploads and Subscribers

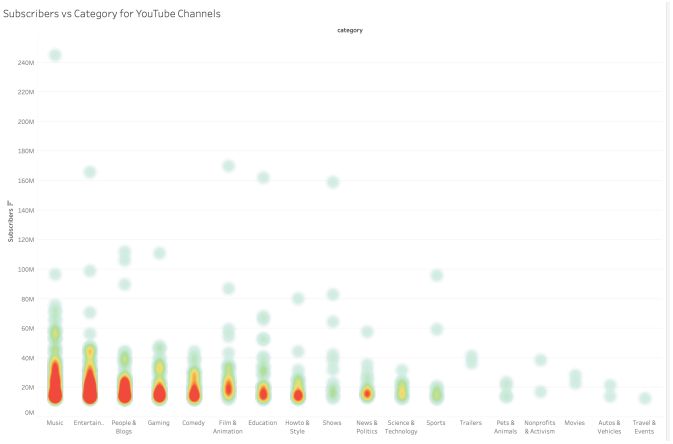


Fig. 3. Density Plot between Category and Subscribers

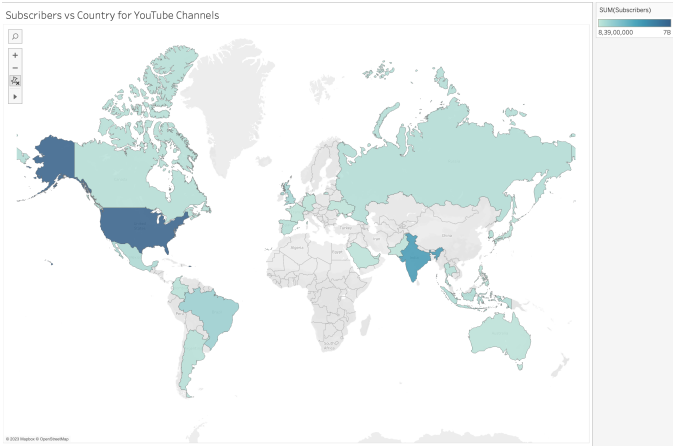


Fig. 4. Cartograph between Country and Subscribers

- Sum of Subscribers is chosen as an attribute to plot in the Cartograph 4 as choosing Average or Median gives an unfair edge to countries with a very small number of channels which are high performing, for example, Latvia and Jordan.

Hypothesis 2: The Average Number of Subscribers and Views for the older channels would be more than that of the newer channels. At the same time, because the newer channels have to have gathered a lot of subscribers quickly to reach the top 1000, their number of Daily Subscribers and Views would be greater.

The idea behind this is that the older channels have had more time on YouTube and thus would tend to have a higher number of Subscribers and Views. This is because they have spent more time on the platform and have had more time to interact with the community to gather more subscribers.

But, we need to acknowledge the fact that the dataset has only the top 1000 channels by subscribers. So, if a channel has to reach this point, in less amount of time, their influence would be concentrated in the recent time, indicated by more number of Daily Subscribers and Views.

The first part is tested by plotting line plots for the Average Number of Subscribers by the Created Year 5 and the Average Number of Views by the Created Year (included in the accompanying folder).

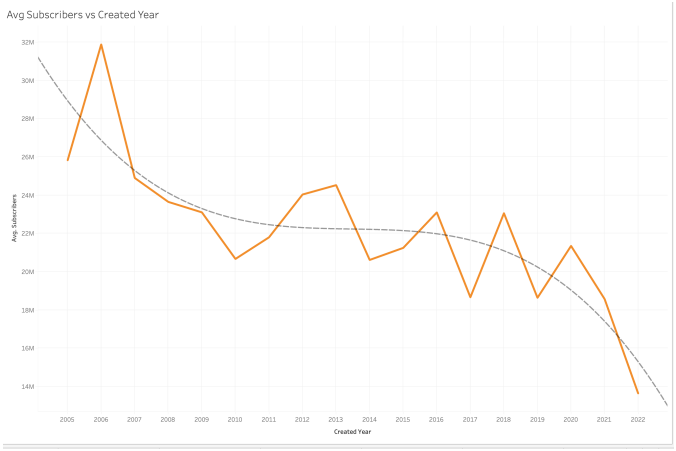


Fig. 5. Line Plot between Subscribers and Created Year

The first part of the hypothesis is supported by the plot and even the trend curve that fits the plot.

To test the second part, a scatter plot between the Daily Subscribers and the Days Since Created columns is made 6. The same is done for Views which is included in the accompanying folder.

The scatter plots heavily support the second part of the hypothesis. So much, so that the trend line is mapped as a power function of the Days Since Created.

These plots (5, 6) heavily confirm the second hypothesis. Some considerations:

- The plots for views are very similar to the plots for subscribers and are therefore skipped for brevity. They are included in the folder.
- The trends for these plots are chosen via trial and error from options like Linear, Polynomial, Logarithmic, Exponential and Power.

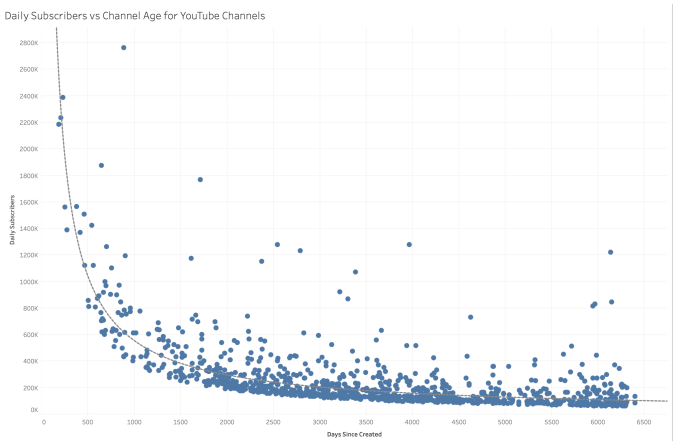


Fig. 6. Scatter Plot between Subscribers and Days Since Created

Trends in the Number of Created Channels: For observing the trends of the number of created channels and Days Since Created, Area Plot for the Country Wise Number of Created Channels and the Days Since Created (Bins) is plotted 7. The same is done for Category Wise plots 8.

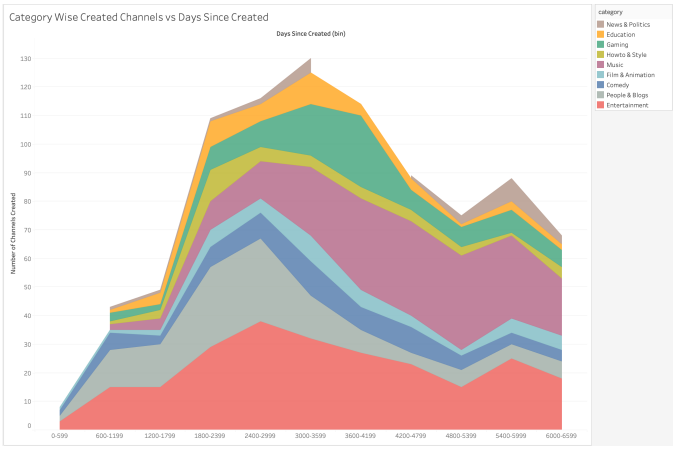


Fig. 7. Area Plot for Number of Channels created vs the Created Year, Category Wise

In this plot 7, we have the following observations:

- All the categories peak around 3000 days since created (around 2016) after the decline as the channel age decreases. This can be attributed to the fact that it is difficult for the newer channels to reach the top 1000.
- The Entertainment Category remains the same more or less, with some noise here and there, after eventually decreasing with the channel age.
- The Music Category initially is attributed to the highest number of created channels, but it eventually decreases.

This can be due to the fact that many big channels related to Music are production and distribution companies, like T-Series, Vevo, etc. These companies publish songs for popular artists, especially in India. The peak is at about 5200 days of age, which corresponds to 2008. Spotify was launched in that year. The decrease in channels can also be attributed to the launch Spotify. A further dip is observed in 2016, when both Apple Music and YouTube Music were launched.

- The People & Blogs Category suddenly flourishes around 2400 days after it eventually decreases according to the global pattern.
- The number of channels created in Gaming peaks at around 2014, when many gaming channels, like PewDiePie were becoming popular, which is also known to have motivated other Gaming channels to start.

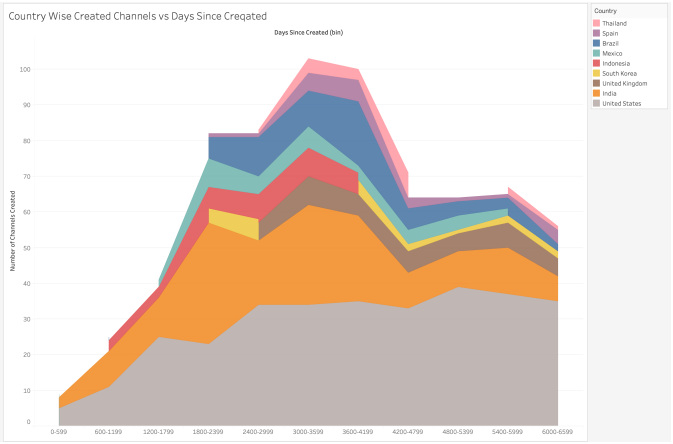


Fig. 8. Area Plot for Number of Channels created vs the Created Year, Country Wise

There are the following observations for 8

- As a global trend, we can see that the number of channels decreased around 2016.
- In this plot, we can see that the channels created in around 2014 in India started to gain popularity. This can be attributed to the boom in the usage of the Internet in India after the introduction of nationwide, affordable 4G internet in 2016. So, the channels created in 2014 were created at the sweet spot, where they gained enough popularity, because of which they were boosted due to the increased usage of the Internet shortly afterwards.
- Channels from Indonesia start appearing around 2014 and show consistent numbers.
- There is a consistent number of channels from Brazil before it vanishes around 2018. Spain and Thailand follow a similar trend.
- Channels from the United Kingdom stopped appearing after 2016.
- There was a consistent number of channels from Mexico between 2008 and 2018.

Some considerations while making these plots:

- A Tableau Colour Palette was chosen to clearly distinguish different fields
- Top Categories and Countries were chosen according to the assumptions stated in the beginning.

1) Recent Trends in Subscribers: In this section, we'll take a look at the trends in the number of Subscribers in recent times. We'll compare the Average Number of Subscribers with the Subscribers in the Last 30 Days, to see what is popular currently. The Plots are made for the Category, Country and the YouTuber fields. Using this we'll observe their current performance to their average performance.

A Line Plot is made for the Subscribers in the Last 30 Days. For different fields in the column we're considering (Category, Country or YouTuber), we make an overlapping Dot Plot for the Average Monthly Subscribers. The two axes for "Subscribers in the Last 30 Days" and "Avg Monthly Subscribers" are shared and synced.

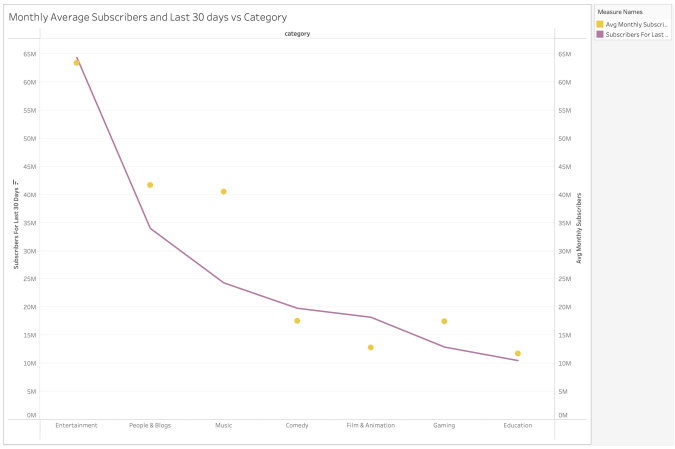


Fig. 9. Plot for Category Wise trends in Subscribers for the last 30 days

We can draw the following conclusions from the above plot 9:

- The categories 'People & Blogs', 'Music', and 'Gaming' are gaining more subscribers than average.
- The categories 'Entertainment' and 'Education' are gaining about the average number of subscribers.
- The categories of 'Comedy' and 'Film & Animation' are gaining fewer subscribers than average.

We can draw the following conclusions from the above plot 10:

- Indian and Indonesian channels are gaining more subscribers than average.
- Channels from Spain are gaining about the average number of subscribers.
- Channels from the US, Brazil, United Kingdom and Mexico are gaining fewer subscribers than usual.

We can draw the following conclusions from the above plot 11:

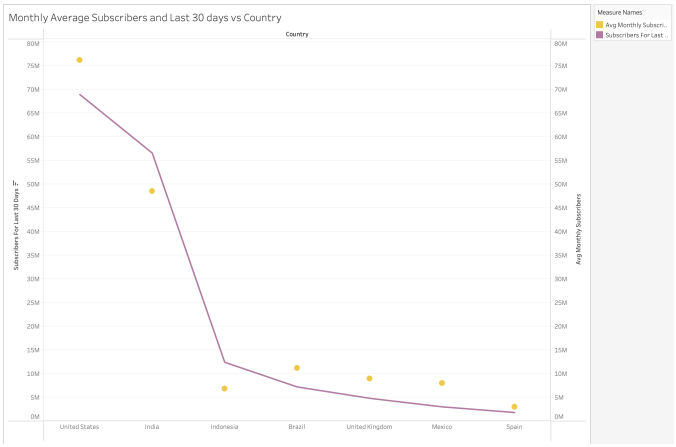


Fig. 10. Plot for Country Wise trends in Subscribers for the last 30 days

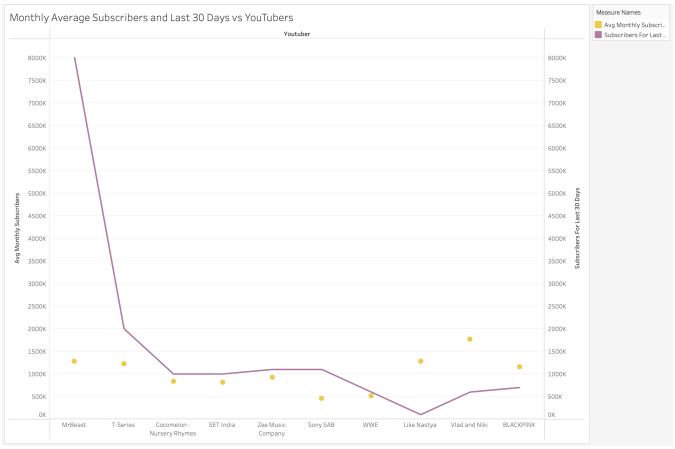


Fig. 11. Plot for YouTuber Wise trends in Subscribers for the last 30 days

- Mr Beast is gaining about 8 times more subscribers than normal. This can be attributed to his high-budget videos and catchy thumbnails and titles.
- The channels 'MrBeast', 'T-Series', 'Cocomelon', 'SET India', 'Zee Music Company', and 'Sony SAB' are gaining more subscribers than average.
- WWE is gaining about the average number of subscribers.
- The channels 'Like Nastya', 'Vlad and Nikki' and 'BLACKPINK' are gaining fewer subscribers than usual.

2) Overview of Categories and Countries: In this section, we'll visualise Categories, group them according to the number of channels in them and look at the group trends in various countries.

For this, we'll plot a Tree Map showcasing the division of groups 12, a Pie Chart on the division of views in these groups 13 and finally a Cartograph with the localised view-based Pie Charts for categories 14.

With these plots, we are able to observe the popularity of different groups of categories in different countries.

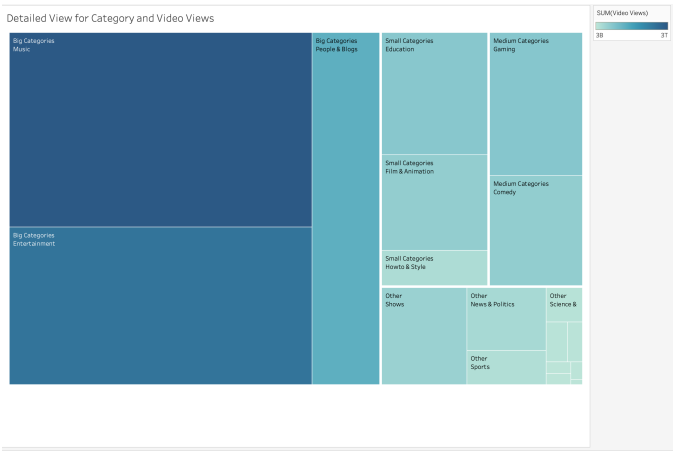


Fig. 12. Division of categories into groups by number of channels

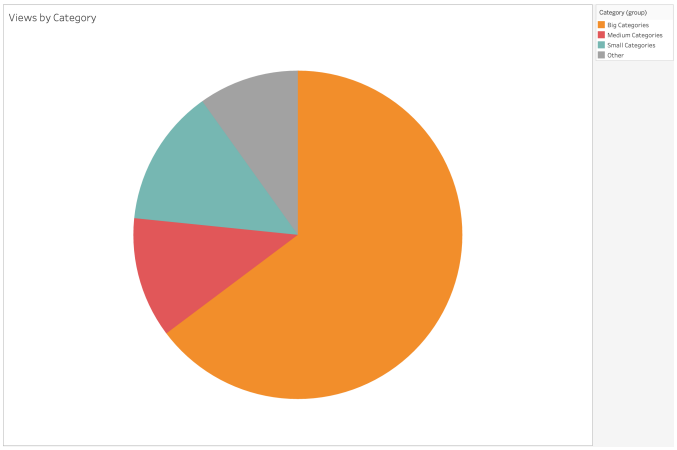


Fig. 13. Views by Category Groups

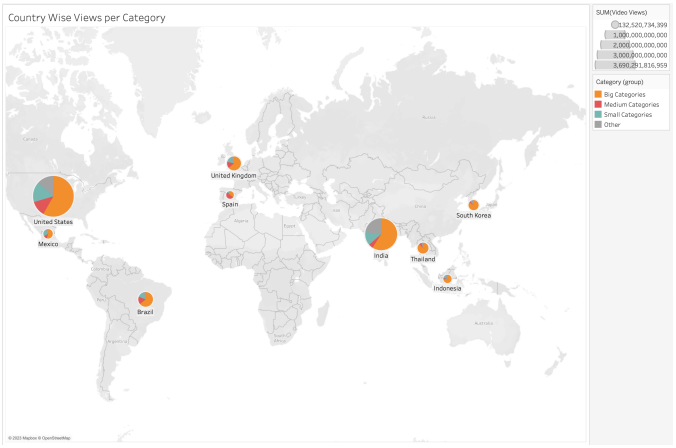


Fig. 14. Popularity of Category groups in different countries

B. Category and Channel Revenue Based

There are 4 columns in the data set that give us an idea of revenue, they are Highest Monthly Earnings, Highest Yearly Earnings, Lowest Yearly Earnings, and Lowest Monthly Earnings. These columns contain the estimated earnings from

the channel. There are 19 categories in the dataset which are determined based on the video content. There were some NaN values in the dataset which we avoided while plotting.

From figure 15, it is clear that the Entertainment category yields the highest income and the rankings of categories remain the same in all 4 graphs.

Fig. 15. Maximum income with respect to different Categories

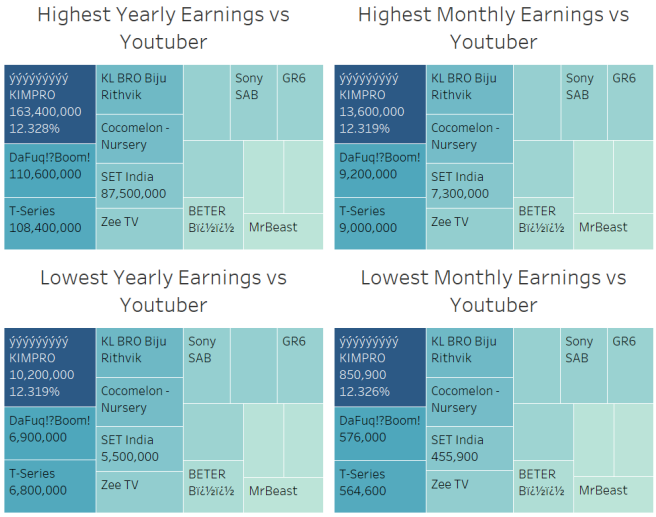


Fig. 16. Maximum Income vs Youtuber

Now, from the figure, 16 i.e., Maximum Income plotted for every YouTuber (Here top 15), again the rankings of YouTubers remain the same. Even their percentage of income is almost the same throughout the graphs. This observation that we see can be clearly seen when we plot the correlation matrix. Correlation is 1 if we try to relate all four of our revenue factors against each other, which is shown in the graphs below.

Next, we tried to visualize Youtubers' earnings in the form of a pie chart and also earnings category-wise.

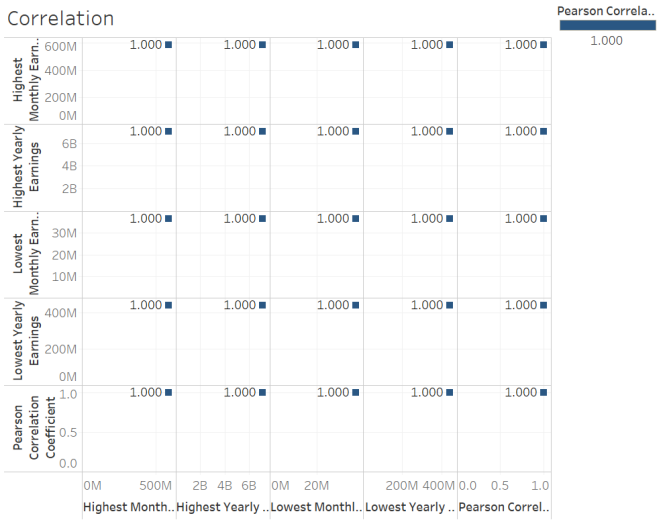


Fig. 17. Correlation between all Revenue related columns

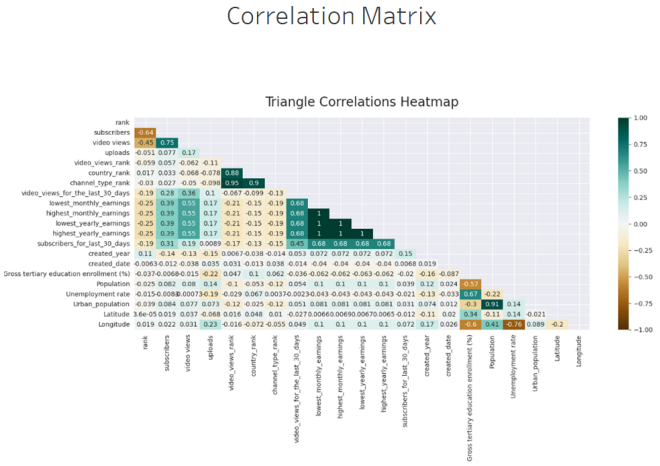


Fig. 18. Correlation Matrix

The figure19, shows us the top Youtubers with their percent of income among the top 20 Youtubers.

And figure 20 shows us the category-wise Highest Yearly Income. This clearly shows us the dominance of entertainment-related content and also that Travel-related content is last in terms of revenue.

The below plot, figure 21 shows us the distribution of revenue category-wise with the percentage income of YouTubers in that particular category.

From the correlation matrix plotted in Python, we can also see that only a very few columns are highly correlated to revenue-related columns. Now when graphs are plotted against all the other columns, only the graph plotted against the views column gives some kind of relationship.

From the data story (Fig 23) we can see that there is no particular pattern in data distribution. But when we plot Views

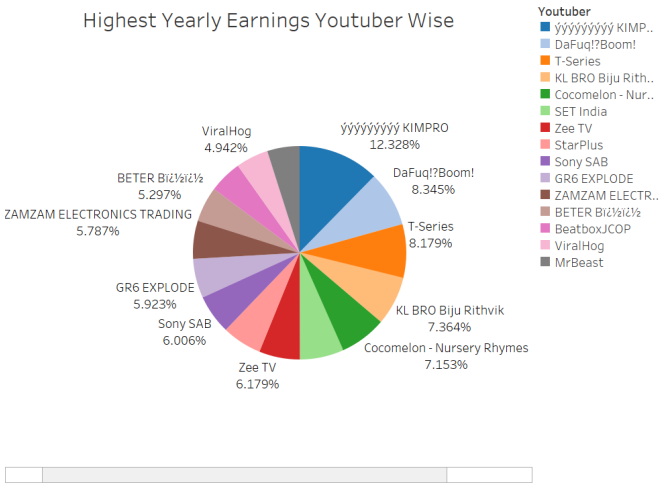


Fig. 19. Highest Yearly Earnings vs Youtuber

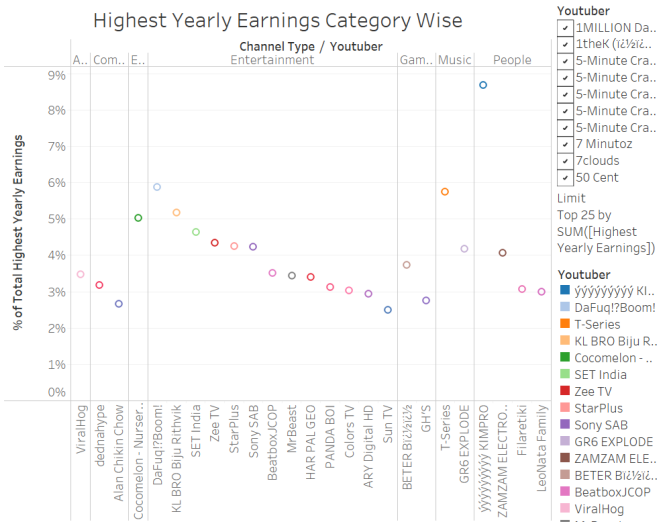


Fig. 21. Highest Yearly Earnings vs Category

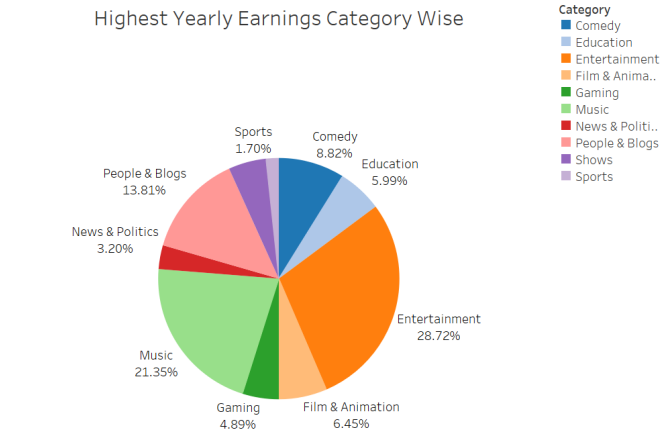


Fig. 20. Highest Yearly Earnings vs Category

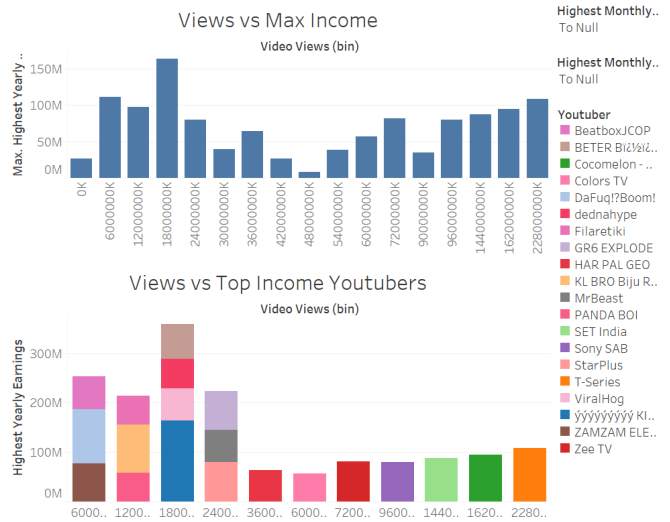


Fig. 22. Views plotted against Maximum Income and also for the Top 20 Youtubers with the highest revenue

against Average Income we get a linearly increasing plot with some outliers.

This plot (Fig 25) contains the highest earning Youtuber Country Wise.

The figure(Fig 26) is the region-wise plot of YouTubers' average revenue and also the count of YouTubers. If you observe both graphs we can see that, USA has the most no of YouTubers and also gets the highest revenue. We can also see that even though no. of Indian YouTubers are less comparatively, they earn more.

C. Unemployment Rate and Education Based

The given data story consists of the following sections:

1) General Overview of the Unemployment rate and trends regarding categories of videos watched and YouTubers involved.
2) Analysis of between subscribers of a YouTube Channel and views with unemployment of a country.
3) What countries with higher unemployment prefer to watch¿
4) What countries with higher tertiary enrollment prefer to watch ?
5) How daily views change with increase with tertiary enrollment.

All 5 pages of the data story can be seen in figure 27,28, 29,30 and 31. We will discuss each of these pages in terms of visualization used and noticeable inferences (if any) in the upcoming part.

1) Page 1: The page starts with a map-based visualization (Figure 32) with a single color color-map that describes the Unemployment rate of the countries. From this, we can infer

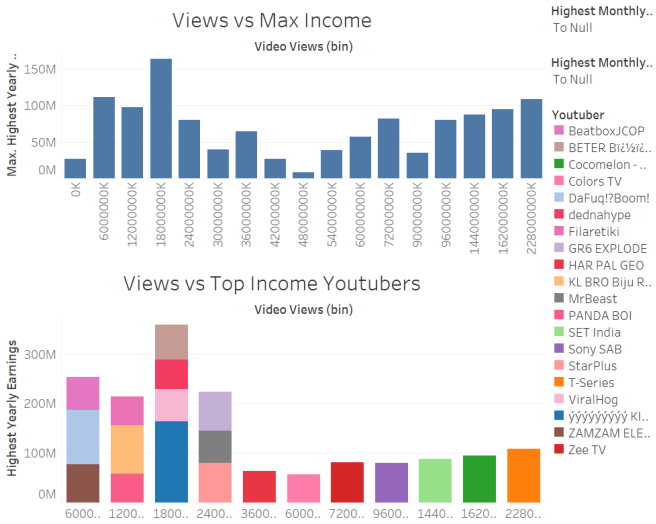


Fig. 23. Views plotted against Maximum Income and also for the Top 20 Youtubers with the highest revenue

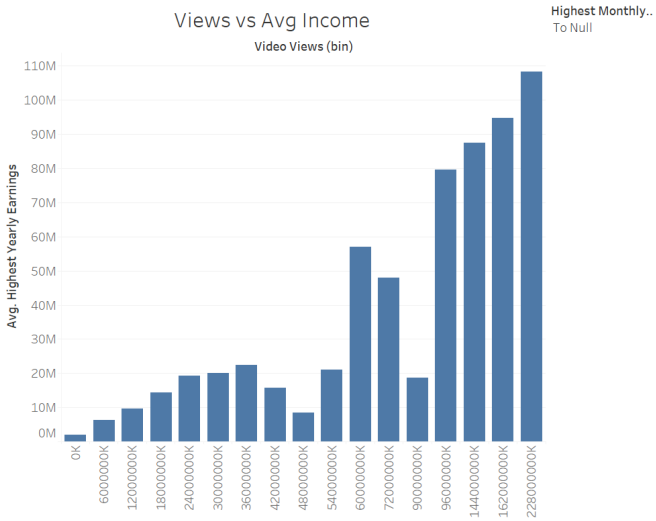
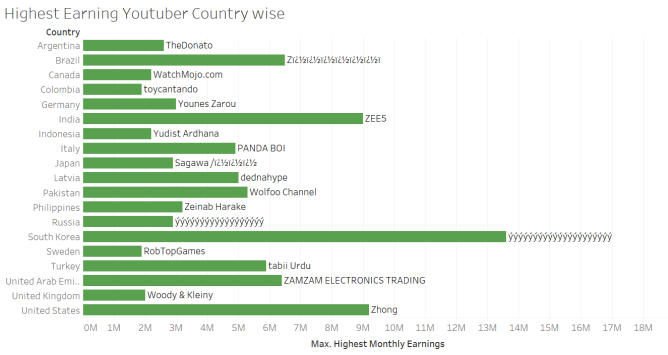


Fig. 24. Views vs Average Income



Maximum of Highest Monthly Earnings for each Country. The marks are labelled by maximum of Youtuber.The view is filtered on maximum of Highest Monthly EarningsandCountry.The maximum of Highest Monthly Earnings filter includes values greater than or equal to 1,795,380.The Country filter keeps 19 members.

Fig. 25. Highest Earning YouTubers Country Wise

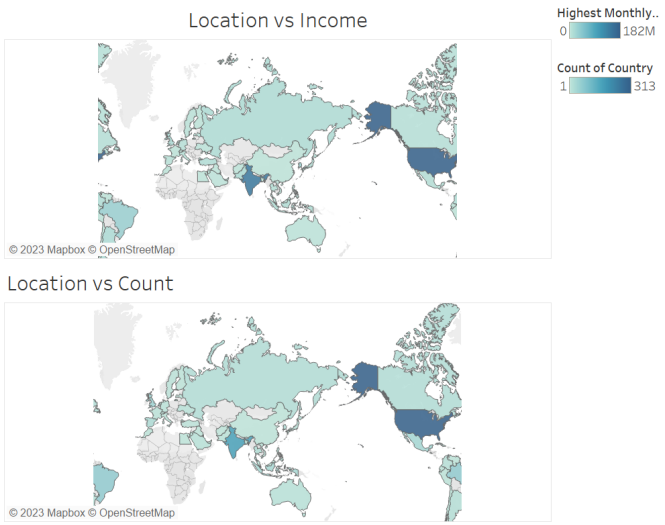


Fig. 26. Countrywise distribution of income and YouTube count

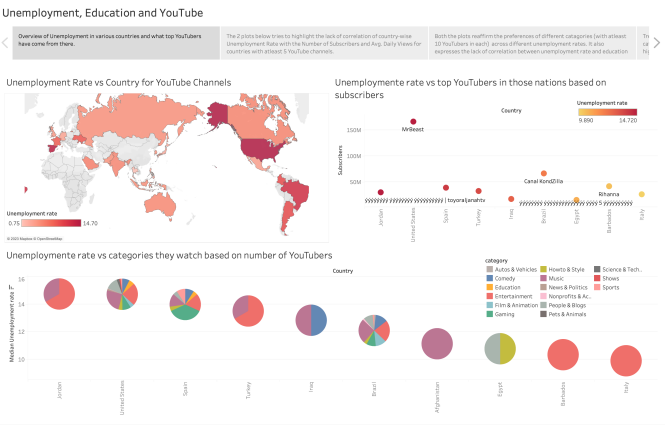


Fig. 27. Story 3 Page 1

that countries like USA, Brazil, and Spain have high unemployment rates. The red color as the base of the color gradient has been taken as its much more prominent and strikes out the large difference of rates easily wrt colors like blue or yellow. However, since a map may not be able to highlight a region a hybrid scatter plot (Figure 33) is plotted ordered on the basis of the unemployment rate where we find a new insight that Jordan is actually the country with the most unemployment rate. The hybrid nature arises from using the mark as a piechart with its channel being the Category of YouTube Channels prominent in those countries. We observe that the countries with a high unemployment rate have more channels in Entertainment and Music with Music comprising a considerable chunk in the top 5 countries ordered by unemployment. Lastly, there is also a scatterplot of the top YouTubers from the countries with the highest unemployment rate based on subscribers. We don't observe any correlation between the properties of the most subscribed YouTuber and the unemployment rate.

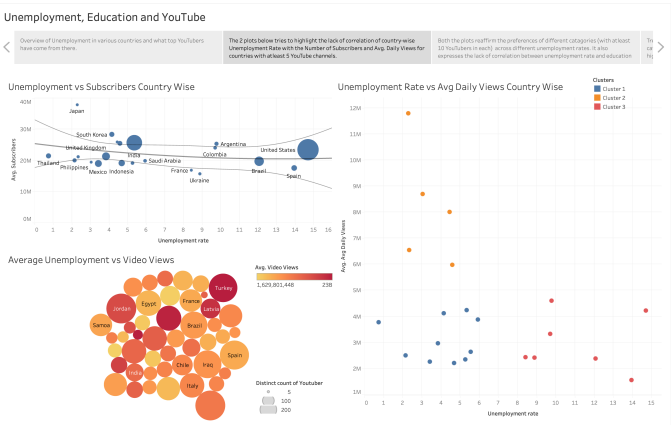


Fig. 28. Story 3 Page 2

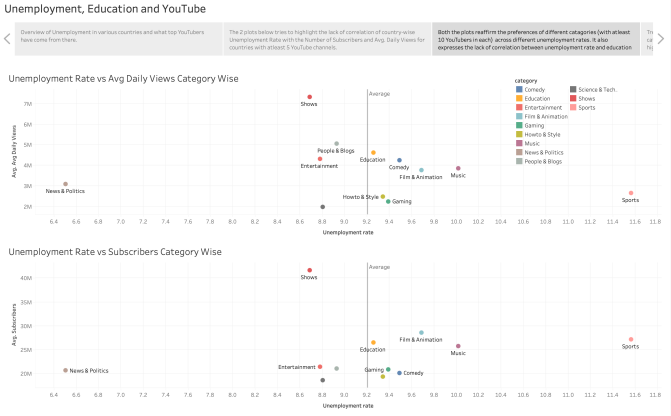


Fig. 29. Story 3 Page 3

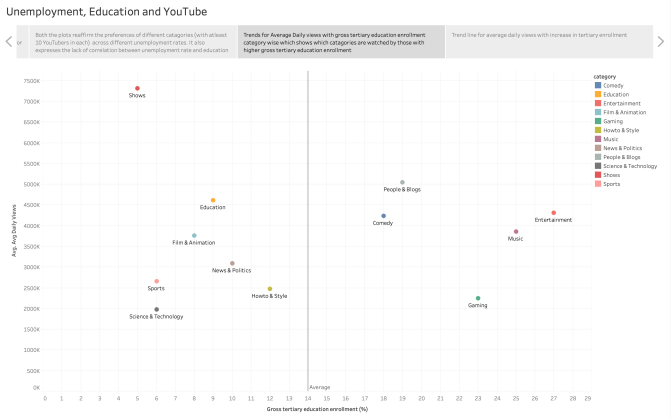


Fig. 30. Story 3 Page 4

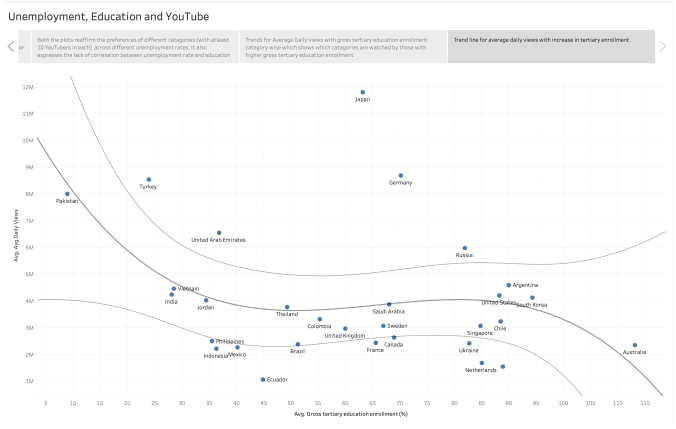


Fig. 31. Story 3 Page 5

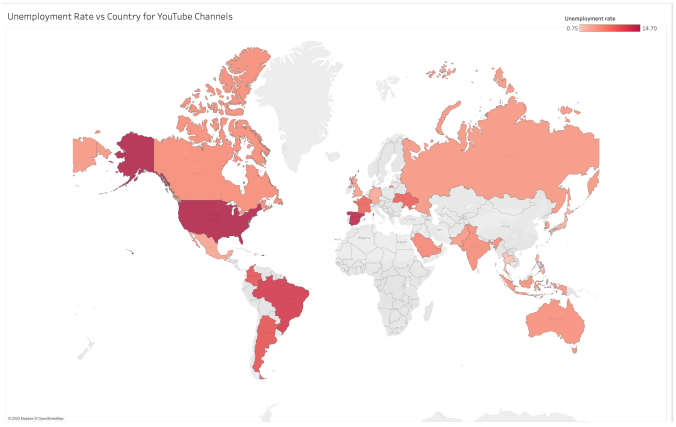


Fig. 32. Map Visualisation for unemployment in countries across the world

2) Page 2: This page tries to infer any correlations possible between the number of subscribers and average daily views of YouTube channels of a country and the corresponding unemployment rates associated with the country. To our dismay, there is no correlation between the above parameters and this is clearly evident from the scatter plot with confidence bounds and trend line as markers shown in figure 34.

For the second plot which is also a scatterplot, we try to visualize the cluster of data points and notice that most countries with high unemployment rates have a lower average daily view. This can be seen in figure35.

We also keep a bubble chart (figure 36) that describes the same as the previous plot but we also allow the outliers (countries with fewer YouTubers (in order of 1-5)). Here we apply a color gradient to highlight the unemployment rate and the size of the average video views. However, as a word of caution, this will not give an accurate description due to the aggregate nature of the data and the number of channels as discussed in the assumption section.

3) Page 3: Our goal here is to understand which video view composition and category preference by countries with high unemployment rates. This takes a deeper look at what we were trying to explore on the previous page. Here given

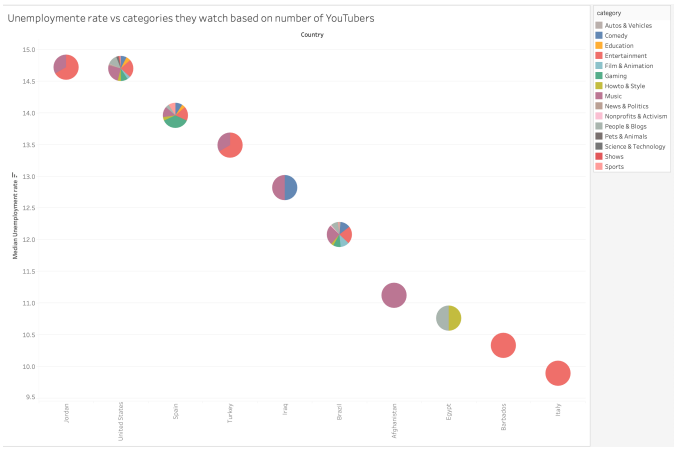


Fig. 33. Scatter plot with pie charts as markers and channel being the categories of video

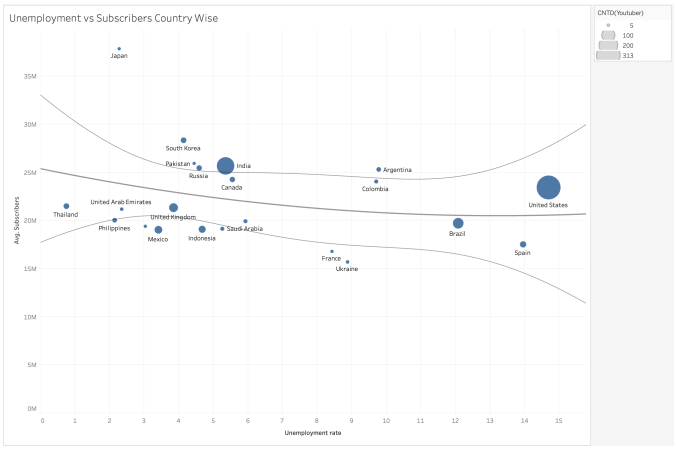


Fig. 34. Proving the lack of correlation between the unemployment rate and subscribers with the help of scatter-plot and confidence intervals

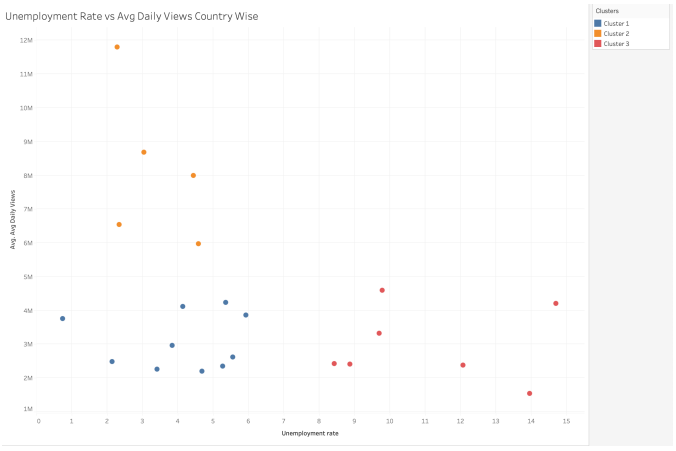


Fig. 35. Clustering plot that describes the low average daily views for countries with high unemployment rate

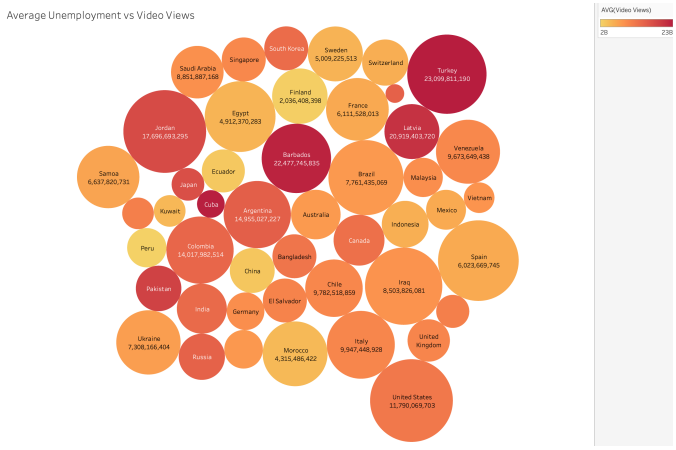


Fig. 36. Bubble chart with color channel highlighting unemployment rate and size channel highlighting average video views.

the smaller view count, we try to understand the composition of those views. We notice using figure 37 that countries with a higher unemployment rate have a very strong preference for watching sports followed by music, film and animation, and gaming.

This is also reaffirmed by the second plot on the unemployment rate and subscribers present on the current page of the story.

4) Page 4: In this section, we try to study the trend of categories and how the viewers of a category have their tertiary enrollment status as. We observe (in figure 38 that people with high tertiary enrollment prefer watching Entertainment, Music, and Gaming while views in shows are followed by a low tertiary enrollment rate which means viewers with low rates prefer to spend more time watching shows.

5) Page 5: Lastly, we conclude our story with a trend analysis of how average daily views change with changes in tertiary enrollment (figure 39). We see a clear indication that on increasing the tertiary enrollment the average daily views for a country decrease. This makes sense as due to enrollment, the citizens are much more occupied with specified tasks rather than spending time watching YouTube. The major outliers noticed are Japan, Germany, and Russia which deviate from the general trend.

VISUALISATIONS

Following are the visualizations that are used and described in detail in the section above.

1) Scatter Plots
2) Pie Charts
3) Cartographs
4) Bubble Chart
5) Heatmap in form of correlation matrix
6) Treemap
7) Density Plot
8) Line Plots
9) Area Plots
10) Bar Plots

Also in each of the types wherever applicable, we have employed various marks and channels for making the visu-

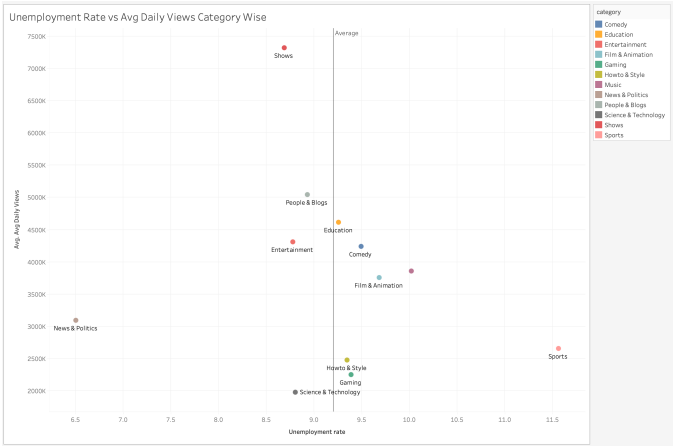


Fig. 37. Unemployment rate vs Daily Average Views with categories color coded

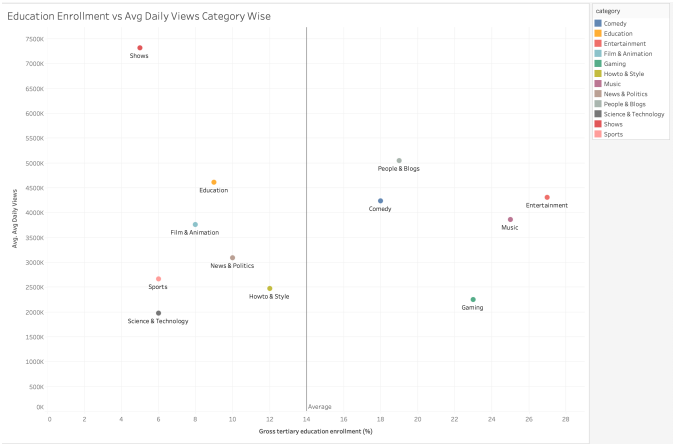


Fig. 38. Unemployment rate vs Daily Average Views with categories color coded

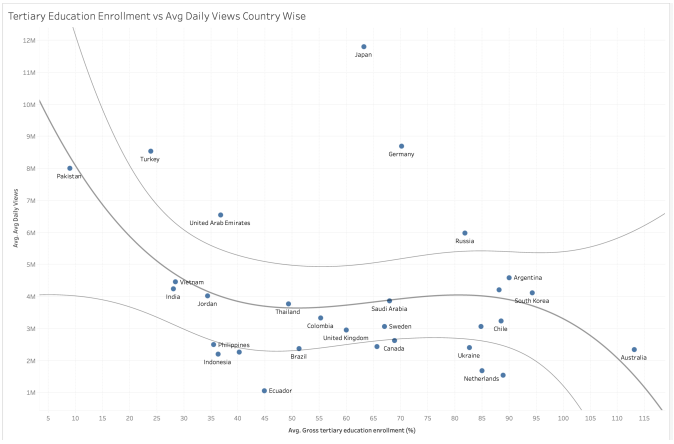


Fig. 39. Trends for tertiary enrollment and average daily views

alizations more expressive for someone to get the maximum insights at the first glance.

Member Wise Contributions

The tasks were initially discussed over a meeting and we came up with the tasks together. For the visualizations, dashboards, and stories, the following distribution was followed:

1) Task 1 and Task 3: Sarthak and Monjoy: Both of us worked together. The tasks weren't divided between us as our productivity and creativity was much better when we worked together. Also, task 3 didn't have many correlating results which required in-depth visualizations while task 1 had a lot of ways to handle and visualize the data and required effort from both.
2) Task 2: Sougandh carried out the analysis of his own and we all combined our findings and analysis for the final project and report.