Analyzing YouTubers' Popularity and Financial Success

Qichuan Zhou

2023-12-08

Motivation:

- As one of the most popular websites in 2023, the successes of YouTube channels rely heavily on the audience worldwide. Therefore, investigating the factors that drive its success in terms of both popularity and earning potential may help us uncover the underlying global trends. In general, this project focuses on the secret of the popularity and financial success of certain YouTube channels both from an individual and country level perspective.
- I believe that this project could provide a deeper understanding of the media and entertainment landscape across various countries and shed light on the elements that make a channel engaging. Moreover, popular YouTube channels may have a significant impact on societal change, potentially facilitating digital diplomacy. By studying the relationship between a channel's popularity and the socioeconomic factors within its originating country, we could find out if these widely believed statements hold true.
- Specifically, this project will focus on the following questions:
  - What are the key individual characteristics that significantly impact a YouTube channel's popularity and financial success?
  - How do country-level characteristics contribute to a YouTube channel's popularity and financial success?
  - How can we segment YouTubers in the dataset into distinctive groups based on their characteristics?

Data Sources:

- There are two data sources in total.
- The first dataset can be downloaded from the Kaggle website with the following link: https://www.kaggle.com/datasets/nelgiriyewithana/global-youtube-statistics-2023* and is sourced from a csv file. It provides a comprehensive overview of YouTube metrics on a global scale for the year 2023. These metrics associated with YouTube creators include the number of subscribers, video views, upload frequency, average yearly earnings, inception date, and more. We retrieved a total of 995 records where each captures the statistics of a prominent YouTuber.
- The secondary dataset was a merged dataset containing country-level variables. It encompasses demographic data for countries worldwide, including GDP per capita, Gross Tertiary Education Enrollment, urban population, English Proficiency Index, country-level censorship score, and internet user population. This dataset was acquired through web scraping

techniques from websites like wikipedia. In this dataset, each record represents a country's statistical data.

Data Manipulation and Visualization

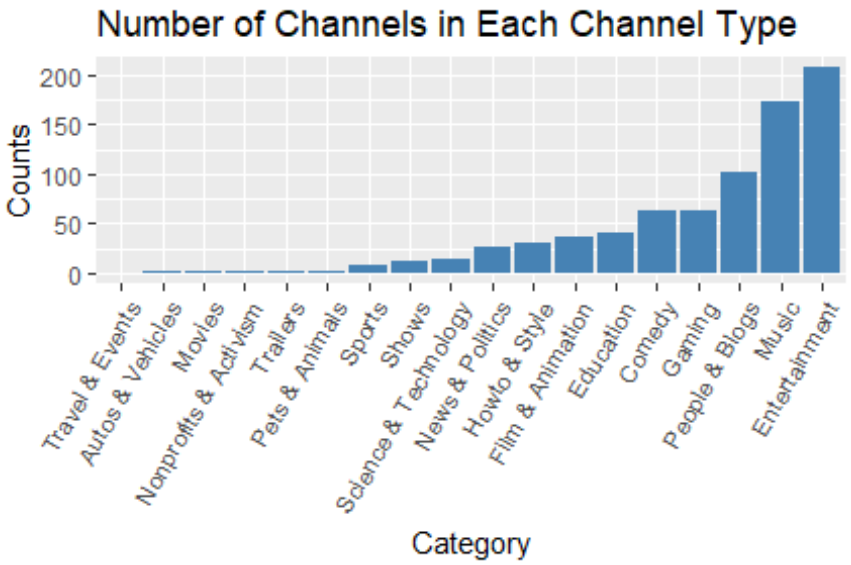
- Prior to merging two datasets, all extraneous variables were eliminated from both data frames I established, streamlining our final combined dataset to the greatest degree. Subsequently, all character-type demographic variables at a country level were transformed into numeric data points given their essential role in the following linear regression analyses and ANOVA tests. The subsequent step involved assessing the instances of 'NA' values in each column and dropping all rows that contained such values. I refrained from using imputation methods given that most missing data pertained to several categorical variables in the "YouTube" dataset, indicating that missing values were not randomly distributed. Consequently, no one imputation method would suitably address this circumstance. Additionally, numeric variables such as `video views` and `subscribers` with a value of 0 were interpreted as unavailable, hence rows consisting of zeros were also omitted from the final refined dataset. Having accomplished these steps, I created a novel variable `avg_yearly_earnings` as an average of `lowest_yearly_earnings` and `highest_yearly_earnings` since it plays a crucial role as a dependent variable in our research questions.

What are the key individual characteristics that significantly impact a YouTube channel's popularity and financial success?

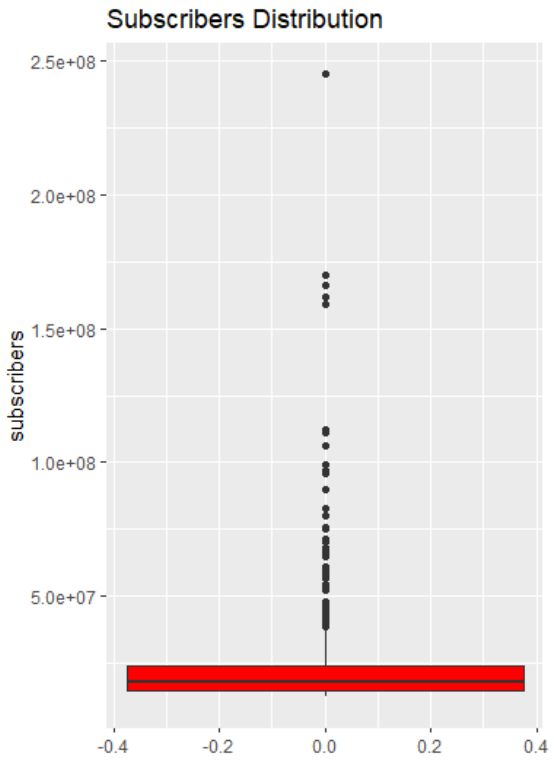
Popularity: `video views` Financial success: `average yearly earnings` Possible Individual Characteristics: `category`, `subscribers`, `uploads`, `created_year`

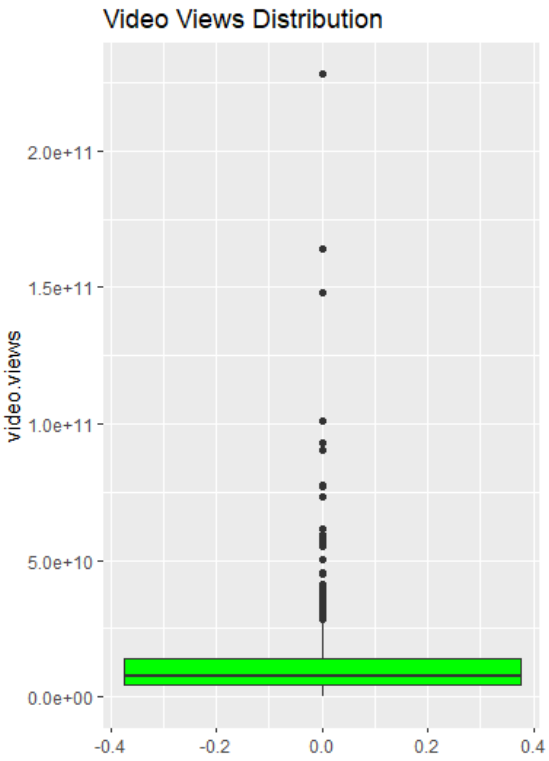
- At the outset, I identified several factors potentially influencing a YouTube channel's popularity and earning capability. The dependent variables here are total video views and average yearly earnings, serving as proxies for popularity and financial success respectively. The examination concentrated on four predictor variables: the number of subscribers, the number of uploads, channel category, and the year of channel creation.

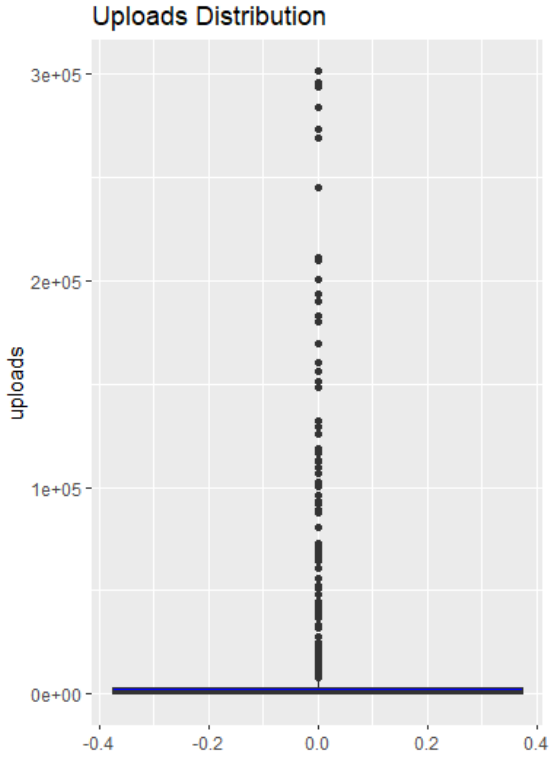
- Subsequently, I initiated a series of exploratory visualizations. Beneath, one can find a bar graph representing the quantity of YouTube channels in each category respectively.

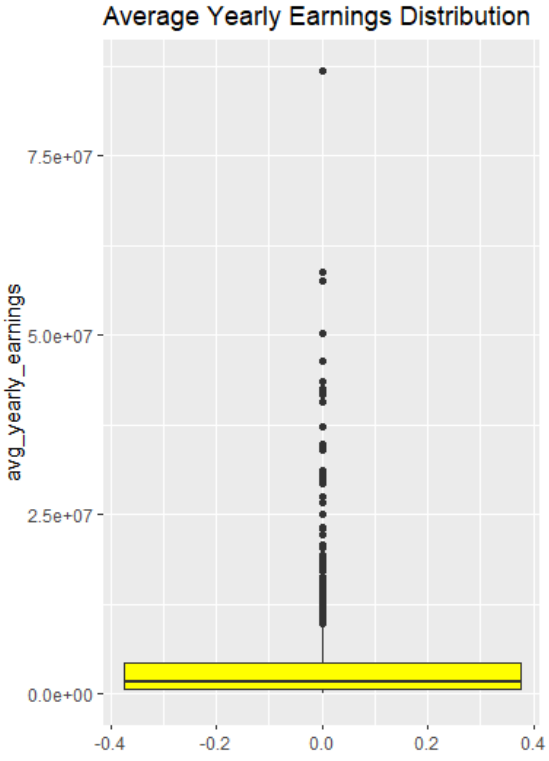


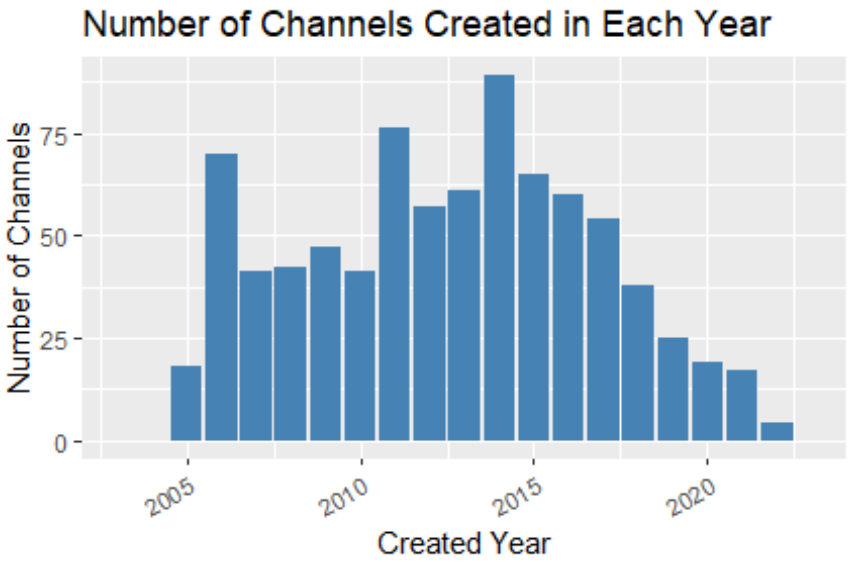
- I proceeded to create boxplots, showcasing a visual representation of the distribution for two dependent variables and two numerical predictor variables. Furthermore, a bar chart illuminating the numerical breakdown of channel creation years is presented below.


Subscribers Distribution
subscribers
2.5e+08
2.0e+08
1.5e+08
1.0e+08
5.0e+07
-0.4
-0.2
0.0
0.2
0.4

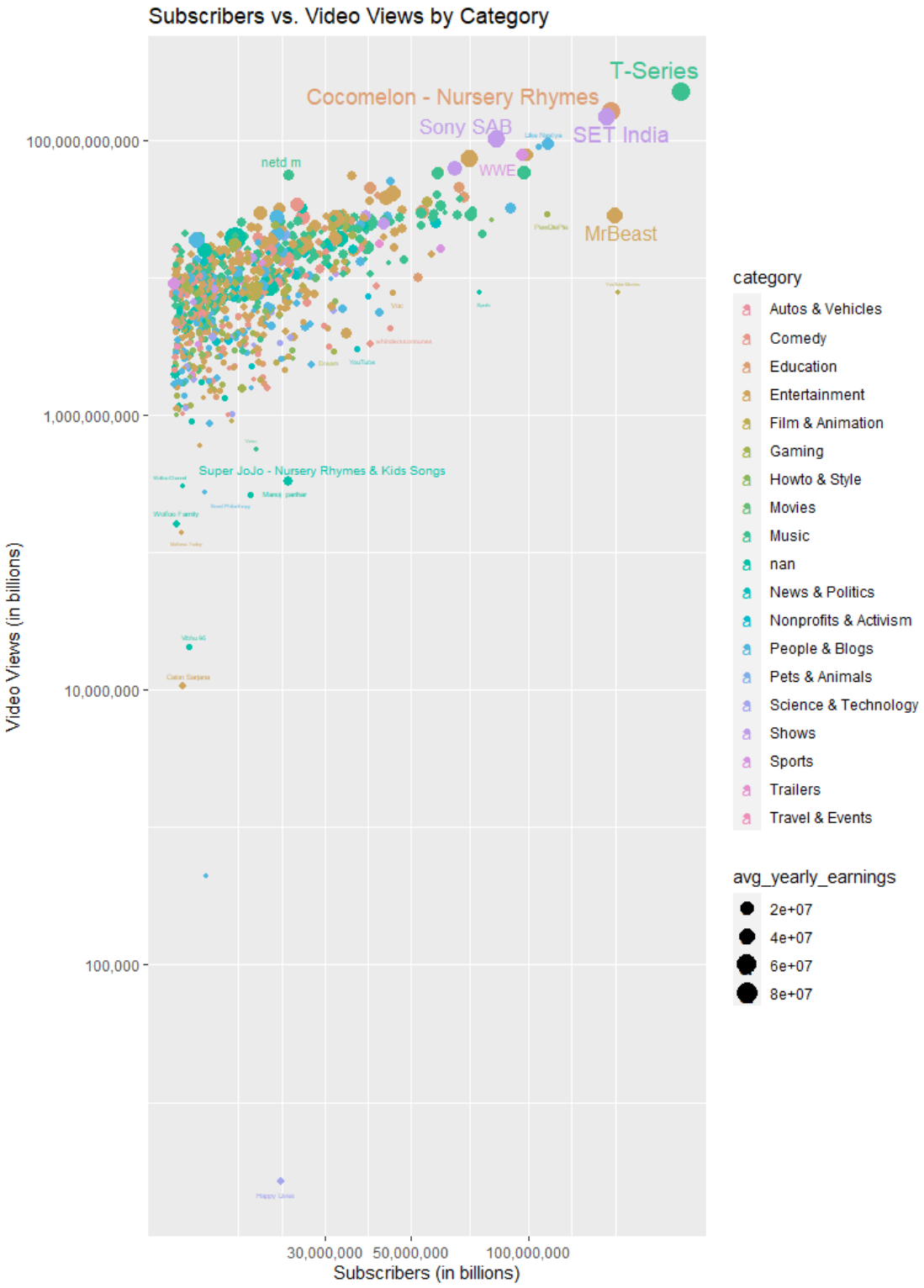

Video Views Distribution
video.views
2.0e+11
1.5e+11
1.0e+11
5.0e+10
0.0e+00
-0.4
-0.2
0.0
0.2
0.4


Uploads Distribution
uploads
3e+05
2e+05
1e+05
0e+00
-0.4
-0.2
0.0
0.2
0.4


Average Yearly Earnings Distribution
avg_yearly_earnings
7.5e+07
5.0e+07
2.5e+07
0.0e+00
-0.4
-0.2
0.0
0.2
0.4

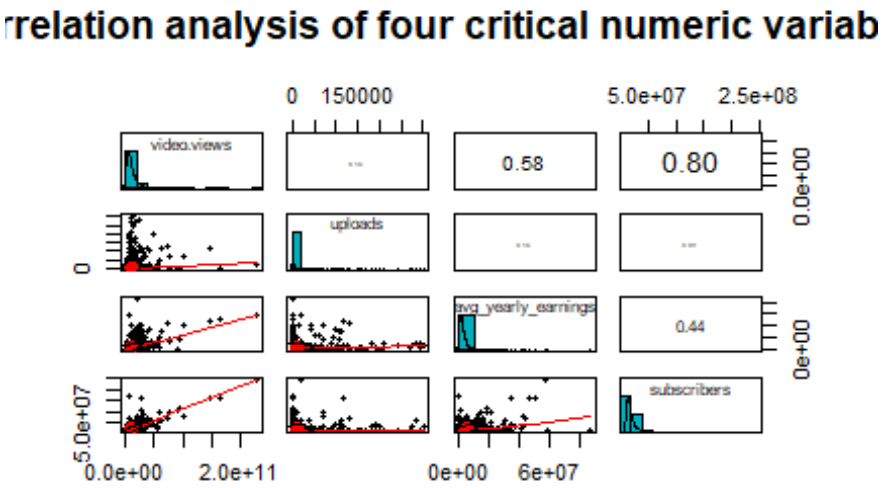


- The scatterplot below vividly highlights the interplay among four essential variables. Each channel category is distinguished by color, while the magnitude of data points mirrors the respective channel's annual average earnings. Notably, a positive correlation between the number of subscribers and the total video view count is observed. Furthermore, an emerging trend indicates an upward trajectory in a channel's average yearly earnings with an increase in the subscriber count.



- The pairplot below highlights the interplay among four essential numeric variables. It demonstrates strong correlation between video views and average yearly earnings, video views and the number of subscribers,

moderate correlation between average yearly earnings and the number of subscribers.



- Given the distinct outliers noticeable in the boxplots of key variables, I employed a Box-Cox transformation on the dependent variables to counterbalance the pronounced "heavy tail" effect. Following this, I executed a linear regression analysis and an Analysis of Variance (ANOVA) test to test the intuitive conclusions drawn from our preliminary visual explorations.
- Drawing from the ANOVA table presented below, it can be inferred that the number of uploads and the number of subscribers substantially and positively correlate with a YouTube channel's popularity and financial prosperity. Yet, there is insufficient evidence to establish a significant connection between the year of the channel's inception and the dependent variables. Moreover, it is noticeable that different categories exhibit different levels of popularity and earning potential.

```
##
## Call:
## lm(formula = log(video.views) ~ subscribers + uploads + created_year,

##     data = glob_yt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.7460  -0.4221   0.1771   0.5828   1.9877
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.088e+01  1.653e+01   3.077 0.002158 **
## subscribers   2.611e-08  1.981e-09  13.176  < 2e-16 ***
## uploads       3.762e-06  9.698e-07   3.880 0.000113 ***
## created_year -1.432e-02  8.212e-03  -1.744 0.081608 .
```

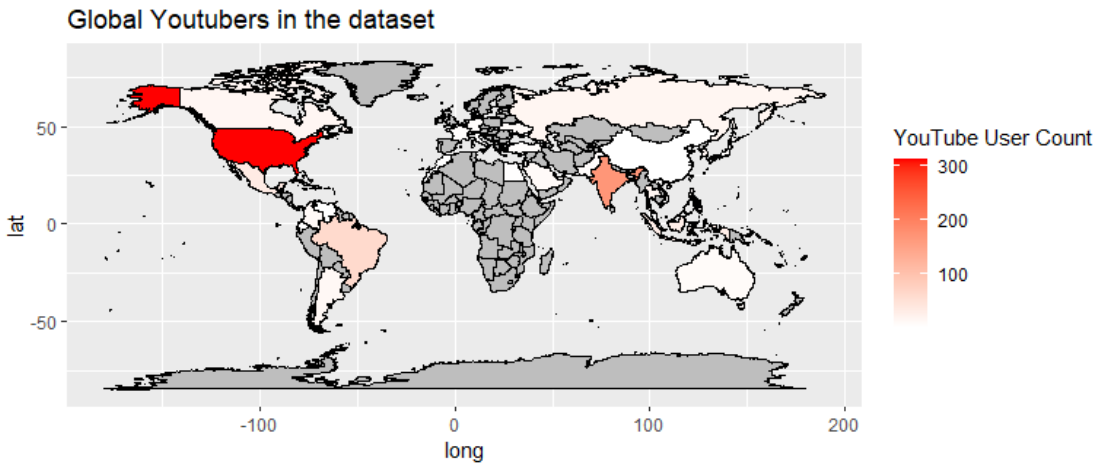
```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.021 on 821 degrees of freedom
## Multiple R-squared:  0.2072, Adjusted R-squared:  0.2043
## F-statistic: 71.54 on 3 and 821 DF,  p-value: < 2.2e-16

##               Df Sum Sq Mean Sq F value   Pr(>F)
## subscribers    1  202.3  202.33  193.93  < 2e-16 ***
## uploads        1   18.4   18.42   17.66 2.94e-05 ***
## created_year   1    3.2    3.17    3.04   0.0816 .
## Residuals    821  856.6    1.04
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##
## Call:
## lm(formula = avg_yearly_earnings ~ subscribers + uploads + created_y
ear,
##     data = glob_yt)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -29663352  -2716620  -1243889    629060  82774205
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -5.071e+08  1.067e+08  -4.754 2.35e-06 ***
## subscribers   1.868e-01  1.278e-02  14.619  < 2e-16 ***
## uploads       3.070e+01  6.256e+00   4.908 1.11e-06 ***
## created_year  2.517e+05  5.298e+04   4.752 2.38e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6589000 on 821 degrees of freedom
## Multiple R-squared:  0.234,  Adjusted R-squared:  0.2312
## F-statistic: 83.59 on 3 and 821 DF,  p-value: < 2.2e-16

##               Df    Sum Sq   Mean Sq F value   Pr(>F)
## subscribers    1 9.134e+15 9.134e+15  210.38  < 2e-16 ***
## uploads        1 7.732e+14 7.732e+14   17.81 2.71e-05 ***
## created_year   1 9.804e+14 9.804e+14   22.58 2.38e-06 ***
## Residuals    821 3.565e+16 4.342e+13
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Analysis of Variance Table
##
## Response: log(video.views)
##                   Df Sum Sq Mean Sq F value    Pr(>F)
## factor(category)  18 144.93  8.0517  6.9367 < 2.2e-16 ***
```

```
## Residuals        806 935.56  1.1607
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

How do country-level characteristics contribute to a YouTube channel's popularity and financial success?

Popularity: `video views` Financial success: `average yearly earnings` Possible country-level predictor variables: `Population`, `Urban_population`, `gdp_percap`, `ter_educ`, `epi_score`, `int_user`, `censor_score`

- Most variables' names are self-explanatory, but some may require further illustration. "ter_educ" stands for Gross Tertiary Education Enrollment; "epi_score" refers to English Proficiency Index score, which attempts to rank countries by the equity of English language skills among those adults who took the EF test; "censor_score" stands for "Freedom on the Net" score, which are a set of numerical ratings regarding the state of Internet freedom for countries worldwide.
- The world map below displays the distribution of the home country of YouTubers in the dataset.



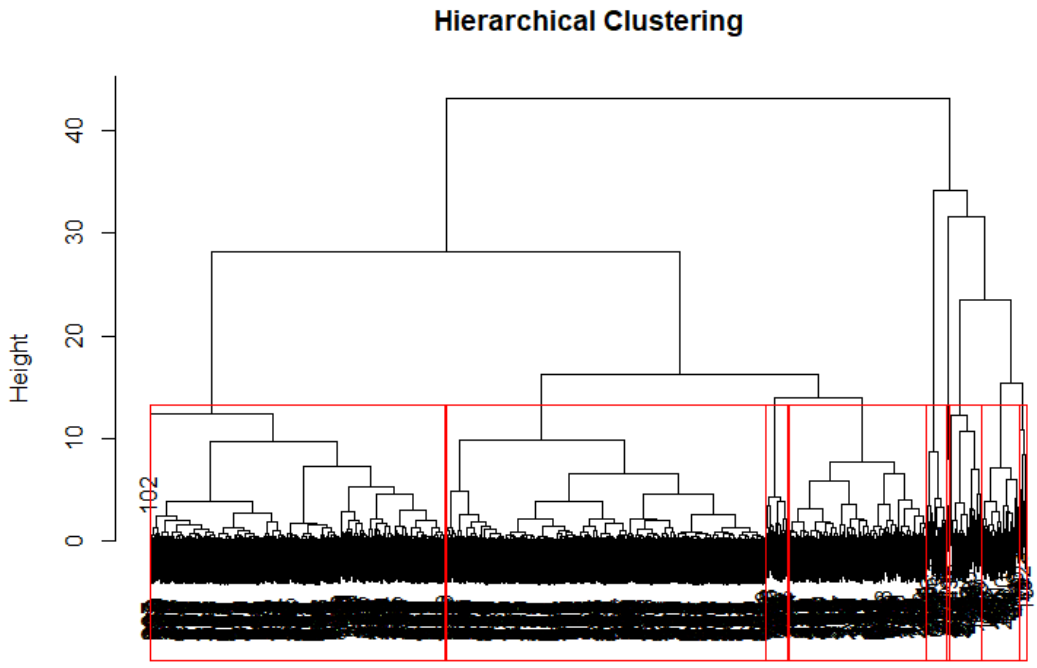
- Through the regression analysis, I surprisingly found that among these country-level factors, only population and GDP per capita of the country of

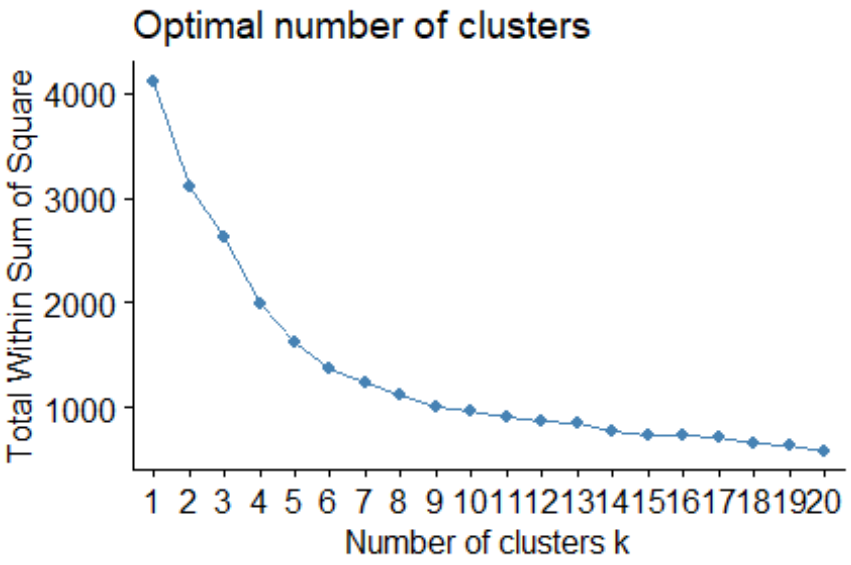
the YouTube channel creator are positively and significantly associated with the channel's popularity and all of the country-level factors fail to contribute significantly to a YouTube channel's financial success.

```
##
## Call:
## lm(formula = log(video.views) ~ Population + Urban_population +
##     gdp_percap + ter_educ + epi_score + int_user + censor_score,
##     data = glob_yt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.9173  -0.5156   0.0641   0.6197   3.4151
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2.286e+01  4.548e-01  50.275  < 2e-16 ***
## Population        8.137e-09  3.955e-09   2.057  0.03997 *
## Urban_population -1.668e-09  2.973e-09  -0.561  0.57487
## gdp_percap        5.257e-08  1.811e-08   2.903  0.00379 **
## ter_educ          3.190e-03  3.438e-03   0.928  0.35379
## epi_score        -2.921e-04  1.096e-03  -0.267  0.78989
## int_user         -1.199e-08  7.400e-09  -1.620  0.10568
## censor_score     -6.873e-04  5.548e-03  -0.124  0.90144
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.138 on 817 degrees of freedom
## Multiple R-squared:  0.02015,    Adjusted R-squared:  0.01176
## F-statistic:   2.4 on 7 and 817 DF,  p-value: 0.01956

##
## Call:
## lm(formula = log(avg_yearly_earnings) ~ Population + Urban_populatio
n +
##     gdp_percap + ter_educ + epi_score + int_user + censor_score,
##     data = glob_yt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.2959  -0.3942   0.5545   1.4298   4.7887
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.436e+01  1.221e+00  11.762   <2e-16 ***
## Population        9.093e-09  1.062e-08   0.856    0.392
## Urban_population  6.629e-09  7.982e-09   0.831    0.406
## gdp_percap        3.792e-08  4.861e-08   0.780    0.436
## ter_educ          1.362e-03  9.230e-03   0.148    0.883
## epi_score        -1.701e-03  2.942e-03  -0.578    0.563
## int_user         -1.675e-08  1.987e-08  -0.843    0.399
```

```
## censor_score     -1.387e-03  1.489e-02  -0.093    0.926
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.056 on 817 degrees of freedom
## Multiple R-squared:  0.01327,    Adjusted R-squared:  0.004815
## F-statistic:  1.57 on 7 and 817 DF,  p-value: 0.141
```

How can we segment YouTubers in the dataset into distinctive groups based on their characteristics?





- Utilizing hierarchical and kmeans clustering, I finalized the number of clusters to be nine and summarized the mean of critical variables grouped by clusters in the column displayed on the slide. According to this column, I could put distinctive and well-characterized tags on each cluster of YouTuber based on theri individual characteristics as below.

```
## # A tibble: 9 × 5
##   cluster mean_videoviews mean_avg_yearly_earnings mean_subscribers
mean_uploads
##   <fct>             <dbl>                    <dbl>            <dbl>
       <dbl>
## 1 1           9485226960.                2061250.        19362130.
       4779.
## 2 2          28548715811.                7514289.        46436508.
       8521.
## 3 3         180000000000                51483333.       188666667.
      45861.
## 4 4          41297662236.                44655000         44680000
      57655.
## 5 5           6688762102.                 2230180.        18270638.
       1880.
## 6 6          16550093117.                20056250         21243750
       6036.
## 7 7           7700096716.                 2106696.        18620863.
       4539.
## 8 8          54835926330.                13492467.       116266667.
       9650.
## 9 9          14502957903.                 5783879.        22603846.
     189699.
```

- The following nine tags correspond to the nine clusters.

- YouTubers boasting a substantial number of video views, significant average yearly earnings, a considerable amount of subscribers, and a prolific number of uploads.
- YouTubers with a modest number of video views, lower average yearly earnings, a limited number of subscribers, and few uploads.
- YouTubers maintaining a balanced combination of video views, average yearly earnings, subscribers, and uploads—each at a moderate level.
- YouTubers featuring a small number of video views, low average yearly earnings, a limited number of subscribers, and a moderate number of uploads.
- Youtubers presenting a balanced mix with a moderate number of video views, high average yearly earnings, a moderate number of subscribers, and a high number of uploads.
- YouTubers demonstrating a modest number of video views, low average yearly earnings, a moderate number of subscribers, and a balanced number of uploads.
- YouTubers displaying a limited number of video views, low average yearly earnings, a moderate number of subscribers, and a low number of uploads.
- Youtubers maintaining a balanced combination with a moderate number of video views, moderate average yearly earnings, a moderate number of subscribers, and a high number of uploads.
- YouTubers featuring a moderate number of video views, moderately high average yearly earnings, a limited number of subscribers, and a low number of uploads.