Analyzing Global YouTube Statistics and Trends

AUTHORS (Youtube Youths):
Serra Choi
Karoline Klan Hansen
Kratika Rathi

Project Focus & Intended Audience

From this project, we hope to learn about some of the factors that may or may not be correlated with each other in the Youtube space. With the large number of creators and video uploads, this dataset can give interesting insights into Youtube video trends, metrics, user behavior, and more. This information can be beneficial for creators, businesses, and users. The main intended audience would be creators, advertising companies and brand makers. Understanding popular content and niches will uncover opportunities for effective advertising and content creation and thus, monetization.

Cohesion

The main goal of our dashboard is to help content creators and youtubers grow their channel and increase the revenue generated and for businesses to decide which channel type would be the most effective way to advertise their products / services.

Task 1 : Helps us see which channel category/type has the most subscribers, video views and thus earn the highest yearly income.

Task 2: Helps content creators see which countries' audience they should be targeting in order to increase their revenue.

Task 3: Helps us see how the highest yearly income, subscribers and channel types interact with each other, ie, do content creators need to have a very large following in order to earn maximum money and does this vary across channel types?

Task 4: How many new channels are being created in each category, i.e. how has the popularity of each category changed over the years, and how saturated is each space.

Task 5: How does the tertiary education enrollment, unemployment rate affect video views in each region.

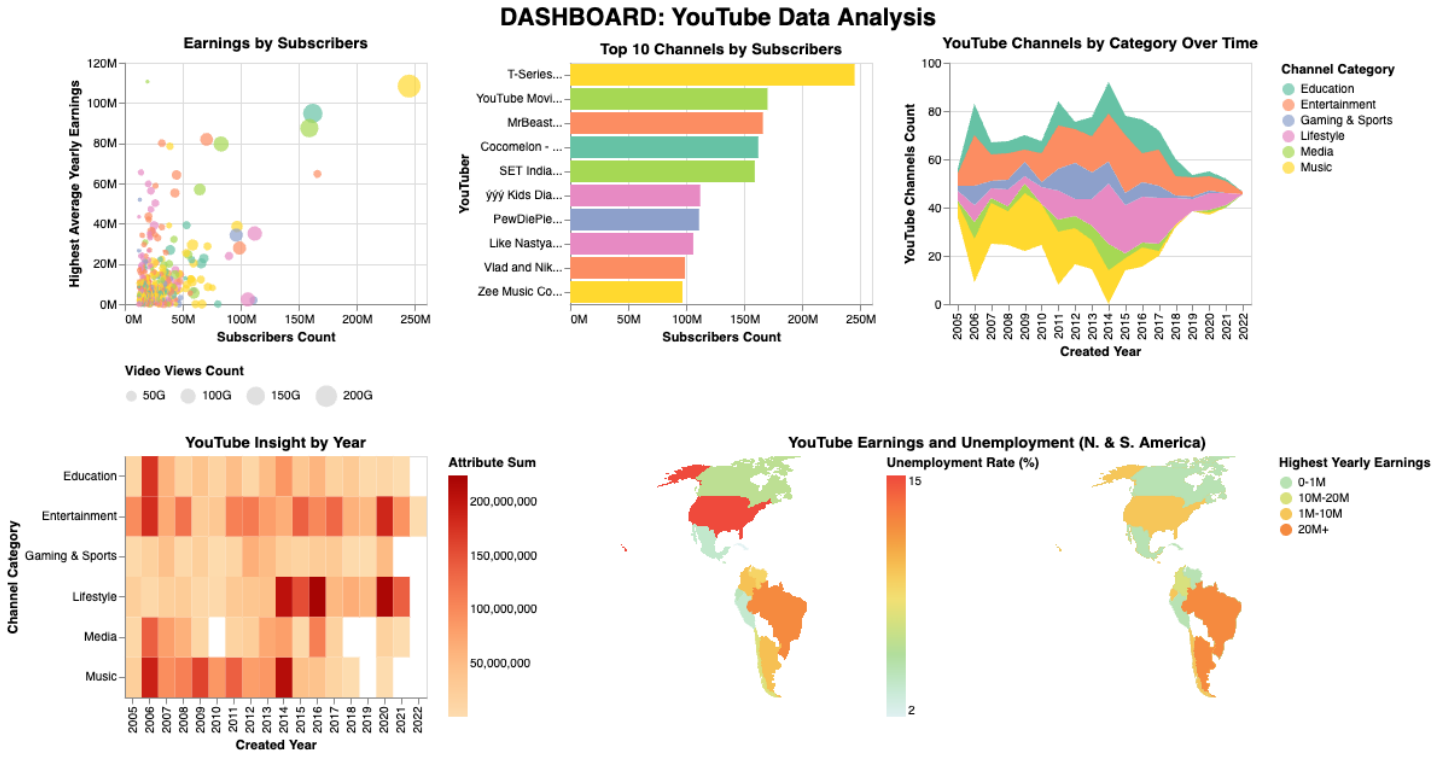
What audience can learn

Together, all 5 of our visualizations have helped us gain invaluable information that would be extremely beneficial to content creators and businesses alike. We have noticed that Lifestyle and Entertainment channel types earn the highest amount of money. Moreover, creating content for countries such as Brazil, which has a high rate of unemployment, could help maximize a channel's monetization.

Further, these channel types do not require a lot of followers to earn significant money, ie, a relatively new channel with ~200,000 subscribers would make a notable amount of money. From our fourth visualization we understand that there are less people entering the youtube space (perhaps due to the prevalence of tik tok and instagram). However, this is actually a blessing in disguise since the youtube space is not as saturated as Tik Tok's and there's more scope for growth and monetization.

Finally, the last visualization helps us to further hone our knowledge about the key demographics we should look for while attempting to create content for a particular audience, i.e. creating content for countries with high unemployment rate and low tertiary education enrollment.

Interactive Dashboard



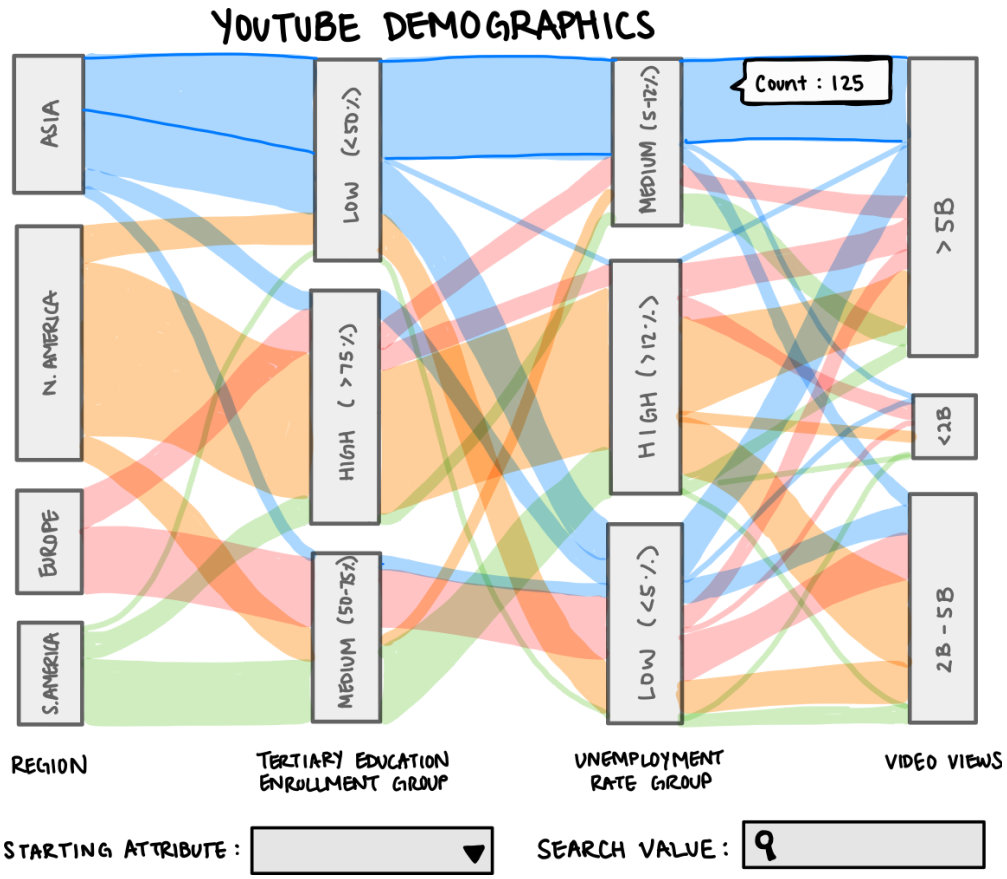
Interaction/ Linking

There is at least one level of interaction in each chart. Individually, we have:

- a dropdown selection for the heatmap where you can select which parameter you wish to judge a channel success on (ex video views, subscribers and highest yearly income) and see whether there is a correlation among them.
- bidirectional filtering selection between the scatter and streamgraph so you can select which channel type or income group you wish to focus on.
- unidirectional selection between the scatter, streamgraph and bar chart
- tooltips on all the charts.

As a dashboard, we ensured that the charts were able to interact with each other as well. As an example, clicking on a single data point on the scatter plot links to both the bar graph and the streamgraph. This highlights the relevant category on each of the charts so that we can gather insight on the particular channel category.

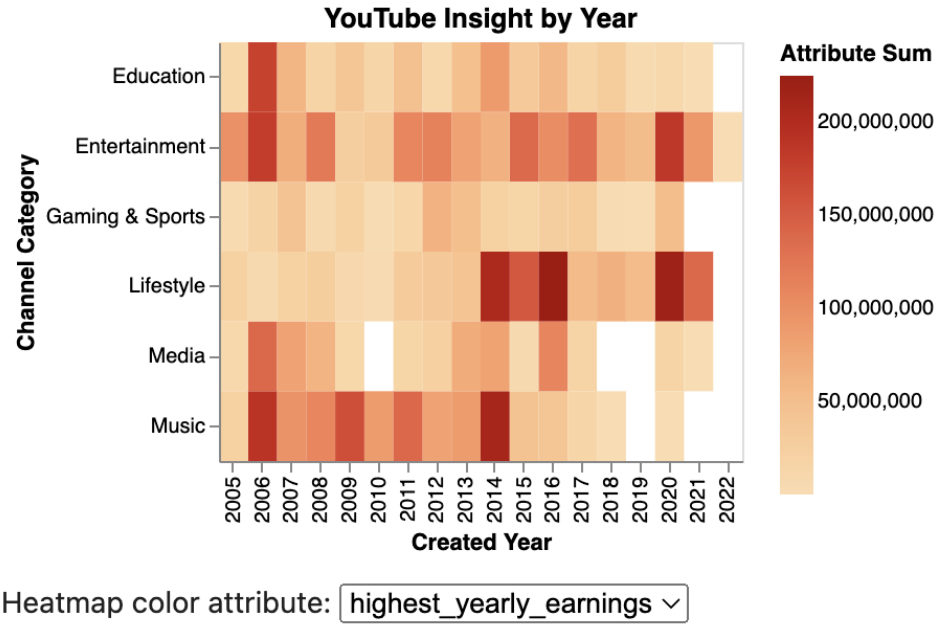
Novel/Hi Fidelity Viz



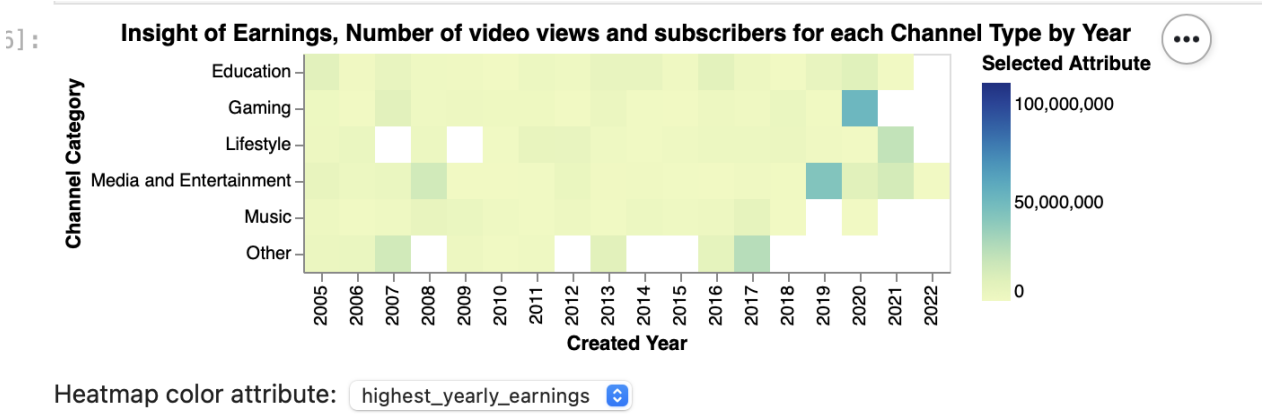
(More information on the hi fidelity viz is located in the Justifications section.)

Justifications

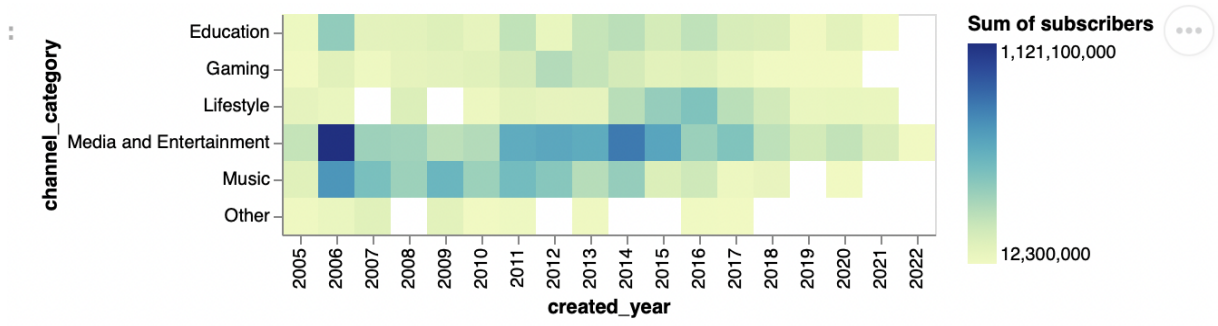
TASK 1: "What are some trends in income, subscribers and views among different YouTube channel types on a yearly basis?"



Previous iteration(s) of viz:



This was our first iteration, where we were encoding the highest yearly income instead of the sum of the highest yearly income. Looking at this heatmap, we clearly knew we were doing something wrong, and hence went back to the drawing board and tried to encode more meaningful attributes. Previous iterations of this visualization also included a line chart, however, we realized that a heatmap provided a more meaningful visualization and was easier to understand.



In our second iteration, while we managed to make our heatmap look better (in terms of color distribution), it still wasn't great. This made us go back to our initial data wrangling, and we changed our categories to create a better, more informative visualization.

Objective (Tasks Addressed):

This visualization would be beneficial to both content creators and businesses looking to advertise their products/services. For youtubers and influencers, this would help them ascertain the popularity of each channel category. Long-time youtubers would be motivated to make more targeted content (ex. Lifestyle or Entertainment) which attracts a wider audience and helps them grow their channel. Businesses would be able to see which categories attract more views and hence use those youtube channels to advertise.

Explanation/Description of Marks & Channels:

Our initial dataset had 14 categories. This would've cluttered the visualization. To solve this problem, we decided to bin the categories into 6 distinct categories. This made our heatmap (and rest of the dashboard) much easier to read and understand.

Since the color channel is encoding a quantitative variable, only the saturation and luminance channel is being changed, and not the color hue channel.The color scheme was changed from the default one provided by Altair, to make it look more cohesive with the rest of the dashboard.

Since we wanted to look at the trend of video views, highest yearly income and subscribers over the years, we chose a heatmap. Using a stacked bar graph or a line chart wouldn't be as effective, since the charts would be cluttered.

From trial and error, we realized that the color channel should encode the sum of an attribute, instead of just the attribute. This not only made more theoretical sense, but also improved the quality of our heatmap.

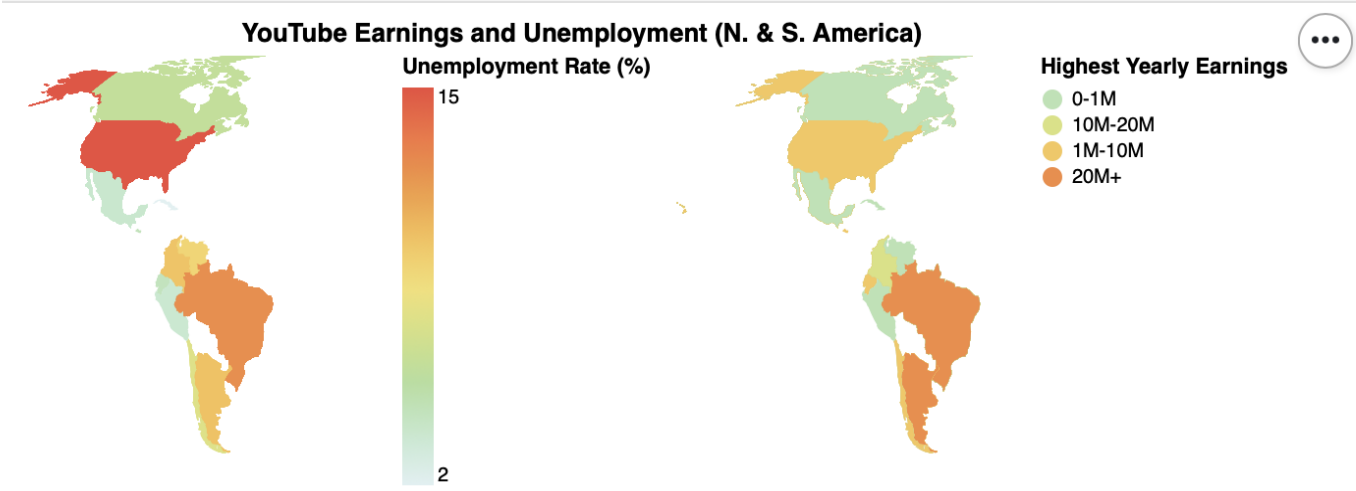
Critique:

Even though we tried our best to make the heatmap better, in terms of color distribution, it still isn't perfect. There are only 3 rectangles with the darkest color, i.e. only 3 rectangles that have the highest value of a given attribute. There are a few ways to fix this issue, however, all of them are beyond the scope of this project. A prominent fix would be collecting more data (including more information

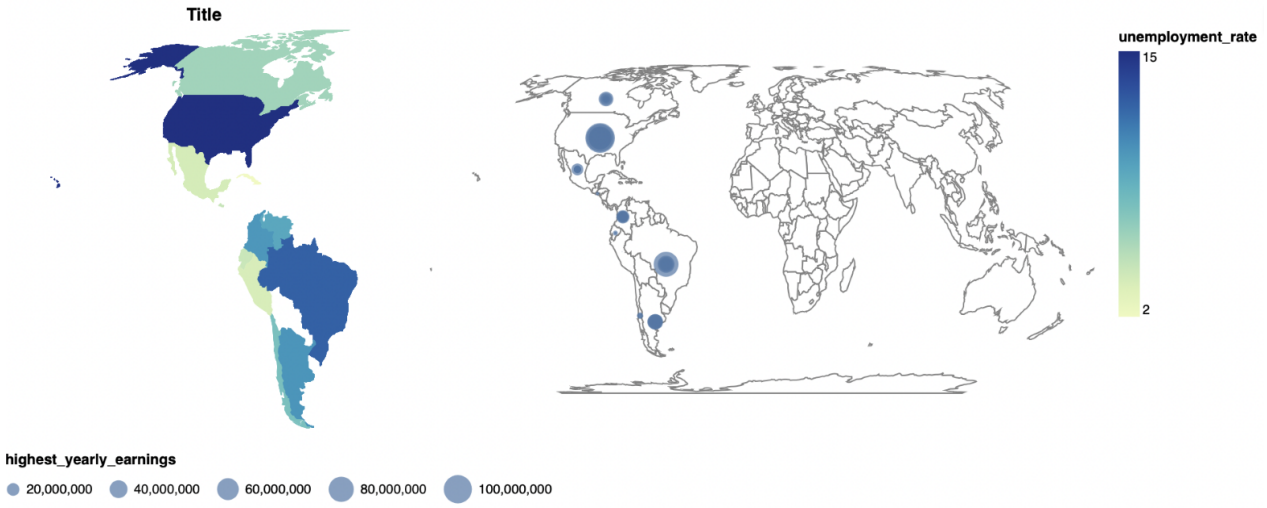
about channels) in our dataset. Another way would be to adjust the scale at the price of excluding some values at the bottom of the scale.
Another major issue we see is the legend title. We were unable to display the name of the selected attribute on the legend. Working with interactions in such a capacity was beyond the scope of our technical abilities.

Interaction(s):
We decided to add a UI dropdown button so that users could understand trends from 3 different standpoints.We chose to have this interaction since users could have different parameters for the success of a channel. Some may consider subscribers to be an accurate parameter, others may think it is the video views. Hence, we have provided options so that the users can make well-informed decisions.
This interaction removes the need for having 3 different static visualizations, and helps us make our dashboard more concise.
Finally, we also included a tool tip to increase readability and understanding of the visualization.

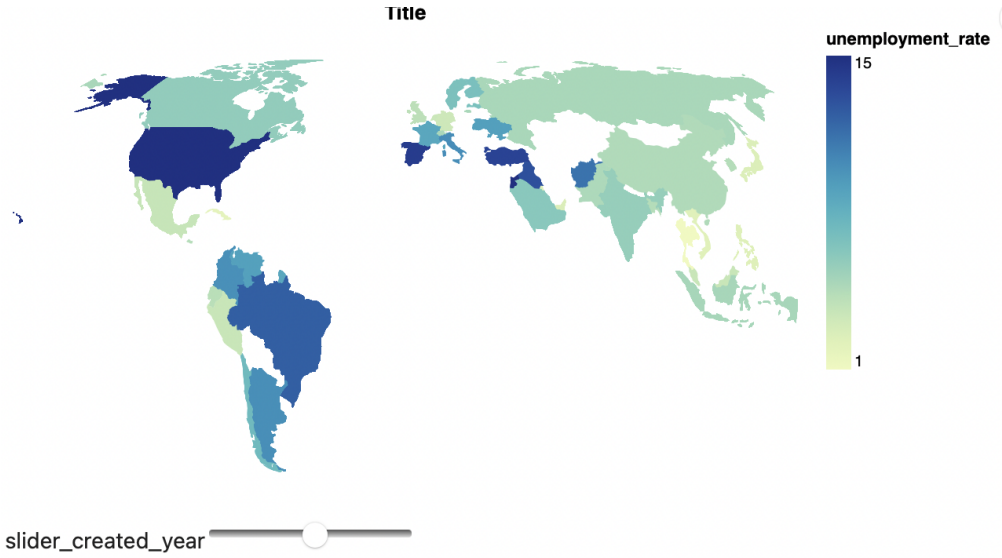
TASK 2: "How does the Highest Yearly Income of Youtube Channels correlate with the Unemployment Rate of the Country they were created in?"



Previous Iterations:



This was our initial draft of task number 2. We rejected this idea because size is not an effective channel to encode population or any other quantitative value. Because of this, we decided to move forward with different channels. Moreover, we weren't sure whether to look at all the countries in our dataset or focus on a select few.



In our second draft, we chose to encode our attributes through color. Moreover, we decided to include a slider in our visualization. However, this was later scrapped, as our data did not support a slider (no meaningful information was being encoded). Later, we also changed the color scheme to make our dashboard seem more cohesive.

Objective (Tasks Addressed):

From our first task, we explored the channel categories that get the most views and subscribers. Now, we want to check whether the country you're based in, or for which country you primarily make content for, has any impact on the total highest yearly income for a channel.

We restricted our analysis to only 2 regions (N & S America) as we had the most data for these countries (significantly more than Asia or Africa or any other region).

Comparing the unemployment rate between countries to check for youtube trends may appear strange, but our analysis has revealed some interesting details. Youtube channels made in countries with higher unemployment rates tend to earn a higher amount of money than channels created in countries with lower rates. Many reasons may contribute to this discrepancy. One valid reason, which is supported by our novel visualization, is that people in countries with higher unemployment rate spend more time watching youtube videos. Content creators and youtubers could benefit from this knowledge by creating content that is more directed towards people from these countries. Higher video views would mean greater monetization and channel growth.

Explanation/Description of Marks & Channels:

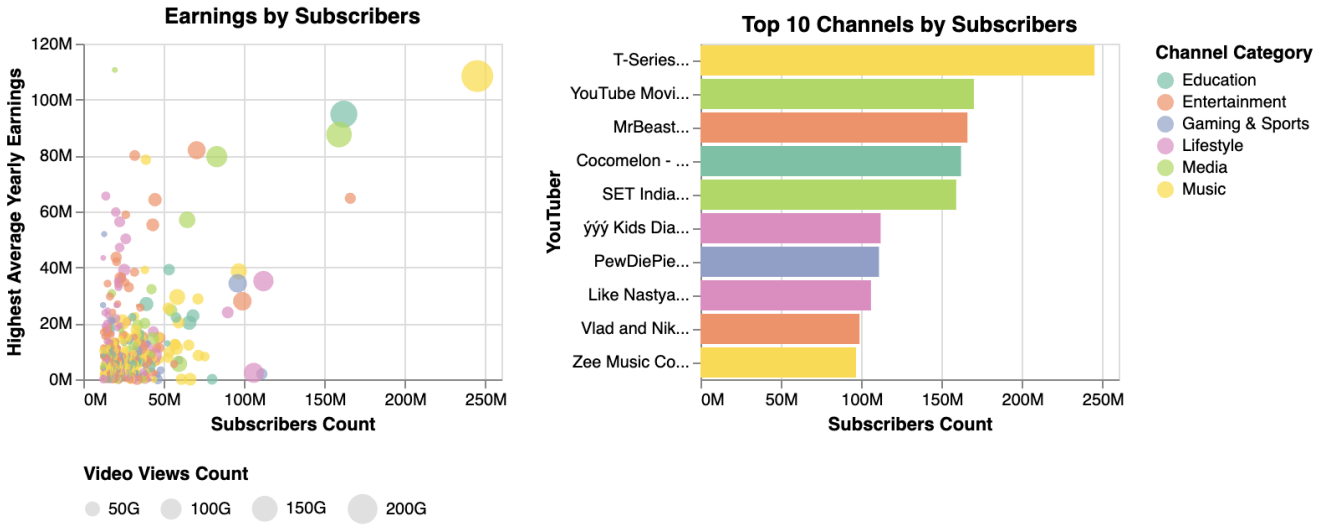
Here, we experimented with a geospatial visualization. Since we wanted to look at youtube trends between different countries, plotting relevant information on a map with those countries would lead to a simple yet effective visualization.

From our previous iterations you can see that we experimented including a size channel in our visualization. However, we learnt in class that size is not a very effective channel to represent data. This is why we switched to a color channel for both our maps. In our first map, as the color channel is representing a quantitative variable (unemployment rate), we chose a continuous scale. On the other hand, to improve understanding, we binned a quantitative variable (highest yearly income) and made the color channel in our second visualization ordinal. This was done so that the 4 distinct categories of income would pop out and quickly help us compare the two maps.
The color scheme was changed from the default Altair scheme to better fit with the rest of the dashboard.

Critique:
The most significant drawback of our visualization is that it is limited to only 2 continents. An effective solution to this problem would be to collect more data about a country's youtube channels and unemployment rate. This would help provide an overall picture and help creators make well-informed business decisions about their youtube channels.
Another drawback is the number of bins we were able to create for the highest yearly income. There is a struggle between making things simple and making them informative. While having more bins would be more informative, it would overburden the color channel and make our visualization messy.

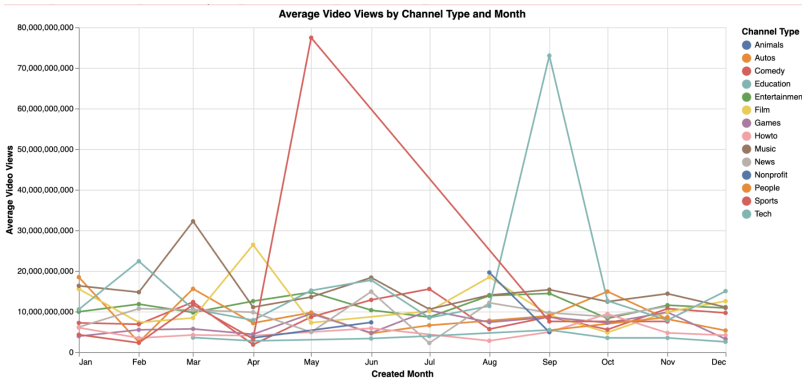
Interaction(s):
We initially decided to include a slider to shift between years, however, our data did not support such a functionality. That is a significant drawback of our selected dataset, and we did not realize this until much later. However, fixing this would have been beyond the scope of this project.
For easy user experience, we decided to add a tooltip so anyone could hover a country and learn about its unemployment rate and highest yearly income a youtube channel makes.

TASK 3: "How does the number of subscribers and video views correlate with the channel's earnings, and what are the top 10 YouTube channels within each category based on subscribers?



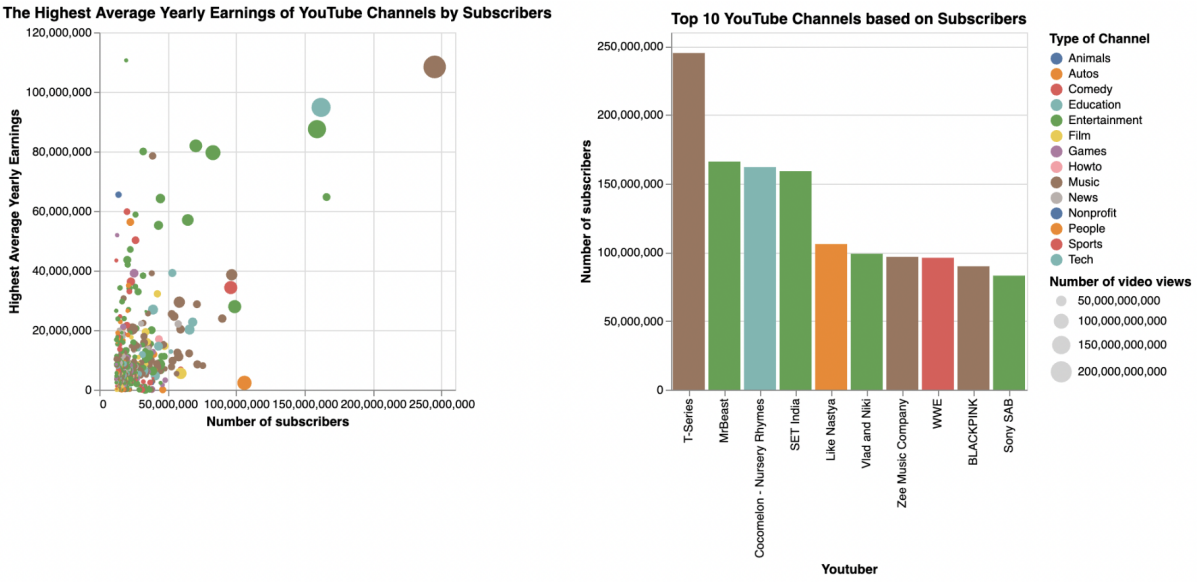
Previous iteration(s) of viz:

The very first iteration of task 3 that we had in Milestone 1 was totally scrapped, and the task question as well as the visualization was totally revamped. The initial task abstraction was: "Is there a trend between average video views by Channel Type over the months?", and the visualization was:



This visualization had a lot of issues in general but we will only mention a few, given that it was completely scrapped from the project. First of all the color channel is not effective when dealing with more than 5 categories. The choice of lines as the mark was based on the idea of visualizing the trend of the categories over the months, but it was not effective because all the lines were on top of each other and was really hard for the user to interpret.

Revamping the task completely we landed on the previously mentioned: "How does the number of subscribers and video views correlate with the channel's earnings, and what are the top 10 YouTube channels within each category based on subscribers?"



Comparing the previous iteration with the final visualization we have edited the axis and size-channels measurements on both charts to be of SI-unit to save space. Additionally the previous version of the barchart had the youtubers on the x-axis, which is not very effective as the user would have to tilt their head in order to read the names, so this was encoded on the y-channel instead thus flipping around the axis. The length of the names of the Youtube content creators varied a lot, making the axis title move around when using the selection, therefore the length of allowed characters in the name was limited to 12.
As previously mentioned the number of categories was scaled from 14 to 6.
Finally the legend for the size channel was moved to be in the bottom instead of the right to make it fit in the dashboard.

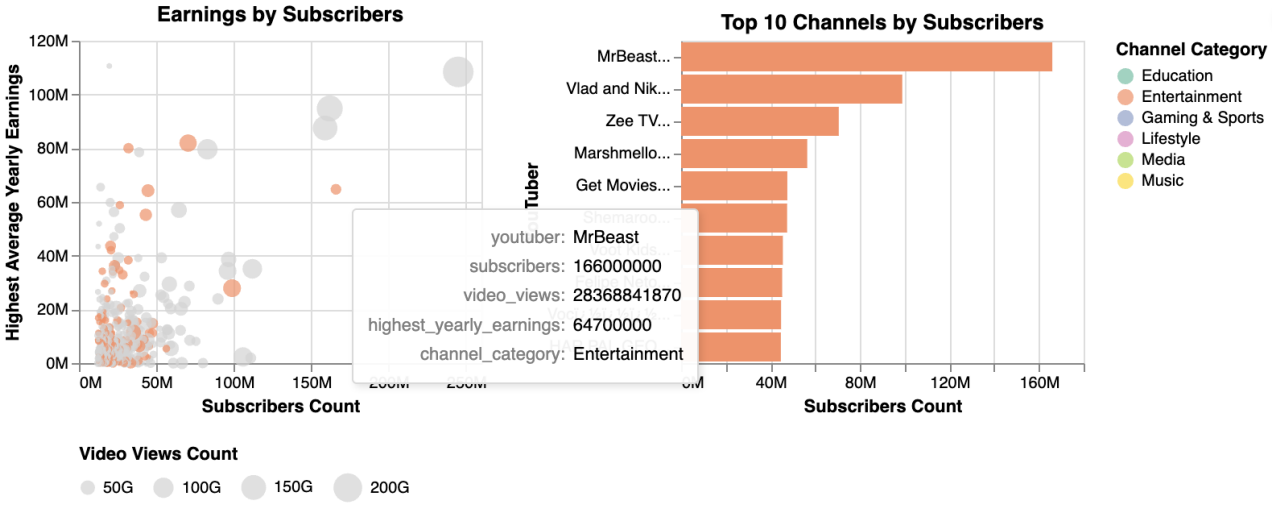
Objective (Tasks Addressed):
Investigate the relationship between the number of subscribers and video views on a YouTube channel and its earnings, while simultaneously identifying the top 10 channels within each category based on subscriber count, aiming to see patterns in viewership and revenue across the different content categories.

Explanation/Description of Marks & Channels:
This visualization addresses the task by using two views consisting of a scatter chart and a barchart. The scatter chart uses the circle mark for each YouTube channel, with the number of subscribers on the horizontal channel, and highest average yearly income on the vertical channel. The size channel encodes the number of video views and the color channels are encoding the channel categories. The Bar Chart represents the top 10 YouTube channels on the vertical channel based on the number of subscribers encoded on the horizontal channel using the bars mark. The channel categories are encoded using the color channel as in the scatter chart.

The choice of a scatter chart for this task enables an understanding of the relationships between subscribers, highest average yearly earnings, video views, and channel categories in a single view. The bar chart specifically focuses on the top 10 channels based on the number of subscribers. This provides a clear overview of the most popular youtubers, allowing for quick identification and analysis of the highest-performing ones which enhances the effectiveness of the visualization as you focus on a subset of the data instead of having to consider all at once.
The bar chart provides clear separability in terms of subscriber count and this type of viz is good when doing comparisons as it is easy to interpret and retrieve direct values from. Channels within the same category of both visualizations share a common color which contributes to the perceptual and visual grouping of related data points.

Critique:
Even though the scatter chart demonstrates separability, in most of the graph, the points are somewhat dense and overlapping when looking in the range of 0M-50M subscribers. This can make it challenging to visually distinguish individual data points and may reduce the precision with which users can analyze and interpret the data. Additionally when points are densely packed the effectiveness of interactive elements such as tooltips or selection tools may be compromised.
Some of the Youtube channels had very long names which we chose to shorten to be no more than 12 characters in order to keep the size of the charts. This could potentially affect the accuracy of the information because users of the viz will not have access to the full information if not using the tooltip.
In general the size channel can be a challenging choice to visualize a quantitative variable since they are hard to compare and is also really influenced by the distance of the attribute items.
Usually the color-channel is sufficient for a maximum of 5 different colors when considering effectiveness. In the initial data wrangling the channel category attribute was grouped from including 14 distinct categories, to capture only 6, in order to justify the use of the color-channel in terms of discriminability. The use of a limited set of channel categories enhances the interpretability of the visualization. Even though a smaller number of categories would enhance the effectiveness, it's important to ensure that the chosen categories effectively capture the diversity of YouTube channels categories, wherefore we couldn’t go lower than 6 categories.

Interaction(s):



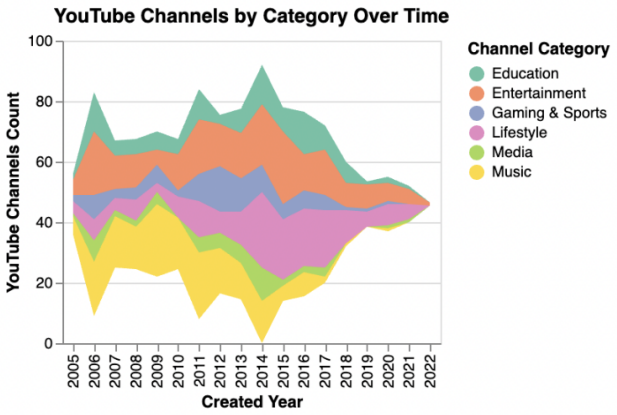
For the visualization we added a tooltip that enables detailed information when hovering over data points, providing intel on the data in this case the channel name, number of subscribers, video views, and highest average yearly earnings. This can be justified because it enhances user engagement by providing context and makes it easier to interpret the cluttered data points.

An interactive zoom element was also added for the user to be able to investigate cluttered data points by zooming in.

We have added a unidirectional action that uses selection by click. This allows for a channel category selection, creating a filter effect on both the scatter plot and histogram. This provides a tailored view based on user interest, which provides a deeper exploration of particular channel types, and answers questions like: “Within the Educational youTube channels, which one has the highest number of subscribers?”.

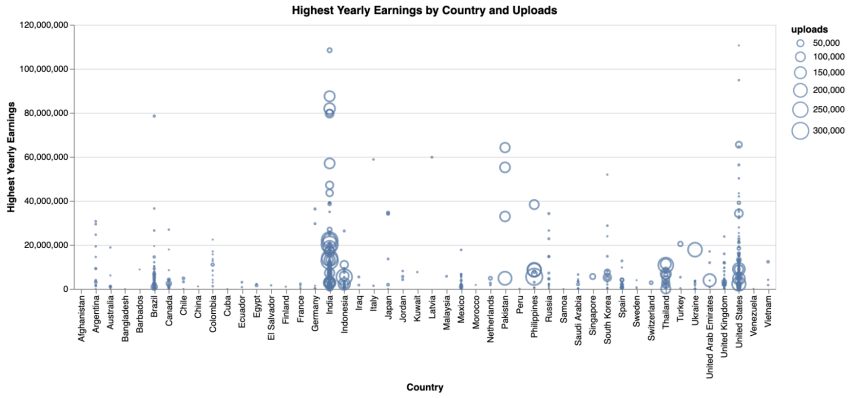
A static version might display all channels at once, making it challenging to focus on specific categories. This creates an indirect focus because clicking on a channel category in the scatterplot filters the histogram, allowing users to focus on specific channel types. The action is implicit because the user does not know it's there before they play around with the viz. An example of a way to make the action explicit is to adjust the stroke width of points when you hover over it, so it becomes clear that it is possible to click on and select. The action is also discrete as the filtering selection happens in one move.

TASK 4: "Are there distinct patterns/trends in creation of YouTube channels within each category over time?"

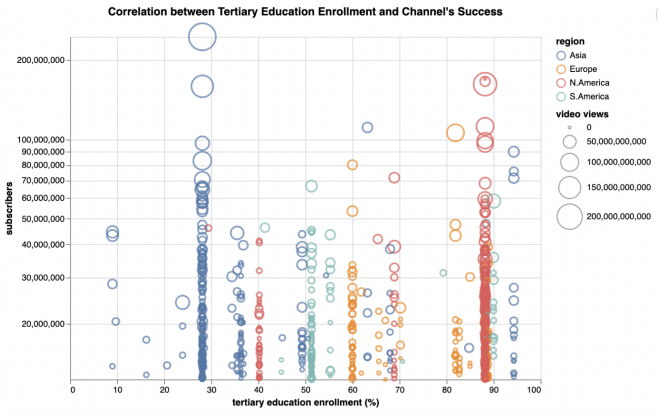


Previous interaction(s) of viz:

Like in the previous iterations mentioned in the section about task three, the visualization used for task 4 in Milestone 1 was also revamped. Initially the task abstraction was: "Is there a cluster of highest yearly income a youtube channel makes by country and number of uploads?" and the visualization was:



The initial chart shows how the earnings of content creators vary by country and number of uploads. On the horizontal axis, the different countries were encoded, and on the vertical axis the highest yearly earnings. The idea with the graph was to investigate clusters in the data, but having a categorical variable on the horizontal channel made the points clutter a lot on top of each other.



We incorporated regions on the color-channel and tried scaling the y-axis to declutter some of the points, as well as using tertiary education enrollment of each country on the x-axis instead of country as a nominal variable. However, we eventually scrapped it because of the overall project scope where our focus is to address content creators and businesses looking to advertise their products/services. We determined that until about the amount of youtubers, would be of more interest than tertiary education enrollment's effect on subscribers, as there didn't seem to be any clear distinct correlation and therefore no effective conclusions from that examination. Therefore we decided to move forward with the streamgraph.

Objective (Tasks Addressed in Final Viz):

To analyze and understand how the number of YouTube channels varies over time within each content category to gain insights of growth or decline of created channels within specific categories on the platform.

Explanation/Description of Marks & Channels:

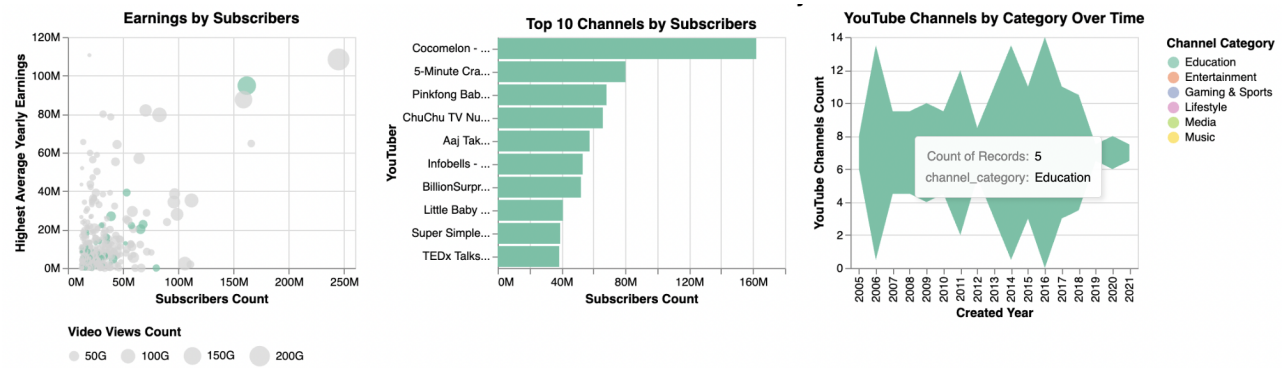
The chart is a streamgraph that explores the temporal evolution of created YouTube channels by category. On the horizontal channel the created year is encoded and on the vertical channel the number of YouTube channels created that year is encoded. The mark used to visualize the count is the area, and the color channel is used to encode the channel category. Encoding year on the horizontal axis as an ordinal variable makes it easy to interpret and follow the change and patterns over time. The use of a stacked area chart provides a cumulative representation of YouTube channels, showing both the total count and the contribution of each category to the overall count.

Critique:

With the compact width of 250 and height of 200, the visualization is designed to fit into the final dashboard, however the x-axis is pretty dense and could be stretched out more to enhance readability by the users. You have to be very delicate with the tooltip to explore exact years. The information showcased in this type of graph could be visualized in many other ways as well. A heatmap could have captured a lot of the same information, but you wouldn't have had an overview of the contribution of each category to the overall count. A grouped or stacked bar chart could also

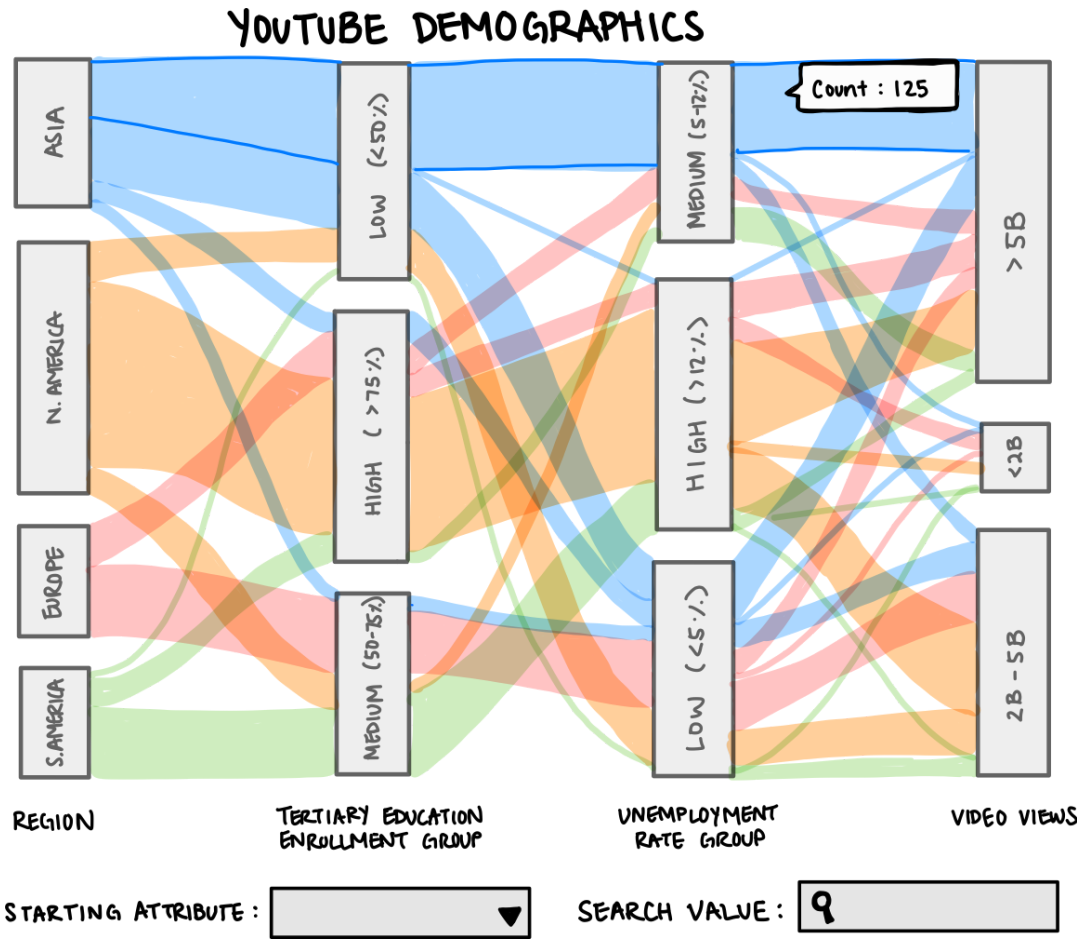
have displayed the information in an effective manner by representing each created year as a group/stack, with channel category encoded on the color channel and count on the vertical axis, allowing for a clear comparison of counts within each year. But eventually the streamgraph was chosen based on the aesthetics and cohesion with the rest of the dashboard as it is more visually pleasing than a barchart. With our target audience in mind we emphasized aesthetics over functionality for this task.

Interaction(s):



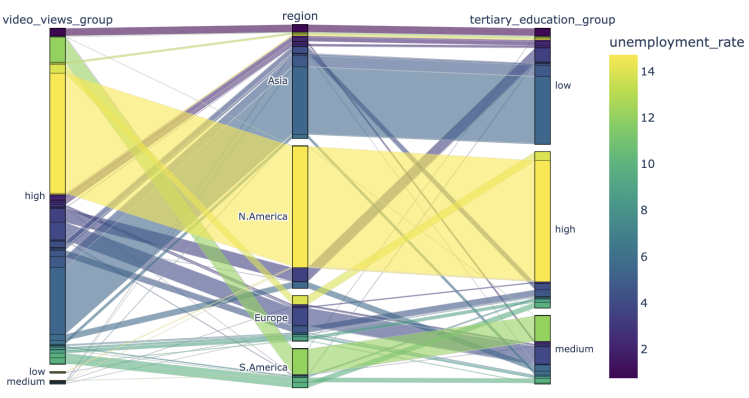
A tool-tip shows the count of records within specific years of the specific category when hovering over it, which enhances the user's ability to explore and interpret the chart. There is a bidirectional link between multiple views of the data including the ones specific for task 3, which enables filtering between all three. The bidirectional interaction is facilitated through a click or tap event type, allowing users to select a channel category in one discrete move. This makes the dashboard more cohesive, as interacting with one visualization makes the filtering appear in the other visualizations as well.

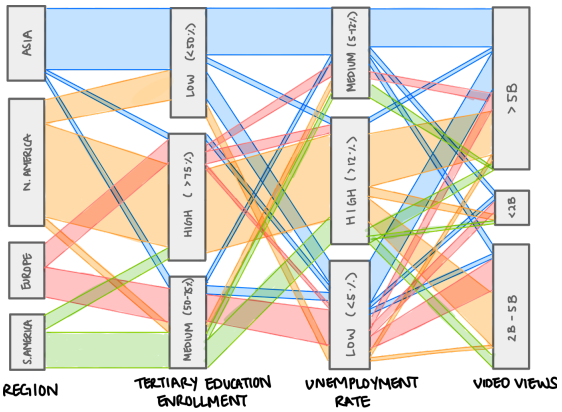
TASK 5: "Are there any patterns or correlations between YouTube demographics and video views?"



Previous iteration(s) of viz:

Our previous interactions were lacking titles, interactions, and aesthetics which we have addressed in the final sketch.





Objective (Tasks Addressed):

With this hi fidelity vis we aim to seek potential correlation between several attributes relevant to the YouTube statistics. We can explore potential links between educational demographics and video popularity, which would provide insights into content preferences and engagement patterns. If there

is such correlation, we would be able to see what the optimal “combination” of values are, for the purpose of targeting different regions, audiences, demographics, etc.

We chose this visualization as we were able to include several attributes that might be of interest to our target audience (businesses, content creators, etc). In terms of cohesiveness with the rest of our project and other tasks/charts, we find that this gives a good overview of the core question we aim to answer with our project: what are some insights that might be of interest/ help to a business advertise on YouTube? How do they interact?

The vis gives insights on characteristics of four broad regions and the number of video views, as well as what might be influencing them or correlates to them. This vis is not limited to the four attributes that we selected, and could also be moved around to display different column orders.

Explanation/Description of Marks & Channels:
The parallel sets chart uses area as the primary mark. Color is used as one of the channels to show different regions - however the exact attribute it encodes can be changed via interaction which will be explained further below.

Critique:
One disadvantage to using a parallel set chart is that it can become messy depending on the chosen attributes, and how they correlate (or do not correlate). For example, the right side section looks a lot messier than the left side in our current display. Without using interaction, the static version of the chart might be a little hard to interpret when looking at it as a whole.

On the other hand, if we want to get specific correlation information about a region, selecting (via interaction) that region makes it possible to highlight just the relevant values. This can be helpful when gathering insight and thus is the main goal of our sketch.

Interaction(s):
A selection tool which helps us choose which column we want to display first (i.e. on the left)
Tooltip which will show the count within a highlight area horizontally across, for a specific sequence of attribute values. The search menu will help us search within the attributes. For example, searching ‘Canada’ should highlight the section for which data points have country = Canada.
These interactions make it possible to answer more detailed questions a user might have, that a static version may not be able to address. A smaller scale business that only wishes to focus on North American YouTube insights may care more about those countries belonging within this region. A search bar makes it possible so that users can key in specific countries etc. An unidirectional link from this viz should enable the selection/filtering to happen in the dashboard as well, and highlight the selected variables in multiple views to gather insights of a number of various tasks that this interactive element enables.

Reflection

Strengths:

1. Project Topic → the topic itself is interesting and applicable to not only businesses but content creators or consumers. Most people know what Youtube is, and post or use the platform to some extent.
2. Coding Skills → This project showcases strong coding skills by effectively employing Altair, demonstrating a somewhat well-structured implementation of data visualization principles.
3. Learning → This project has been a valuable learning experience, providing opportunities to enhance skills in data analysis, visualization, and collaboration.

Weaknesses:

1. Chosen Dataset → There were a lot of issues with our dataset, that we realized after we had already decided to base our project off of it. For example, according to this dataset, France was in South America (based on the latitude and longitude provided)
2. Efficiency vs simplicity - in many parts of this project we struggled with making things simpler versus making them informative. We wanted to create a clear picture in the users' mind without overburdening them with unnecessary information.

Things we would have done differently:

1. Planning/Strategy → Rather than separately taking on tasks and viz, start together first so that the overall project is cohesive etc. (milestone 1)
2. Communication → We could have improved interpersonal communication throughout the project by establishing regular check-ins, and create a more collaborative atmosphere with a more constructive and positive dialogue.

Work Distribution

HIGH LEVEL OVERVIEW (numbers are in %)				
Project milestone	**Kratika**	**Serra**	**Karoline**	
Milestone 1	33	33	33	
Milestone 2	33	33	33	
BREAKDOWN				
Milestone Description				
M1: Project Scope	33	33	33	
M1: Visualization Ideas	33	33	33	
M1 coding	33	33	33	
M1 writing	33	33	33	
M2: Cohesion Dashboard coding		60	40	
M2: Cohesion Dashboard writing	100			
M2: Tasks 1-5:				
Task 1 coding & writing	75		25	
Task 2 coding & writing	75	25		
Task 3 coding & writing			100	
Task 4 coding & writing	33	33	33	
Task 5 drawing & writing		100		
M2: Reflections	33	33	33	