CS732: Data Visualisation Assignment 3 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

Dataset and Data Mining Options

We continue to use the same dataset as before i.e the Global YouTube Statistics 2023 dataset. This comprises 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. To get a detailed description of each of the columns, please refer to the Appendix. Initially, we planned on scraping data using the YouTube API as suggested by the TAs and prepared the necessary script for the same. However, the following are some of the challenges that we faced:

1) The free API has daily, weekly, and monthly limits which work in a counter-intuitive way. Upon experimenting with the API itself we ran out of credits granted. Repeatedly using the API put a block on our API key generated which was tagged to our gmail accounts.
2) The data from the API didn't consider any additional columns and also many other fields like the yearly earnings were not provided for privacy reasons. Alongside that, the API output itself rounded off figures after a certain digit. This made a lot of measurements like subscriber count or video views unreliable.
3) Further search of data on Kaggle revealed that the dataset provided for A1 is the largest source of YouTube data and all other associated data are subsets of the same dataset.
4) Augmenting data from other sources would help increase the number of columns but would dive away from the main theme of discussing YouTube using visual analytics.

We therefore decided to stick with the dataset we had before and try to do progressive fine-grain analytics to identify interesting insights using a structured visual analytics pipeline.

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) Analysis on Category, Time, and Countries: Sarthak and Monjoy: Both of us worked together. A blurry distinction can be made in the work done. Monjoy worked more on the initial Data Mining and the analysis of Categories, while Sarthak worked more from the 4^{th} iteration of the visual analytics workflow.
2) Analysis on revenue streams: Sougandh carried out the analysis of his own and we all combined our findings and analysis for the final project and report.

I. Visual Analytics Workflow

As mentioned, we will be using the Visual Analytics Workflow as proposed by Keim et al. [1]. This workflow is represented as a flow diagram shown in 1. We will be referring to this

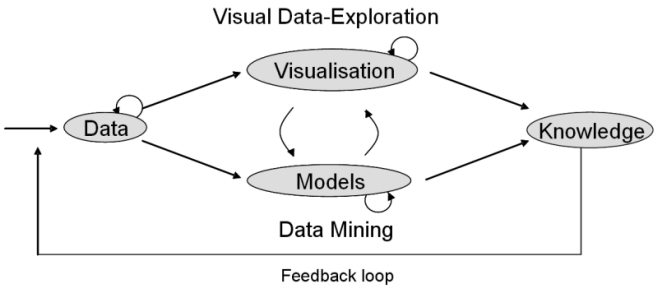

Fig. 1. Visual Analytics Workflow

figure subsequently throughout the report. The following steps are defined as part of the workflow:

- Data: Changes made to data eg. Data Transformations
- Visualization: Plots or figures made using the data eg. histograms
- Models: Models constructed on the available data eg. KMeans clustering
- Knowledge: Inferences or observations drawn from the visualizations and/or Models

Submission Structure

The submission to the assignment consists of this report, a copy of A1 titled 'Appendix_A1_report', and the demo video titled 'demo'. Along with this, the images of the plots created can be found in the Images folder with the filename prefixed with the figure number as mentioned here in the report. Inside the code folder, 2 respective folders with code contributions from both tasks can be found in their respective folder titled the task name along with all required csv and requirements file that cite the packages required for running the notebook successfully.

[1]Keim, D., Andrienko, G., Fekete, J. D., Görg, C., Kohlhammer, J., and Melançon, G. (2008). Visual analytics: Definition, process, and challenges (pp. 154-175). Springer Berlin Heidelberg.

Analysis on Category, Time, and Countries

Loop 1 - Assignment 1

The analysis of the categories, time, and countries was started in Assignment - 1. This can be seen in Task 1 and Task 3 (mentioned in the Appendix). The following steps are followed:

- Data: Additional columns like 'Days Since Created' were created using the available attributes. Grouping of the 'Category' column into bigger categories was also done.
- Visualizations: Multiple visualizations like Area Plots, Juxtaposed Pie Charts and Cartographs, Scatterplots with Trendlines, etc were created.
- Models: Clustering was done as part of a couple of visualizations.
- Knowledge: Several conclusions and possible reasons for anomalies and patterns in the data were given. Some examples include the correlation between video views and subscribers, and other attributes.

Although the content in Assignment-1 can be unrolled into multiple loops, we conclude the work regarding the same as the first loop for the sake of brevity.

Loop 2 - Deciding Categories to Perform Study on

Data Table Creation: For our data to be useful in visualization in the pipeline we perform some basic preprocessing of data and convert it into Data Table. Some instances of preprocessing involve cleaning text names of YouTubers, and removing names (about 16) that are corrupted and cannot be made sense of as we cannot backtrack these points meaningfully. Based on the channel type (category) we try to partition the existing data table.

Visualization 1

To get an initial overview we use our data table to create a donut chart using Plotly to see the composition of channel types in Fig. 2

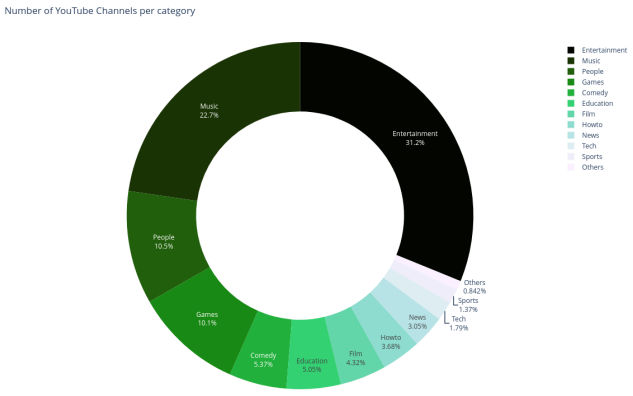

Fig. 2. Donut Chart to view composition of channel types

Knowledge 1: We notice that Entertainment, Music, People, and Games have the highest category share and should naturally show up more in any aggregate analysis as they have more channels contributing to that statistic. We have already plotted the Highest and Lowest Earners per category(refer A1 in Sougandh's contribution) but not the same for views in the previous Loop 1, Let us analyze that to understand whether Quantity dominates or Quality of channels in a category

Visualization 2: A dumbell plot has been created as shown in Fig. 3 to visualize the maximum and minimum values range in the case of Total Video Views and Subscribers generated in the categories.

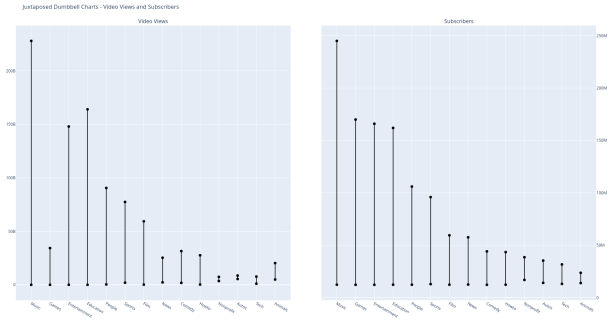

Fig. 3. Dumbell Chart to show the range of values in each case

Knowledge 2: We notice from this that Music dominates in the case of both subscribers and video views across all categories even if it doesn't have the most number of channels. However, to get a further conclusion on the hypothesis of "Quantity Dominance over Quality" one needs to check the average metric of each of the categories.

Data Filtration 1: As we are studying specific details of the data we filter out our data table to take in the required columns of 'Youtuber', 'channel_type', 'video views', and 'subscribers'.

Visualization 3: We create a juxtaposed scatterplot to analyze the average subscribers and video views. The output can be seen in Fig. 4

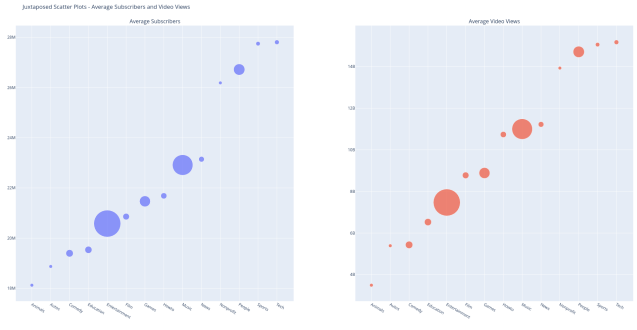

Fig. 4. Juxtapose Visualization of Scatterplots

Knowledge 3: The average metric reveals a different story that Tech has the most average subscribers and video views, however, there are only 17 channels in the data present. The size of the scatter shows the count and on hovering the exact value can be seen. Music having the most number of subscribers and video views lies somewhere in the middle of this trend plot. However, using the plot before and this we can hypothesize that on the basis that in cases where the range between max and min values in categories like Music is high, the average might be skewed towards the lower part of the dumbell. To confirm this we need to come up with the distribution of some of the categories to understand. We will try to get distribution plots for categories like Music, Education, People, and Games. This design decision is purely based on the insights gained from the plot on the data given so far.

Visualization 4: To view a distribution in a detailed way, we chose a violin plot as it expresses both the underlying data distribution shape as well as the corresponding boxplot. Plotly's violin plot also allows the displaying of critical information about distributions like upper fence, and whiskers, and on hover a user skilled in statistics can gain more insights using the values along with the basic visual insights. This can be seen in Fig. 5

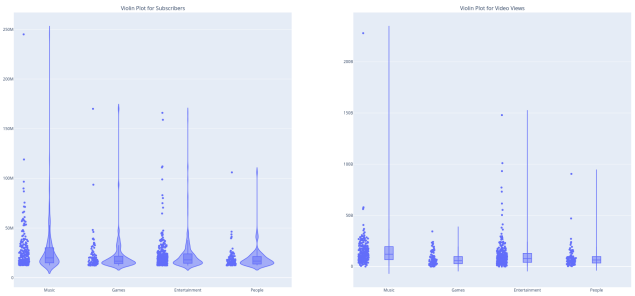

Fig. 5. Combined View of Violin plot on Subscribers and Video Views

Knowledge 4: We notice that for the three categories, the first 2 Music and Entertainment suffer from a wide range of values and most of them are concentrated on the side lesser than the (maximum + minimum)/2 mark. This results in an overall lesser actual sample mean for subscribers. While violin plot for video views confirms the same for this but also brings an interesting insight regarding the view values in general. There is a very thin violin (more apparent on zooming in) spread in case of video views. A thin violin suggests that the data points are concentrated in a narrow range, indicating lower variability or spread. There is no particular concentration of values at any region which suggests that these video channels may provide content that differs and they have their unique fanbases or people see some channels to watch some specific content on them while viewing the other channels to view the other content.

This brings us to the end of loop 2 and helped us to decide on which categories we can perform further detailed analysis as part of our zoom + context visualization methodology that can be employed in the case of data analytics using visualization.

Loop 3

As concluded from the analytics present in loop 2. We now start our analysis making our data selection even more specific by studying only some of the categories. As discussed above we select the categories of Music, Entertainment, People, and Games. A user might be interested in studying only some of the domains where there are a good number of YouTube channels and try to study which are the highly related features in these categories to understand how the YouTube content creation landscape has evolved.

Data Filtration - 1: We filter out the data based on the channel type to be only 'Entertainment', 'Music', 'People', 'Games'.

Visualization - 1: To get an overview of how numerical features interact with each other we go ahead with a correlation plot. This can be seen in Fig. 6

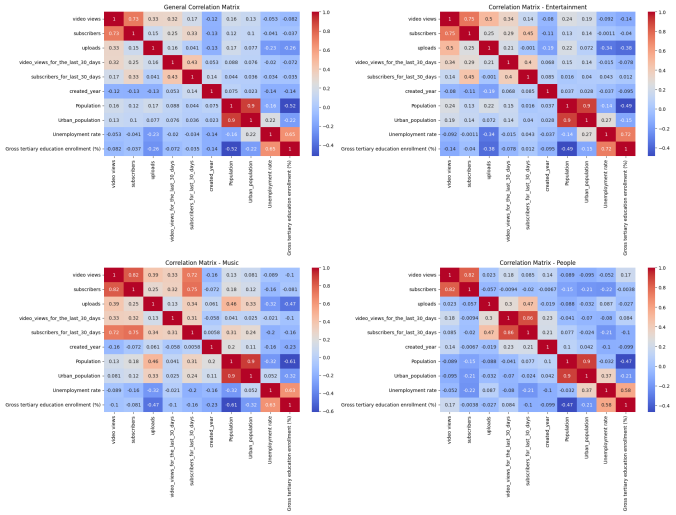

Fig. 6. Correlation Matrices visualized for 3 categories along with Entertainment, Music, and People along with General Plot for reference and comparison provided in a single main plot

Knowledge 1: Some insights that can be seen here are:

1) In general, the covariance between Subscribers and Video Views is 0.73. But, in the case of Music and People, this number is greater and reaches about 0.82. On the other hand, even after being a bigger category (with more channels), the covariance value for Entertainment is close to the general value of 0.75.
2) In the case of uploads w.r.t to subscribers in general there is a no standout correlation. However, in Music and Entertainment, this value is significantly larger. This can help to infer that a large number of subscribers

get added to music and entertainment channels only if they are putting on new music regularly. For the rest of the categories in general spamming uploads as a means to get subscribers won't help to grow the channel unless your channel is in music or entertainment. On the other hand, the trend for the People category is far from the general. The covariance is slightly negative (although negligible). This shows that the number of uploads makes no difference to the subscribers for the channels in the People category.

3) Another case similar to 2 can be seen in the case of the population of a place and subscribers in the last 30 days where the general trend suggests that there is near zero correlation (or negative correlation), however in the case of categories like Music has a large enough correlation which means that addition of population can help in the growth of these categories automatically as recent data suggests.

The overall conclusion from this loop is that from the above aggregate statistics for the specialized video types we see that the Entertainment Category is a little different than the other categories. The correlation between the number of subscribers and video views is weak for the Entertainment Category as compared to the other bigger categories. In the upcoming section and loop, we will take a deeper look at the Entertainment Category.

Loop 4 - A Closer look into the Entertainment Channel Type

Data 1: Transformation of the data is done to obtain the Video Views Rank and Subscriber Rank attributes for both the Entertainment channel type. These attributes indicate the rank of a particular YouTube channel in terms of the number of total video views. So, the channel with the highest number of video views is given rank 1, the channel with the second highest number of video views is given rank 2 and so on. The same is applied to obtain the Subscriber rank for each channel under the channel type 'Entertainment'. The Video Views rank is obtained by first taking a subset of the dataset containing only channels of type 'Entertainment' and then taking the index of the channel after arranging the channels in the descending order of the video views. Subscriber rank is obtained similarly.

To put it simply, a lower Video Views rank equals more Video Views.

Visualization 1: The scatterplot for the Subscriber Rank versus the Video Views Rank is created to observe the distribution of the channels on their Video Views Rank and their Subscriber Rank. This plot for Entertainment can be seen in 7

A green identity line (x=y) is also drawn on the plot.

Model 1: Regression analysis can be done on this data. We plot a linear regression (least squares) on this data. A slope of less than $45\,\mathrm{deg}$ is obtained and the same is plotted back. This line is visible as the blue line in 7.

Knowledge 1: The points are widely scattered about the green identity line. The blue regression line has a slope of less than $45\,\mathrm{deg}$ and deviates from the identity line. Ideally,

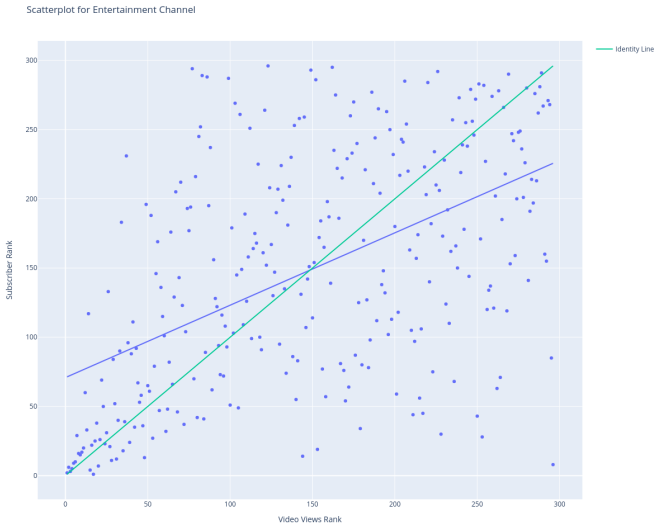

Fig. 7. Scatter Plot for the channels of type Entertainment. The X-axis denotes the Video Views Rank and the Y-axis denotes the Subscribers Rank

if the subscribers and video views were perfectly correlated, the subscriber rank and the video views rank for each channel would be the same. In that case, all the points would lie on the identity line. The regression line indicates the amount of deviation from this line.

Some channels lie very far from the identity line and lie on the opposite sides of the spectrum. They either have a high subscriber rank and a low video views rank (upper triangle) or a high video views rank and a high subscriber rank (lower triangle). The points lying in the upper triangle have more video views, but fewer subscribers and the opposite is applicable for the lower triangle.

Additional clustering can help in understanding the data better

Model 2: K Means clustering with number of clusters equal to 4 is done on the data. Another attribute with the cluster number is created for each channel and added to the existing data table.

Visualization 2: Using the additional cluster information, the scatterplot is made again, but this time, different colours are given to different clusters. The resulting plot can be seen in 8

Knowledge 2: With this plot, we can see the four different types of clusters as follows:

- Cluster 1 (Green): The channels with low video views rank and high subscribers rank.
- Cluster 2 (Magenta): The channels with high video views rank and high subscriber rank
- Cluster 3 (Red): The channels with low video views rank and low subscriber rank
- Cluster 4 (Blue): The channels with high video views rank and low subscriber rank

Fig. 8. Scatter Plot for the channels of type Entertainment. The X-axis denotes the Video Views Rank and the Y-axis denotes the Subscribers Rank. The clusters are denoted with different colours

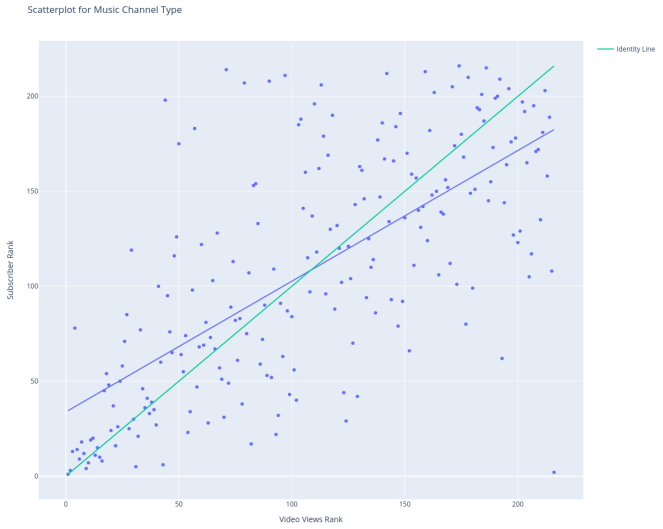

Fig. 9. Scatter Plot for the channels of type Music. The X-axis denotes the Video Views Rank and the Y-axis denotes the Subscribers Rank

The orange identity line separates Cluster 1 and Cluster 4, supporting our earlier conclusion of the channels which deviate from the identity line.

Although this visualization is useful to conclude a lot about the Entertainment channel types, further conclusion with other channel bigger channel type like Music is required.

Loop 5 - Comparison with the Music Channel Type

Data 1: Transformations similar to those done for the Entertainment channel type are performed to obtain the Video Views Rank and the Subscribers Rank for the Music channel type as well. The semantics of these attribute remain the same as they were for the Entertainment channel type.

Visualization 1: Similar to the Entertainment channel type, a scatterplot between the Video Views rank and Subscribers rank is made. In a similar fashion, a green identity line is drawn. This plot can be seen in

Model 1: Similar to the Entertainment channel type, a blue linear regression line is calculated and drawn on the plot.

Knowledge 1: The spread for the Music channel type is much less and near to the green identity line as compared to the Entertainment channel type. Along with this, the blue regression deviates much less than that of the Entertainment channels.

A cluster of analysis of the Music channel type is also warranted to compare the patterns observed.

Model 2: Similar to Entertainment, 4 clusters are fit using KMeans. These are added to the data table as attributes.

Visualization 2: Using the additional cluster information, the scatterplot is made again, a color scheme similar to that of Entertainment is used. The resulting plot can be seen in 10

Fig. 10. Scatter Plot for the channels of type Entertainment. The X-axis denotes the Video Views Rank and the Y-axis denotes the Subscribers Rank. The clusters are denoted with different colours

Knowledge 2: Even though the spread of the Music channels is much less, the patterns observed in the clusters are similar and clusters similar to that in Entertainment are obtained. For the Music channels, the separability of Cluster 1 (Green) and Cluster 4 (Blue) from the identity line is a bit less than Entertainment. This can be because of the decrease in deviation from the identity line.

The visualizations indicate that there are a significant number of channels of the Entertainment type which have less subscribers, but more video views and vice versa as compared

to other bigger channel types like Music. The reason behind this can be that there are channels which have less views, but more subscribers, this is observed more in the newer channels. On the other hand in the older, less active channels there are more subscribers and less video views as the channel activity is reduced. To verify this, we can look at the channel age spread in these channel types.

Loop 6 - A Look at Channel Age and other Important Factors

Data 1: To obtain the Channel Age attribute (in days), the attributes Created Date, Created Month and Created Year are used along with $31^{st} December 2022$. $31^{st} December 2022$ is used because the original dataset was collected with the information limited to 2022 only. This attributed is added the data table.

Visualization 1: A juxtaposed visualization is created with two violin plots, one for the Entertainment channel type and the other for the Music channel type. These plots share the same y axis. Entertainment is shown in orange and Music is shown in Blue. Box plots corresponding to each violin are also plot. The visualization can be seen in 11

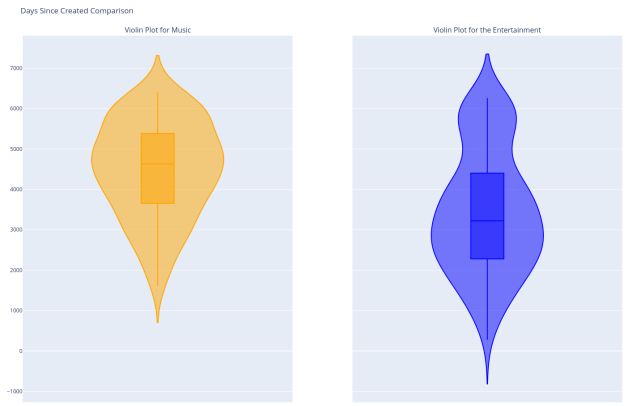

Fig. 11. Juxtaposed Violin Plots for the channels of type Entertainment and Music. The Y-axis denotes Days Since Created.

Knowledge 1: A curios observation can be made in the visualization in 11 which might explain the spread of the scatterplot in 8 and 10.

In the Music plot, we see a higher probability distribution focused near one place and hence there is only one global maxima in the distribution. This means that there are more channels which are of a particular age. As lesser new music channels have emerged, these older channels are active and have more video views along with more subscribers.

On the other hand, in the Entertainment plot, we can see that the probability distribution is such that two local maximas are formed. This indicates the possibility of two types of deviating channels (from the identity line), older ones with more subscribers which are less active and the newer active ones (hence more views) and less subscribers.

The inclusion of time introduces a possinility of a lot of new attributes. These attributes can also be studied.

Loop 7 - Supervised Learning to study Feature Importance

Data 1: Using the age of the channel, the following attributes are created and added to the data table:

- Average Daily Views
- Average Monthly Views
- Average Yearly Views

Attributes like Average Daily Subscribers can also be created, but using them for Supervised learning is basically cheating as its product with the age directly gives the number of subscribers.

Model 1: A light gradient boosting regressor is used to fit on the data table. This model is used because of its resilience against null values in the data and also its high explainability which in turn provides us with the feature importances for prediction.

After fitting the model, a feature correspoding corresponding to each feature is obtained, which are normalized, such that the sum of all these values comes out to 1. A new data table is created using these values.

Visualization 1: A pie chart showcasing the feature importances of each features is plot using the newly created feature importance table. This plot can be seen in 12. Different colours are given to each feature, the angle channel of the pie chart denotes the percent of the importance of each feature.

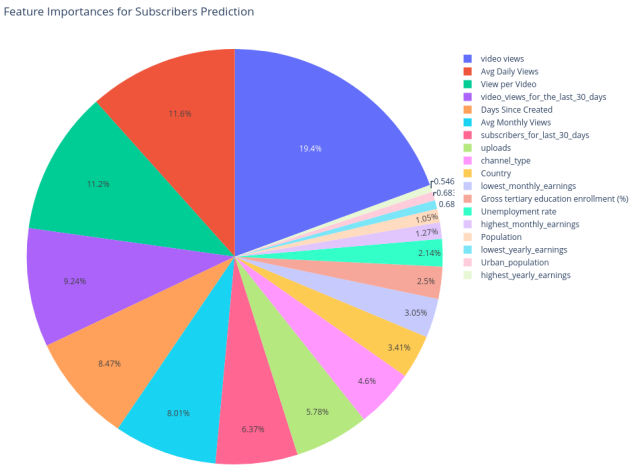

Fig. 12. Pie Chart for the feature importance

Knowledge 1: Unsurprisingly, total video views is the most important attribute to predict the number of subscribers. Following this, 3 out of the top 5 most important features are based directly or indirectly on the channel age. This indicated the importance of being consistent in quality (having high Average Daily Views and Views per Video) and being consistent in uploading (having a greater channel age) to get more subscribers.

But the plot suggests that country is not an important factor. To study this more, we will take a look at the Country attribute.

Loop 8 - A Closer look at Country

Data 1: Due to earlier observed problems in the Channel Type attribute where channel types with less number of channels adding noise to the data, we will remove the countries with less than 20 channels. This leaves us with a Data Table with 7 countries.

Visualization 1: To observe the effects of the Country attribute along with the others, we plot a Parallel Coordinates Plot using the new data table and limited attributes. This can be seen in 13

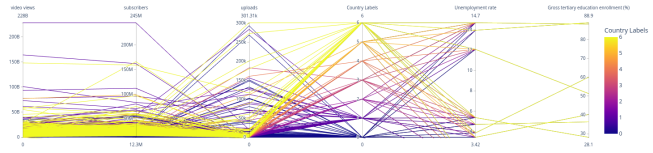

Fig. 13. Raw Parallel Coordinates Plot for the feature interdependence.

Using interactions (brushing) only one of the countries (US) is selected and can be seen in 14. This then corresponds to multiple values of Unemployment Rate and Gross Tertiary Enrollment (%) attributes, indicating spurious values in these attributes. Hence these attributes are discarded.

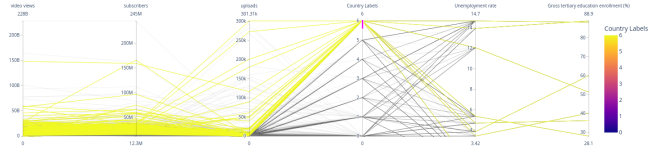

Fig. 14. Spurious values indicated in the Parallel Coordinates Plot.

Using brushing on the Subscribers attribute, we highlight the channels with extremely high number of subscribers (more than 50 million). This can be seen in 15

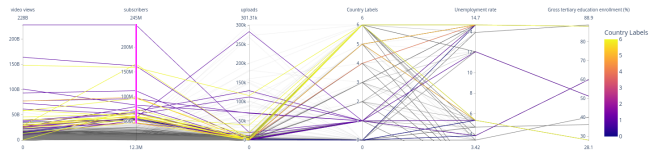

Fig. 15. High Subscriber channels indicated in the Parallel Coordinates Plot.

Knowledge 1: Countries with label values 6 and 1 are highlighted. These countries are the US and India. Even after being very different culturally and in various other aspects like population, these countries are the ones with the highest number of channels with exceptionally high number of subscribers. This warrants a closer look at these countries.

Visualization 2: To study the differences in the channels belonging to India and the US, the following plots are drawn:

- Juxtaposed Pie Charts indicating the number of channels belonging to different channel types. Can be seen in 16
- Juxtaposed Violin Plots with shared y axis to compare the distribution of channels with respect to their ages. Can be seen in 11
- Juxtaposed Violin Plots with shared y axis to compare the distribution of channels with respect to their subscribers. Can be seen in 18

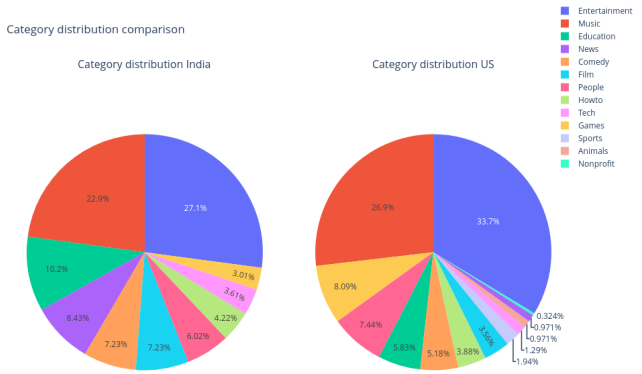

Fig. 16. Pie Charts indicating the distribution of channels in different channel types.

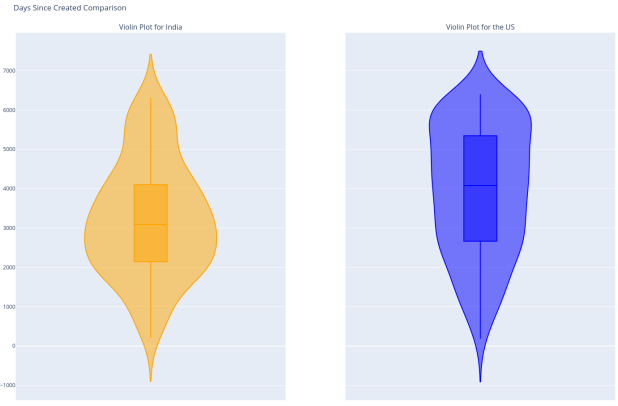

Fig. 17. Violon plots indicating the distribution of channels with respect to their ages.

Knowledge 2: The pie charts in 16 shows that While the two biggest channel types for both India and the US are the same (Entertainment and Music respectively), the other channel types in the top 5 differ for both. These channel types are sometimes very differnt. For example the 3rd most popular

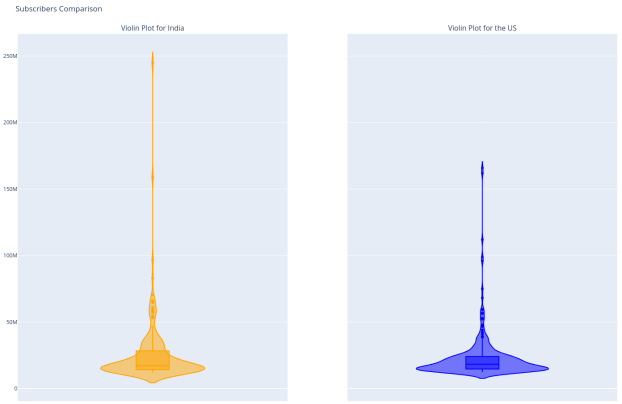

Fig. 18. Violon plots indicating the distribution of channels with respect to their subscribers.

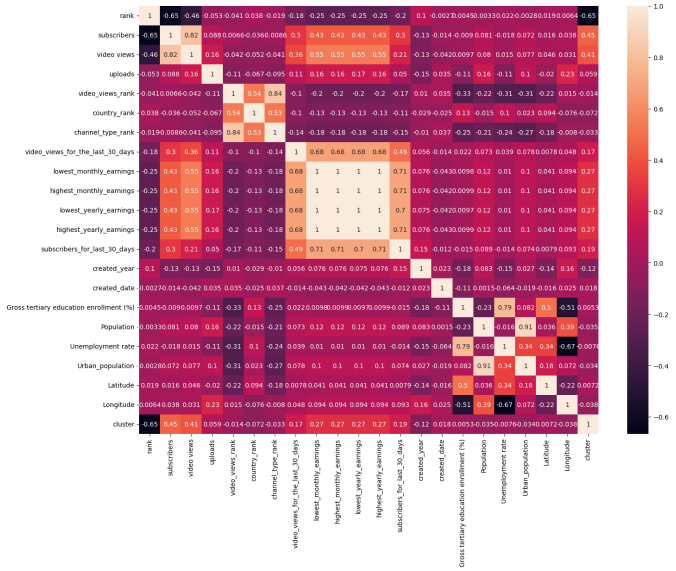

Fig. 19. Correlation Matrix Heatmap

channel type in India is Education, while for the US it is Games. This shows us how the culture in different countries has shaped the popular channels. Also, the chances of a channel type being successful is different in these countries.

Next, the violin plots in 17 indicate the emergence of channels from these countries. While even older channels from the US have high subscribers, and the number is more or less consistent with time, the distribution in India is different. Channels originating around 2013-14 have more subscribers. This can be attributed to the boost in the availability of high speed internet during that time.

Next, violin plots in 18 shows that even after being different, the distribution in the channels versus the subscribers is very similar. There are some outlying channels with very high number of subscribers as shown by the long tail of the violins. But most channels lie around 16-18 million subscribers as indicated by the median lines of the box plots.

Analysis on Revenue Streams

Before creating visualizations, the dataset underwent pre-processing steps to ensure data cleanliness. Rows with minimal null values were retained, and missing values in ”object” columns were replaced with ’not_specified’, while numerical columns received either 0 or -1. Non-alphanumeric characters in the ’Youtuber’ and ’Title’ columns were replaced, and trailing/leading whitespaces were removed. Rows with 0 video views and those with a blank ’Youtuber’ name were filtered out. The index was reset, and a rounded numerical summary of the dataset was generated. These actions aimed to optimize the dataset for clear visual representation and subsequent analysis.

Observation from Correlation Heatmap

Subscribers and Video Views:

- There is a strong positive correlation between subscribers and video views.
- Channels with more subscribers tend to have higher view counts.
- Both these variables exhibit moderately weak correlations with other variables, with no correlation observed with video uploads.

Earnings and Subscriber/Video Views:

- Earnings exhibit a low to moderate positive correlation with both the number of subscribers and video views.
- Factors such as ad revenue variability, audience engagement, and monetization methods could contribute to this correlation.

Earnings and Video Uploads:

- There is a negligible positive correlation between earnings and the number of video uploads.
- Simply uploading more videos does not necessarily lead to increased earnings.

Correlation of Subscribers:

- The variables highly correlated with subscribers are video views and average earnings.
- Video uploads show a very low correlation with subscribers.

Correlation Heatmap Observations:

- The variables ”Lowest Monthly Earnings” and ”Highest Monthly Earnings” show a perfect positive correlation (absolute correlation of 100%), indicating a strong relationship between these two variables (We have seen these in our bar plots from Assignment 1 too).
- Even other income variables like “Lowest Yearly Earnings” and “Highest Yearly Earnings show a very high positive correlation which is almost equal to 100%.

Next Steps: Visualizing the correlation through scatter plots can provide a more detailed exploration of the relationships between these variables. We use correlation mapping with scatter plots to get a better understanding.

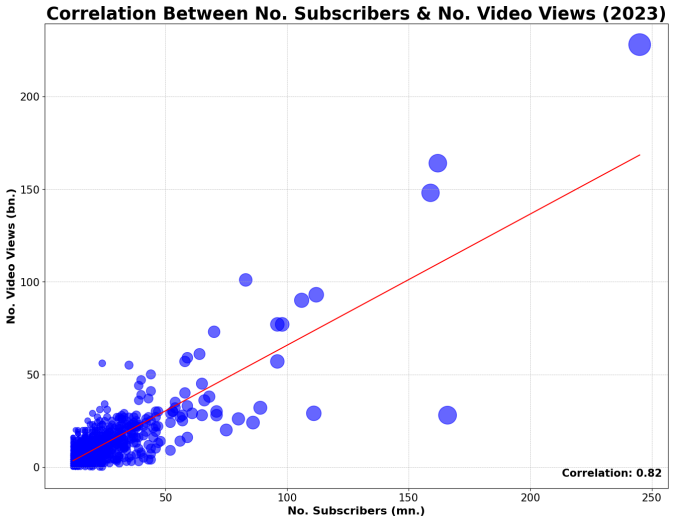

Fig. 20. Correlation between number of subscribers and video views

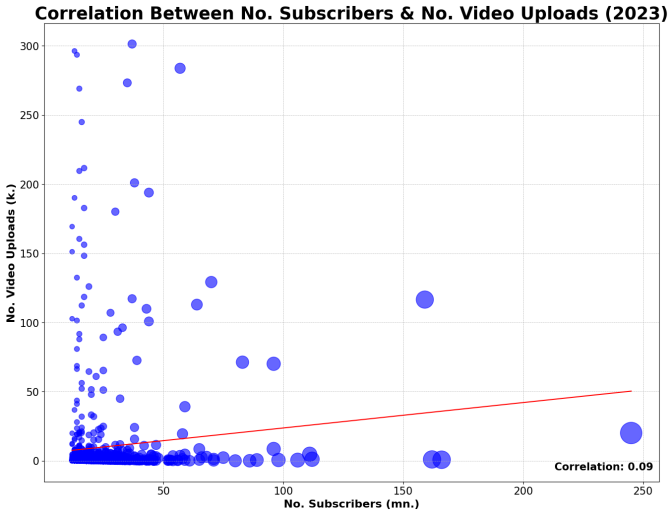

Fig. 21. Correlation between number of subscribers and video uploads

Reiterating the initial observations (From Heatmap), there is a clear and strong positive correlation between the number of subscribers and video views. This underlines the fact that channels with a larger base of subscribers tend to garner higher view counts. The majority of data points on the scatter plot are concentrated in the lower ranges for both subscribers and video views. However, there are a few outlier channels that stand out with notably larger audiences.

Combining Results from these graphs and previous observations (Assignment 1 and 2):

- The top 10 channels, with a correlation rate of up to 82%, affirmed the positive relationship between subscribers and views.
- However, when examining the broader scatter plot, a substantial majority of channels fell within lower subscriber and view ranges compared to the top 10, with only a small fraction showcasing significantly higher numbers.
- 95% of YouTube channels have fewer than 46 million subscribers and less than 28 billion views. This indicates a concentration of channels in the lower range and suggests that only a few channels stand out significantly in terms of both subscribers and views.

Conclusion:

There is a notable dependency between subscribers and views, and the data suggests that only a handful of channels truly excel in terms of success, as measured by both subscribers and views. This leads to the pivotal question: What factors contribute to the exceptional success of a YouTube channel? Further exploration is needed to uncover the specific elements influencing a channel's performance in the dynamic landscape of YouTube.

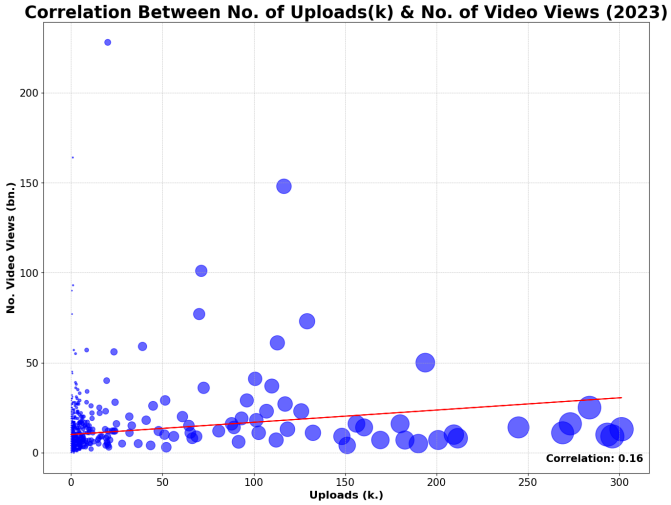

Fig. 22. Correlation between number of uploads(k) and number of video views

There is a minimal positive correlation between a channel's subscriber count and the number of videos they've uploaded. We have seen this in the correlation heatmap. This suggests that the quality and content of videos may have a more significant impact on attracting subscribers than the sheer quantity of uploads. Certain channels stood out with a unique pattern of low subscribers but an exceptionally high volume of videos. This could indicate the presence of dedicated content teams consistently producing short-duration videos. Such channels are likely in sectors like news, entertainment, or music, known for frequent updates or short clips. Correlation plots for upload views showed extremely low correlation rates (0.9% and 17%) with inconsistent data dispersion. The conclusion is that having many uploaded videos does not necessarily correlate with higher subscribers and views. The emphasis on video quality

and content is crucial for attracting subscribers and viewers, rather than solely focusing on the quantity of uploads.

Exceptional Cases:

Exceptional cases were observed through scatter plots, highlighting channels with low subscribers and views but a significantly high number of uploaded videos. This could suggest content categories heavily exploited by organizations with ample resources and long-term operations, possibly in entertainment, news, or music genres known for frequent short-duration uploads.

Conclusion: The data indicates that the correlation between subscriber count and the number of uploaded videos is minimal. Quality and content play a crucial role, as evidenced by exceptions where channels with high video counts don't necessarily have more subscribers or views. Categories explored with high frequency include those likely to produce short, frequent videos, such as news updates or interview excerpts.

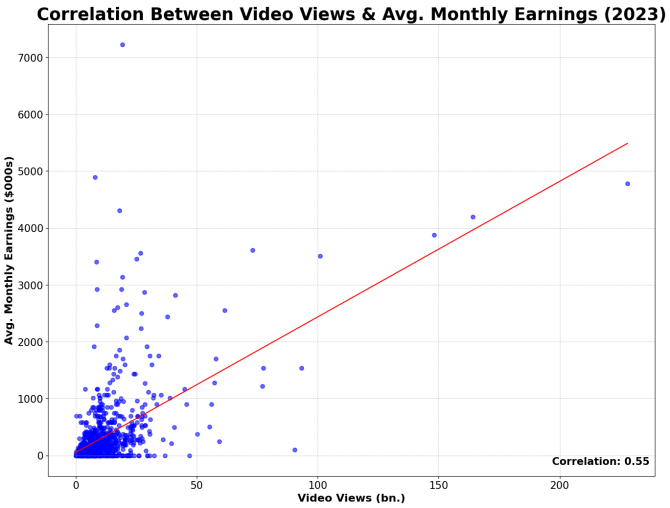

Fig. 23. Correlation between video views and average monthly earnings

The correlation between views and average monthly earnings, calculated from the mean of minimum and maximum monthly earnings, is positive but not exceptionally strong. This implies that accumulating a high number of views doesn't guarantee a proportionate increase in earnings. Other factors like video length, user engagement metrics, and strategic ad placements play a substantial role in influencing a channel's revenue. A moderate correlation emphasizes that the quality of views, considering factors like viewer retention and ad click-through rates, may be more crucial than sheer volume. Channels prioritizing engaging, high-quality content can potentially earn more, even with a lower number of views. The correlation aligns with the idea that successful YouTubers often diversify their revenue streams. Sales of merchandise, sponsored content, affiliate marketing, and fan donations contribute significantly to earnings. This reduces dependence on traditional ad revenue generated solely from views.

Key Observations: High views alone don't guarantee high income, indicating the importance of additional revenue-influencing factors. Quality of views, focusing on viewer retention and engagement metrics, maybe more crucial than sheer volume. Successful YouTube channels often rely on diverse revenue sources, reducing dependency on traditional advertising revenue tied solely to user views.

The correlation analysis highlights the multi-faceted nature of YouTube channel revenue. While views play a role, factors such as video quality and diversified revenue streams significantly contribute to a channel's overall financial success

I have also tried out k means clustering on the dataset and could come up with a few conclusions. After selecting and scaling the relevant features, the elbow method is used to find the optimal number of clusters. Subsequently, K-Means clustering is applied, and the resulting cluster assignments are visualized through a scatter plot. This analysis helps identify inherent patterns and groupings based on the specified features, aiding in the exploration of distinct audience engagement levels among YouTubers.

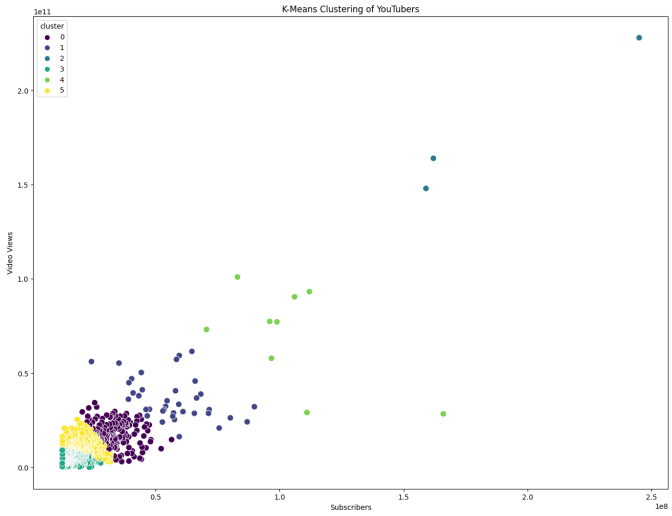

Fig. 24. K-Means Clustering performed on YouTubers

Based on the k-means clustering results for the 'video views' and 'subscribers' columns, the dataset has been divided into six clusters. Each cluster is characterized by its centroid values for 'subscribers' and 'video views.' Here are the centroids for each cluster:

1) 1. Cluster 0:
   - Subscribers: 34.1 million
   - Video Views: 18.3 billion
2) Cluster 1:
   - Subscribers: 56.5 million
   - Video Views: 36.1 billion
3) Cluster 2:
   - Subscribers: 188.7 million
   - Video Views: 180 billion
4) Cluster 3:

- Subscribers: 15.7 million
- Video Views: 4.7 billion

5) Cluster 4:
   - Subscribers: 104.5 million
   - Video Views: 69.8 billion
6) Cluster 5:
   - Subscribers: 20.2 million
   - Video Views: 12 billion

Why 6 clusters? The idea is to choose the value of k at the point where adding more clusters doesn't significantly improve the explanation of the variance within the data. The point where the rate of improvement slows down (the elbow) is often considered a good choice for the number of clusters.

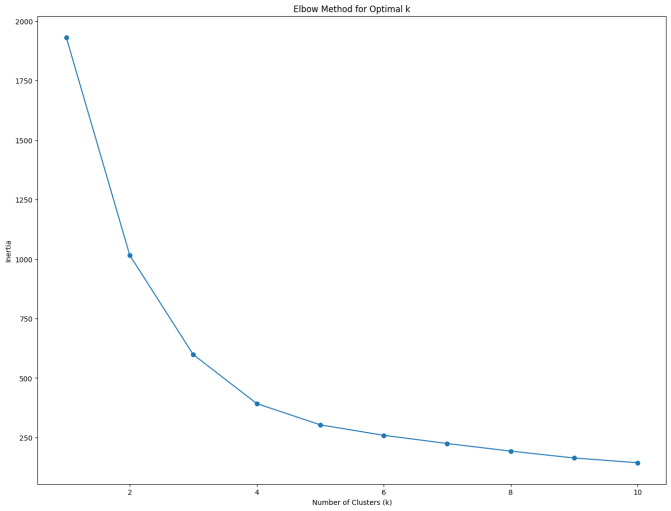

Fig. 25. Elbow Plot to determine K

So here we see that after 6 clusters, the rate of improvement doesn't change much. Hence chose no of clusters as 6

The k-means clustering results provide insights into distinct segments within the dataset, based on the features 'subscribers' and 'video views.' Here are some potential inferences:

Cluster Size and Engagement:

- Cluster 2: This cluster stands out with a significantly higher number of subscribers and video views, indicating a group of highly engaged and popular YouTube channels.
- Clusters 1 and 4: These clusters also have substantial subscribers and video views, suggesting a level of popularity but potentially less than Cluster 2.
- Clusters 0, 3, and 5: These clusters exhibit moderate engagement levels, with a balance between subscribers and video views. Channels in these clusters may have a steady and consistent audience.

Diverse Engagement Patterns:
The existence of multiple clusters implies diversity in engagement patterns among YouTube channels. Some may have high subscriber counts with relatively lower video views, while others might achieve high views despite a smaller subscriber base.
Content Strategies:
Channels in Cluster 2 might have effective content strategies that appeal to a broad audience, leading to both high subscribers and views. Channels in other clusters may have specific strategies, such as focusing on a niche audience, leveraging trends, or prioritizing quality over quantity. Channels in clusters with lower engagement may explore strategies to increase either their subscriber base or video views to enhance overall performance.

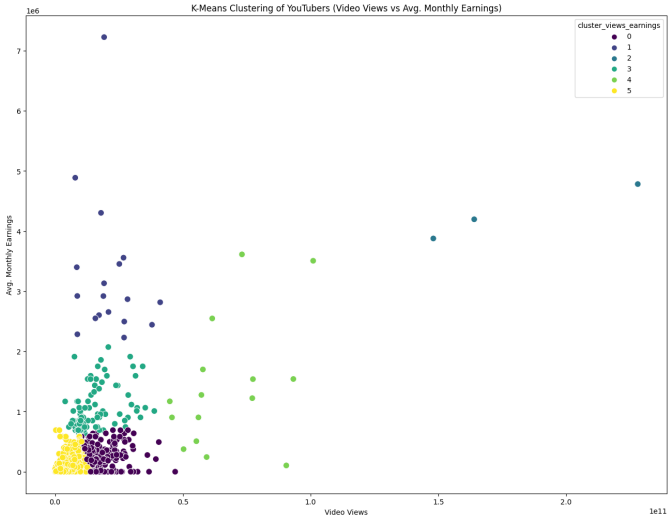

Fig. 26. K-Means clustering for Video Views vs Average Earning

The provided k-means clustering results for the "Video Views (bn.)" and "Avg. Monthly Earnings" columns reveal distinct characteristics of each cluster:

- Cluster Characteristics:
  a) Cluster 0: Moderate video views (around 17.79 billion) and relatively low average monthly earnings (around $466,956).
  b) Cluster 1: Higher video views (around 35.63 billion) and higher average monthly earnings (around $737,754).
  c) Cluster 2: Very high video views (180 billion) and the highest average monthly earnings (around $4,285,717).
  d) Cluster 3: Low video views (around 4.21 billion) and the lowest average monthly earnings (around $128,616).

e) Cluster 4: High video views (around 69.44 billion) and above-average average monthly earnings (around $1,797,789).
f) Cluster 5: Moderate video views (around 11.49 billion) and relatively low average monthly earnings (around $438,137).

- Interpretation:
  a) Cluster 2 stands out with very high video views and the highest average monthly earnings, indicating a group of highly successful and lucrative YouTube channels.
  b) Clusters 1 and 4 also have high video views, but Cluster 4 has higher average monthly earnings, suggesting potentially more effective monetization strategies or premium content.
  c) Clusters 0, 3, and 5 represent different tiers of channels with varying levels of video views and average monthly earnings.
- Segmentation Insights: This segmentation provides insights into different categories of YouTube channels based on their video views and revenue generation.
  a) Channels in Cluster 2 might be considered the most successful and lucrative, while Clusters 0, 3, and 5 represent different levels of success.
  b) The clustering results help identify patterns and trends in terms of video views and revenue, enabling a more targeted analysis of successful channels.

There are numerous factors determining the success of a YouTube channel. Some factors influencing a YouTube channel's success and key trends that can be drawn from this dataset are as follows:

- Subscriber Engagement and Niche Selection:
  1) Building a strong subscriber base is crucial for increasing views.
  2) Popular categories include Entertainment, Music, Gaming, and Comedy, with Education and How-to Style also performing well.
  3) Shows with loyal fans have higher views per subscriber, and Travel Events get more views, potentially due to viral content.
- Monetization Isn't Just About Views:
  1) Earnings are not solely tied to views; factors like engagement, video length, and ad placement matter.
  2) The "Show" category stands out for strong earning opportunities beyond the popularity of Music and Entertainment.
- Quality Over Quantity:
  1) Making more videos doesn't guarantee higher earnings; the focus should be on creating high-quality and relevant content.
  2) Channels with fewer videos but higher earnings may effectively leverage trends.
- Global Trends and Branding:
  1) YouTube is primarily dominated by the U.S. and India, but a diverse audience presents unique opportunities.
  2) Choosing a channel name aligned with content enhances discoverability and brand recall.
- Diversify Revenue Streams:
  1) Creators are encouraged to explore alternative revenue sources like merchandise sales, sponsorships, affiliate marketing, and fan donations.
  2) Successful YouTubers utilize multiple income streams, reducing dependence on ad revenue alone.

These takeaways collectively provide comprehensive insights for content creators seeking success on YouTube, emphasizing the importance of subscriber engagement, strategic content creation, and diversified revenue streams for sustainable growth.

TECHNOLOGIES USED

1) We use Python as the language for writing code in Jupyter Notebook to generate visualization
2) For libraries we use a mix of tools like numpy and pandas for data manipulation and plotly, matplolib, and seaborn to create visualization.
3) We mostly use Plotly as this is a pipeline-based approach where a user is expected to interact with the visualization and hover over to see information associated with the points and then write some more instructions on working with data and repeat (in a theoretical sense) while unrolling the loop. Plotly allows us to write less code to create interactive visualizations and hence the decision.

APPENDIX

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 27 and Uploads 28 fields. This is also easily verified by the correlation between these two (Corr(Uploads, Subscribers) = 0.077, Corr(Views, Subscribers) = 0.752)

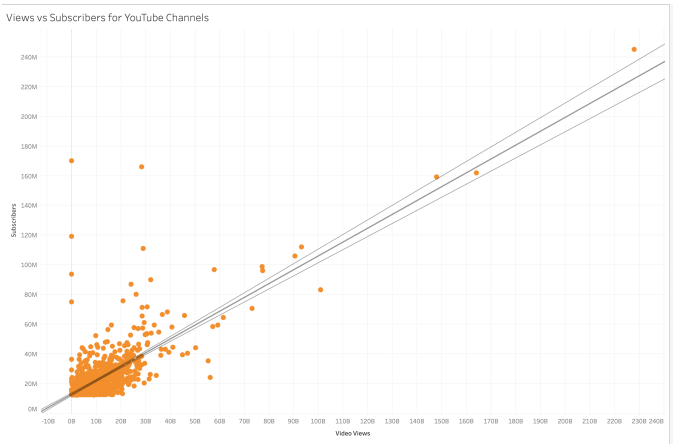

Fig. 27. Scatter Plot between Views and Subscribers

For testing out the correlation between Categories and Country, a Density Plot 29 and Cartograph 30 are plotted, respectively.

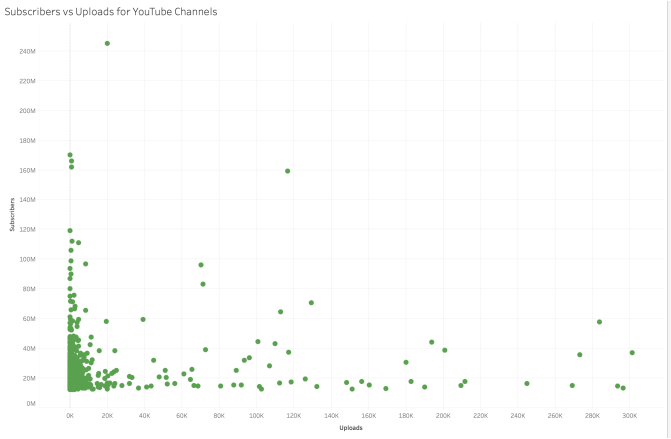

Fig. 28. Scatter Plot between Uploads and Subscribers

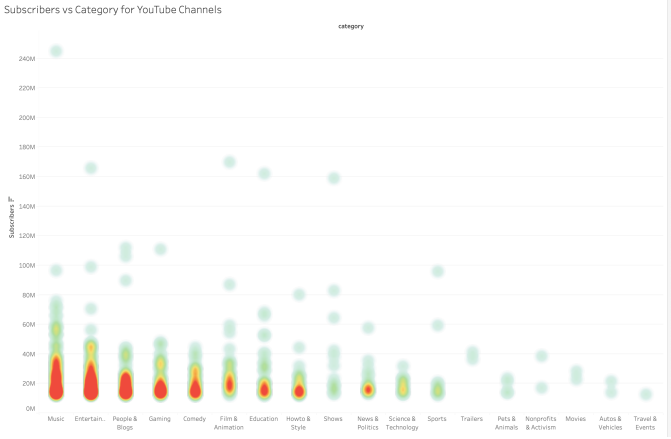

Fig. 29. Density Plot between Category and Subscribers

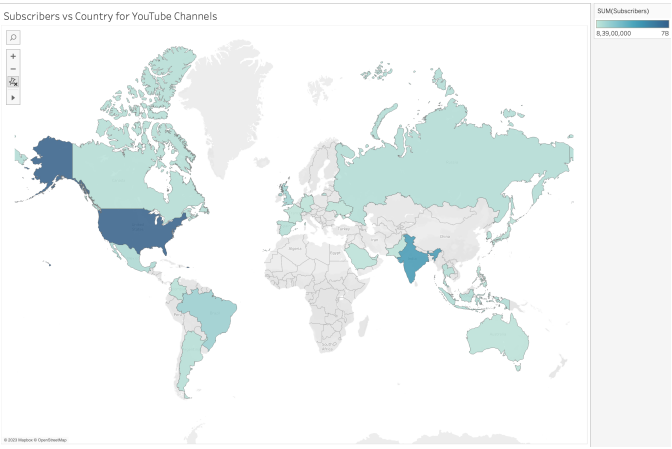

Fig. 30. Cartograph between Country and Subscribers

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 27 denotes the trend as the line fits to the data, along with its confidence interval.
- The colours for the Views 27 and Uploads 28 plot are chosen to look aesthetically pleasing.
- The colour map in the density plot 29 is such that red signifies a higher density of points. (The legend for the same is not available in Tableau).
- Sum of Subscribers is chosen as an attribute to plot in the Cartograph 30 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 31 and the Average Number of Views by the Created Year (included in the accompanying folder).

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 32. 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 (31, 32) 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.

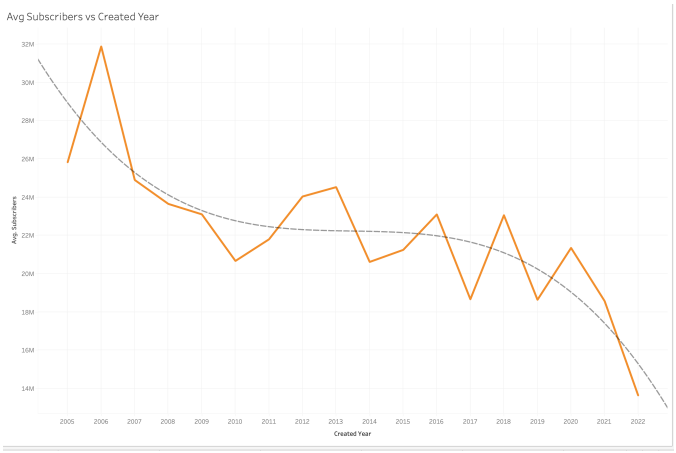

Fig. 31. Line Plot between Subscribers and Created Year

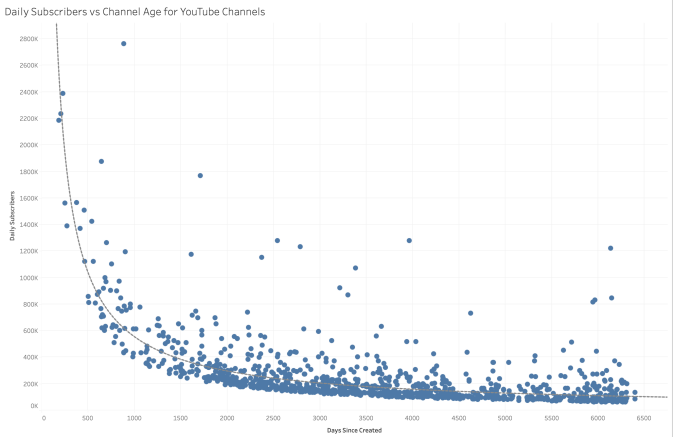

Fig. 32. Scatter Plot between Subscribers and Days Since Created

- The trends for these plots are chosen via trial and error from options like Linear, Polynomial, Logarithmic, Exponential and Power.

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 33. The same is done for Category Wise plots 34.

In this plot 33, 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.

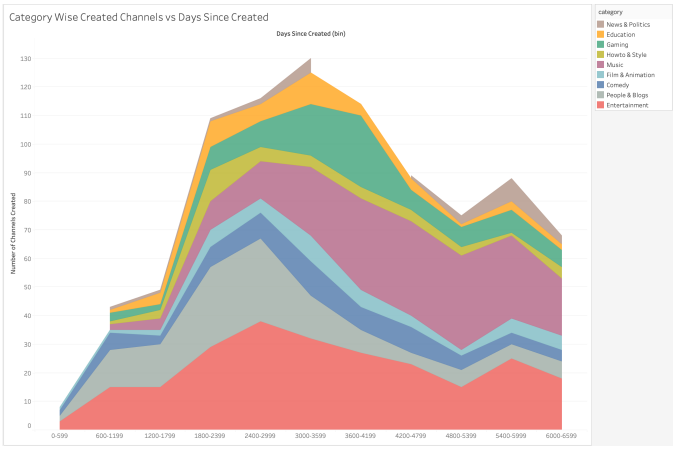

Fig. 33. Area Plot for Number of Channels created vs the Created Year, Category Wise

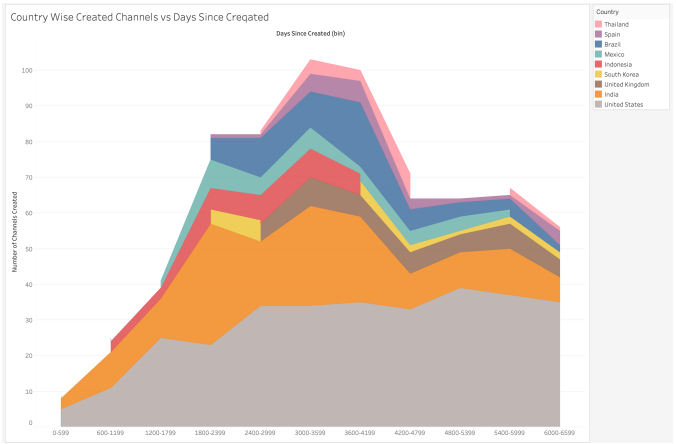

Fig. 34. Area Plot for Number of Channels created vs the Created Year, Country Wise

There are the following observations for 34

- 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.

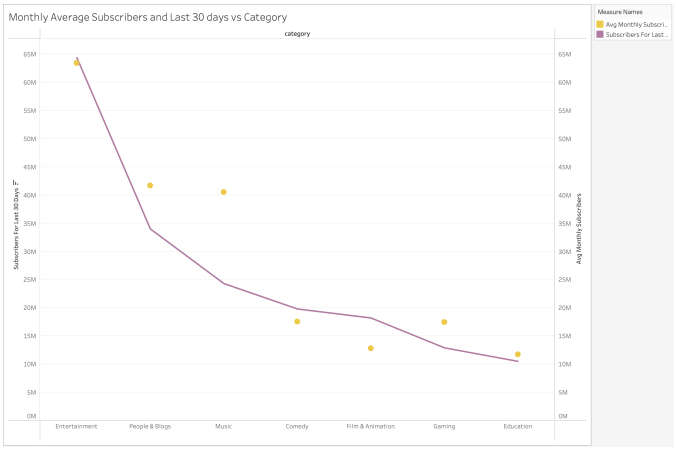

Fig. 35. Plot for Category Wise trends in Subscribers for the last 30 days

We can draw the following conclusions from the above plot 35:

- 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.

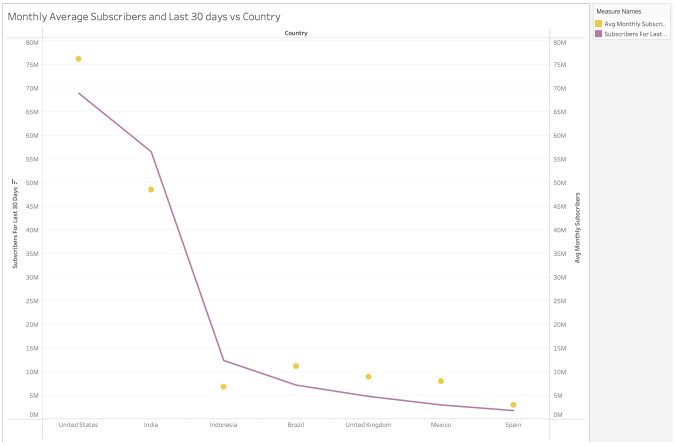

Fig. 36. Plot for Country Wise trends in Subscribers for the last 30 days

We can draw the following conclusions from the above plot 36:

- 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.

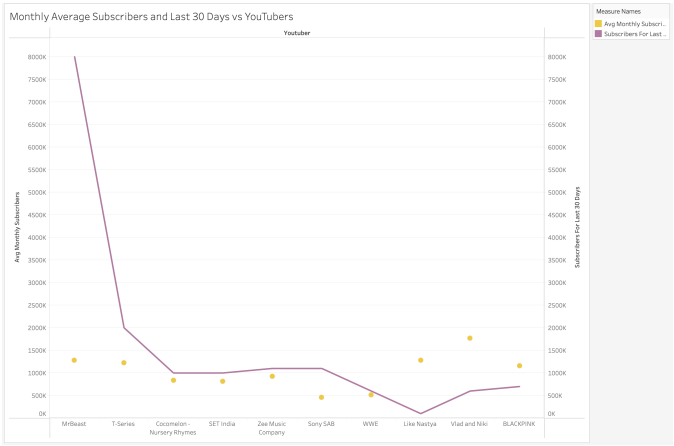

Fig. 37. Plot for YouTuber Wise trends in Subscribers for the last 30 days

We can draw the following conclusions from the above plot 37:

- 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 38, a Pie Chart on the division of views in these groups 39 and finally a Cartograph with the localised view-based Pie Charts for categories 40.

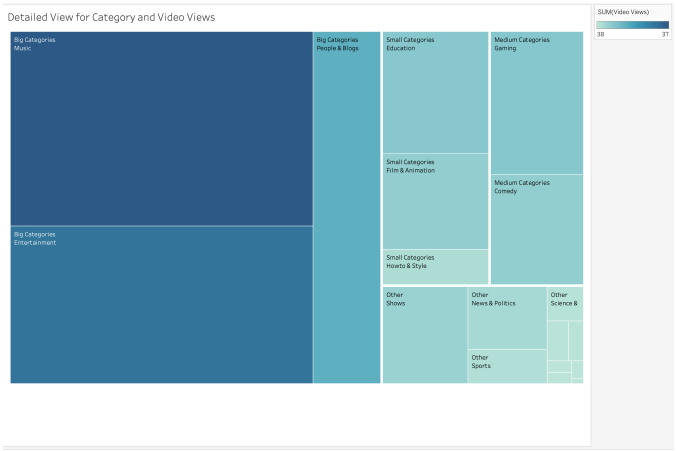

Fig. 38. Division of categories into groups by number of channels

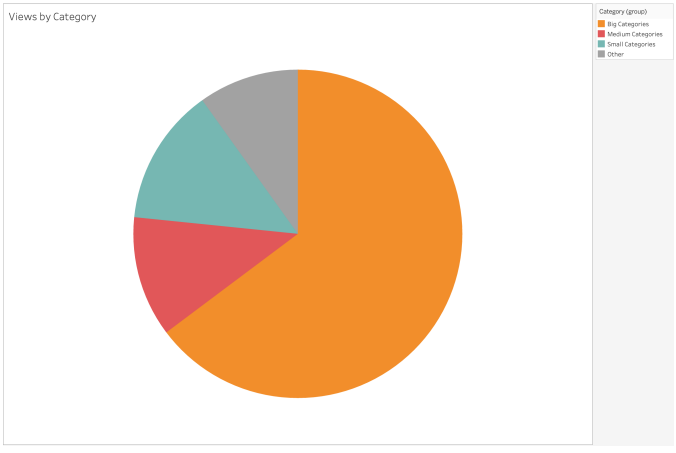

Fig. 39. Views by Category Groups

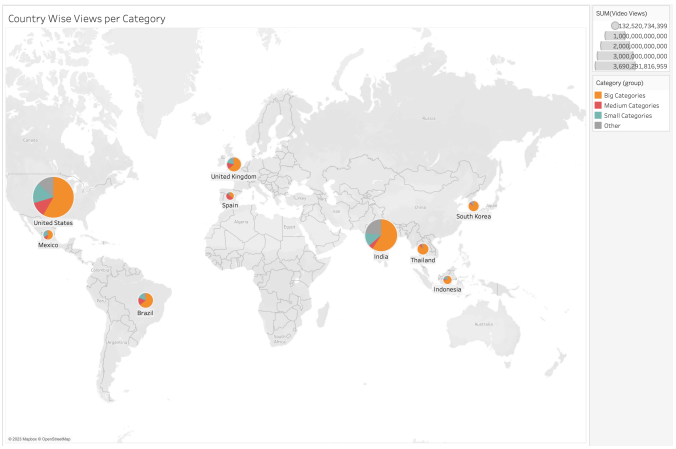

Fig. 40. Popularity of Category groups in different countries

With these plots, we are able to observe the popularity of different groups of categories 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 41, it is clear that the Entertainment category yields the highest income and the rankings of categories remain the same in all 4 graphs.

Fig. 41. Maximum income with respect to different Categories

Now, from the figure, 42 i.e., Maximum Income plotted for every YouTuber (Here top 15), again the rankings of

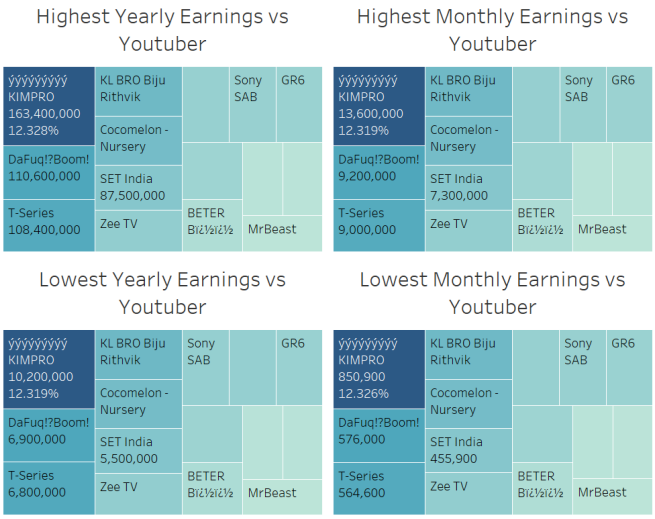

Fig. 42. Maximum Income vs Youtuber

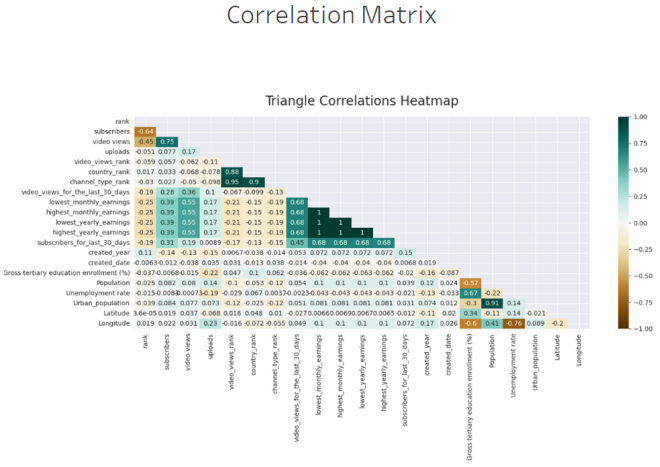

Fig. 44. Correlation Matrix

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.

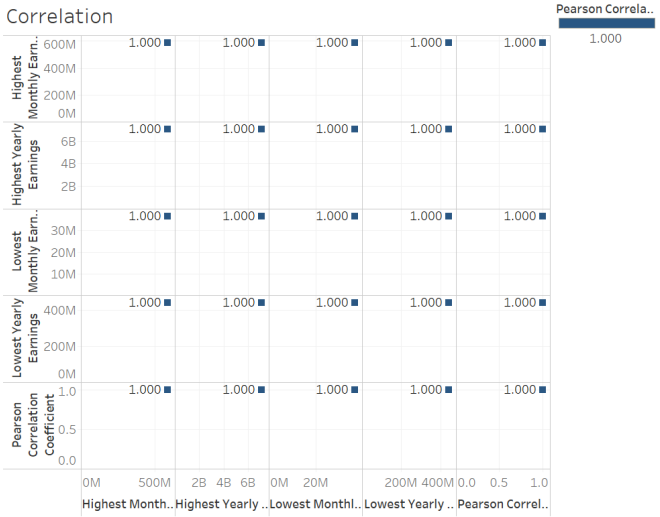

Fig. 43. Correlation between all Revenue related columns

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

The figure45, shows us the top Youtubers with their percent of income among the top 20 Youtubers.

And figure 46 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.

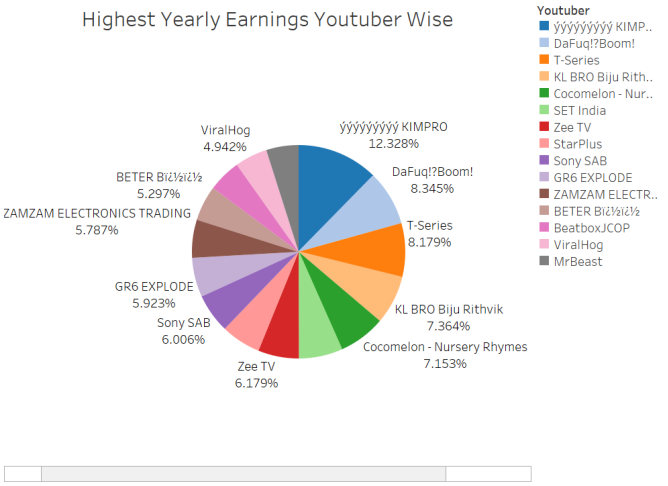

Fig. 45. Highest Yearly Earnings vs Youtuber

The below plot, figure 47 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 49) we can see that there is no particular pattern in data distribution. But when we plot Views against Average Income we get a linearly increasing plot with some outliers.

This plot (Fig 51) contains the highest earning Youtuber Country Wise.

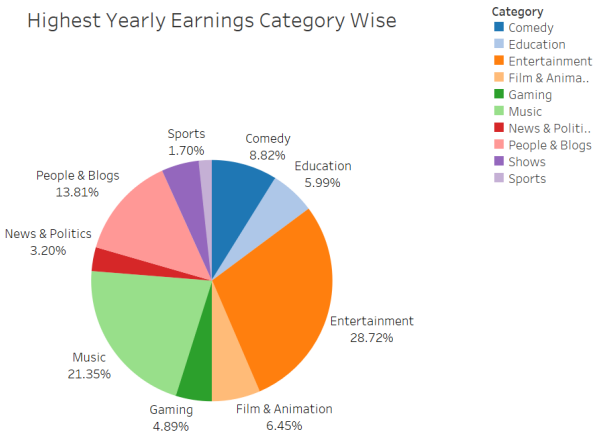

Fig. 46. Highest Yearly Earnings vs Category

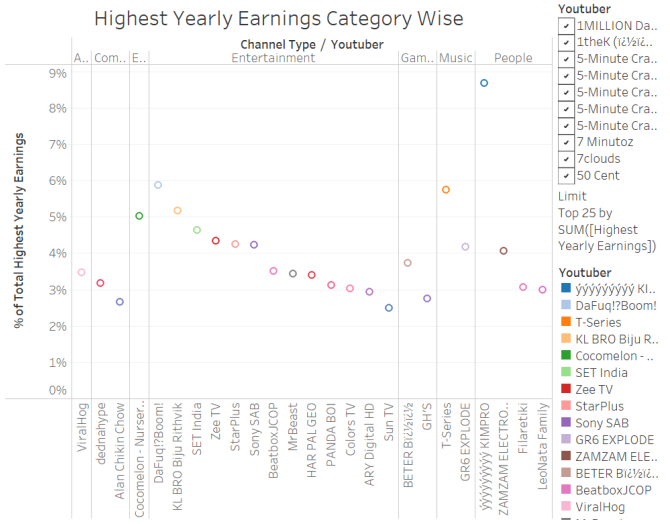

Fig. 47. Highest Yearly Earnings vs Category

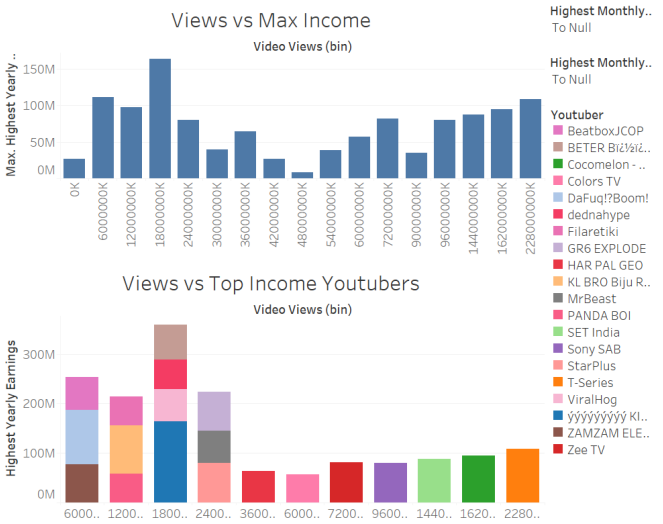

Fig. 48. Views plotted against Maximum Income and also for the Top 20 Youtubers with the highest revenue

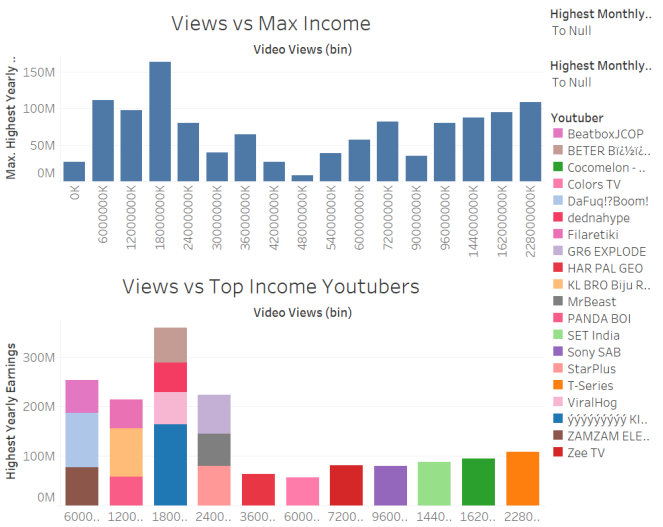

Fig. 49. Views plotted against Maximum Income and also for the Top 20 Youtubers with the highest revenue

The figure(Fig 52) 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 53,54, 55,56 and 57. 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 58) with a single color color-map that describes the

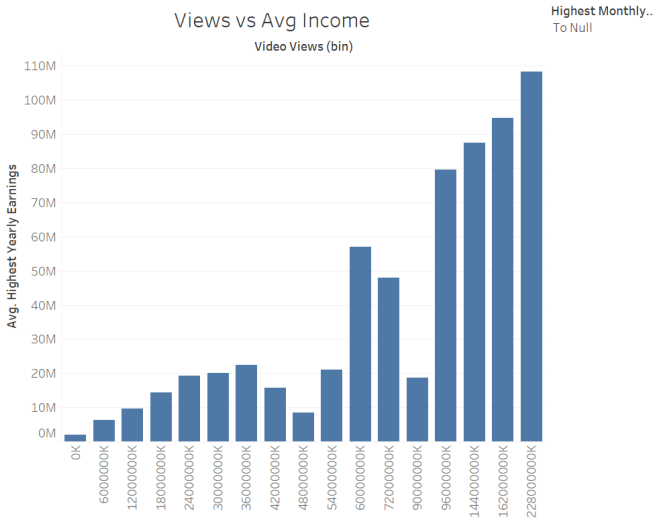

Fig. 50. Views vs Average Income

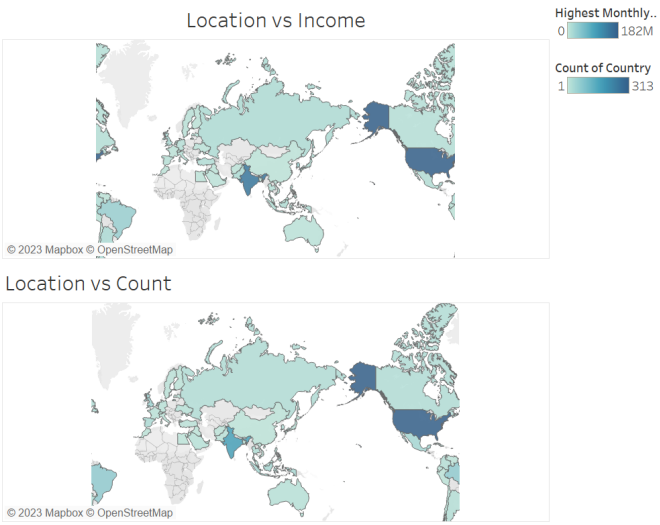

Fig. 52. Countrywise distribution of income and YouTube count

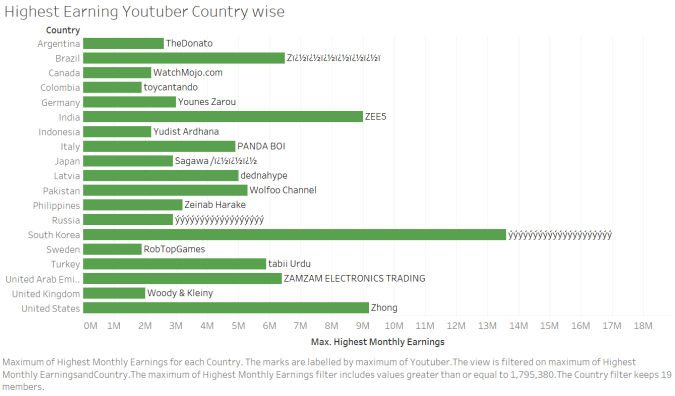

Fig. 51. Highest Earning YouTubers Country Wise

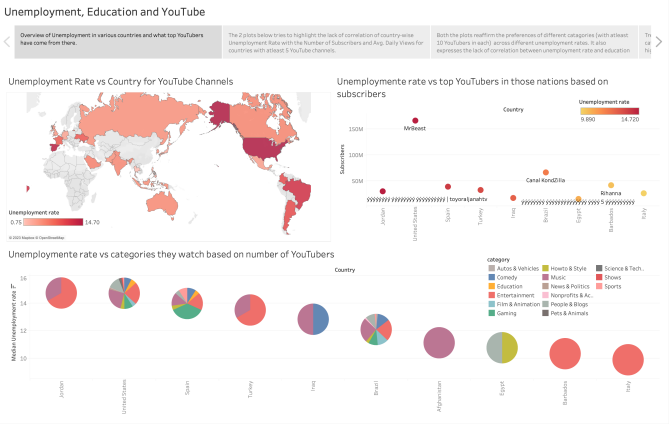

Fig. 53. Story 3 Page 1

Unemployment rate of the countries. From this, we can infer 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 59) 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.

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 60.

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 figure61.

We also keep a bubble chart (figure 62) 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

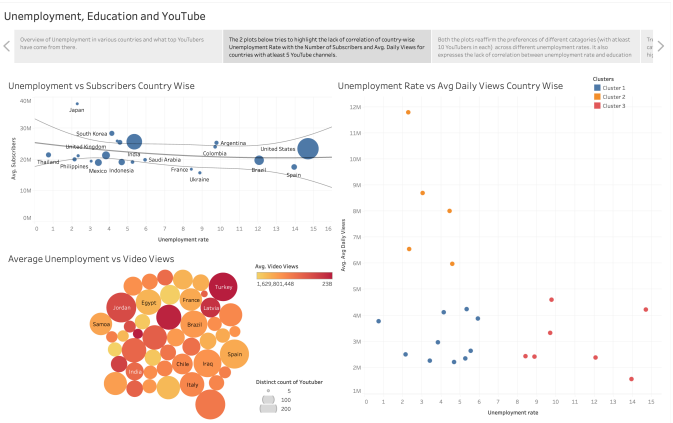

Fig. 54. Story 3 Page 2

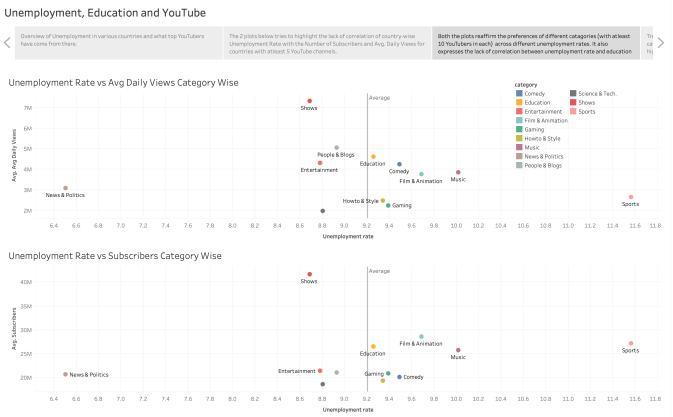

Fig. 55. Story 3 Page 3

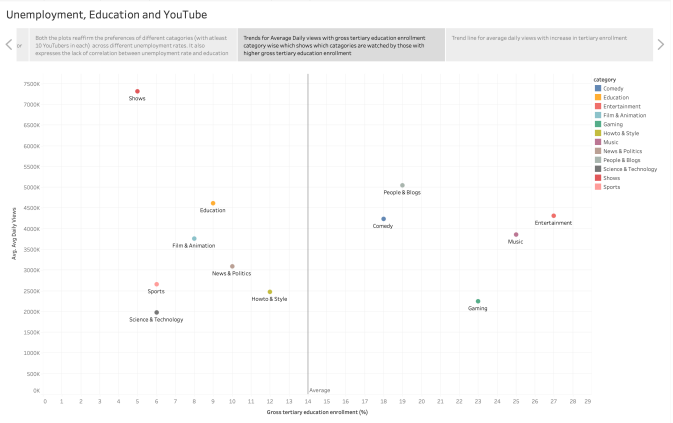

Fig. 56. Story 3 Page 4

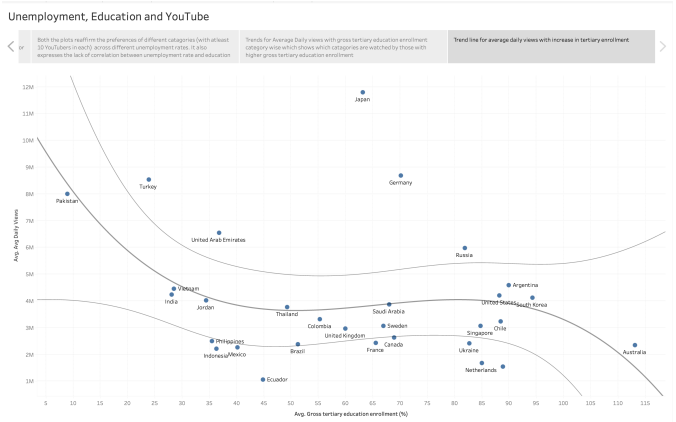

Fig. 57. Story 3 Page 5

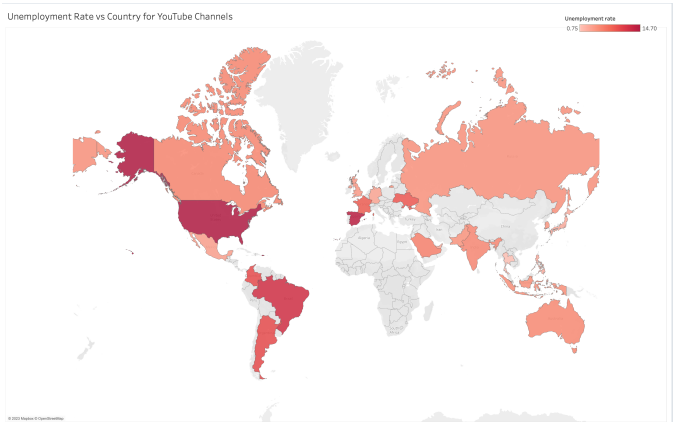

Fig. 58. Map Visualisation for unemployment in countries across the world

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 the smaller view count, we try to understand the composition of those views. We notice using figure 63 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 64 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

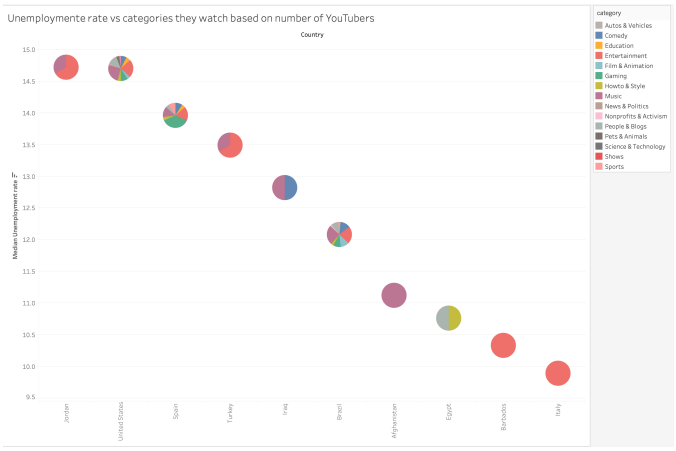

Fig. 59. Scatter plot with pie charts as markers and channel being the categories of video

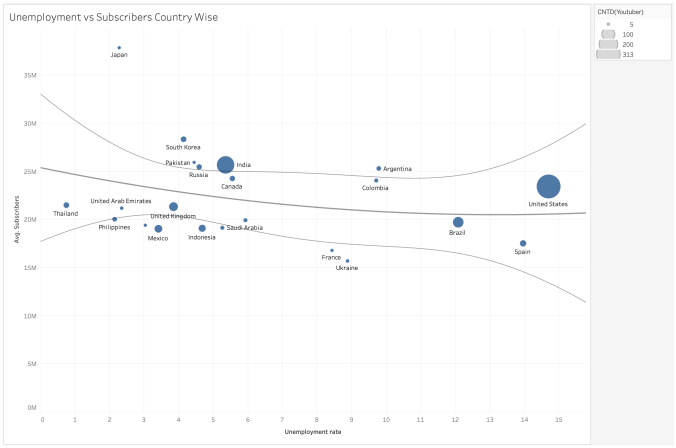

Fig. 60. Proving the lack of correlation between the unemployment rate and subscribers with the help of scatter-plot and confidence intervals

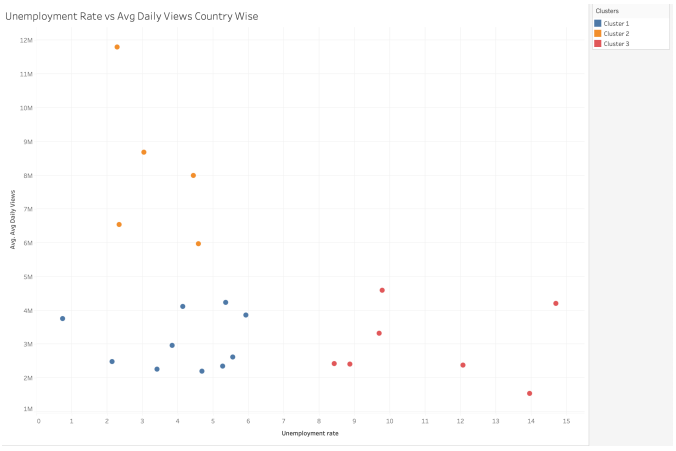

Fig. 61. Clustering plot that describes the low average daily views for countries with high unemployment rate

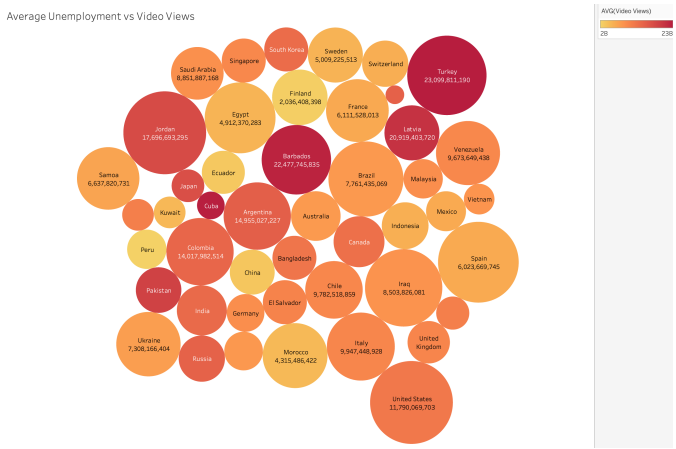

Fig. 62. Bubble chart with color channel highlighting unemployment rate and size channel highlighting average video views.

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 65). 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 visualizations more expressive for someone to get the maximum insights at the first glance.

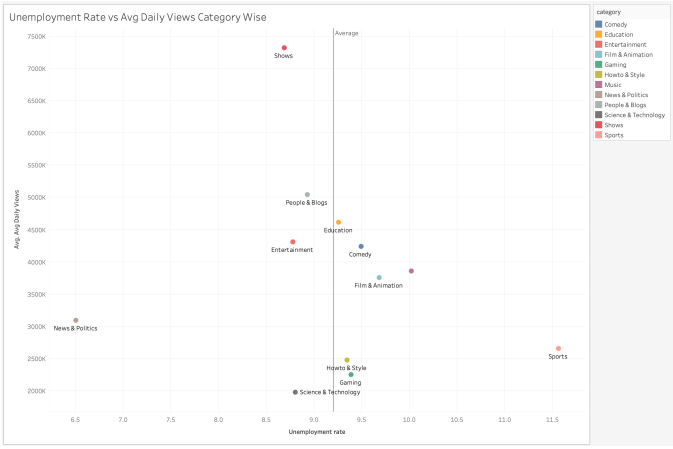

Fig. 63. Unemployment rate vs Daily Average Views with categories color coded

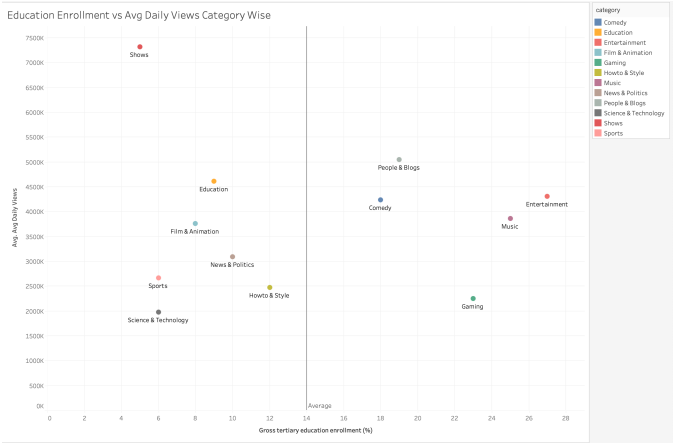

Fig. 64. Unemployment rate vs Daily Average Views with categories color coded

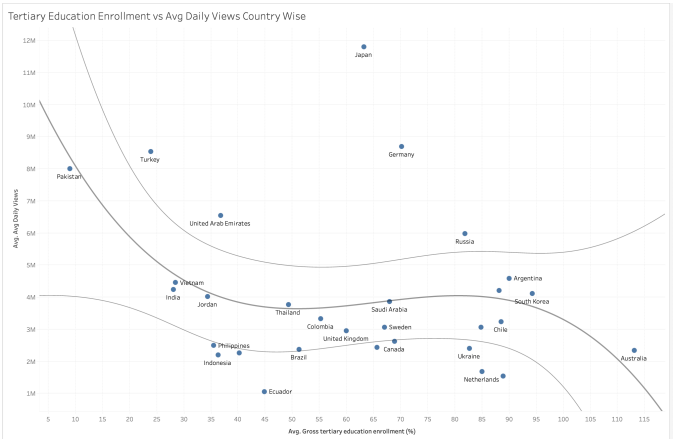

Fig. 65. Trends for tertiary enrollment and average daily views