DSCI 320 PROJECT
MILESTONE 1

Proposal, Data and Task Abstraction and Initial Sketches

Serra C.
Karoline H.
Kratika R.

DSCI 320: Project Milestone 1

PART I: Initial Exploration

Data Abstraction

The dataset we chose for this project includes 25+ attributes. From this, we chose a subset of 13 relevant attributes for our initial exploration. These are listed in the table below:

data attribute	semantics	data attribute type	cardinality	notes
subscribers	number of subscribers to a channel	quantitative	Min: 12300000, Max: 245000000	
video views	total views across all videos on channel	quantitative	Min: 0, Max: 2.28e+11	
category	category/niche of channel	nominal	18	
uploads	total number of videos uploaded on channel	quantitative	Min: 0, Max: 301308	
country	country where YouTube channel originates	nominal	48	
channel type	type of YouTube channel (e.g., individual, brand)	nominal	14	
lowest yearly earnings	lowest estimated yearly earnings from channel	quantitative	Min: 0, Max: 6900000	
highest yearly earnings	highest estimated yearly earnings from channel	quantitative	Min: 0, Max: 110600000	
created year	year when YouTube channel was created	temporal	Min: 1970, Max: 2022	The minimum value of this attribute is 1970, but YouTube as a platform was released in 2005, so this is not a valid datapoint.
created month	month YouTube channel was created	temporal	12	this attribute could be considered both ordinal or temporal
gross tertiary education (%)	percentage of population enrolled in tertiary education	quantitative	Min: 7.6, Max: 113.1	It should be noted that this attribute is a %, but max value exceeds 100%
population	total population of country	quantitative	Min: 2.025060e+05, Max: 1.397715e+09	
unemployment rate	unemployment rate in the country	quantitative	Min: 0.75, Max: 14.72	

DSCI 320: Project Milestone 1

Introduction:

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 amount of creators and video uploads, this dataset can give interesting insights into Youtube video trends, metrics, user behaviour, 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 thus, monetization.

EDA:

We chose to work with a subset of the data, and filtered our dataset accordingly.

We removed all null-values as the size of the data-set stays sufficient even after doing so. In the original dataset there were 28 attributes and 995 observations. After filtering for the desired attributes and removing all null-values we have 13 attributes and 824 observations

```
print(data.shape)
print(data.info())
```

```
(995, 28)
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 995 entries, 0 to 994
Data columns (total 28 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   rank                                  995 non-null    int64
 1   Youtuber                              995 non-null    object
 2   subscribers                           995 non-null    int64
 3   video views                           995 non-null    float64
 4   category                              949 non-null    object
 5   Title                                 995 non-null    object
 6   uploads                               995 non-null    int64
 7   Country                               873 non-null    object
 8   Abbreviation                          873 non-null    object
 9   channel_type                          965 non-null    object
 10  video_views_rank                      994 non-null    float64
 11  country_rank                          879 non-null    float64
 12  channel_type_rank                     962 non-null    float64
 13  video_views_for_the_last_30_days      939 non-null    float64
 14  lowest_monthly_earnings               995 non-null    float64
 15  highest_monthly_earnings              995 non-null    float64
 16  lowest_yearly_earnings                995 non-null    float64
 17  highest_yearly_earnings               995 non-null    float64
 18  subscribers_for_last_30_days          658 non-null    float64
 19  created_year                          990 non-null    float64
 20  created_month                         990 non-null    object
 21  created_date                          990 non-null    float64
 22  Gross tertiary education enrollment (%)  872 non-null  float64
 23  Population                            872 non-null    float64
 24  Unemployment rate                     872 non-null    float64
 25  Urban_population                      872 non-null    float64
 26  Latitude                              872 non-null    float64
 27  Longitude                             872 non-null    float64
dtypes: float64(18), int64(3), object(7)
memory usage: 217.8+ KB
None
```

```
data.isnull().sum()
```

```
rank                                    0
Youtuber                                0
subscribers                             0
video views                             0
category                               46
Title                                   0
uploads                                 0
Country                               122
Abbreviation                          122
channel_type                           30
video_views_rank                        1
country_rank                          116
channel_type_rank                      33
video_views_for_the_last_30_days       56
lowest_monthly_earnings                 0
highest_monthly_earnings                0
lowest_yearly_earnings                  0
highest_yearly_earnings                 0
subscribers_for_last_30_days          337
created_year                            5
created_month                           5
created_date                            5
Gross tertiary education enrollment (%)  123
Population                            123
Unemployment rate                     123
Urban_population                      123
Latitude                              123
Longitude                             123
dtype: int64
```

DSCI 320: Project Milestone 1

```
clean_data = filtered_data.dropna()
print(clean_data.shape)
print(clean_data.info())
```

```
(824, 13)
<class 'pandas.core.frame.DataFrame'>
Index: 824 entries, 0 to 994
Data columns (total 13 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   subscribers                           824 non-null    int64
 1   video views                           824 non-null    float64
 2   category                              824 non-null    object
 3   uploads                               824 non-null    int64
 4   Country                               824 non-null    object
 5   channel_type                          824 non-null    object
 6   lowest_yearly_earnings                824 non-null    float64
 7   highest_yearly_earnings               824 non-null    float64
 8   created_year                          824 non-null    float64
 9   created_month                         824 non-null    object
 10  Gross tertiary education enrollment (%)  824 non-null  float64
 11  Population                            824 non-null    float64
 12  Unemployment rate                     824 non-null    float64
dtypes: float64(7), int64(2), object(4)
memory usage: 90.1+ KB
None
```

Cardinality:

	subscribers	video views	category	uploads	Country	channel_type	lowest_yearly_earnings	highest_yearly_earnings
min	12300000.0	0.000000e+00	NaN	0.0	NaN	NaN	0.0	0.0
max	245000000.0	2.280000e+11	NaN	301308.0	NaN	NaN	6900000.0	110600000.0
nunique	NaN	NaN	18.0	NaN	48.0	14.0	NaN	NaN

	created_year	created_month	Gross tertiary education enrollment (%)	Population	Unemployment rate
min	1970.0	NaN	7.6	2.025060e+05	0.75
max	2022.0	NaN	113.1	1.397715e+09	14.72
nunique	NaN	12.0	NaN	NaN	NaN

DSCI 320: Project Milestone 1

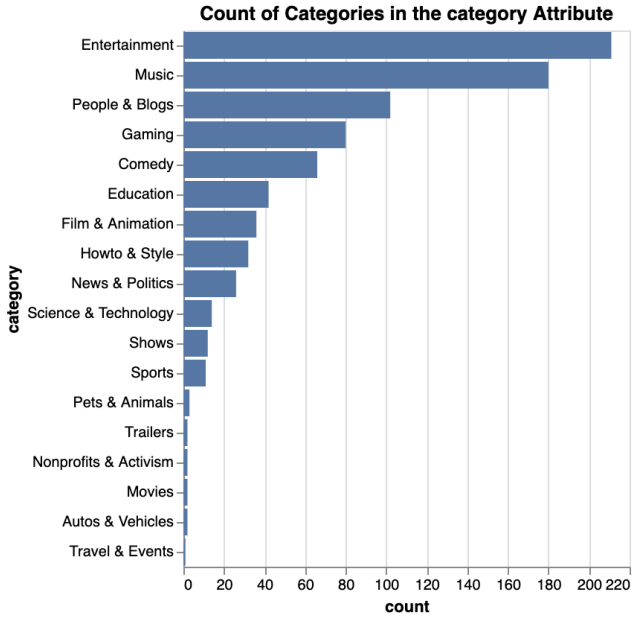
UNIVARIATE SUMMARIES

Numeric

	subscribers	video views	uploads	lowest_yearly_earnings	highest_yearly_earnings	created_year	Gross tertiary education enrollment (%)	Population	Unemployment rate
count	824.0	8.240000e+02	824.0	824.0	824.0	824.0	824.0	8.240000e+02	824.0
mean	23069296.1	1.149078e+10	10964.2	456751.2	7313899.7	2012.2	63.4	4.327673e+08	9.2
std	18172673.0	1.509424e+10	37279.4	794091.9	12723050.5	4.3	25.9	4.755086e+08	4.9
min	12300000.0	0.000000e+00	0.0	0.0	0.0	1970.0	7.6	2.025060e+05	0.8
25%	14500000.0	4.351711e+09	316.2	64050.0	1000000.0	2009.0	36.3	8.313280e+07	5.4
50%	17600000.0	7.881591e+09	966.5	192000.0	3100000.0	2013.0	67.0	3.282395e+08	8.9
75%	24225000.0	1.385086e+10	3597.2	489800.0	7825000.0	2015.0	88.2	3.282395e+08	14.7
max	245000000.0	2.280000e+11	301308.0	6900000.0	110600000.0	2022.0	113.1	1.397715e+09	14.7

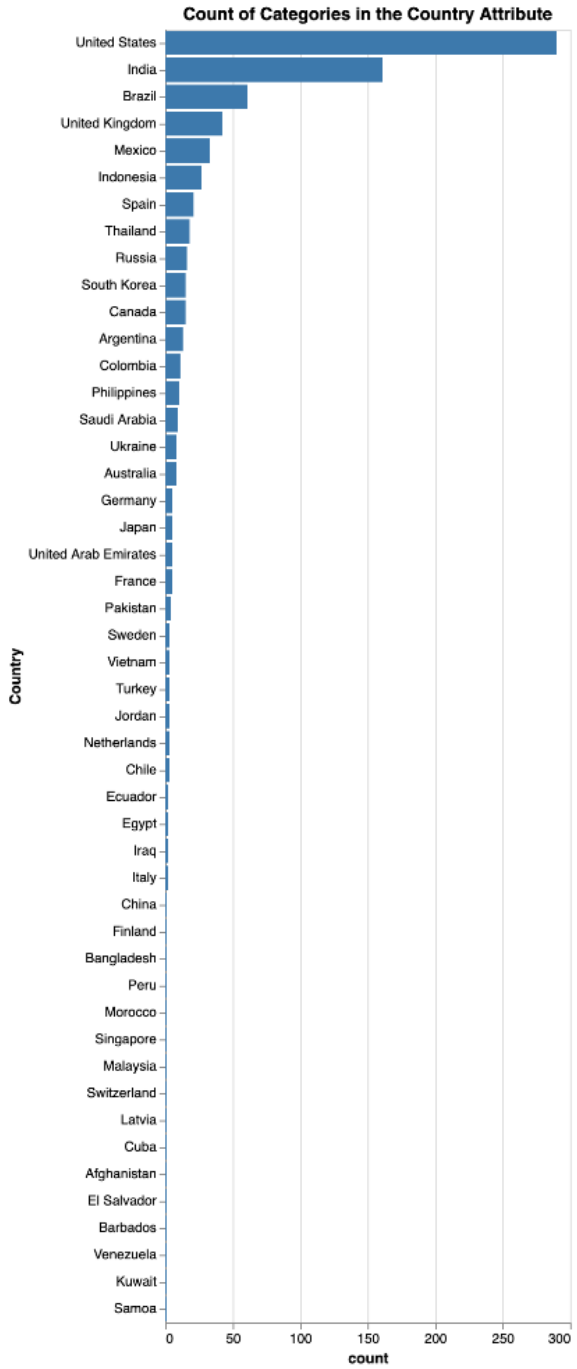
Categorical

```
category
Entertainment            211
Music                    180
People & Blogs           102
Gaming                    80
Comedy                    66
Education                 42
Film & Animation          36
Howto & Style             32
News & Politics           26
Science & Technology      14
Shows                     12
Sports                    11
Pets & Animals             3
Trailers                   2
Nonprofits & Activism      2
Movies                     2
Autos & Vehicles           2
Travel & Events            1
Name: count, dtype: int64
```



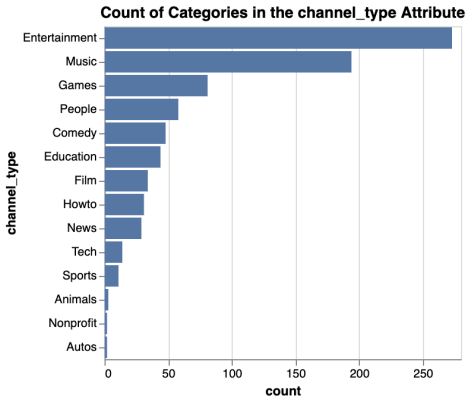
DSCI 320: Project Milestone 1

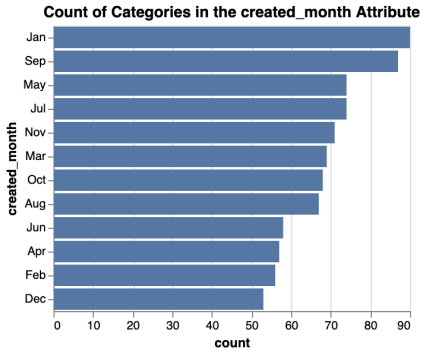
```
Country
United States           290
India                   161
Brazil                   61
United Kingdom           42
Mexico                   33
Indonesia                27
Spain                    21
Thailand                 18
Russia                   16
South Korea              15
Canada                   15
Argentina                13
Colombia                 11
Philippines              10
Saudi Arabia              9
Ukraine                   8
Australia                 8
Germany                   5
Japan                     5
United Arab Emirates      5
France                    5
Pakistan                  4
Sweden                    3
Vietnam                   3
Turkey                    3
Jordan                    3
Netherlands               3
Chile                     3
Ecuador                   2
Egypt                     2
Iraq                      2
Italy                     2
China                     1
Finland                   1
Bangladesh                1
Peru                      1
Morocco                   1
Singapore                 1
Malaysia                  1
Switzerland               1
Latvia                    1
Cuba                      1
Afghanistan               1
El Salvador               1
Barbados                  1
Venezuela                 1
Kuwait                    1
Samoa                     1
Name: count, dtype: int64
```



DSCI 320: Project Milestone 1

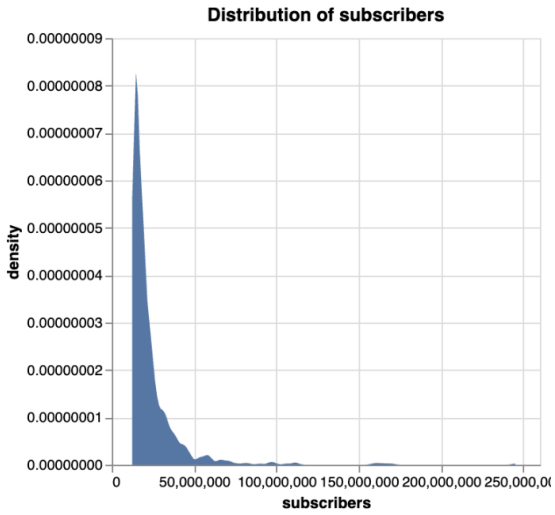
```
channel_type
Entertainment    273
Music            194
Games             81
People            58
Comedy            48
Education         44
Film              34
Howto             31
News              29
Tech              14
Sports            11
Animals            3
Nonprofit          2
Autos              2
Name: count, dtype: int64
created_month
Jan    90
Sep    87
May    74
Jul    74
Nov    71
Mar    69
Oct    68
Aug    67
Jun    58
Apr    57
Feb    56
Dec    53
Name: count, dtype: int64
```

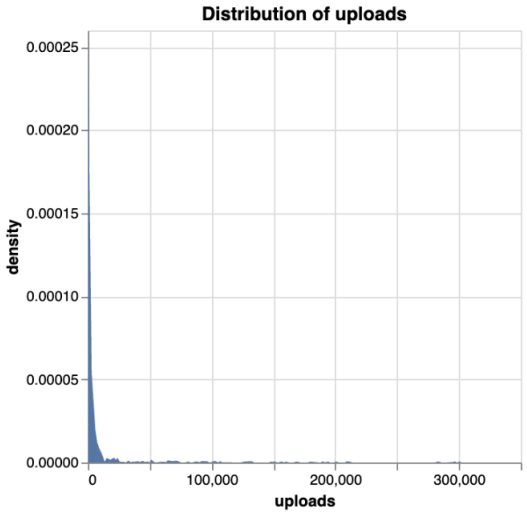


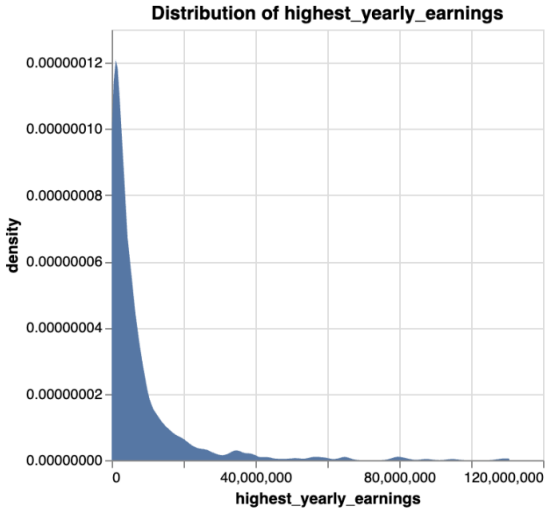


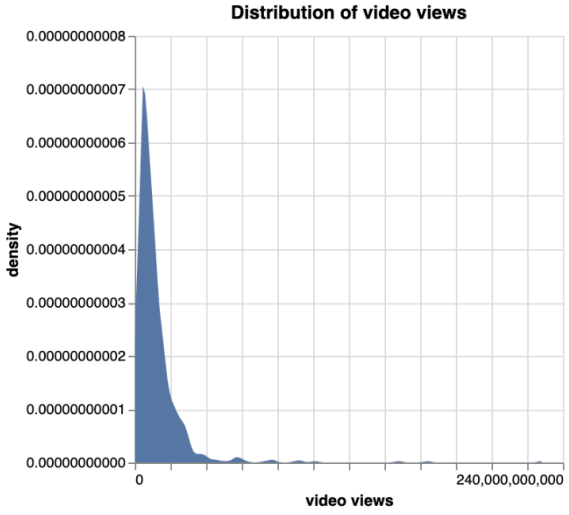
DSCI 320: Project Milestone 1

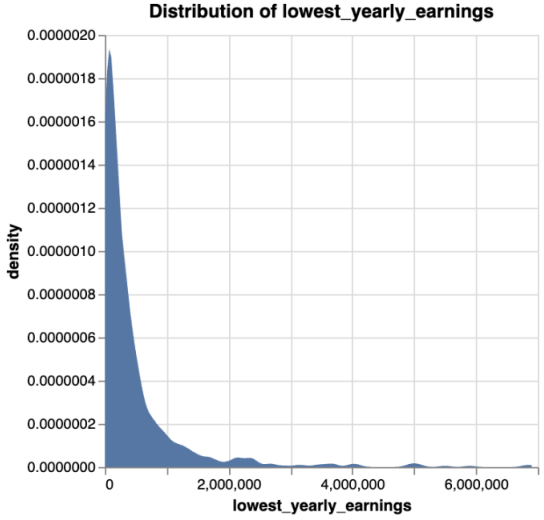
Quantitative

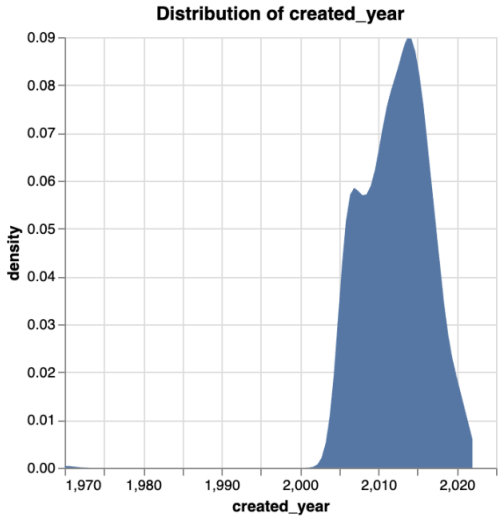




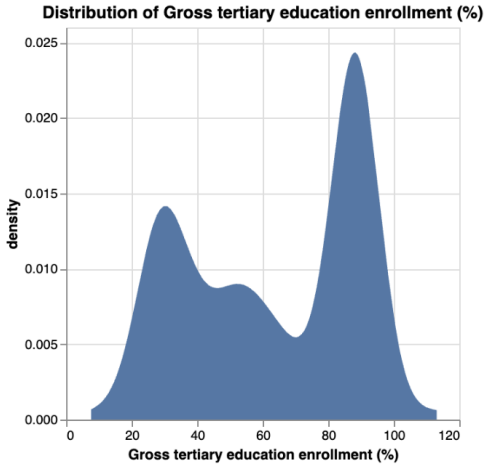


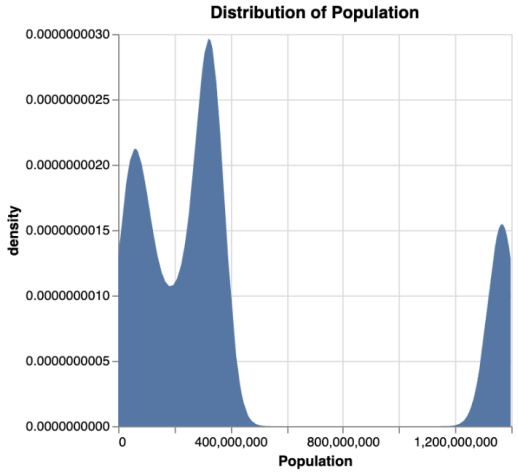


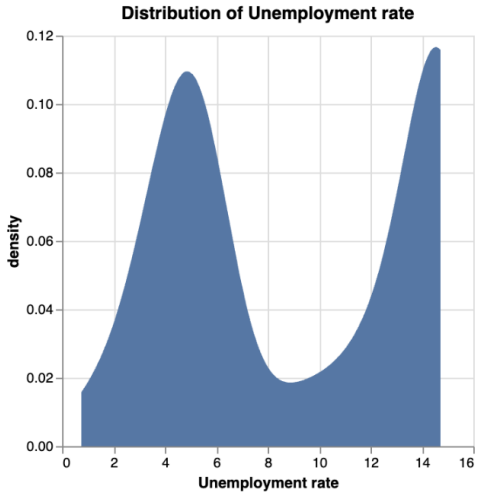




DSCI 320: Project Milestone 1







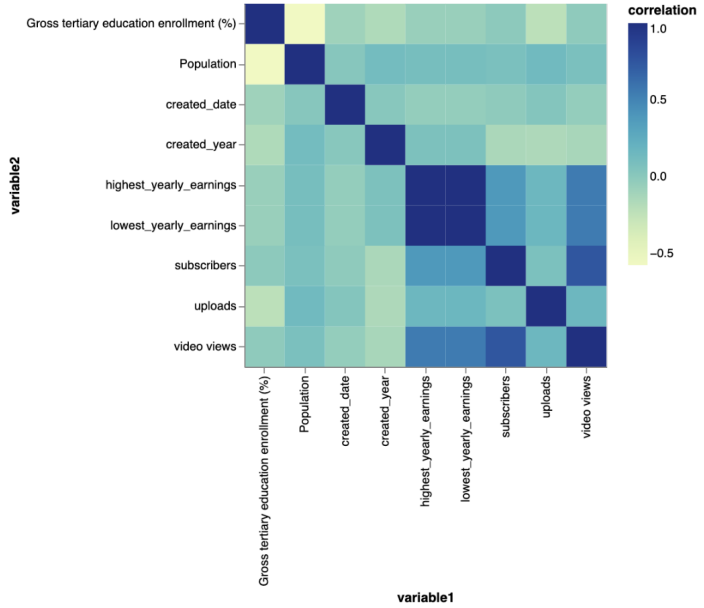
DSCI 320: Project Milestone 1

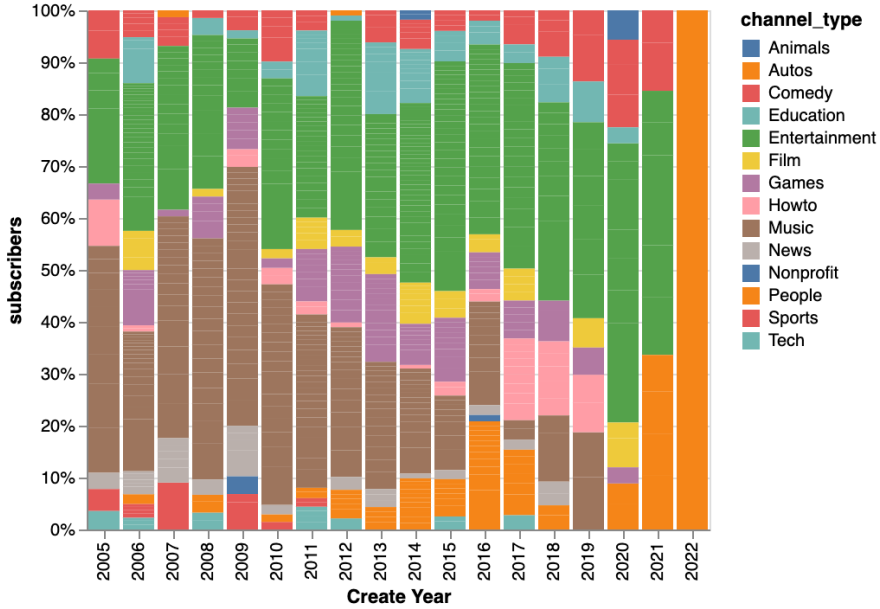
Insights:

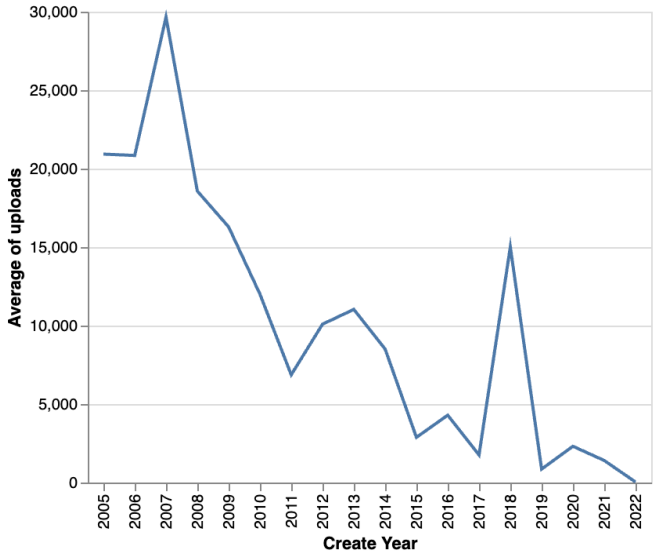
- Created Year: We can see from initial data exploration and density distribution that the ranges are focused after the year 2000. We also noticed that there is data at the start of the axis (1970) - which is not possible as YouTube was released in 2005. Thus we decided to modify our data wrangling steps to drop this value.
- Unemployment Rate: has a bimodal distribution - this seems interesting, so we decided to explore this attribute further.
- Country: There are a lot more counts of USA compared to other countries. This is important to keep in mind.
- Subscribers: Very left-skewed distribution - indicates that lower number of subscribers are more common than a very high number of subscribers.
- Uploads: Also very left-skewed distribution. Like Subscribers, we should further investigate this type of skewed distribution and properly scale the axes if we are using these attributes.

DSCI 320: Project Milestone 1

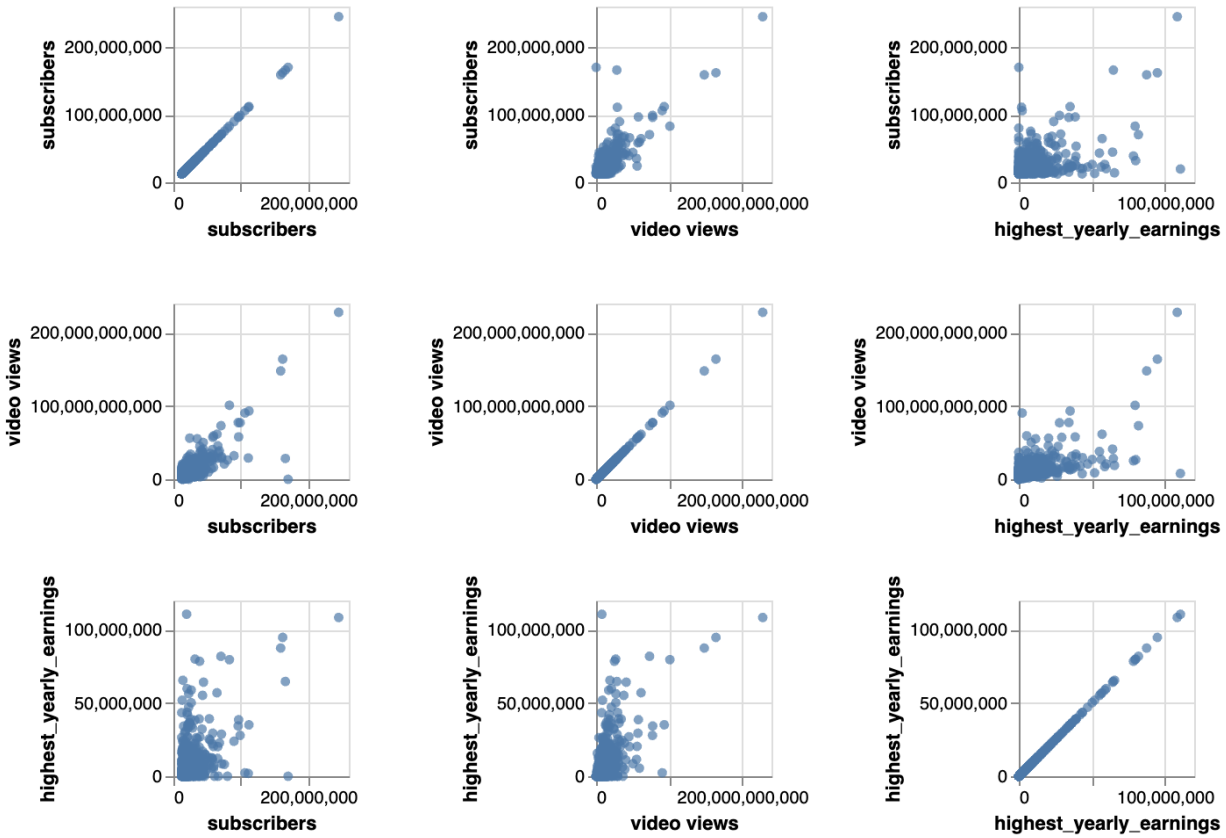
MULTIVARIATE SUMMARIES







DSCI 320: Project Milestone 1



Insights for Multivariate:

- Correlation Matrix - we don't see significant correlation between attributes. However, it is obvious that the highest and lowest yearly earnings correlate with each other which also makes sense since they are the opposite of each other.
- Scatter Plot - when further investigating Video views vs. subscribers, as well as the highest earnings vs. subscribers, we see a very weak positive correlation.
- We plot average upload count with the created year to see if there are any trends

DSCI 320: Project Milestone 1

PART II: Project Scope

Introduction

Title: Youtube Statistics Analysis

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 amount of creators and video uploads, this dataset can give interesting insights into Youtube video trends, metrics, user behaviour, 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 thus, monetization.

Task Analysis

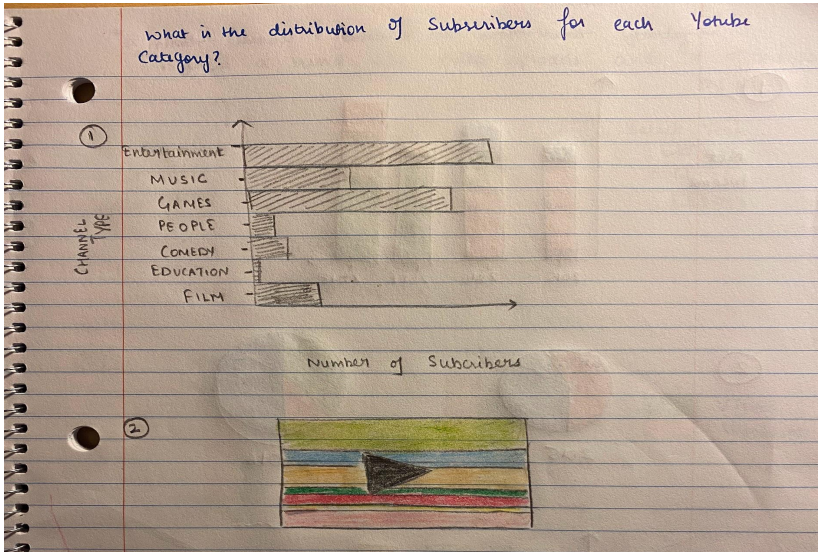
1. Distribution → What is the distribution of subscribers for each YouTube channel type based on Country?
2. Filter → What are the top Earners Among YouTube Channels by category on a Yearly Basis?
3. Correlate → Is there a trend between average video views by Channel Type over the months?
4. Cluster → Is there a cluster of highest yearly income a youtube channel makes by country and number of uploads?
5. Compute derived value → What is the average unemployment rate in the countries with the highest amount of uploads.

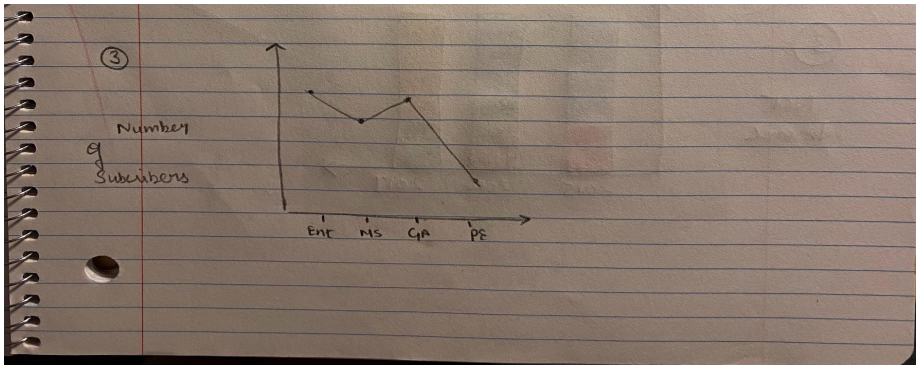
DSCI 320: Project Milestone 1

Part III: Visualization Idea

"What is the distribution of subscribers for each Youtube Channel type based on Country?"

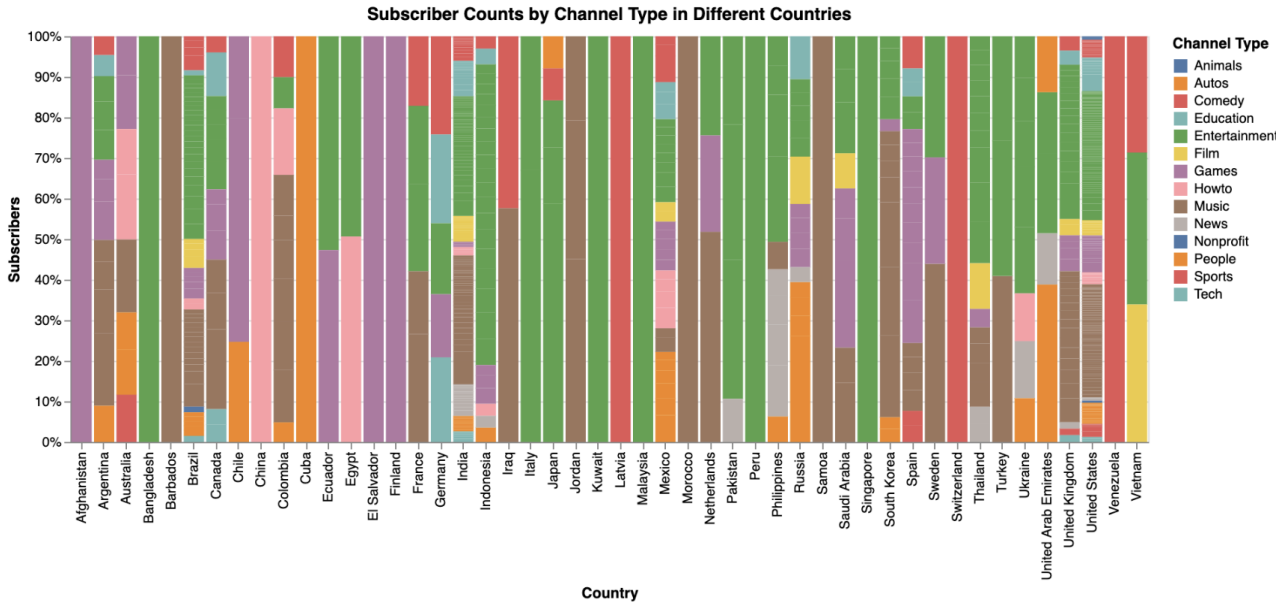
The first visualization is a normalized bar graph, coloured by the channel type. This seems like an effective and convenient way to represent out information. Second, we have a pie chart made for each country. While, we can see the distribution of subscribers, we would have to create pie charts for each country would be a hassle. Thus, this is not the most effective way. Third, we have the box plot. While a box plot provides a lot of information such as the mean, median, upper and lower quantiles, we would have to create a box plot for each country and each channel type which also seems cumbersome.





DSCI 320: Project Milestone 1

High Fidelity: This stacked normalized bar chart offers a comprehensive view of the subscriber distribution across various countries, with the vertical axis representing the number of subscribers and the horizontal axis displaying different countries. The chart's color encoding enables the differentiation of channel types, facilitating a clear understanding of the relative contributions of each channel type to the overall subscriber counts. The stacked bar format aids in visualizing the total subscriber count for each country and at the same time allows the viewer to discern the contribution of each channel type within the countries.



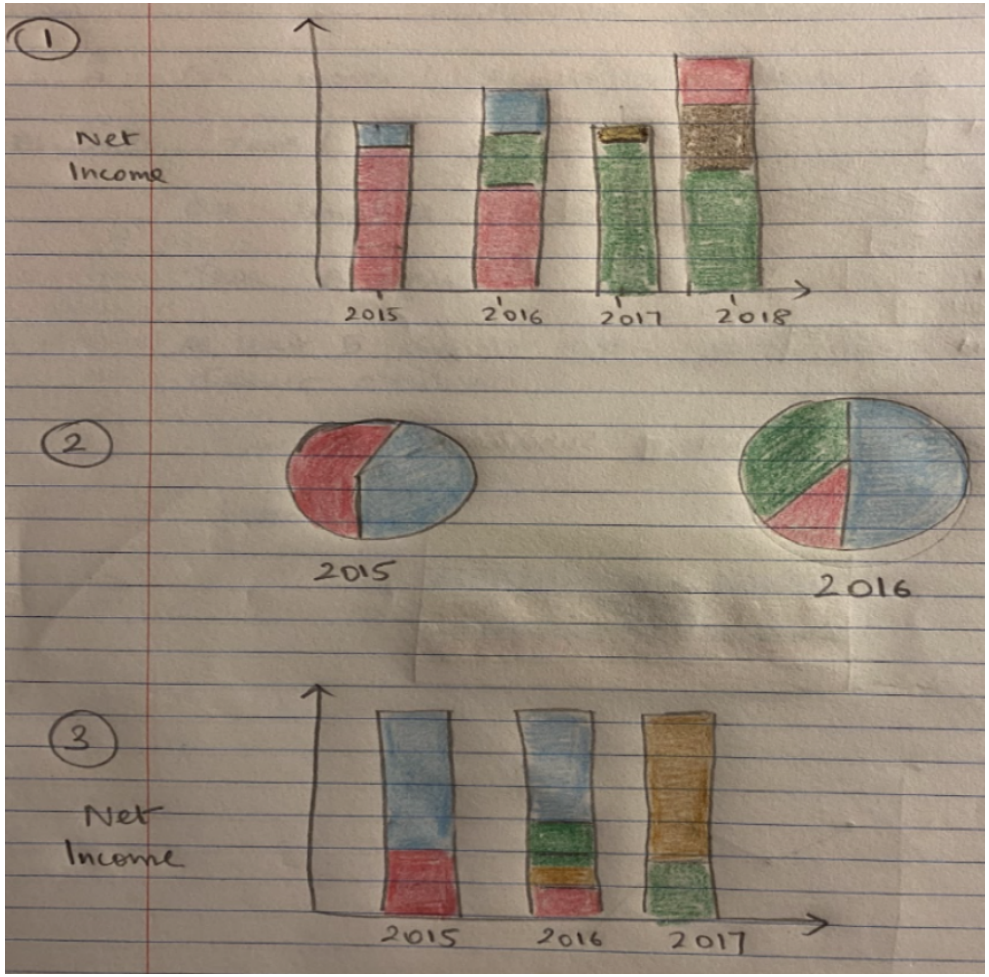
Insights:

- From the Viz it is apparent that In the United States more than half of the subscribers fall withing the channel type "Music" and "Entertainment".
- The data suggests that Vietnam is one of the only countries where "film" is a channel type that holds above 30% of the subscribers
- The data suggests that all subscribers from China falls within the category of "Howto".
- We should be careful using the normalized bar-chart as we only have one data-attribute from the country of China, whereas we in comparison have 290 observations from the United states. Therefore we can't make any generalizations of the country, we can only describe what our data suggests.

DSCI 320: Project Milestone 1

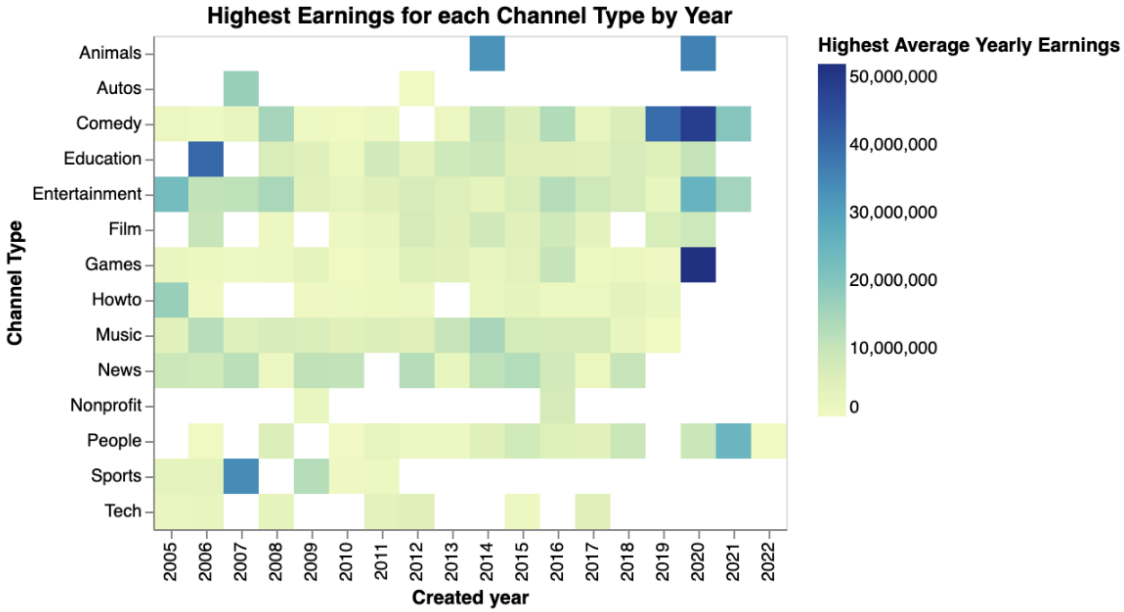
"What are the top Earners Among YouTube Channel Types on a Yearly Basis?"

The first visualization is a stacked bar graph. It is great for comparing the net income of youtube channels over the years. The second is a pie chart for every year. You will notice that the area for 2016 is larger than 2015, since Youtube channels made more money in 2016 than 2015. Lastly, we have the normalized bar graph. This is great for comparing the proportion of income for every channel by year. This visualization provides the most amount of information in a simple and understandable manner.



DSCI 320: Project Milestone 1

High Fidelity: Despite not having a heatmap in our initial sketches, we decided to choose it for our high fidelity sketch as it provides more information in a more concise manner. As opposed to using color for channel type in a bar graph (there are a lot more than a few different channel types), we set the types on the vertical axis. To visualize the difference in yearly earnings, we added this to the color-channel. Rather than getting the actual value of the earnings of the channel type over the years, we focus on differences between earnings of different channel types - this provides a more nuanced view. In this way we prioritize the comparative analysis over the absolute measurements and the color hue overview of the size of the earnings compared to other channel types for each year.



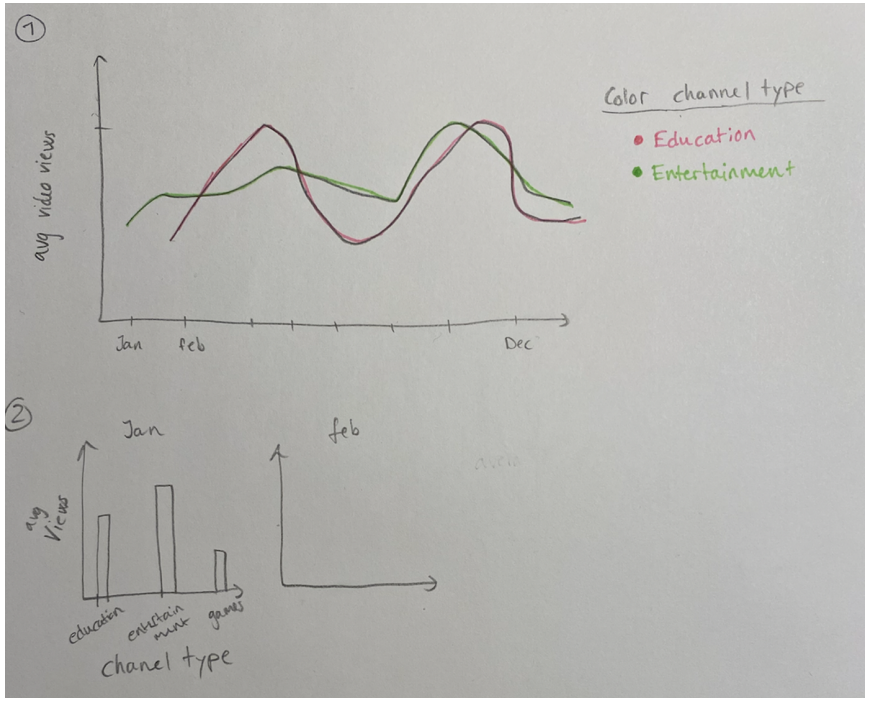
Insights:

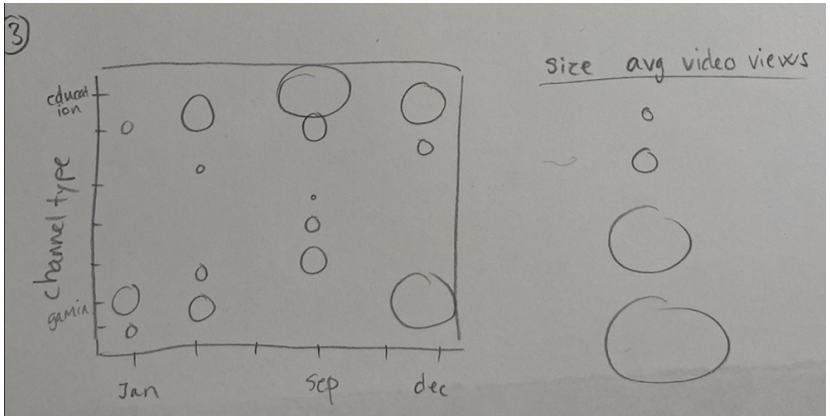
- The viz suggests that in 2020 the highest average yearly earnings fall within the categories of "Comedy" and "Games"
- The year that channel types of the category education earned the most was in 2006 according to the available data.
- Comments: there is a lot of missing information in the viz, which makes it hard to compare all scenarios against each other.

DSCI 320: Project Milestone 1

<u>"Is there a trend between average video views by Channel Type over the months?"</u>

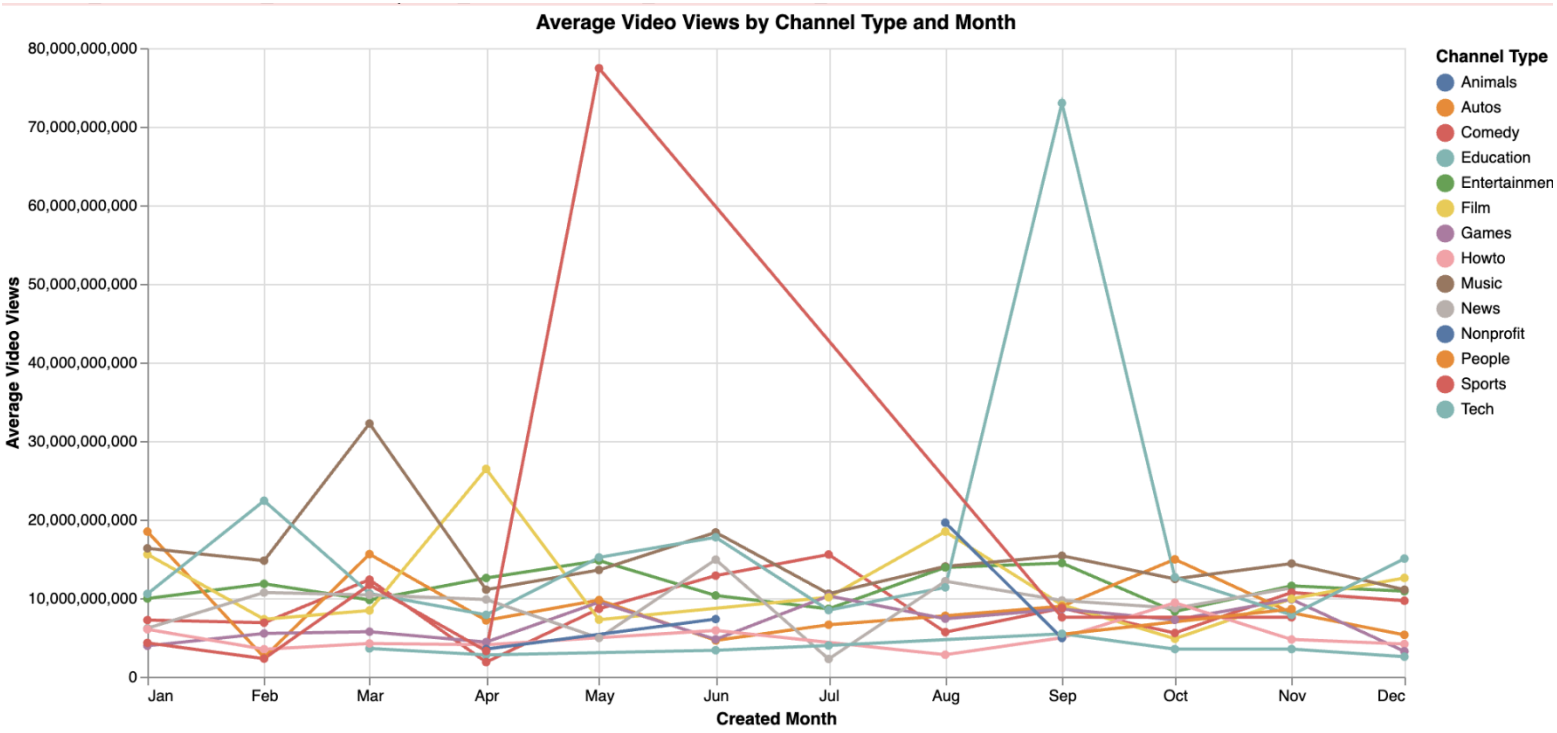
The first sketch is a line-chart with months on the horizontal channel and average video views on the vertical channel. The channel type is encoded on the color channel, enabling a view of the trend for the different channel types. The second is a bar-chart with avg views on the y-channel and channel type on the x-channel faceted on each month. This could show the distribution of views for each channel-type within each month, but facetting on the month attributes means that the viewer has to compare 12 different visualizations. This might not be as effective. In the last sketch the month is encoded on the x-axis and the channel type is on the y-axis. The size channel encodes the average video views. This choice might not be the best as we would like to see if there is a trend between the number of views and over the months, and the circles as the mark and using size channel to encode the quantitative value of views, it can be hard to compare the values and even see if there is a trend or not.





DSCI 320: Project Milestone 1

High Fidelity: For the high fidelity sketch we chose the line chart which shows how the average video views change throughout the average year. The vertical axis tells us the average number of views, while the horizontal axis represents the months. Each line stands for a different type of channel, helping us compare how they perform over the months. By hovering over the lines, you can see the specific average view counts for each month. This chart makes it easy to spot trends and compare the performance of different channel types over time. However the the channel-type has 14 distinct categories meaning that the color-channel might not be as effective.



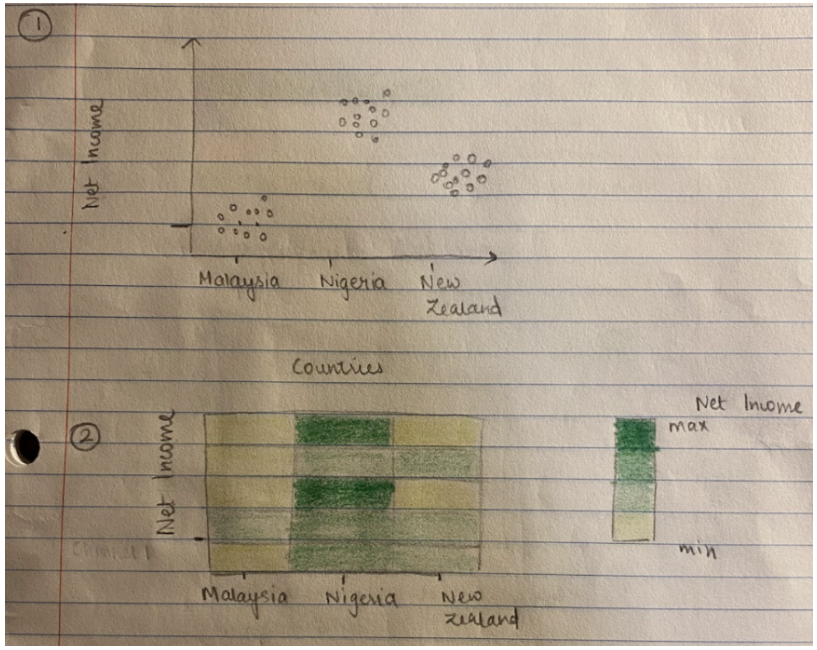
Insights:

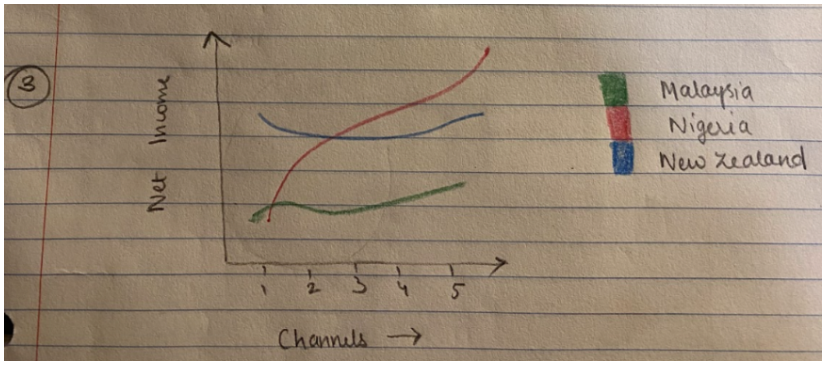
- It could be apparent from the data that videos falling within the category of 'Education' have the most views in September (which could make sense as this is when the school year typically begins).
- It might be that the months where the views are the lowest is in the summer-months. (when people prefer to stay outdoors?)

DSCI 320: Project Milestone 1

<u>“Is there a cluster of highest yearly income a youtube channel makes by country and number of uploads?“</u>

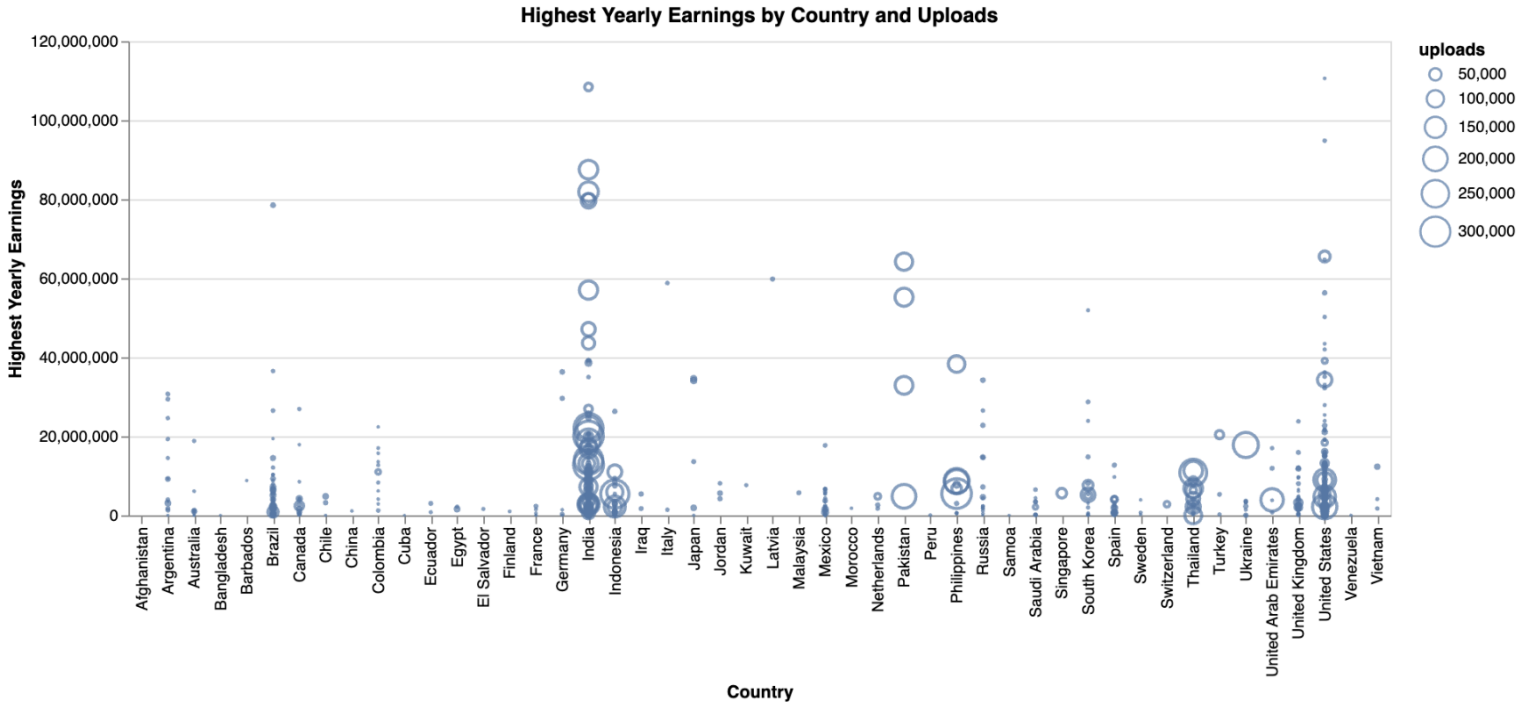
The first visualization is a bubble chart with countries on the x axis and net income on the y. This is a very effective visualization to observe a cluster of incomes across countries. However since both spatial encodings are not quantitative variables, this might be difficult to make. Second, we have a heatmap, with each rectangular box depicting the net income of a youtube channel from a given country. This is also an effective and quick way of seeing income disparity between countries. Lastly, we have a line chart with channels on the x - axis. Here, we treat channels as numbers (1,2,3…) and treat it as an ordinal variable. Each line depicts a country. This visualization may not work well with a lot of countries (as in our case)





DSCI 320: Project Milestone 1

High Fidelity: The chosen chart shows how the earnings of content creators vary by country and number of uploads. On the horizontal axis, the different countries are encoded, and on the vertical axis the highest yearly earnings. The size chanel encode how many uploads the corresponding you-tube channel have made. A tool-tip is added so, you can see the exact number of uploads for each point. We have chosen this as the best viz because in the low fidelity sketches, we had a line-chart each line encoding a different country, which made our visualization very cluttered and difficult to interpret. We couldn't choose the heatmap because different countries had different number of youtube channels which made comparisons very difficult. Finally the scatter-plot was our best option.



Insights:

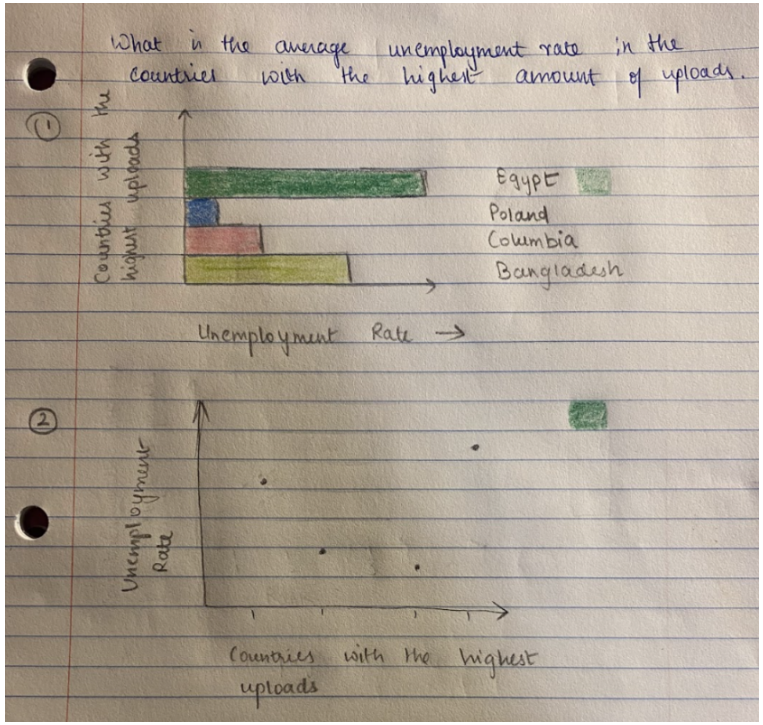
- It seems from the viz that there is 1 youTuber from India who didn't have a lot of uploads, but still made one of the highest yearly earnings
- Comment: This helps us see that a higher number of uploads does not necessarily mean a higher yearly income for a youtube channel. Moreover, the highest yearly income of a channel could also depend on the country a channel is based out of.

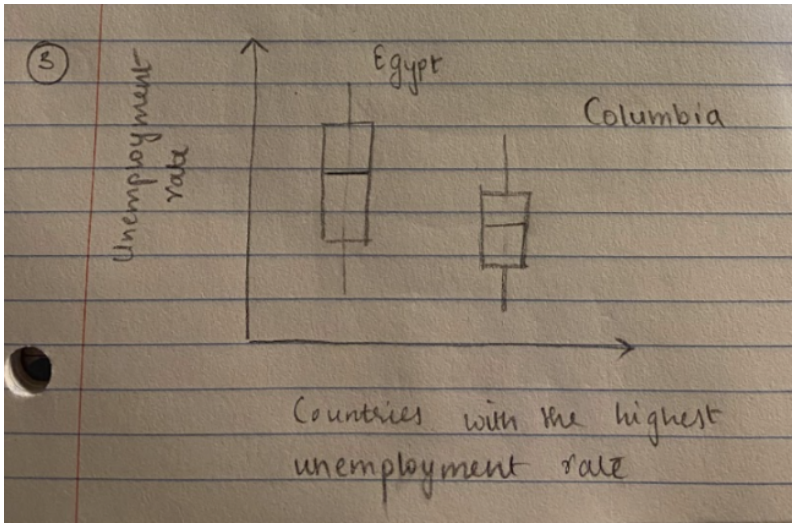
DSCI 320: Project Milestone 1

<u>"What is the average unemployment rate of the top 5 countries with the highest amount of uploads?"</u>

The first visualization is a bar graph, with different bars representing different countries. A bar graph is a simple way to display all the required information, with minimal distractors. The second is a scatterplot. This is just as effective as the bar graph. The third is a boxplot. This not only provides information about the average unemployment rate, but also the median and the upper and lower quantiles. This helps us see the bigger picture.

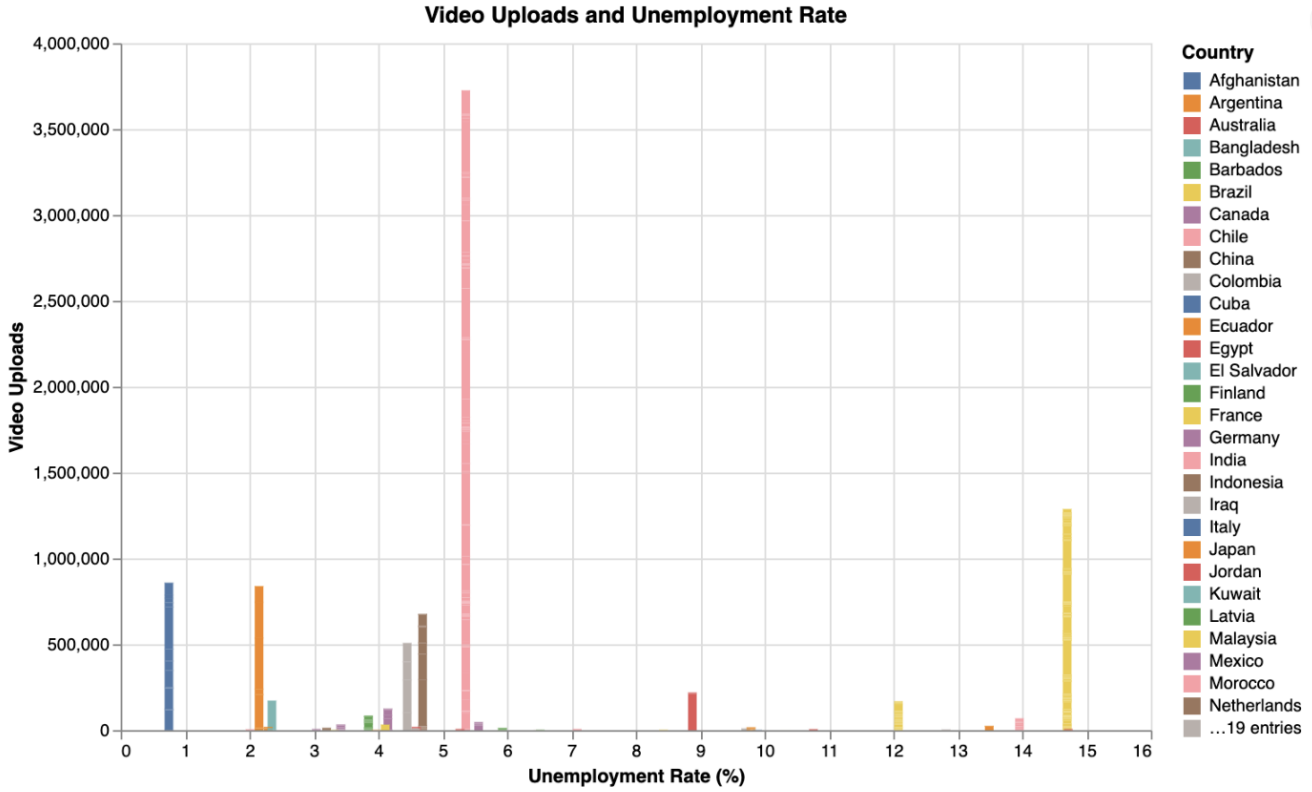
Note: In the official data visualization/wrangling process we should filter for the top 5 countries first, in order to avoid using the Color channel on the Country attribute.





DSCI 320: Project Milestone 1

High Fidelity: This bar chart compares the 'Unemployment Rate to the 'Video Uploads' for different countries. The horizontal channel represent the unemployment rate, while the vertical channel encodes the number of video uploads. The color-channel represents the country. We had the notion that the countries with the highest unemployment rate would have the largest number of uploads (unemployed persons seeking a secondary source of income). We selected the bar graph as it was the most effective way of representing this data. We are able to make precise conclusions with minimal distractors (ie, we are able to compare bars which may be far apart from each other easily since the in-between bars do not create distractions.)



DSCI 320: Project Milestone 1

PART IV: Next Steps

1. Meet and discuss whether additional data cleaning and preprocessing is needed to ensure the dataset is sufficient for the analysis we are interested in going in depth with.
2. Become more familiar with the data by expanding the current exploratory data analysis and do more multivariate summaries to find other interesting connections between the attributes. (this should be done as a group to discuss all possible interesting tasks within the data-set chosen)
3. Perform essential data wrangling to facilitate the visualization of the required task components.
4. Use the feedback from milestone 1 to refine the visualization prototypes and tasks to make sure that they effectively meet the needs and expectations of the project.
5. Make guidelines for maintaining visual consistency and aesthetics across all visualizations to ensure a unified and professional presentation for the final product.