EXPLORING THE IMPORTANT FACTORS BEHIND YOUTUBE VIDEO SELECTION AND THEIR EFFECTS IN DIFFERENT CATEGORIES

VIDYASAGAR UNIVERSITY

SEMESTER: 6th Sem

PAPER: DSE 4 (project)

REG. NO : VU2211605306 of 2022-23

ROLL NO: 1126116220152

Haldia Government College

Department of Statistics

ACKNOWLEDGMENT

The success and final outcome of this project depended on guidance and help from many people, and I am extremely grateful to have received this support throughout my project. I sincerely thank them for giving me the opportunity to work under the Department of Statistics at Haldia Government College and for providing all the support and guidance I needed, which enabled me to complete the project on time. I am also deeply grateful to Dr. Shyamsundar Sahoo, Mr. Sibsankar Karan, Mr. Tanmay Kumar Maity, and Mr. Bijitsh Halder for their excellent guidance, instructions, continuous encouragement, and keen interest in my project's progress. Lastly, I want to thank all my friends who helped and supported me in a timely manner, which contributed to the successful completion of this project.

SOURISH HALDER

CONTENTS

INTRODUCTION :-

In the era of digital media, YouTube has become one of the most widely used platforms for watching videos. From entertainment and education to news and personal development, people of all age groups and backgrounds rely on YouTube for a wide variety of content. With millions of videos uploaded every day, users are constantly choosing what type of content to watch. These choices are not always random; they are often influenced by personal traits and viewing habits. Understanding the factors that influence people's selection of YouTube video categories can offer valuable insights into human behavior in the digital age.

This project explores the relationship between several user characteristics and their preferences for different YouTube video categories. Specifically, it focuses on how the independent variables, such as **age**, **gender**, **occupation**, **watch frequency**, **use of subtitles**, and **preferred video length,** are related to the types of videos people choose to watch. These variables represent a mix of demographic factors and personal viewing habits, which together may shape user preferences in meaningful ways.

To examine these relationships, I used a combination of statistical methods. The **Chi-Square Test of Independence** was applied to determine whether there are significant associations between the independent variables and the chosen video categories. Additionally, I conducted a **test of homogeneity of proportions** to compare preferences across different subgroups, followed by **pair-wise tests** to pinpoint specific differences between them. Although I also attempted to use **multinomial logistic regression** to model and predict video category choices based on these variables, the results were not strong enough to draw reliable conclusions likely due to data limitations or complexity in user preferences that could not be captured fully by the model.

Overall, this project is aimed at identifying the most influential factors in YouTube video category preferences and understanding how personal and behavioral characteristics relate to online content consumption. These findings can have practical implications for content creators, educators, platform designers, and others interested in the intersection of media, technology, and human behavior.

OBJECTIVE :-

1. **To identify whether there is a significant association** between user characteristics (age, gender, occupation, etc.) and their preferred YouTube video categories using the Chi-Square Test of Independence.

2. **To compare proportions of video category preferences** across different user groups using the Test of Homogeneity, and determine whether preferences vary significantly between these groups.

3. **To investigate which specific groups significantly differ in their video category preferences** using pairwise comparison tests and to rank them according to these differences.

DATA SOURCE :-

The data used in this study was collected through a **Google Form Survey** designed specifically for this project. The survey was distributed online to a diverse group of participants to gather information about their YouTube viewing habits and preferences. Respondents were asked to provide details on both demographic and behavioral factors believed to influence their choice of YouTube video categories.

DATA DESCIPTION :-

A total of **349** valid responses were collected, ensuring a range of perspectives across different age groups and occupations. The responses were prepared for analysis using appropriate statistical tools.

This primary data collection method helped my project to ensure that the dataset was relevant, up-to-date, and specifically suited to the research objectives of the project.

❖ FACTORS DESCRIPTIONS :

- **Age**
  Categorizes respondents based on their age group. Age can influence content preferences, such as younger users leaning toward entertainment and older users favoring educational or news content.

- **Gender**
  Refers to the gender identity of the respondent. Gender may play a role in video preferences due to differences in interests or content engagement styles.

- **Occupation**
  Indicates the current role or employment status of the respondent. Occupation may reflect lifestyle and time availability, which can affect what type of content is consumed.

- **Watch Frequency**
  Measures how often the respondent watches YouTube videos (e.g., Daily, Weekly, Occasionally). Higher frequency users may have broader exposure to various categories compared to occasional viewers.

- **Subtitle**
  Identifies whether the respondent usually watches videos with subtitles (Yes/No). This can reflect language preferences or accessibility needs and may influence content selection.

- **Preferred Video Length**
  Refers to the typical length of videos that a respondent prefers (e.g., less than 5 minutes, 5–15 minutes, more than 15 minutes). Video length preference may indicate the viewer's attention span, purpose of watching, or available free time.

- **Preferred YouTube Video Category** *(Dependent Variable)* Represents the primary type of content the respondent chooses to watch on YouTube. The video categories are Music , Education , Entertainment , News , Gaming , Vlogs , Comedy , Technical , Creative , and Cooking. This is the main outcome variable that the study aims to understand and explain based on the other factors.

METHODOLOGY :-

❖ Pearson's chi-square for independence :

The Chi-square test of independence checks whether two variables are likely to be related or not. We have counts for two categorical or nominal variables. We also have an idea that the two variables are not related. The test gives us a way to decide if our idea is plausible or not.

This is the motivation behind the hypothesis for the Chi-Square Test of Independence:

H_0: In the population, the two categorical variables are independent.

H_1: In the population, the two categorical variables are dependent.

The Chi-Square test statistic is calculated as follows :

$$\chi^2 = \sum_{i=1}^{rc} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{(r-1)(c-1)}, \quad \text{under null hypothesis}$$

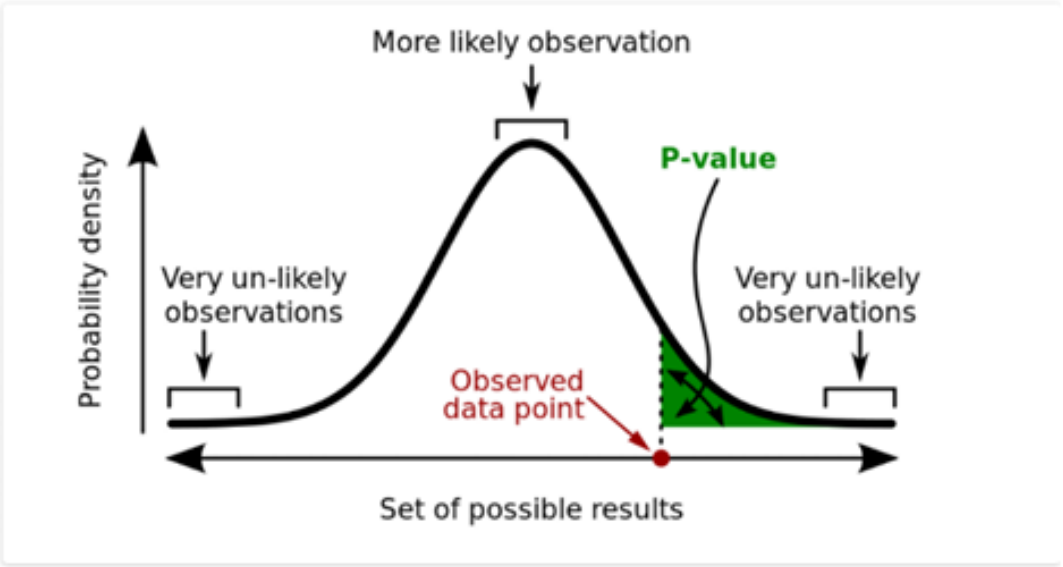
where O_i = the observed frequency
E_i = the expected frequency
r = no. of rows
c = no. of columns.

Under the null hypothesis and certain conditions, the test statistic follows a Chi-Square distribution with degrees of freedom equal to $(r-1)(c-1)$, where r is the number of rows and c is the number of columns. We leave out the mathematical details to show why this test statistic is used and why it follows a Chi-Square distribution. As we have done with other statistical tests, we make our decision by either comparing the value of the test statistic to a critical value (rejection region approach) or by finding the probability of getting this test statistic value or one more extreme (p-value approach).

❖ P-Value :

In statistics, the p-value is the probability of obtaining results at least as extreme as the observed results of a statistical hypothesis test, assuming that the null hypothesis is correct. The p-value is used as an alternative to rejection points to provide the smallest level of significance at which the null hypothesis would be rejected. A smaller p-value means that there is stronger evidence in favour of the alternative hypothesis. We know that the P-value is a statistical measure that helps to determine whether the hypothesis is correct or not. P-value is a number that lies between 0 and 1. The level of significance(α) is a predefined threshold that should be set by the researcher. It is generally fixed as 0.05.

P-value	Decision
P-value > 0.05	The result is not statistically significant and hence don't reject the null hypothesis.
P-value < 0.05	The result is statistically significant. Generally, reject the null hypothesis in favor of the alternative hypothesis.
P-value < 0.01	The result is highly statistically significant, and thus reject the null hypothesis in favor of the alternative hypothesis.



❖ Test of Homogeneity of Proportions :

The test of homogeneity of proportions checks whether different populations (or groups) share the same proportion for a categorical variable or not.

Here we test,

H_0: In the population, all the proportions are equal, i.e $P_1=P_2=\ldots=P_k=P_0$ (say)

H_1: At least one population proportion is different, i.e $P_i \neq P_0$ for at least one i.

Under the null hypothesis and certain conditions, the test statistic follows a Chi-Square distribution with degrees of freedom equal to k-1, where k is the number of mutually exclusive and exhaustive classes. We leave out the mathematical details to show how and why the test statistic is used and why it follows a Chi-Square distribution. As we have done with other statistical tests, we make our decision by either comparing the value of the test statistic to a critical value (rejection region approach) or by finding the probability of getting this test statistic value or one more extreme (p-value approach).

❖ Test for Proportion (Z-test for the equality of two proportions) :

Suppose we have two random samples of size n_1 and n_2 respective proportion p_1 and p_2 calculated.

Here we test,

H_0: $p_1 = p_2$ against H_1: $p_1 > p_2$ or $p_1 < p_2$

The test statistic for testing the hypothesis is calculated as follows

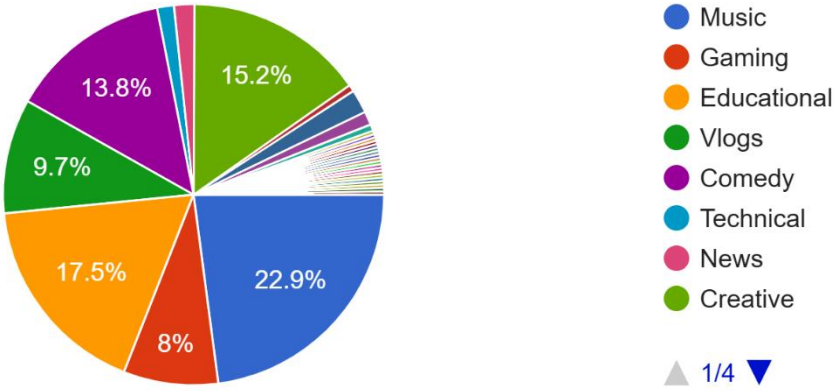
$$\mathbf{Z} = \frac{\boldsymbol{p_1 - p_2}}{\{\boldsymbol{p}\,(\,\mathbf{1}-\boldsymbol{p}\,)\left(\frac{\mathbf{1}}{\boldsymbol{n_1}}+\frac{\mathbf{1}}{\boldsymbol{n_2}}\right)\}}$$

Where, $\mathbf{p} = \frac{\boldsymbol{p_1 n_1 + p_2 n_2}}{\boldsymbol{n_1 + n_2}}$

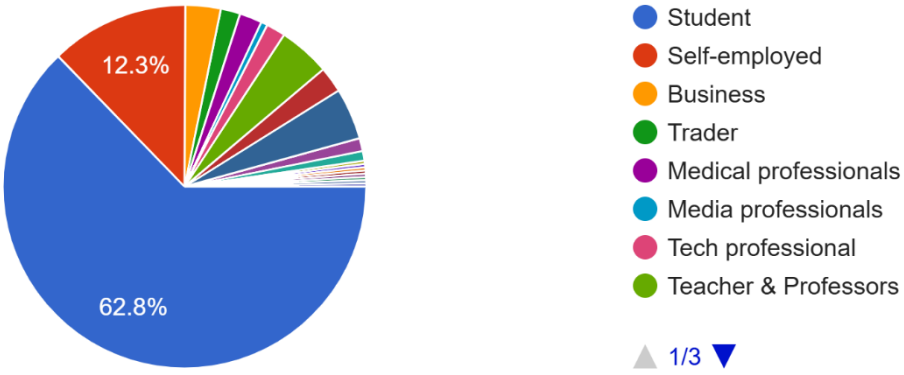
Under the null hypothesis and certain conditions, the test statistic follows a standard normal distribution. As we have done with other statistical tests, we make our decision by either comparing the value of the test statistic to a critical value (rejection region approach) or by finding the probability of getting this test statistic value or one more extreme (p-value approach).

RESULT AND ANALYSIS :-

❖ Graphical Representation :



From the above pie chart we can see that the maximum responses of the respondents for my project chose music, education, gaming , vlogs, creative, comedy, and technical videos. So, for not getting enough responses on the other video categories, I continue my analysis specifically on these video categories.



From the above pie chart we can observe that the maximum respondents for my project are students and self-employed persons. So I analysed this project giving importance to students and self-employed.

❖ Chi-square for independence :

FOR MUSIC

I've checked the association between the dependent variable 'Music' and the other independent factors by using the Chi-square test. The result is given as follows :

Table 1: The results from the Chi-square test for Music videos

FACTORS	CHI-SQUARE VALUE	DF	P-VALUE
AGE	0.82079	2	0.6634
GENDER	2.059	1	0.1513
OCCUPATION	21.899	11	0.1105
WATCH FREQUENCY	14.088	3	< 0.01
SUBTITLE	3.7876	1	0.05163
VIDEO LENGTH	43.122	3	< 0.01

Here we can see that the P-values, for the factors age , gender , occupation are greater than $\alpha = 0.05$. Hence, there is no association of these factors with the dependent variable, denoted as 'Music'.

On the other hand, the P-values for the factor watch frequency and video length are less than $\alpha = 0.05$. And hence, these factors have an association with the dependent variable, 'Music'. While the factor subtitle has a p-value nearly 0.05, from which we can comment that there is a moderate association with the dependent variable, 'Music'.

FOR EDUCATION

Now I've checked the association between the dependent variable 'Education' and the other independent factors by using Chi-square test. The result is given as follows:

Table 2 : The results from Chi-square test for Education videos

FACTORS	CHI-SQUARE VALUE	DF	P-VALUE
AGE	2.3781	2	0.3045
GENDER	0.22056	1	0.6386

FACTORS	CHI-SQUARE VALUE	DF	P-VALUE
OCCUPATION	21.57	11	0.1196
WATCH FREQUENCY	10.488	3	0.01484
SUBTITLE	1.6352	1	0.201
VIDEO LENGTH	4.5434	3	0.0429

Here we can clearly see that the P-values, for the factors age , gender , occupation , subtitle are greater than $\alpha = 0.05$. Hence, there is no association of these factors with the dependent variable, denoted as 'Education'.

On the other hand, the P-values for the factor watch frequency and video length are less than $\alpha = 0.05$. And hence, these factors have an association with the dependent variable, 'Education'.

FOR CREATIVITY

At last, I've checked the association between the dependent variable 'Education' and the other independent factors by using Chi-square test. The result is given as follows:

Table 3 : The results from Chi-square test for Creativity videos

FACTORS	CHI-SQUARE VALUE	DF	P-VALUE
AGE	1.7737	2	0.412
GENDER	6.4079	1	0.01136
OCCUPATION	72.418	11	< 0.01
WATCH FREQUENCY	59.072	3	< 0.01
SUBTITLE	15.764	1	< 0.01
VIDEO LENGTH	8.0983	3	0.04402

Here we can clearly see that the P-values for the factor age are greater than $\alpha = 0.05$. Hence, there is no association of the factor age with the dependent variable, denoted as 'Creativity'.

On the other hand, the P-values for the factors gender, occupation, subtitle , watch frequency, and video length are less than $\alpha = 0.05$. And hence, these factors have an association with the dependent variable, 'Creativity'.

❖ Homogeneity Test of Proportions for different groups :

Defining proportions for different groups

AGE

P1= proportion of young persons

P2= proportion of middle aged persons

P3= proportion of old persons

GENDER

P1= proportion of male

P2= proportion of female

WATCH FREQUENCY

P1= proportion of watching a few times in a week

P2= proportion of watching daily

P3= proportion of watching once a week

P4= proportion of watching rarely

SUBTITLE

P1= proportion watching with subtitle

P2= proportion watching without subtitle

VIDEO LENGTH

P1= proportion of length between 10-20 min

P2= proportion of length between 5-10 min

P3= proportion of length less than 5 min

P4= proportion of length more than 20 min

OCCUPATION

P1= proportion of artist

P2= proportion of business

P3= proportion of govt. service

P4= proportion of house wife

P5= proportion of media professionals

P6= proportion of medical professionals

P7= proportion of student

P8= proportion of teacher and professors

P9= proportion of tech professionals

P10= proportion of trader

P11= proportion of self employed

P12= proportion of private sector employee

FOR MUSIC

I've tested homogeneity of proportions for the dependent variable 'Music' videos with the other independent factors by using the homogeneity test of proportions. The result is given as follows:

Table 4 : The results from the homogeneity test for Music videos

FACTORS	No. Of Proportions	DF	P-VALUE
AGE	3	2	0.6634
GENDER	2	1	0.1182
OCCUPATION	12	11	0.1105
WATCH FREQUENCY	4	3	< 0.01
SUBTITLE	2	1	0.05163
VIDEO LENGTH	4	3	< 0.01

Here we can see that the P-values, for the factors age , gender , occupation are greater than $\alpha = 0.05$. Hence, we can say that the proportions belonging to these factors are equal, which implies that the effect of music videos on the different sub-factors of the factors is equal.

On the other hand, the P-values for the factors "watch frequency" and "video length" are less than $\alpha = 0.05$. And hence, the proportions belonging to these factors are not all equal, which implies that the effect of music on the different sub-factors of the factors is not all the same. While the factor subtitle has a p-value nearly 0.05, from which we can comment that the effect of music videos on the different sub-factors of the factor subtitle are closely equal.

FOR EDUCATION

Now I've tested homogeneity of proportions for the dependent variable 'Education' videos with the other independent factors by using the homogeneity test of proportions. The result is given as follows:

Table 5 : The results from homogeneity test for Education videos

FACTORS	No. Of Proportions	DF	P-VALUE
AGE	3	2	0.3045
GENDER	2	1	0.5419
OCCUPATION	12	11	0.1196
WATCH FREQUENCY	4	3	0.01484
SUBTITLE	2	1	0.1559
VIDEO LENGTH	4	3	0.2085

Here, we can see that the P-values for the factors age, gender, occupation, subtitle, and video length are greater than $\alpha = 0.05$. Hence, we can say that the proportions belonging to these factors are equal, which implies that the effect of education videos on the different sub-factors of the factors is equal.

On the other hand, the P-value for the factor watch frequency is less than $\alpha = 0.05$. Hence, the proportions belonging to this factor are not all equal, implying that the effect of education videos on the different sub-factors of the factor is not uniform.

FOR CREATIVITY

I've tested homogeneity of proportions for the dependent variable 'Creativity' videos with the other independent factors by using the homogeneity test of proportions. The result is given as follows:

Table 6 : The results from the homogeneity test for Creativity videos

FACTORS	No. Of Proportion	DF	P-VALUE
AGE	3	2	0.412
GENDER	2	1	< 0.01
OCCUPATION	12	11	< 0.01
WATCH FREQUENCY	4	3	< 0.01
SUBTITLE	2	1	< 0.01
VIDEO LENGTH	4	3	0.04402

Here we can clearly see that the P-value for the factor age is greater than $\alpha = 0.05$. Hence, we can say that the proportions belonging to this factor are equal, which implies that the effect of creativity videos on the different sub-factors of the factor is equal.

On the other hand, the P-values for the factors gender , occupation , watch frequency , subtitle, and video length are less than $\alpha = 0.05$. And hence, the proportions belonging to these factors are not all equal which implies that the effect of creativity videos on the different sub-factors of the factors is not the same.

FOR GAMING

I've tested homogeneity of proportions for the dependent variable 'Gaming' videos with the other independent factors by using the homogeneity test of proportions.The result is given as follows:

Table 7 : The results from the homogeneity test for Gaming videos

FACTORS	No. Of Proportion	DF	P-VALUE
AGE	3	2	0.1008

FACTORS	No. Of Proportion	DF	P-VALUE
GENDER	2	1	< 0.01
OCCUPATION	12	11	0.6417
WATCH FREQUENCY	4	3	0.02607
SUBTITLE	2	1	0.09751
VIDEO LENGTH	4	3	0.081

Here we can see that the P-values, for the factors age , occupation , subtitle and video length are greater than $\alpha = 0.05$. Hence, we can say that the proportions belonging to this factor are equal, which implies that the effect of gaming videos on the different sub-factors of the factor is equal.

On the other hand, the P-values for the factors gender , watch frequency are less than $\alpha = 0.05$. And hence, the proportions belonging to these factors are not all equal, which implies that the effect of gaming videos on the different sub-factors of the factors is not all the same.

FOR NEWS

I've tested homogeneity of proportions for the dependent variable 'News' videos with the other independent factors by using the homogeneity test of proportions.The result is given as follows:

Table 8 : The results from the homogeneity test for News videos

FACTORS	No. Of Proportion	DF	P-VALUE
AGE	3	2	< 0.01
GENDER	2	1	0.3804
OCCUPATION	12	11	0.1564
WATCH FREQUENCY	4	3	0.7497
SUBTITLE	2	1	0.348
VIDEO LENGTH	4	3	0.9058

Here we can clearly see that the P-values, for the factors gender , occupation , watch frequency , subtitle, video length are greater than $\alpha = 0.05$. Hence , we can say that the proportions belonging to this factor are equal which implies that the effect of news videos on the different sub-factors of the factor is equal.

On the other hand, the P-values for the factors age is less than $\alpha = 0.05$. And hence, the proportions belonging to these factor are not all equal which implies that the effect of news videos on the different sub-factors of the factors are not all same.

FOR COOKING

I've tested homogeneity of proportions for the dependent variable 'Cooking' videos with the other independent factors by using the homogeneity test of proportions.The result is given as follows:

Table 9 : The results from the homogeneity test for Cooking videos

FACTORS	No. Of Proportion	DF	P-VALUE
AGE	3	2	< 0.01
GENDER	2	1	0.1193
OCCUPATION	12	11	< 0.01
WATCH FREQUENCY	4	3	< 0.01
SUBTITLE	2	1	0.7104
VIDEO LENGTH	4	3	0.4001

Here we can clearly see that the P-values, for the factors gender , subtitle, video length are greater than $\alpha = 0.05$. Hence, we can say that the proportions belonging to this factor are equal which implies that the effect of cooking videos on the different sub-factors of the factor is equal.

On the other hand, the P-values for the factors age , occupation , watch frequency are less than $\alpha = 0.05$. And hence, the proportions belonging to these factors are not all equal which implies that the effect of cooking videos on the different sub-factors of the factors is not the same.

FOR HEALTH

I've tested homogeneity of proportions for the dependent variable 'Health' videos with the other independent factors by using the homogeneity test of proportions.The result is given as follows:

Table 10 : The results from the homogeneity test for Health videos

FACTORS	No. Of Proportion	DF	P-VALUE
AGE	3	2	0.1001
GENDER	2	1	0.9703
OCCUPATION	12	11	< 0.01
WATCH FREQUENCY	4	3	0.9439
SUBTITLE	2	1	0.6166
VIDEO LENGTH	4	3	0.04362

Here we can see that the P-values for the factors gender, watch frequency, and subtitle are greater than $\alpha = 0.05$. Therefore, we can conclude that the proportions associated with these factors are equal, implying that the effect of health videos on the different subfactors of these factors is also equal. On the other hand, the P-values for the factors occupation and video length are less than $\alpha = 0.05$. Consequently, the proportions associated with these factors are not all equal, indicating that the effect of health videos on the different subfactors of these factors is not the same.

LIMITATION OF TEST OF HOMOGENEITY OF PROPORTIONS

Since the test of homogeneity for k proportions is used to determine whether multiple groups share the same proportion for a particular characteristic, it does not indicate which specific groups differ from each other. While this global test is useful for detecting overall differences among groups, its interpretation becomes limited, especially when the goal is to understand the nature or direction of these differences.

To overcome this limitation, I perform pairwise tests of proportions, comparing two groups at a time. These pairwise comparisons helped me identify whether there was a significant difference and allowed me to rank the proportions meaningfully. From the full set of factor levels, I performed pairwise comparisons only for those where the global null hypothesis was rejected, indicating that not all proportions are equal. This focused approach ensures that follow-up testing is meaningful and avoids unnecessary comparisons. By arranging the proportions in increasing order based on the results of the pairwise tests, I obtained a clearer picture of the relative standing of each group.

❖ Pairwise test of proportions for different groups :

For Music

Table 11 : The results from pair-wise test for Music videos

FACTORS	PROPORTIONS	RANK	SUMMARY
WATCH FREQUENCY	P1= proportion of people watching a few times in a week	1	P1=P2>P3=P4
	P2= proportion of people watching daily	1	
	P3= proportion of people watching once a week	2	
	P4= proportion of people watching rarely	2	
VIDEO LENGTH	P1= proportion of length between 10-20 min	1	P1>P2>P3>P4
	P2= proportion of length between 5-10 min	2	
	P3= proportion of length less than 5 min	3	
	P4= proportion of length more than 20 min	4	

For Education

Table 12 : The results from pair-wise test for Educational videos

FACTORS	PROPORTIONS	RANK	SUMMARY
WATCH FREQUENCY	P1= proportion of people watching a few times in a week	1	P1=P2=P3>P4
	P2= proportion of people watching daily	1	
	P3= proportion of people watching once a week	1	
	P4= proportion of people watching rarely	2	

For Creativity

Table 13: The results from the pair-wise test for Creative videos

FACTORS	PROPORTIONS	RANK	SUMMARY
GENDER	P1= proportion of male	1	P1>P2
	P2= proportion of females	2	
OCCUPATION	P1= proportion of artists	1	
	P2= proportion of business	2	
	P3= proportion of the government. service	2	
	P4= proportion of housewife	1	
	P5= proportion of media professionals	4	
	P6= proportion of medical professionals	3	
	P7= proportion of students	2	

FACTORS	PROPORTIONS	RANK	SUMMARY
OCCUPATION	P8= proportion of teachers and professors	3	P1=P4>P2=P3=P7>P6=P8>P5=P9=P10=P11=P12
	P9= proportion of tech professionals	4	
	P10= proportion of trader	4	
	P11= proportion of self employed	4	
	P12= proportion of private sector employee	4	

For Gaming

Table 14 : The results from pair-wise test for Gaming videos

FACTORS	PROPORTIONS	RANK	SUMMARY
GENDER	P1= proportion of male	2	P2>P1
	P2= proportion of female	1	
WATCH FREQUENCY	P1= proportion of people watching a few times in a week	1	P1=P3=P4>P1
	P2= proportion of people watching daily	2	
	P3= proportion of people watching once a week	1	
	P4= proportion of people watching rarely	1	

For News

Table 15 : The results from pair-wise test for News videos

FACTORS	PROPORTIONS	RANK	SUMMARY
AGE	P1= proportion of young persons	2	P2=P3 >P1
	P2= proportion of middle aged persons	1	
	P3= proportion of old persons	1	

For Cooking

Table 16 : The results from pair-wise test for Cooking videos

FACTORS	PROPORTIONS	RANK	SUMMARY
AGE	P1= proportion of young persons	1	P1=P2>P3
	P2= proportion of middle aged persons	1	
	P3= proportion of old persons	2	
OCCUPATION	P1= proportion of artist	3	
	P2= proportion of business	4	
	P3= proportion of govt. service	4	
	P4= proportion of house wife	1	
	P5= proportion of media professionals	4	

FACTORS	PROPORTIONS	RANK	SUMMARY
OCCUPATION	P6= proportion of medical professionals	3	P4>P7=P8>P1=P6>P2=P3=P5=P9=P10=P11=P12
	P7= proportion of student	2	
	P8= proportion of teacher and professors	2	
	P9= proportion of tech professionals	4	
	P10= proportion of trader	4	
	P11= proportion of self employed	4	
	P12= proportion of private sector employee	4	
WATCH FREQUENCY	P1= proportion of people watching a few times in a week	2	P2>P1=P3=P4
	P2= proportion of people watching daily	1	
	P3= proportion of people watching once a week	2	
	P4= proportion of people watching rarely	2	

For Health

Table 17 : The results from pair-wise test forvideos

FACTORS	PROPORTIONS	RANK	SUMMARY
OCCUPATION	P1= proportion of artist	3	P6>P4=P7=P8>P1>P2>P3>P5>P9>P10>P11>P12
	P2= proportion of business	3	
	P3= proportion of govt. service	3	
	P4= proportion of house wife	2	
	P5= proportion of media professionals	3	
	P6= proportion of medical professionals	1	
	P7= proportion of student	2	
	P8= proportion of teacher and professors	2	
	P9= proportion of tech professionals	3	
	P10= proportion of trader	3	
	P11= proportion of self employed	3	
	P12= proportion of private sector employee	3	
VIDEO LENGTH	P1= proportion of length between 10-20 min	1	P1=P4>P2>P3
	P2= proportion of length between 5-10 min	2	
	P3= proportion of length less than 5 min	3	
	P4= proportion of length more than 20 min	1	

CONCLUSION

This study aimed to analyze the main factors influencing the preferred categories of YouTube videos among users, focusing on demographic and behavioral variables such as age, gender, occupation, watch frequency, video length preference, and the presence of subtitles. Using a combination of the Chi-square test of independence, the test of homogeneity of proportions, and pair-wise proportion tests, I tried to examine how these variables relate to different video categories.

The results indicate that no single factor consistently determines video category preference across all categories. However, age, gender, and occupation emerged as the most frequently significant variables, though not in every category. Interestingly, in several cases where a strong association was expected, such as between certain occupations and video types, the statistical results did not support these assumptions, suggesting a more complex or weaker relationship than initially thought.

Furthermore, the test of homogeneity showed that for some factors, the proportions across groups were similar, while for others, there were significant differences. Pair-wise comparisons of those with different proportions revealed similar ranking patterns in most cases, reinforcing the idea that while differences exist, they often follow a consistent pattern.

Overall, the findings highlight the complex and varied nature of user preferences on YouTube. While demographic and behavioral factors like age, gender, and occupation do influence preferences, their effects are neither consistent nor universally significant across all video categories. This suggests that user preferences are shaped by a combination of factors rather than any single dominant variable.

REFERENCES :-

1. Fundamental of Statistics,Vol-1, Vol-2 by A.M Gun , M.K Gupta, B. Dasgupta.

2. An Introduction to Categorical Data Analysis by ALAN AGRESTI

3. An introduction to probability and statistics : A. K. Md. Ehsanes Saleh, Vijay K. Rohatgi

4. Wikipedia

APPENDIX :-

For Chi-Square Test

```
#chi-square test for dependent_variable vs independent variable
#creating a binary indicator variable that identifies whether a video belongs to the dependent video category.
data$dependent_flag <- ifelse(data$Video.Category == "dependent_variable", 1, 0)
table_data <-table(data$independent_variable , data$dependent_flag)
chisq.test(table_data)
```

For Test Of Homogeneity

```
#test for proportions
#for dependent_variable vs independent variable
#creating a variable for successes in each group
successe <- table_data1[, "1"]
#creating a variable for totals in each group
total <- rowSums(table_data)
prop.test(x = successe , n = total)
```

For Pair-wise Test

```
#for dependent_variable vs independent variable
pvals <- c()
for (i in 1:(length(successe)-1))
{
  for (j in (i+1):length(successe))
{
  test <- prop.test(x = c(successe[i], successe[j]),n = c(total[i], total[j]),alternative = "greater")
pvals <- c(pvals, test$p.value)
 }
}
```

Analysis Of Viewer Preferences In YOUTUBE Video Consumption

This survey is conducted for a research project aimed at understanding what kind of YouTube videos viewers prefer and why. Your information will remain confidential and will only be used for academic purposes. The form will take less than 5 minutes to complete. Thank you for your participation.

* Indicates required question

1. **Email Address** *

_____

2. **Age** *

_____

3. **Gender** *

Mark only one oval.

- ◯ Male
- ◯ Female
- ◯ Others

4. **Occupation** *

Mark only one oval.

- Student
- Self-employed
- Business
- Trader
- Medical professionals
- Media professionals

Tech professional

Teacher & Professors

Government service

Artist

Other: _____

5. **How often do you watch YouTube ?** * *Mark only one oval.*

Daily

A few times in a week

Once a week

Rarely

6. **What type of YouTube videos do you watch the most ?** *

Mark only one oval.

Music

Gaming

Educational

Vlogs

Comedy

Technical

News

Creative

Health & wellness Other: _____

7. **What is your preferred video length ?** *

Mark only one oval.

- Less than 5 minutes
- 5 - 10 minutes 10 - 20
- minutes more than 20
- minutes

8. **Do you prefer videos with subtitles ?** *

Mark only one oval.

- Yes
- No