Final report

Group 2

1. Topic and question

As a social media company, it sponsors the video makers in the platform of YouTube to make more profits. It helps video makers control the quality, and help them get more views, forwards, and subscription. If the Lethal India company can discover a better pattern to make videos and launch videos, it will have a better reputation in the entertainment industry, and then it can sign more video makers and expand its affect.

In Lethal India company, we are going to set this problem in three content creators to explore more details in the video-making and video-launch process. This analysis will be applied in Lethal India's musical video-making process ultimately for those musical video makers in Lethal India company.

Our mission was to plan a nationwide tour for our producers to give their work the widest possible exposure. We landed on this subject because we're always on the lookout for new ways to expand our client brands' exposure to the widest possible audience. Is it crucial that we have the ability to pitch brands for their 2023 marketing strategies? As we will be able to demonstrate the results of reach achieved via social media on a variety of platforms, our research will be of interest to brands.

On our planned cross-country tour, we hope to make pitches to different companies about marketing their products and services.

2. Data

1) Two data sets as the secondary data

   Since all of our creators have millions of followers, it was simple to get Default analytics from YouTube Studio, so we decided to start there. We got two datasets from YouTube Analytics.

   - The first dataset is about the shares, subscribers, likes, views, and impressions for each creator, in this data set, we have three subsets, each subset represents each creator's video information. This data set is to help us to filter the most popular videos from each creator, and then we can analyze the insight pattern of these successful videos.
   - Another dataset is about the geography information about each creators' viewers, to figure out where to help these creators to hold the fan convention/concert in the future. In this dataset, we also have three subsets, each subset represents the geography information of each creator's viewers.

2) One data set as the primary data

   We want to dig out the insight successful elements in most successful videos, we decided to see how these creators' videos are presented and what's viewers' opinions towards these elements.

   - We decide to collect some data manually in terms of comments, video types, thumbnails designs, and guests invited for the most popular videos for each creator. These directions can help us to figure which elements are most significate to produce a specific type of video.
   - Besides, we also created a Google Survey to ask about these creator's video contents, like types, thumbnails, duration, guests, and fan conventions. Based on the survey results, we got the third dataset.

3) URL of Data Sources:

   YouTube Analytics is the back stage of the Lethal India company, and data about each creator is intangible asset for the company, we couldn't show entrance to the back stage.

Google Surveys: https://forms.gle/dcSgWE151BxcfS7f8

3. Information quality

The information quality mainly focuses on secondary data sets

- Data accessibility problem of survey

  We used a variety of sources, including surveys, Social Blade, and YouTube Studio, to gather the data for this project. Each of them provides us with data in a different format and in a different order, which raises issues with data quality such as inconsistent data in surveys where respondents may have lied about their opinions or provided answers that were completely unrelated to the survey's questions. while accessing the data, a great deal of privacy and confidentiality could be compromised for users who are following those creators, such as user ID, email address, etc. So, to solve this problem, we got data from users who really follow these YouTube creators, organized the data, and then replaced the missing and irrelevant data with the answer given by the majority of users.

  In addition, surveys require judgment because respondents may be biased or subjective, which can significantly affect the analysis and quality of the data. Therefore, if we receive responses from a large number of users, we can include a disclaimer on the form to request only objective opinions without becoming biased, and we can also compare the actual result to the anticipated result. If there is an excessive amount of variation, we can repeat the survey with a different group of users to verify the outcome.

- Judgement from person who sort videos is involved

  Also while sorting out the top videos for each creator, we have used the criteria like most number of views, likes, comments and shares which may have an effect on information quality if the same users are using different accounts to increase views, likes and comments on the specific videos.

  And we team sort videos to different types, the judgement from the person who sort videos is involved. If the person think it is Funny video, it is Funny, if the person think it is Roast video, it is Roast video. And sometime, a video may have a lot of elements in it, it is hard to decide their main style.

  For this problem, we team tried to sort these video types together to eliminate the personal judgement in the whole process.

4. Methods and Tools

The project involves generating a score from secondary data and collecting relevant data from the primary source.

1) Dealing with Primary data

   - Methods

     Collect Data Manually: We collected and recorded the comments, video types, duration, thumbnails designs, and guests invited of most popular videos for each creator.

     Sentimental Analysis: Comments have been picked for sentimental analysis from each category of videos in order understand how viewers receive each category of videos from each creator.

     Comparison By Graphs: We added various graph types, such as bar charts and pie charts, to perform visual analysis, which significantly improved the quality and analysis of the data. Tools & Technologies

   - Tools & Technologies:

     Google Survey: We used Google Survey to collect data about people's opinions about the characteristics added in each creator's videos.

     IBM Watson Engine: Sentimental analysis of comments has been performed using the IBM tool.

Excel: For the results of Google Survey, it is quite formatted, the main thing is creating methods to compare and analyze by graphs.

- Challenges

  Lack of proper categorization of videos in the channel which demanded the team to categorize the videos manually for analysis. We have picked top-performing videos from the channel for analysis manually. (A compromise has been made to take the project ahead).

2) Dealing with secondary data

- Methods:

  Linear Regression: Regression has been performed twice for each creator taking 'Watch Time' & 'Views' as Target Variables considering all other attributes as predictors to understand the significance of each attribute and its impact on Watch time, count of views or rather the overall engagement of a channel of a particular creator.

  Logistic Regression: This regression has been performed with 'Watch Time' as target variable with display (Category of content) to understand the performance of each category under consideration for each creator.

  Generating a score for each creator: Weightage has been assigned to each attribute based on their significance that has been estimated from the above two regressions and with common sense (Judged by humans).

  Generating a score for each type of act: The mean values of each category have been calculated from the output obtained after Logistic regression.

  Generating a score for each city: This component of the final Formula has been derived as a fraction or a ratio of views from a city to the total number of views a channel has from all parts of the planet and inculcated into the final evaluation.

  Generating a single score from all the influencing factors: Weighted attributes have been used to generate a score to each creator which imply the engagement of a creator. The Engagement score has been multiplied with the subscriber count (Taking 1 Million subscribers as a unit. For instance, 0.211 is the factor multiplied to the engagement score of creators with 211k subscribers). The next factor (Category significance) has been derived as the mean value of watch time for that particular category. The following and the final factor is obtained from the geographical data obtained from YouTube Analytics data.

- Tools & Technologies:

  Python: Both Linear and Logistic Regression have been performed using Python Programming language.

  GitHub Co-Pilot: Co-Pilot is an AI tool that helps us in coding effectively, efficiently and quickly.

  Tableau: For a better visual representation of the results obtained, several visualizations have been obtained using Tableau.

  Excel: Collected data has been organized, transformed and made useable using excel (Further detail in the next section of the report)

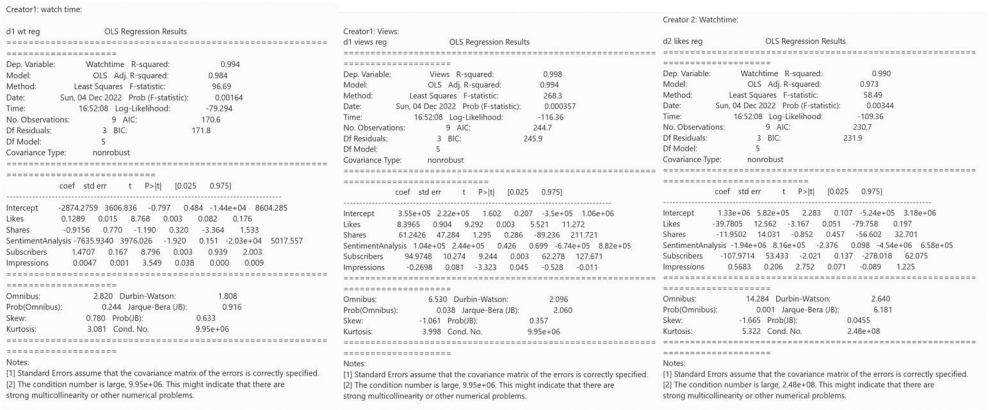
  Word: To document each step and track progress during the entire process, MS Word has been used.

- Final Formula

((0.625*(likes)) + (0.25*(Sentimental Analysis score) + (12.5*Duration)) * (Subscriber Score) * (Category Score) * (Geographical Score)

- Challenges and our attempts to tackle them
  - Human judgment involved in deriving the mathematical formula
  - Performing Sentimental analysis for one comment at a time was mundane work. Each comment's sentiment has been analyzed but the reliability of the tool used to do so is questionable, we intend to use better tools for future analysis of this kind.
  - Managing several datasets simultaneously is a tedious task, we handled it by dividing the workload optimally among the team.
- The results of the above-mentioned methods:

Regression:



Creator1: watch time:

d1 wt reg OLS Regression Results

Dep. Variable:	Watchtime	R-squared:	0.994		
Model:	OLS	Adj. R-squared:	0.984		
Method:	Least Squares	F-statistic:	96.69		
Date:	Sun, 04 Dec 2022	Prob (F-statistic):	0.00164		
Time:	16:52:08	Log-Likelihood:	-79.294		
No. Observations:	9	AIC:	170.6		
Df Residuals:	3	BIC:	171.8		
Df Model:	5				
Covariance Type:	nonrobust				

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-2874.2759	3606.836	-0.797	0.484	-1.44e+04	8604.285
Likes	0.1289	0.015	8.768	0.003	0.082	0.176
Shares	-0.9156	0.770	-1.190	0.320	-3.364	1.533
SentimentAnalysis	-7635.9340	3976.026	-1.920	0.151	-2.03e+04	5017.557
Subscribers	1.4707	0.167	8.796	0.003	0.939	2.003
Impressions	0.0047	0.001	3.549	0.038	0.000	0.009

Omnibus:	2.820	Durbin-Watson:	1.808
Prob(Omnibus):	0.244	Jarque-Bera (JB):	0.916
Skew:	0.780	Prob(JB):	0.633
Kurtosis:	3.081	Cond. No.	9.95e+06

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 9.95e+06. This might indicate that there are strong multicollinearity or other numerical problems.

Creator1: Views:

d1 views reg OLS Regression Results

Dep. Variable:	Views	R-squared:	0.998		
Model:	OLS	Adj. R-squared:	0.994		
Method:	Least Squares	F-statistic:	268.3		
Date:	Sun, 04 Dec 2022	Prob (F-statistic):	0.000357		
Time:	16:52:08	Log-Likelihood:	-116.36		
No. Observations:	9	AIC:	244.7		
Df Residuals:	3	BIC:	245.9		
Df Model:	5				
Covariance Type:	nonrobust				

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	3.55e+05	2.22e+05	1.602	0.207	-3.5e+05	1.06e+06
Likes	8.3965	0.904	9.292	0.003	5.521	11.272
Shares	61.2426	47.284	1.295	0.286	-89.236	211.721
SentimentAnalysis	1.04e+05	2.44e+05	0.426	0.699	-6.74e+05	8.82e+05
Subscribers	94.9748	10.274	9.244	0.003	62.278	127.671
Impressions	-0.2698	0.081	-3.323	0.045	-0.528	-0.011

Omnibus:	6.530	Durbin-Watson:	2.096
Prob(Omnibus):	0.038	Jarque-Bera (JB):	2.060
Skew:	-1.061	Prob(JB):	0.357
Kurtosis:	3.998	Cond. No.	9.95e+06

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 9.95e+06. This might indicate that there are strong multicollinearity or other numerical problems.

Creator 2: Watchtime:

d2 likes reg OLS Regression Results

Dep. Variable:	Watchtime	R-squared:	0.990		
Model:	OLS	Adj. R-squared:	0.973		
Method:	Least Squares	F-statistic:	58.49		
Date:	Sun, 04 Dec 2022	Prob (F-statistic):	0.00344		
Time:	16:52:08	Log-Likelihood:	-109.36		
No. Observations:	9	AIC:	230.7		
Df Residuals:	3	BIC:	231.9		
Df Model:	5				
Covariance Type:	nonrobust				

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.33e+06	5.82e+05	2.283	0.107	-5.24e+05	3.18e+06
Likes	-39.7805	12.562	-3.167	0.051	-79.758	0.197
Shares	-11.9502	14.031	-0.852	0.457	-56.602	32.701
SentimentAnalysis	-1.94e+06	8.16e+05	-2.376	0.098	-4.54e+06	6.58e+05
Subscribers	-107.9714	53.433	-2.021	0.137	-278.018	62.075
Impressions	0.5683	0.206	2.752	0.071	-0.089	1.225

Omnibus:	14.284	Durbin-Watson:	2.640
Prob(Omnibus):	0.001	Jarque-Bera (JB):	6.181
Skew:	-1.665	Prob(JB):	0.0455
Kurtosis:	5.322	Cond. No.	2.48e+08

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.48e+08. This might indicate that there are strong multicollinearity or other numerical problems.

Creator2: Views:

d2 views reg OLS Regression Results

Dep. Variable:	Views	R-squared:	0.975		
Model:	OLS	Adj. R-squared:	0.933		
Method:	Least Squares	F-statistic:	23.23		
Date:	Sun, 04 Dec 2022	Prob (F-statistic):	0.0133		
Time:	16:52:08	Log-Likelihood:	-90.607		
No. Observations:	9	AIC:	193.2		
Df Residuals:	3	BIC:	194.4		
Df Model:	5				
Covariance Type:	nonrobust				

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	2.014e+05	7.25e+04	2.777	0.069	-2.94e+04	4.32e+05
Likes	-1.6545	1.564	-1.058	0.368	-6.631	3.322
Shares	-3.6282	1.747	-2.077	0.129	-9.187	1.930
SentimentAnalysis	-2.761e+05	1.02e+05	-2.717	0.073	-6e+05	4.73e+04
Subscribers	-16.0420	6.652	-2.412	0.095	-37.211	5.127
Impressions	0.0682	0.026	2.653	0.077	-0.014	0.150

Omnibus:	0.676	Durbin-Watson:	1.873
Prob(Omnibus):	0.713	Jarque-Bera (JB):	0.531
Skew:	0.039	Prob(JB):	0.767
Kurtosis:	1.812	Cond. No.	2.48e+08

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.48e+08. This might indicate that there are strong multicollinearity or other numerical problems.

Creator 3: Watchtime:

d3 likes reg OLS Regression Results

Dep. Variable:	Watchtime	R-squared:	0.995		
Model:	OLS	Adj. R-squared:	0.988		
Method:	Least Squares	F-statistic:	130.6		
Date:	Sun, 04 Dec 2022	Prob (F-statistic):	0.00105		
Time:	16:52:08	Log-Likelihood:	-90.295		
No. Observations:	9	AIC:	192.6		
Df Residuals:	3	BIC:	193.8		
Df Model:	5				
Covariance Type:	nonrobust				

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-6.733e+04	1.79e+04	-3.770	0.033	-1.24e+05	-1.05e+04
Likes	0.5176	0.418	1.238	0.304	-0.813	1.849
Shares	2.0472	0.994	2.060	0.131	-1.115	5.210
SentimentAnalysis	7.866e+04	1.93e+04	4.078	0.027	1.73e+04	1.4e+05
Subscribers	6.2577	1.376	4.547	0.020	1.878	10.638
Impressions	0.0005	0.001	0.575	0.606	-0.002	0.004

Omnibus:	1.158	Durbin-Watson:	1.789
Prob(Omnibus):	0.561	Jarque-Bera (JB):	0.512
Skew:	0.554	Prob(JB):	0.774
Kurtosis:	2.627	Cond. No.	1.79e+08

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.79e+08. This might indicate that there are strong multicollinearity or other numerical problems.

Creator 3: Views:

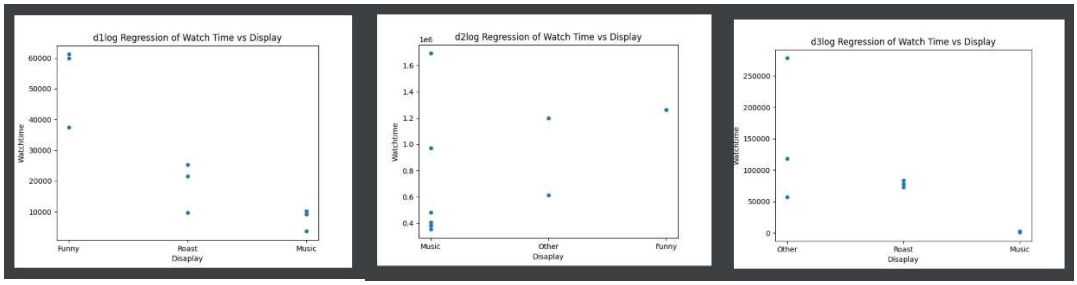
d3 views reg OLS Regression Results

Dep. Variable:	Views	R-squared:	0.997		
Model:	OLS	Adj. R-squared:	0.993		
Method:	Least Squares	F-statistic:	227.6		
Date:	Sun, 04 Dec 2022	Prob (F-statistic):	0.000457		
Time:	16:52:08	Log-Likelihood:	-108.69		
No. Observations:	9	AIC:	229.4		
Df Residuals:	3	BIC:	230.6		
Df Model:	5				
Covariance Type:	nonrobust				

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-5.441e+05	1.38e+05	-3.945	0.029	-9.83e+05	-1.05e+05
Likes	5.9804	3.230	1.851	0.161	-4.299	16.260
Shares	8.2729	7.675	1.078	0.360	-16.153	32.699
SentimentAnalysis	6.958e+05	1.49e+05	4.671	0.019	2.22e+05	1.17e+06
Subscribers	67.2262	10.630	6.324	0.008	33.398	101.055
Impressions	0.0039	0.007	0.543	0.625	-0.019	0.027

Omnibus:	1.489	Durbin-Watson:	1.837
Prob(Omnibus):	0.475	Jarque-Bera (JB):	0.498
Skew:	0.569	Prob(JB):	0.780
Kurtosis:	2.816	Cond. No.	1.79e+08

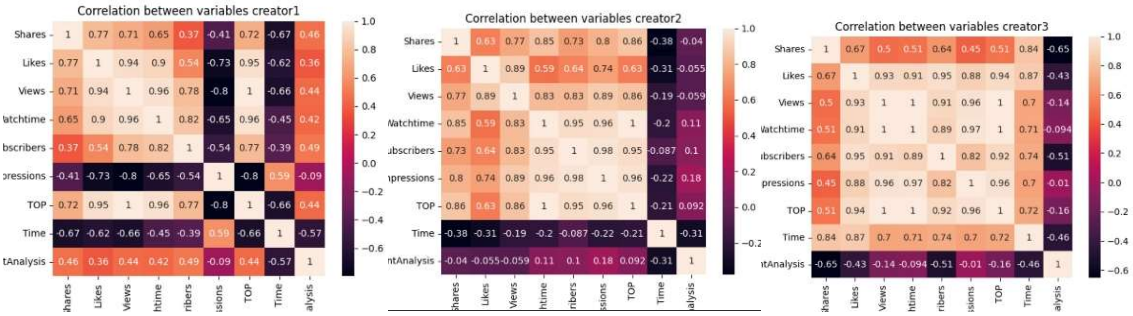
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.79e+08. This might indicate that there are strong multicollinearity or other numerical problems.



5. Data Wrangling Process

Since data has been obtained from YouTube analytics, the data was clean and reliable, but not ready for the analysis intended to perform. Whereas the data obtained from the survey is raw data obtained directly from people (Primary Data), it had to be wrangled more in order to be made useable, comparable and reliable for the project.

1) Wrangling Secondary Data
    - Data Transformation: The duration of each video has been given in minutes, it has been transformed into seconds.
    - Data Pre-Processing: Redundant columns have been removed using the regression results.



    - Data Profiling: The score obtained for each cell in the final dataset using the formula mentioned in the previous section was a huge number. In order to make it interpretable and useful, it has been scaled. The logarithmic function was applied to the scores and the resulting number was multiplied by 100.
    - Data Enrichment: Each Component in the formula required a quantified measure of each attribute. Additional data has been generated from existing data to make each attribute useful in the final evaluation. (For instance, the ratio of views from each city to the total views is calculated for each cell. Similar methods have been used to make each attribute useful)

2) Wrangling Primary Data
    - Data formatting

      Structured Data: we obtained parts of the survey in which no description or writing is required, which are presented in a pre-defined data model or format.

      Unstructured data
        - Such as short-answer survey questions about “Who do you want him to invite some guests into the type of his videos you like”, people will see the question if they answer “Yes” for the question of “Do you want him to invite some guests into the type of his videos you like?”. It can pose a significant data quality issue. To ensure that the data is consistent and relevant, all required fields must be made mandatory, and we can replace missing values and irrelevant data with the values entered by the majority of survey respondents.
        - In addition, we analyzed the sentiment of the top 20 comments on each video using the top 20 comments. So while collecting the comments, it included all the subjective and objective emotions of the users, along with emoticons, which also come in unstructured data. To resolve this, we used IBM Watson to do the sentiment analysis, with the score varying from -1 to +1, where 0 is a neutral score.
    - Data Enrichment

      After getting the dataset about all videos information and filtering the top videos for each creator, it is necessary to incorporate additional data, such as a sentiment analysis score, thumbnails, figures, guests, time, and display, which

improved the quality of the data set by making it easier to analyze and by adding more valuable output to the final analysis.

After receiving the data in the form of survey responses, the data was in raw form, which can be a quality issue. To improve the data, we added more tables based on the number of responses to each creator based on the type of content each creator produces, as well as the percentage for each category, making the data more relevant and easier to analyze. In the end, we have utilized the data sets that we have obtained from a variety of sources in order to get the final score table data set based on each city and content type. This was accomplished through the utilization of regression analysis techniques such as linear regression and logistic regression, which enabled us to predict which cities and which creators will be most profitable to host live shows for which type of category. For the purposes of analysis, this score data set has contributed the most significant value additions to the overall quality of the information.

6. Analysis and results

We collect and analyze data in an objective way from YouTube Analytics, which focuses on the real-time records about the videos' views. And we also collect and analyze data in a subjective way from questionnaire results, which focuses on the viewers' preference for these three creators' video content.

1) Analysis for YouTube Analytics data

For the most popular cities with better fundamental infrastructure, these creators all have possibilities to perform on the stage, and we are going to figure out the most profitable business plan for the company and sponsors, they can refer this score to make a strategic planning.

We predict and score the creators' performance style in each city, to be specific, in Delhi, we should hold a fan convention for Vasu, and let him perform roast and standing comedy together, because he has the highest score with these two styles in Delhi. And in Mumbai, the company can also hold a fan convention for Vasu with performance of roast and standing comedy can bring higher profit to the company.

What's more, the company can cooperate with some domestic sponsors from Hyderabad, Bengaluru and Ahmedabad to hold the concerts for the creators with higher score in this region and let them perform in the style which represents higher score.

Cities	Creators								
	Vanshaj Singh			Vasu Kainth			Adit Minocha		
	Funny	**Roast**	**Music**	**Music**	**Funny**	**Roast**	**Music**	**Roast**	**Funny**
Delhi	1832.825	1729.697	1639.443	2029.632	2086.652	2053.477	1391.571	1767.232	1833.374
Mumbai	1770.402	1667.273	1577.019	1983.168	2040.188	2007.013	1391.224	1766.884	1833.026
Hyderabad	1644.741	1541.612	1451.358	1867.817	1924.838	1891.663	1269.547	1645.207	1711.349
Bengaluru	1673.97	1570.841	1480.587	1886.233	1943.254	1910.079	1321.3	1696.96	1763.102
Ahmedabad	1723.059	1619.93	1529.676	1968.127	2025.148	1991.973	1266.407	1642.067	1708.209

2) Analysis for questionnaire result

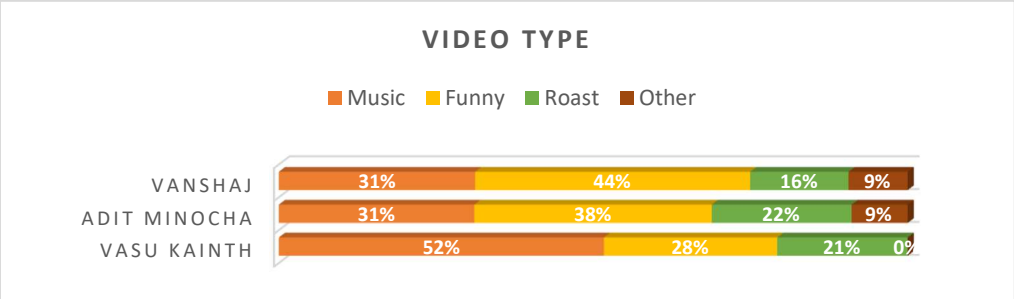
In the questionnaire, we asked a lot of question about the content in terms of Types, Duration, Thumbnails, Guests, and Fan convention/Concert. To ask about what is their direct feel of videos from these three video creators.

- Types

  For VASU KAINTH, the most famous video type is Music, among people who like Vasu's videos, 52% of the people chose Music as their favorite video type.

  For ADIT MINOCHA, the most famous video type is Funny, among people who like Adit's videos, 38% of the people chose Funny as their favorite video type.

For VANSHAJ, the most famous video type is Funny, among people who like Vanshajs's videos, 44% of the people chose Funny as their favorite video type.



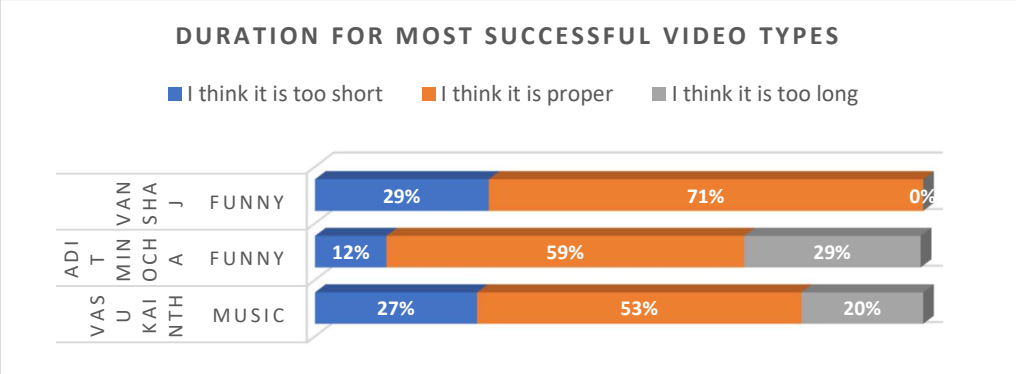
- Duration

Based on the analysis about popular video type for each creator, we analyze video duration from viewers' opinions.

For Vasu's Music videos, 53% people's choices are "I think it is proper", 27% people's choices are" I think is too short".

For Adit's Funny videos, 59% people's choices are "I think it is proper", 29% people's choices are" I think is too long".

For Vanshajs's Funny videos, 71% people's choices towards Vanshajs's video duration focuses on "I think it is proper".



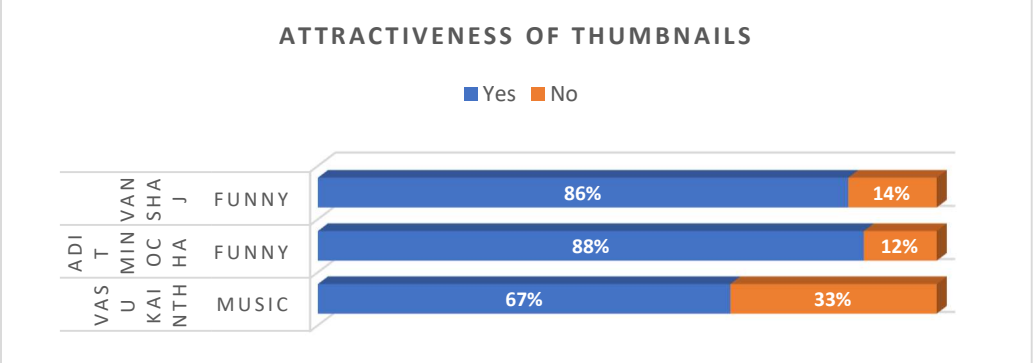
- Thumbnails

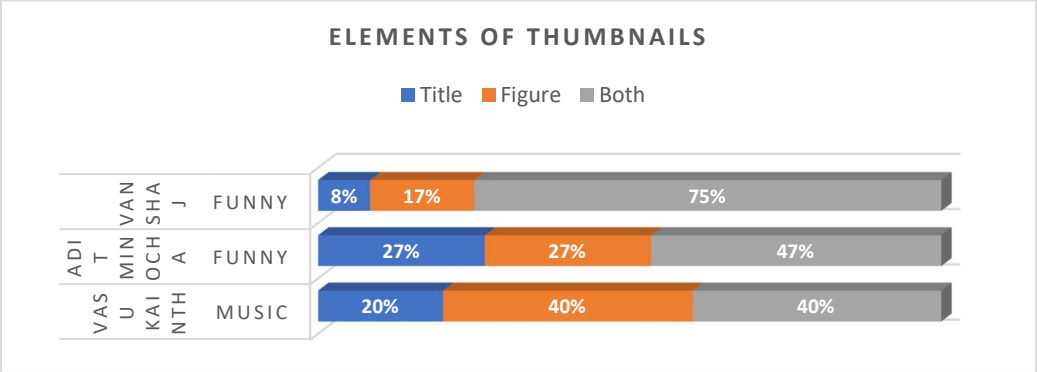
Based on the analysis about popular video type for each creator, we analyze thumbnails from viewers' opinions.

For Vasu's Music videos, 67% people choose they are attracted by the thumbnails design, and 33% people choose they are not. Among people who choose they are attracted by the thumbnails design, 40% people choose they are attracted by title and figure together, and 40% people choose they are attracted by figure only.

For Adit's Funny videos, 88% people choose they are attracted by the thumbnails design, and 12% people choose they are not. Among people who choose they are attracted by the thumbnails design, 47% people choose they are attracted by both.

For Vanshajs's Funny videos, 86% people choose they are attracted by the thumbnails design, and 14% people choose they are not. Among people who choose they are attracted by the thumbnails design, 75% people choose they are attracted by both.





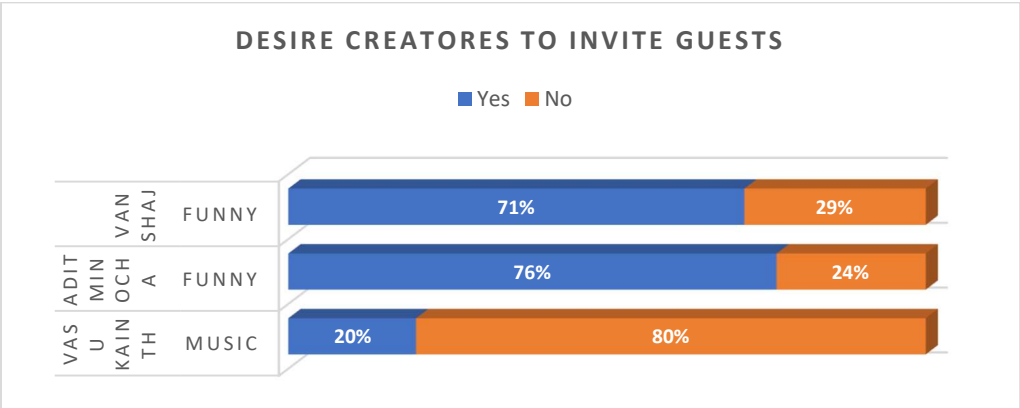
- Guests

  Based on the analysis about popular video type for each creator, we analyze if audience desire them to invite guests in their videos or not.

  For Vasu's Music videos, 80% show they don't want Vasu to invite some guests in his music video to make a show together.

  For Adit's Funny videos, 76% show they want Adit to invite some guests in his music video to make a show together. And they give some advises of the guests, Virat Kohli and Yash Tyagi.

  For Vanshajs's Funny videos, 71% show they want Vasu to invite some guests in his funny video to make videos together. And they give some advises of the guests, Ashish Chanchlani ,Vasu ,and other female guests.

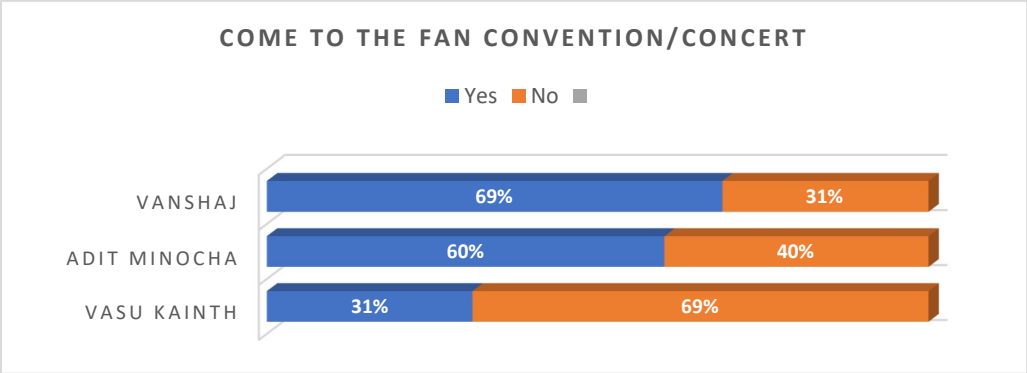


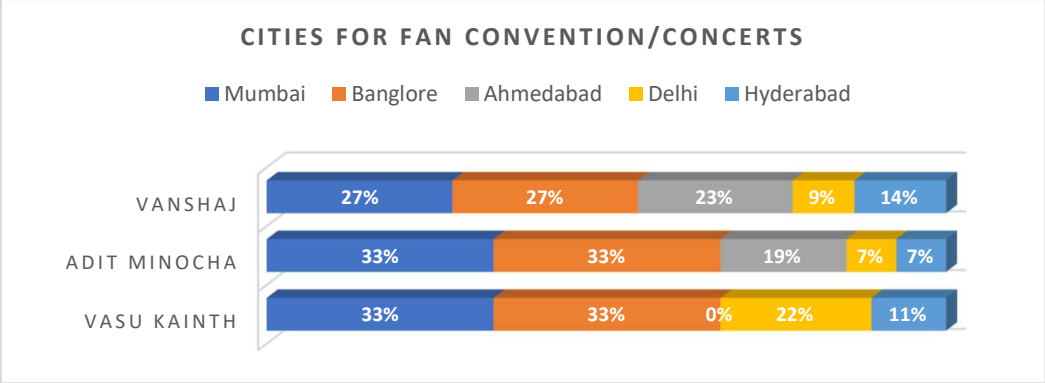
- Fan convention/Concert

  If Vasu holds a fan convention, 69% show they won't come to it, and 31% people show they will come, and most people choose Mumbai and Banglore as their desire places to hold the fan convention.

  If Adit holds a fan convention, 60% show they will come to it, and 40% people show they won't come, and most people choose Mumbai and Banglore as their desire places to hold the fan convention.

If Vanshajs holds a fan convention, 69% show they will come to it, and 31% people show they won't come, and most people choose Mumbai and Banglore as their desire places to hold the fan convention.





3) Comparison

By comparing how creators present their videos and how viewers want the videos to be, we can try to figure out the difference between the sides, and then provide some clear suggestions to these creators to improve the quality, attract more viewers, and expand their influence

- Thumbnails

  By analyzing are shown in their thumbnails and what viewers are attracted by, we figure out proper duration for successful videos.

  For the most popular videos' which are selected based on data, their thumbnails include both titles and figures. And for the favorite videos chosen by viewers, the attractive thumbnails should include both titles and figures. And for Vasu's music videos, a lot of people show they are attracted by the figures only. So we can infer that in successful videos' thumbnails, both figures and titles are important, and sometimes figures can catch viewers' eyes more.

- Video duration

  By analyzing how long creators make their videos, and viewers' opinions towards videos' duration, we figure out that the length of popular videos and viewers' tastes towards them are the same.

  For Vanshaj's Funny videos , he always makes some shorts, and the duration of the shorts is around 60s, and most people think the duration of his Funny videos is proper. For Adit's Funny videos, he always produces some long videos over 10 mins, while most of viewers think it is proper.

  We can infer that the duration for their videos right now is relatively proper, and they do not have to imitate from others.

4) Suggestions

- Considering that these video producers have some industry status and are at a stage where they need fan growth, we suggest that they need try to improve the quality of their successful video type, to keep a unique style. Like for Vasu should keep its Funny videos in the pattern shorts, not long videos.
- And if company decides to hold fan convention/concerts for these popular creators to increase the profit, we suggest the company holds for Vanshaj and Adit, because based the survey, 69% people don't want to come to Vasu's converts even it get good score based on the analysis from YouTube Analytics. We need listen to all viewers' opinions.

  For Vanshaj and Adit, we compare the score of each type for every creator and decide the best performance combination for each city.

  - For Delhi, we suggest Adit and Vanshaj perform funny shows together.
  - For Mumbai, we suggest Adit and Vanshaj perform funny shows together.
  - For Hyderabad, we suggest Adit perform Funny and Roast style shows on the stage.
  - For Bengaluru, we suggest Adit perform Funny and Roast style shows on the stage.
  - For Ahmedabad, we suggest Adit and Vanshaj perform funny shows together.

Cities	Creators								
	Vanshaj Singh			Vasu Kainth			Adit Minocha		
	Funny	**Roast**	**Music**	**Music**	**Funny**	**Roast**	**Music**	**Roast**	**Funny**
Delhi	1832.825	1729.697	1639.443	2029.632	2086.652	2053.477	1391.571	1767.232	1833.374
Mumbai	1770.402	1667.273	1577.019	1983.168	2040.188	2007.013	1391.224	1766.884	1833.026
Hyderabad	1644.741	1541.612	1451.358	1867.817	1924.838	1891.663	1269.547	1645.207	1711.349
Bengaluru	1673.97	1570.841	1480.587	1886.233	1943.254	1910.079	1321.3	1696.96	1763.102
Ahmedabad	1723.059	1619.93	1529.676	1968.127	2025.148	1991.973	1266.407	1642.067	1708.209

For the business strategy for the country host, the company and sponsors can earn more profit and lower the risk of no audience come to the concerts/fan convention.

7. External material

https://support.google.com/youtube/answer/9088722?hl=en
https://www.shopify.com/blog/6763696-youtube-analytics-10-ways-to-track-video-performance
https://www.youtube.com/watch?v=-ile8LK5juw
https://www.youtube.com/watch?v=MWLLdw6YH_0

1) Since our group decided to focus our project on analyzing YouTube creators' videos, the most important thing is to find the data source. We found out that YouTube have an organized system to collect data of posted videos in each creator's back stage called YouTube Analytics. Through these supporting articles and teaching videos, we were able to not only discover our main data source, but also export and download the datasets for further analysis.
   https://sproutsocial.com/insights/youtube-analytics/
   https://blog.hootsuite.com/youtube-analytics/
   https://clipchamp.com/en/blog/how-to-see-youtube-analytics-other-channels/
   https://zapier.com/blog/youtube-metrics/
2) As we looked into the datasets, we needed to select the attributes that are related to our project. We assumed to host a concert for the creator, and to help figure out what content of his videos is the most popular and can attract people to come. Having more insight of different metrics not only let us realize the meanings of the attributes, but also help us decide the top videos of each creator. For example, the more watch time, the more popular. However, videos have different footage, watch time is not

a good predictor. As a result, we combined Shares, Likes, Views, and Subscribers to choose the top videos in different displays, because these metrics are measured by times.

https://www.analyticsvidhya.com/blog/2021/06/linear-regression-in-machine-learning/

https://towardsdatascience.com/understanding-linear-regression-output-in-r-7a9cbda948b3

https://www.managementstudyguide.com/regression-scoring.htm

https://www.scribbr.com/statistics/linear-regression-in-r/

3) In order to choose the right method for our analysis, we need to have some basic understanding of different regression models. We use Linear regression in R for our project, because it's easier to see the relationships between multiple variables. After calculating, we discovered the most relevant metrics to the videos in our datasets. With data regarding age, gender and city the final dataset can be generated, and we also added scores to better compare top videos in different cities. These attributes are also the elements for our survey that the information allows us to ask questions related to our observation.

8. Additional data and analysis

If we had more time, we would refine the granularity of data collected manually for the most popular videos and expanding the scope of the questionnaire, to enhance the data resources and analysis process. Here are some ways to do the additional data for the further analysis.

1) Refine the granularity of data about content of the most popular videos

   For the top popular videos' content for each creator, we just collected the types, video duration, guests, thumbnails about the videos. For these attributes, we can refine the granularity of them. What's more, we can add some more attributes to analyze the contents.

   - Add more choices for types and thumbnails.

     For additional analysis, we will add more elements in types and thumbnails. For a video, it contains a lot of elements, it may be both a funny video and a music video, or may be both a funny video and a roast video. We will list all types included in the video as many as possible to dig out a pattern for successful video.

     And for thumbnails, we will refine the categories, not just title, figure, and both. We will add more categories for that, such as colors of the title, size of the title, size of the figure, the number of figures in the thumbnails.

     That way, we can increase the accuracy of the categories of the displays and thumbnails.

   - Analyze more attributes about content of videos

     For digging out a general pattern for successful videos and apply it in some video creators with similar styles, we need figure out more attributes to measure the contents of videos besides types, video duration, guests, thumbnails about the videos.

     We will collect and analyze the effect used in the video, some videos use special voice changer, some videos use image processing, and some videos add other people's video clips, they just use special ways to express the emotion and ideas. And video viewers may be attracted by these special effects used in videos.

     Additionally, we will collect and analyze the background music used in videos. The languages of music can affect viewers feeling, some people prefer English songs to express emotions, while some other people prefer Indian songs. And some creators love to use ambient as their background music. More in-depth, we will collect and analyze the styles of background music, pop music, country music, idyllic style… to analyze matched music style for a specific type of video.

2) Expanding the scope of the questionnaire

   To match with the increased choices of some attributes and more attributes added to analyze the content, we will redesign the survey to ask about viewers' opinions about them. We will add some questions about the background music and video effects.

And we will expand the answer choices for some existing question, for example, the multiple choices question “how are you be attracted by the thumbnail?”, we will provide choices of texts of title, colors of the title, size of the title, size of the figure, and the number of figures.

3) Refine the model

For the final formular to score the combination of creators and geography, now we just multiply creator score, subscribers score, displays score, and geography score together. For the further analysis, we can refine the model to increase the accuracy rate, and simplify the distributed computing steps.

For the additional data, we will analyze the content of videos from more dimensions, if we can dig out some most common elements in most popular videos for a specific video type, we can let some other video makes with similar styles use the most common elements in their videos to increase their videos’ views, like and watch time, in order to increase the revenues of videos.