Savitribai Phule Pune University

K.T.H.M. COLLEGE, Nashik

(K.R.T. Arts, B.H. Commerce and A.M. Science College.)

A

Project Report On

"Streaming Insights: A Comprehensive Study of OTT Platform Usage"

T. Y. B. Sc. (Statistics) (2023-2024)

Submitted by

Sonawane Aditya Balkrushna (9007)

Under Guidance Of

Mr. Digambar B. Uphade

MARATHA VIDYA PRASARAK SAMAJ, NASHIK

K.R.T. Arts, BH Commerce & AM Science (KTHM) College, Nashik

Shivaji Nagar, Gangapur Road Tal & Dist. -Nashik
422002 0253-2571376 www.kthmcollege.ac.in

Department of Statistics

CERTIFICATE

This is to certify that, Mr. **Sonawane Aditya Balkrushna** of T.Y.B.Sc. (Statistics) have successfully completed project titled "**Streaming Insights: A Comprehensive Study of OTT Platform Usage"** satisfactorily as partial fulfilment of curriculum for T.Y.B.Sc. (Statistics) Semester-VI during academic year 2023-2024.

Place: Nashik

Date:

Mr. D. B. Uphade Project Guide	Examiner	Dr. G.S. Phad Head

Acknowledgement

We would like to express our heartfelt thanks to all those who have supported us throughout this statistics project. First and foremost, we extend our gratitude to our instructor who has provided us with valuable guidance, feedback, and motivation throughout the project. We are grateful for their insights, expertise, and support that have helped us to better understand the concepts and principles of Statistics.

We wish to thank **Dr. G. S. Phad,** Head, Department of Statistics and project guide **Mr. D. B. Uphade** for the valuable support and guidance. Thanks to the teaching faculty of our college for their support in completing the work successfully.

We would also like to thank our classmates for their helpful feedback and constructive criticism during our group discussions. Their insights and ideas have been invaluable in shaping our understanding of the project's goals and objectives.

Thank you all for your support, guidance, and encouragement throughout this project.

Index –

Introduction

OTT ('Over the top') is a technology that delivers streamed content to devices via the internet. Initially launched in India in late 2000s, OTT has gained popularity huge popularity over the past few years, and its use has grown exponentially over that period. With the increase in the availability of low-cost internet data and affordable smartphones, OTT platforms and its content have become available to a larger audience.

The Indian OTT market has been under rapid transformational growth, and the number of people who subscribe for an OTT platform increased by a staggering 70 per cent during the COVID-19 Pandemic. Indian OTT industry is brimming with local and international players such as Amazon Prime (13 million users), Netflix (11 million users), Disney+ Hotstar (300 million users), and many more, witnessing heavy competition. In such an environment, analysis of the OTT streaming patterns of users can provide value insights about different aspects of this technology.

With this motivation, a statistical analysis of OTT using patterns among citizens is performed in this project. Through this, we can gain information about the most popular OTT platforms and their use by people. Such an analysis can be of great help to the OTT service providing platforms to analyze streaming patterns among users and device strategies to increase the viewership which will in turn help them to boost their revenues. It will also be beneficial for content creators as it will help them understand which content is consumed more than other.

Through this project, we have made an effort to carry out such an analysis. Though the survey has been conducted on a relatively small space of 278 individuals, certain observations made in this study can be related to overall viewing patterns in country.

Objectives

- To compare various OTT platforms on the basis of viewership.
- To study use of OTT among different professions, age groups and income groups.
- To find most commonly viewed content on different platform.
- To analyze time spent by users streaming content.
- To study subscription patterns.
- To gain an insight into user perspective on subscription pricings.
- To analyze user experience and satisfaction on basis of usability and variety of content offered.

Questionnaire

Survey on OTT using habits among citizens.

'Over the top' (OTT) is a technology that delivers streamed content to devices via Internet. OTT has gained popularity over the past few years & it's use has grown exponentially over that time.

We, the students of T.Y.B.Sc. Statistics, HPT Arts & RYK Science College, Nashik are interested in performing a statistical analysis of the use of OTT by the citizens. We request you to fill the below questionnaire to the best of your knowledge. The data gathered through this form will be used for academic purposes only. We will also be happy & honoured to share the project report with you at the end of this survey.

* Indicates required question

Q.1. Email

Your answer

Q.2. Age*

Your answer

Q.3. Gender*

☐ Male
☐ Female ☐
Other

Q.4. Employment Status*

☐ Employed
☐ Self-employed
☐ Unemployed
☐ Student
☐ Homemaker
☐ Other: Click or tap here to enter text.

Q.5. Monthly Income (or Monthly Allowance for students) *

☐ Under ₹10000
☐ ₹10000 - ₹30000
☐ ₹30000 - ₹50000
☐ ₹50000 - ₹75000
☐ Above ₹75000

Q.6. No. of working hours per week (College & Class hours for students)*

Your answer

Q.7. How many OTT platforms do you use?*

Your answer

Q.8. Do you have any of these OTT subscriptions? Please write in 'Other' if you have subscribed to any other platform.

☐ Netflix
☐ Disney+ Hotstar
☐ Sony Liv
☐ Amazon Prime Video
☐ Jio Cinema
☐ Other: Click or tap here to enter text.

Q.9. Do you use the basic or premium subscription?

Please select responses only for the platforms to which you have subscribed.

	Basic	Premium
Netflix	☐	☐
Disney+ Hotstar	☐	☐
Sony Liv	☐	☐
Amazon Prime Video	☐	☐
Jio Cinema	☐	☐
Other	☐	☐

Q.10. Which of these do you stream most often using these platforms?

Please select responses only for the platforms to which you have subscribed.

		Movies	Web Series	Sports	Daily Soaps	Cooking Shows	Dramas	Others
Netflix	☐	☐	☐	☐	☐	☐	☐	☐
Disney+ Hotstar	☐	☐	☐	☐	☐	☐	☐	☐
Sony Liv	☐	☐	☐	☐	☐	☐	☐	☐
Amazon Prime Video	☐	☐	☐	☐	☐	☐	☐	☐
Jio Cinema	☐	☐	☐	☐	☐	☐	☐	☐
Other	☐	☐	☐	☐	☐	☐	☐	☐

Q.11. Do you use these platforms to view Indian shows or International shows? Please select responses only for the platforms to which you have subscribed.

	Indian	International	Both
Netflix	☐	☐	☐
Disney+ Hotstar	☐	☐	☐
Sony Liv	☐	☐	☐
Amazon Prime Video	☐	☐	☐
Jio Cinema	☐	☐	☐
Other	☐	☐	☐

Q.12. Do you renew your subscription every month?

☐ Yes ☐
No

Q.13. Which device do you use to stream content most often?

Please select responses only for the platforms to which you have subscribed.

	Mobile	Laptop/PC	TV	Tablet
Netflix	☐	☐	☐	☐
Disney+ Hotstar	☐	☐	☐	☐
Sony Liv	☐	☐	☐	☐
Amazon Prime Video	☐	☐	☐	☐
Jio Cinema	☐	☐	☐	☐
Other	☐	☐	☐	☐

Q.14. How much time do you use these OTT platforms per week?

	0-2 hours	2-4 hours	4-6 hours	6-8 hours	More than 8 hours
Netflix	☐	☐	☐	☐	☐
Disney+ Hotstar	☐	☐	☐	☐	☐
Sony Liv	☐	☐	☐	☐	☐
Amazon Prime Video	☐	☐	☐	☐	☐
Jio Cinema	☐	☐	☐	☐	☐
Other	☐	☐	☐	☐	☐

Q.15. How many hours do you spend weekly viewing content on OTT? (All platforms combined)*

Please enter the value in no of hours.

Your answer

Q.16. Which of these do you use more to watch movies?*

☐ OTT
☐ Theatre
☐ TV (DTH)

Q.17. Do you share your subscription(s) with friends or family?

Please select responses only for the platforms to which you have subscribed.

	Yes, I lend it to someone	Yes, I borrow it from someone	No, I use it myself
Netflix	☐	☐	☐
Disney+ Hotstar	☐	☐	☐
Sony Liv	☐	☐	☐
Amazon Prime Video	☐	☐	☐
Jio Cinema	☐	☐	☐
Others	☐	☐	☐

Q.18. What do you think about the price point of the subscription(s) that you have bought?

Please select responses only for the platforms to which you have subscribed.

	Reasonable	Cheap	Expensive
Netflix	☐	☐	☐
Disney+ Hotstar	☐	☐	☐
Sony Liv	☐	☐	☐
Amazon Prime Video	☐	☐	☐
Jio Cinema	☐	☐	☐
Others	☐	☐	☐

Q.19. How would you rate the below platforms on the basis of variety of content available?

Rate out of 5. Please select responses only for the platforms to which you have subscribed.

	5	4	3	2	1
Netflix	☐	☐	☐	☐	☐
Disney+ Hotstar	☐	☐	☐	☐	☐

Sony Liv	☐	☐	☐	☐	☐
Amazon Prime Video	☐	☐	☐	☐	☐
Jio Cinema	☐	☐	☐	☐	☐
Others	☐	☐	☐	☐	☐

Q.20. Are you satisfied with your use of the subscription(s) for the price which you paid?

Please select responses only for the platforms to which you have subscribed.

	Yes	No
Netflix	☐	☐
Disney+ Hotstar	☐	☐
Sony Liv	☐	☐
Amazon Prime Video	☐	☐
Jio Cinema	☐	☐
Others	☐	☐

Data Codes –

Q.3 Gender	
Male	1
Female	2
Other	3

Q.4 Employment Status	
Employed	1
Self-employed	2
Unemployed	3
Student	4
Homemaker	5
Others	6

Q.5 Monthly Income	
Under ₹10000	1
₹10000 - ₹30000	2
₹30000 - ₹50000	3
₹50000 - ₹75000	4
More than ₹75000	5

Q.8 OTT Platform
Netflix
Disney+ Hotstar
Sony Liv
Amazon Prime Video
Jio Cinema
Others

Q.10 Content Streamed	
Movies	1
Web Series	2
Sports	3
Daily Soaps	4
Cooking Shows	5
Dramas	6
Other Shows	7

Q.9 Subscription Type	
Basic	1
Premium	2

Q.11 Nationality of Content	
Indian	1
International	2
Both	3

Q.12 Subscription Renewal	
No	0
Yes	1

Q.13 Device Used
Mobile
Laptop/PC
TV
Tablet

Q.14 Use of OTT (in hrs/week)	
0 – 2 hours	1
2 – 4 hours	2
4 – 6 hours	3
6 – 8 hours	4
More than 8 hours	5

Q.16 Preference for Movies	
Theatre	1
OTT	2
TV (DTH)	3

Q.17 Subscription Sharing	
Yes, I lend it to someone	1
Yes, I borrow it from someone	2
No, I use it myself	3

Q.18 Price Point
Reasonable
Cheap
Expensive

Q.20 Customer Satisfaction	
No	0
Yes	1

Sample Data –

Q.2	Q.3	Q.4	Q.5	Q.6	Q.7	Q.8	Q.12	Q.15	Q.16
21	1	4	2	15	2	N, A	0	7	1
20	1	4	1	35	5	N, D, S, A, J	0	20	2
20	2	4	5	35	5	N, D, S, A, J	0	10	2
20	2	4	1	40	2	N, D	1	10	2
20	1	4	1	40	2	N, D	0	10	3
20	2	4	2	36	5	N, D, S, A, J	1	4	2
25	1	1	5	25	3	N, D, A	1	8	1
19	2	4	1	56	2	N, D	0	25	2
22	1	1	3	40	5	N, D, S, A, J	1	6	2
24	1	3	1	30	6	N, D, S, A, J, O	0	15	2
26	1	1	1	50	1	N	0	4	1
21	1	1	1	40	1	N	0	3	2
20	2	4	1	40	1	N	0	5	2
20	1	4	1	30	3	D, S, J	0	4	2
21	1	3	1	48	2	D, A	1	5	2
20	2	4	1	30	1	O	0	2	3
23	1	1	3	30	1	D	0	30	2
35	2	1	3	40	2	D, A	0	4	2
25	2	2	4	25	1	D	1	8	3
25	1	1	4	48	2	D, J	0	14	3
25	1	1	3	40	4	N, D, S, J	1	9	2
23	1	4	1	72	6	N, D, S, A, J, O	1	24	2
24	2	1	2	48	3	N, D, A	1	8	2
20	2	1	1	40	2	N, D	1	2	2
27	2	1	2	45	2	N, D	1	3	2
17	1	4	1	20	4	N, D, S, J	1	20	2
23	2	4	4	20	3	N, D, A	1	6	2
22	1	4	5	40	2	N, D	1	2	2
36	1	2	3	48	2	S, J	0	2	3
33	2	2	2	20	5	D, S, A, J, O	1	36	2
35	1	1	4	60	1	J	0	6	3
31	1	1	3	40	1	N	0	30	1
23	1	1	1	48	1	N	0	8	2
19	2	4	1	20	5	N, D, S, A, J	1	67	1
20	2	4	1	45	1	N	0	12	1
18	2	4	1	40	3	N, D, J	0	3	2
19	2	4	1	40	5	N, D, S, A, J	1	12	2
50	2	1	3	30	3	N, D, A	0	2	1

43	2	2	2	36	3	D, S, A	1	4	3
33	2	2	1	30	2	S, J	1	5	3
59	1	2	2	48	2	D, A	1	12	2
40	1	1	5	72	2	N, J	1	5	3
55	1	1	4	48	2	N, O	1	4	2
57	1	1	5	45	3	N, S, A	1	5	2
26	2	1	1	40	5	N, D, S, A, J	1	1	2
18	2	4	2	25	5	N, D, S, A, J	1	28	1

Methodology –

1. Graphical Representations –

A] Simple Column Chart

Column charts are pictorial representation of data in the form vertical bars, where the height of the bar is proportional to the measure of the data. The horizontal axis denotes the various categories and the vertical axis denotes the value of the variable.

B] Pie Chart

Pie chart is another type of graphical representation of data. The pie chart is a circular chart and is divided into different parts. Each part represents the fraction of whole.

C] Stacked Column Chart

A stacked column chart is also called as the composite column chart, which divides the aggregate into different parts. In a stacked chart, each bar represents the whole and each segment represents the different parts of the whole.

D] Multiple Column Chart

A multiple column chart is similar to a regular column chart with the exception that there are two or more bars in each category, one for each subdivision. A multiple chart helps us to compare different categories across 2 or more levels.

E] Sub-divided Column Chart

Sub-divided column charts are diagrams which simultaneously present, total values as well as part values of a set of data. Different parts of a bar are shown in the same order for all bars of a diagram.

2.Shapiro-Wilk Test –

Shapiro-Wilk test is a hypothesis test that evaluates whether a data set is normally distributed. It evaluates data from a sample with the null hypothesis that the data set is normally distributed. The test is able to detect departure from normality due either skewness or kurtosis, or both. A large p-value indicates that the data set is normally distributed while a low p-value indicates that it is not normally distributed.

Here we wish to test,

H_0: Given observations are normally distributed.

v/s

H_1: Given observations are not normally distributed.

Command in R-software: shapiro.test(x)

Criteria: Reject H_0 at α% level of significance if,

p-value $<$ α (l.o.s.)

otherwise accept it.

3.Fitting of Exponential Distribution –

A continuous random variable X taking non-negative values is said to follow exponential distribution with parameter 'θ' if it's probability distribution function (p.d.f.) is given by,

$$f(x) = \begin{cases} \theta e^{-\theta x}, & x \geq 0, \theta > 0 \\ 0, & \text{otherwise} \end{cases}$$

It is denoted as $X \rightarrow Exp(\theta)$.

Now, mean of exponential distribution is given as,

$$E[X] = Mean = \frac{1}{\theta}$$

We can find mean from the data using,

$$Mean = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i}$$

Hence, we can estimate the parameter as,

$$\hat{\theta} = \frac{1}{Mean}$$

Now, we know that cumulative distribution function (c.d.f.) of exponential distribution with parameter θ is given as,

$$F(x) = P(X \leq x)$$
$$= 1 - e^{-\theta x}$$

Thus, we can calculate the less than type cumulative probabilities using the above formula. Then, we can calculate the actual probabilities by subtracting the cumulative probability of previous class from the cumulative probability of the current class.

Also, expected frequencies can be calculated as,

$$E_x = N \times P(X)$$

4.Chi Square Test for goodness of fit –

Since there are several probability distributions, which distribution will fit properly to a certain data is a question of interest. In such cases, we want to test the appropriateness of the fit. Hence, we desire to test,

H_0: Fitting of the probability distribution to the given data is proper (good).

v/s

H_1: Fitting of the probability distribution to the given data is not good.

The test based on χ^2- distribution is used to test this H_0.

It is called the χ^2-test of goodness of fit.

Suppose $(O_1, O_2, \ldots O_j, \ldots O_k)$ be a set of observed frequencies and $(E_1, E_2, \ldots E_i, \ldots E_k)$ be corresponding expected frequencies obtained under H_0.

$$\sum_{i=1}^{k} O_i = N = \sum_{i=1}^{k} E_i$$

p = Number of parameters estimated for fitting the probability distribution.

Under H_0, test statistic used is,

$$\chi_{cal} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{k-p-1,\alpha}$$

$$\chi^2_{cal} = \left(\sum_{i=1}^{k} O_i^2\right) - E_i N \sim \chi^2_{k-p-1,\alpha}$$

Criteria: Reject H_0 at α% level of significance if,

$$\chi^2_{cal} \geq \chi^2_{k-p-1,\alpha}$$

otherwise accept it.

5.Testing equality of two population proportion –

Here, we discuss the cases when the two samples are taken from two distinct materials or populations. Suppose, a sample is drawn from each of the population. However, the test statistic is based on both the samples. Suppose these samples give proportion of specific items as p1 and p2 respectively. One may be interested in knowing that the population proportions from which these samples are chosen are same. In other words, we want to know whether difference between two sample proportion is negligible and it has arisen merely due to sampling variations.

Let, n_1 = Size of sample drawn from the first population. n_2

= Size of sample drawn from the second population. x_1

= Number of items of specific type in first sample. x_2 =

Number of items of specific type in second sample.

$p_1 = \frac{x_1}{n_1}$ = Proportions of specific items in first sample.

$p_2 = \frac{x_2}{n_2}$ = Proportion of specific items in second sample.

P_1 = Proportion of specific items in first population.

P_2 = Proportion of specific items in second population.

Thus, the hypotheses for the problem is,

$H_0: P_1 = P_2$ vs $H_1: P_1 \neq P_2$

The test statistic will be,

$$Z = \frac{p_1 - p_2}{\sqrt{\bar{P}\bar{Q}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$ where, $\bar{P} = \frac{x_1+x_2}{n_1+n_2}$

$Z \rightarrow N(0,1)$ for large n_1, n_2

Criteria: Reject H_0 at $\alpha\%$ level of significance if,

$$Z_{cal} \geq Z_{\frac{\alpha}{2}}$$

otherwise accept it.

6.Chi- square test for independence of attributes –

The chi- square test of independence checks whether two variables are likely to be related or not.

Let attribute "A" is classified in 'r' levels and attribute "B" is classified into 's' levels as below.

A \ B	B_1	B_2	.....	B_j	.....	B_s	Total
A_1	O_{11} (E_{11})	O_{12} (E_{12})	.....	O_{1j} (E_{1j})	.....	O_{1s} (E_{1s})	(A_1)
A_2	O_{21} (E_{21})	O_{22} (E_{22})	.....	O_{2j} (E_{2j})	.....	O_{2s} (E_{2s})	(A_2)
. . . .			.....		.....		. . . .
A_i	O_{i1} (E_{i1})	O_{i2} (E_{i2})	.....	O_{ij} (E_{ij})	.....	O_{is} (E_{is})	(A_i)
. . . .			.....		.....		. . . .
A_r	O_{r1} (E_{r1})	O_{r2} (E_{r2})	.....	O_{rj} (E_{rj})	.....	O_{rs} (E_{rs})	(A_r)
Total	(B_1)	(B_2)	.....	(B_j)	.....	(B_s)	N

The above table contains 'r' rows and 's' columns and is called as r × s contingency table.

O_{ij} = observed frequency corresponding to $(i,j)^{th}$ cell.

E_{ij} = expected frequency of $(i,j)^{th}$ cell.

$N = \sum_{i=1}^{r} \sum_{j=1}^{s} O_{ij}$ = Total observed frequency.

$(A_i) = \sum_{i=1}^{r} O_{ij}$ = Total of observed frequency in i^{th} row. $i=1,2,...r$

$(B_j) = \sum_{j=1}^{s} O_{ij}$= Total of observed frequency in j^{th} column. j=1,2,…s

Here we want to test,

H_0: Attributes A and B are independent. v/s

H_1: Attributes A and B are not independent.

Under H_0, the expected frequencies corresponding to the given observed frequencies are obtained as;

$$E_{ij} = \frac{\text{(Total observed frequency in the } i^{th}\text{row)}\times \text{(Total observed frequency in the } j^{th}\text{column)}}{\text{Grand total of all observed frequencies}}$$

$$E_{ij} = (\underset{j=1,2,\ldots s}{\quad\quad A^{i})\times(B_j)} \qquad \forall \qquad i=1,2,\ldots r \quad N$$

Under H_0, test statistic used is,

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{\left(O_{ij} - E_{ij}\right)^2}{\quad\quad} \sim \chi^2_{(E_{ijr}-1)(s-1),\alpha}$$

$$\chi_{cal} = \left({}_2O_{ij} \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{\quad^2}{\quad E_{ij}} \right) - N$$

Criteria: Reject H_0 at α% level of significance if,

$$\chi^2_{cal} \geq \chi^2_{(r-1)(s-1),\alpha}$$

otherwise accept it.

7.Naive Bayes Algorithm –

Naïve Bayes is a supervised non-linear classification algorithm in Rprogramming naïve bayes classifiers are a family of sample probabilistic classifiers based on applying Baye's theorem with strong (Naïve) independence assumptions between the features or variables.

The Naïve bayes algorithm is called "Naïve" because it makes the assumption that the occurrence of a certain feature is independent of the occurrence of other features.

Naïve Bayes algorithm is based on Bayes theorem. Bayes theorem gives the conditional probability of an event A given another event B has occurred.

$$P(A|B) = P(B|A) \frac{P(A)}{P(B)}$$

where, $P(A|B)$ = conditional probability of A given B.

$P(B|A)$ = conditional probability of B given A.

$P(A)$ = probability of event A.

$P(B)$ = probability of event B.

8. Clustering using k-means algorithm –

Clustering is a technique in machine Learning that attempts to find groups or clusters of observations within a data set.

The goal is to find clusters such that the observations within each cluster are quite similar to each other, while observations in different clusters are quite

different from each other. Clustering is a form of unsupervised learning because we are simply attempting to find clusters within a data set rather than predicting the value of some response variable.

K-Means clustering:

K-means clustering algorithm computes centroids, and repeats until the optimal centroid is found. It is presumptively how many clusters there are. It is known as the flat clustering algorithm. The number of clusters found from the data by this method is denoted by the letter 'k' in k-means.

In this method, data points are assigned to clusters in such a way that the sum of the squared distances between the data points and the centroid is as small as possible. It is essential to note that the reduced diversity within clusters leads to more identical data points within the same cluster.

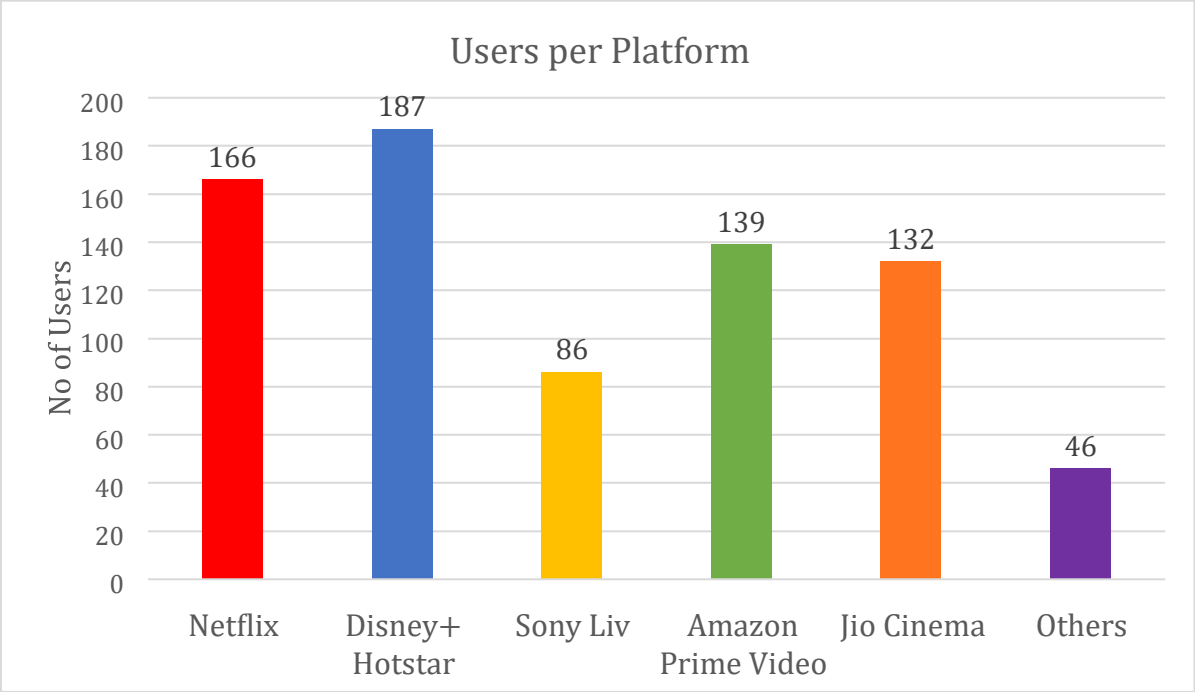
The Manhattan Distance formula is used in k-means algorithm. The distance 'D' between 2 data points (x_1, y_1) and (x_2, y_2) is calculated using the formula,

$$D = |x_2 - x_1| + |y_2 - y_1|$$

Statistical Analysis

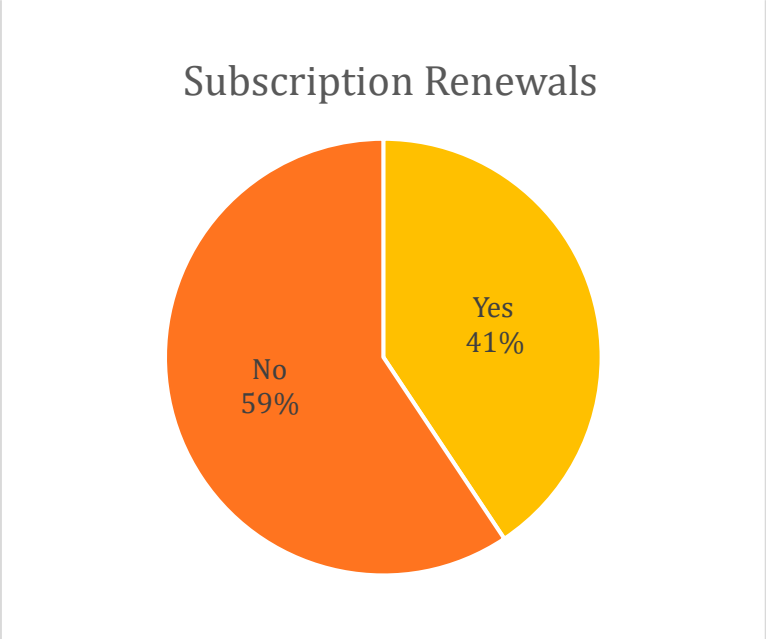
A] Graphical representations –

a) Simple Column Chart –



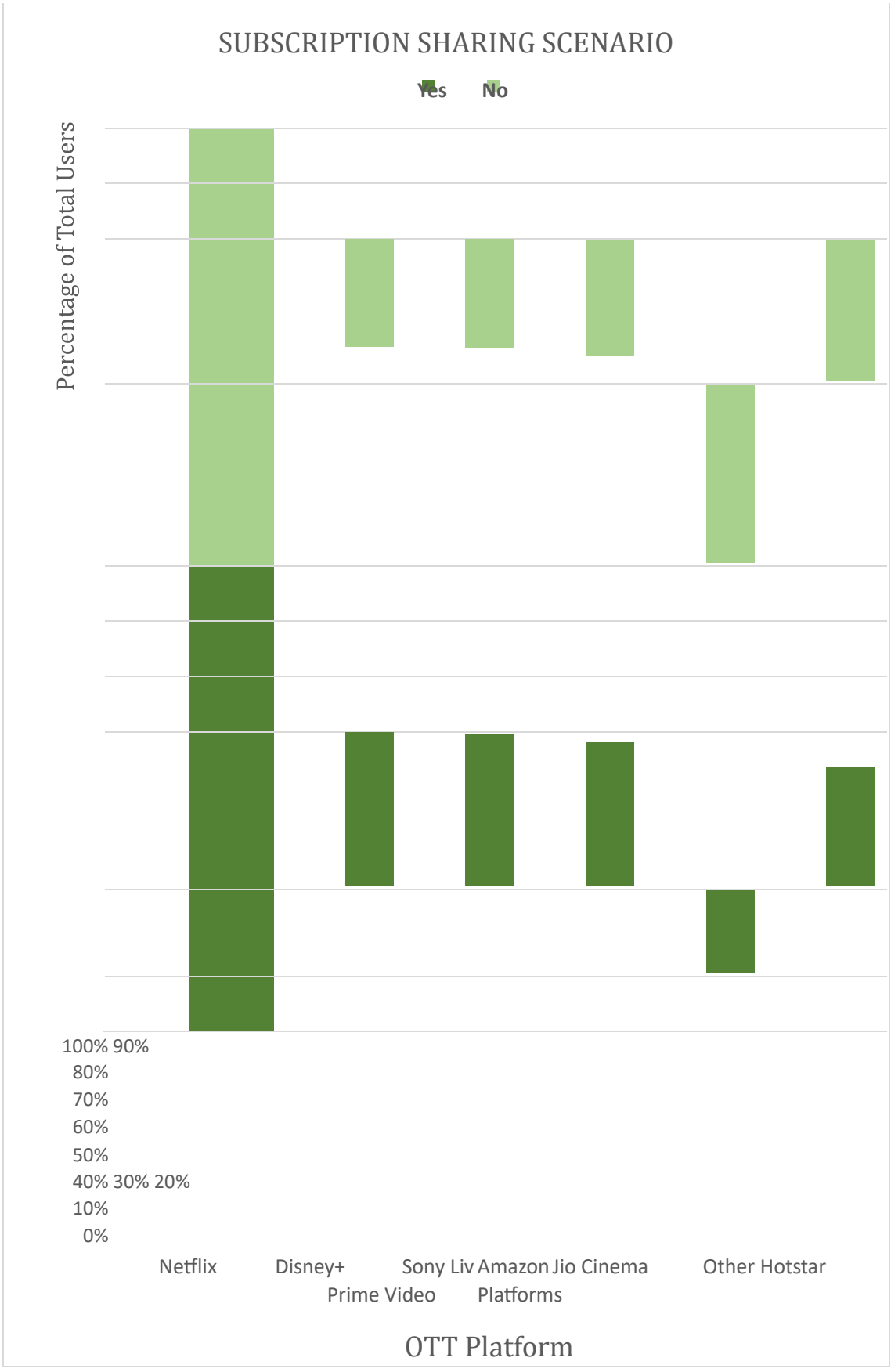
Interpretation: We can observe that Disney+ Hotstar is the most used OTT platform among citizens and Netflix is in the second position. Amazon Prime Video is the third most used OTT platform with Jio Cinema is a close fourth. Sony Liv is the least used OTT platform among the top 5. Also, 46 respondents use other OTT platforms.

b) Pie Chart –



Interpretation: We can observe from the above Pie chart that just over 40% of the OTT users renew their subscriptions regularly. On the other hand, almost 60% of the users do not renew their subscriptions regularly.

c) 100% Stacked Column Chart

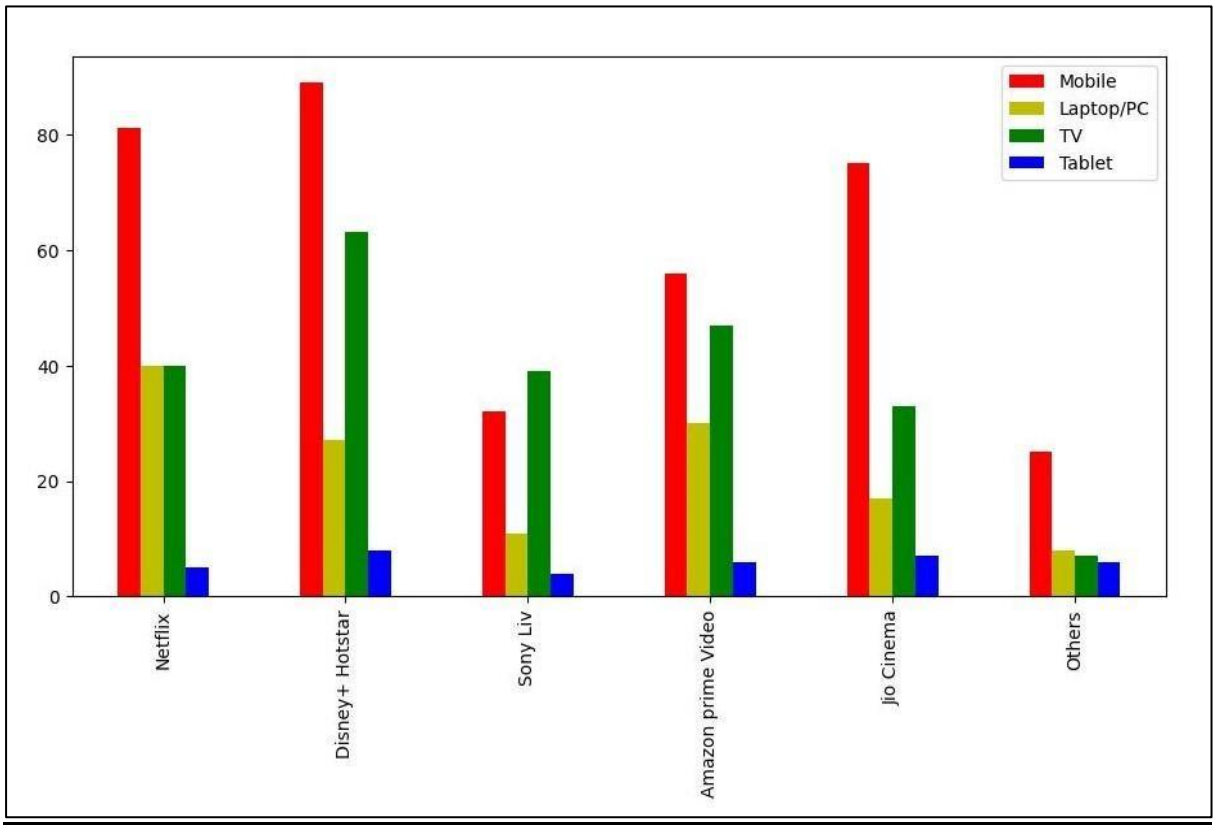

SUBSCRIPTION SHARING SCENARIO
Yes
No
Percentage of Total Users
100% 90%
80%
70%
60%
50%
40% 30% 20%
10%
0%
Netflix
Disney+ Prime Video
Sony Liv
Amazon Jio Cinema Platforms
Other Hotstar
OTT Platform

Interpretation: We can observe from the above graph that apart from Jio Cinema, the proportion of people sharing subscriptions is almost similar among all OTT platforms (between 45% - 60%). Among Jio Cinema users, however, majority (almost 70%) people prefer to use individual subscriptions.

d) Multiple Column Chart –

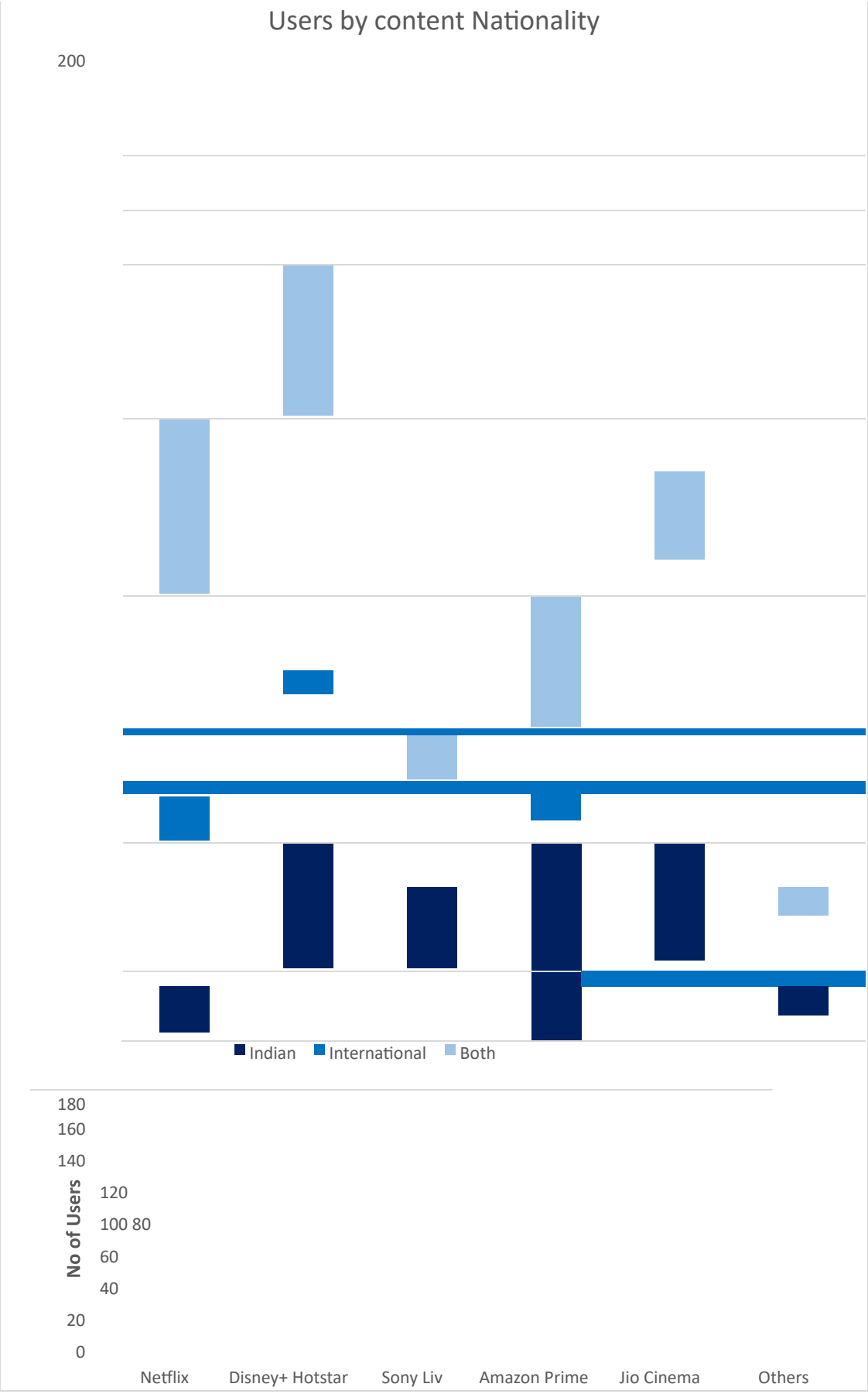
Code for graph using Python –

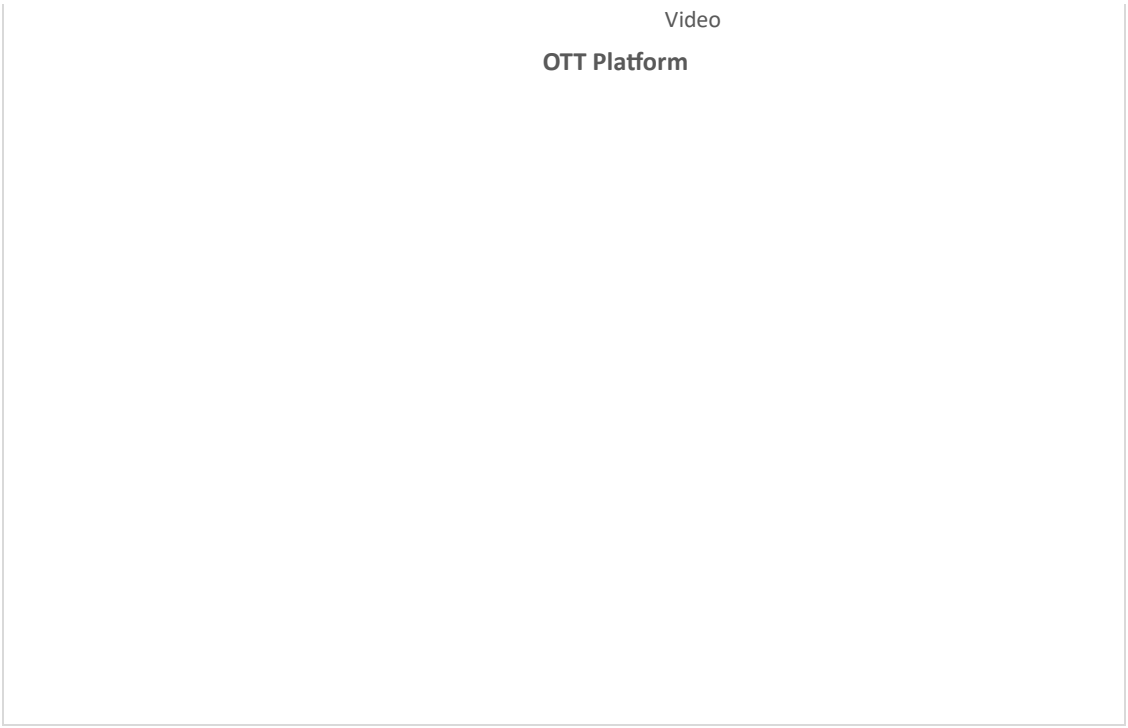
```
import matplotlib.pyplot as plt import
pandas as pd
plat=["Netflix","Disney+ Hotstar","Sony Liv","Amazon prime
Video","Jio Cinema","Others"]
mob=[81,89,32,56,75,25]
lap=[40,27,11,30,17,8]
tv=[40,63,39,47,33,7]
tab=[5,8,4,6,7,6]
data={"Mobile":mob,"Laptop/PC":lap,"TV":tv,"Tablet":tab}
df=pd.DataFrame(data,index=plat,columns=["Mobile","Laptop/PC","TV","
Tablet"]) print(df)
fig=df.plot.bar(color=['r','y','g','b']) plt.show()
```



Interpretation: As evident from the above graphical representation, Mobile is the most commonly used device by most OTT users. Although TV is the most preferred device among Sony Liv users, when we consider all the platforms, it is second after Mobile. Also, we can observe that Tablet is the least used device to stream OTT content.

e) Subdivided Column Chart –


Users by content Nationality
200
Indian
International
Both
180
160
140
120
100 80
60
40
20
0
No of Users
Netflix
Disney+ Hotstar
Sony Liv
Amazon Prime
Jio Cinema
Others



Interpretation: From the above chart, we can conclude that majority people using OTT Platforms view both; Indian as well as International content. However, in case of Sony Liv and Jio Cinema maximum people view Indian shows. Apart from this, the proportion of people using Netflix to International content is highest across all the platforms.

B] Shapiro Wilk Normality Test

Test to verify whether the average times spent by citizens using OTT per week (in no. of hours) are Normally distributed.

Times spent by users using OTT per week (in number of hours):

7,20,10,10,10,4,8,25,6,15,4,3,5,4,5,2,30,4,8,14,9,24,8,2,3,20,6,2,2,36,6,30,8,
67,12,3,12,2,4,5,12,5,4,5,1,28,18,8,10,10,12,10,6,2,10,3,4,20,3,1,15,21,4,2,7,
4,36,6,20,3,12,2,18,2,2,10,10,5,4,40,5,1,10,8,14,10,5,3,20,4,2,4,12,6,4,48,3,6

,10,5,21,6,6,15,4,4,20,2,6,10,5,30,2,35,10,6,3,10,8,4,8,12,6,14,4,6,30,20,6,8,3,15,2,21,10,7,1,6,6,5,30,6,4,69,12,10,5,6,2,4,10,10,5,4,5,2,6,2,4,20,4,12,5,28,8,24,6,15,3,7,8,6,12,6,24,20,6,16,2,7,4,6,4,4,4,5,4,18,8,2,2,5,4,3,9,4,5,8,10,8,5,4,15,5,24,12,8,10,7,7,12,3,8,2,6,2,3,4,2,5,2,2,12,6,28,6,5,1,1,5,3,14,24,7,7,14,10,14,20,2,7,14,14,7,3,4,4,3,2,10,20,36,6,3,40,2,2,3,2,14,5,2,2,2,13,2,48,8,3,5,4,2,2,2,15,2,14,20

Here we want to test

H_0: Average weekly OTT using times are normally distributed.

v/s

H_1: Average weekly OTT using times are not normally distributed.

Test on R-software

```
>
x=c(7,20,10,10,10,4,8,25,6,15,4,3,5,4,5,2,30,4,8,14,9,24,8,2,3,20,6,2,2,36,6,30,8,67,12,3,12,2,4,5,12,5,4,5,1,28,18,8,10,10,12,10,6,2,10,3,4,20,3,1,15,21,4,2,7,4,36,6,20,3,12,2,18,2,2,10,10,5,4,40,5,1,10,8,14,10,5,3,20,4,2,4,12,6,4,48,3,6,10,5,21,6,6,15,4,4,20,2,6,10,5,30,2,35,10,6,3,10,8,4,8,12,6,14,4,6,30,20,6,8,3,15,2,21,10,7,1,6,6,5,30,6,4,69,12,10,5,6,2,4,10,10,5,4,5,2,6,2,4,20,4,12,5,28,8,24,6,15,3,7,8,6,12,6,24,20,6,16,2,7,4,6,4,4,4,5,4,18,8,2,2,5,4,3,9,4,5,8,10,8,5,4,15,5,24,12,8,10,7,7,12,3,8,2,6,2,3,4,2,5,2,2,12,6,28,6,5,1,1,5,3,14,24,7,7,14,10,14,20,2,7,14,14,7,3,4,4,3,2,10,20,36,6,3,40,2,2,3,2,14,5,2,2,2,13,2,48,8,3,5,4,2,2,2,15,2,14,20)

> shapiro.test(x)

    Shapiro-Wilk normality test

data:  x

W = 0.71625, p-value = 2.2e-16
```

Criteria: Reject H_0 at α% level of significance if,

p-value $<$ α (l.o.s.)

otherwise, accept H_0

Decision: Here, p-value $= 2.2\times10^{-16}$ and $\alpha = 0.05$.

i.e. p-value $<$ α (l.o.s.)

Hence, we may reject H_0 at 5% level of significance.

Conclusion: We may conclude that the average times spent by citizens using OTT per week (in number of hours) are not normally distributed.

C] Fitting of Exponential Distribution

Fitting of Exponential Distribution to average weekly times spent by citizens using OTT.

LCB	UCB	Mid Pt. (X_i)	Frequency (F_i)	F_iX_i	Cumulative Probability	Probability	Expected Frequency
0	9	4.5	181	814.5	0.582938	0.582938	162.0568
10	19	14.5	58	841	0.842166	0.259228	72.06529
20	29	24.5	24	588	0.940269	0.098103	27.2726
30	39	34.5	9	310.5	0.977395	0.037126	10.32112
40	49	44.5	4	178	0.991445	0.01405	3.905956
50	59	54.5	0	0	0.996763	0.005317	1.478181
60	69	64.5	2	129	0.998775	0.002012	0.559407
70	∞		0	0	1	0.001225	0.340602
$\sum$ =			278	2861		1	278

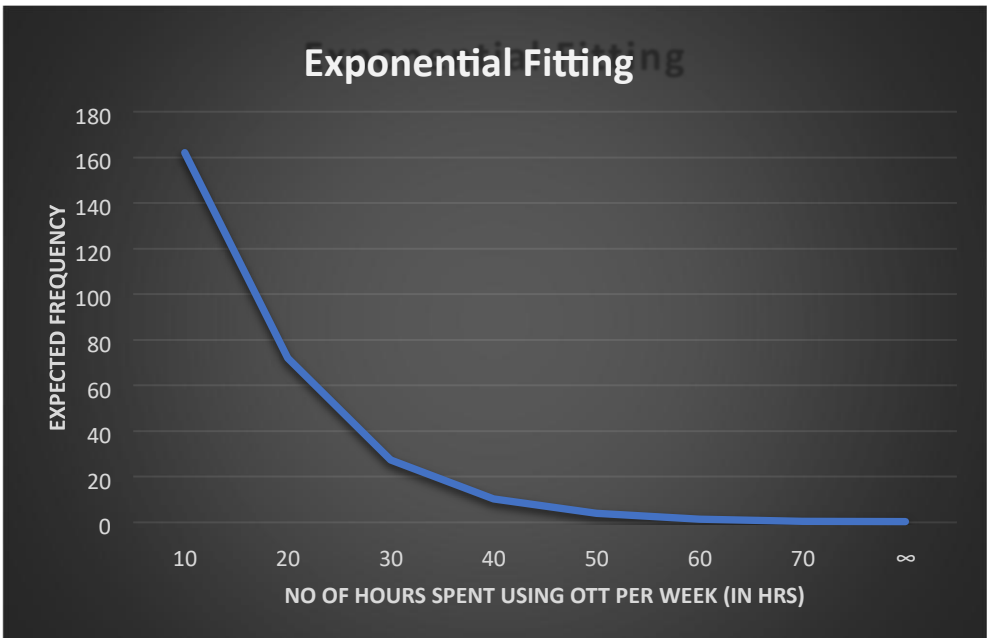
Mean $= \frac{1}{N}\sum F_iX_i = \frac{2861}{278} = 10.291367$

Estimate of $\theta = \hat{\theta} = \frac{1}{}$

$\frac{\quad}{\quad} = \frac{\quad}{\quad}^{1} = = 0.097169$ Mean

10.291367

Variance = Mean^2 = 10.291367^2 = 105.912235



D] Chi Square Test for Goodness of fit

Here we want to test,

H_0: There is no significant difference between observed and expected frequencies. i.e. Fitting of exponential distribution is good.

v/s

H_1: There is significant difference between observed and expected frequencies. i.e. Fitting of exponential distribution is not good.

Observed Frequency (O_i)	Expected Frequency (E_i)	$\frac{O_{2i}}{E_i}$
181	162.0568	202.1575
58	72.06529	46.67989
24	27.2726	21.1201
9	10.32112	7.847983

6	6.284147	5.728701
$\sum = 278$	278	283.5341

N = Total Frequency = 278 k = No of classes (after pooling) = 5 p = No of parameters estimated = 1

$$\chi_{cal} \sum_{}^{kk} \frac{}{} = \sum_{=}^{2} \frac{(O_i - E_i)^2}{E_i} O_i$$

$$\left(\frac{}{}\right) - E_i \quad N \; i=1 \; i=1$$

$$\therefore \chi^2 = 283.5341 - 278 = 5.5341_{cal}$$

$$\chi^2_{tab} = \chi^2_{(k-p-1),\alpha} = \chi^2_{(5-1-1),0.05}$$

$$\therefore \chi^2_{tab} = \chi^2_{3,0.05} = 7.815$$

Criteria: Reject H_0 at α % level of significance if

$$\chi^2_{cal} \geq \chi^2_{(k-p-1),\alpha}$$

otherwise accept it.

Decision: Here, level of significance $(\alpha) = 0.05$

$$\chi^2_{cal} = 5.5341 \quad \& \quad \chi^2_{tab} = \chi^2_{(k-p-1),\alpha} = 7.815$$

We can observe that,

$$5.5341 < 7.815$$

i.e. $\chi 2_{cal} < \chi 2_{(k-p-1),\alpha}$

So, we may accept H_0 at 5% level of significance.

Conclusion: We can conclude that the given data follows Exp(θ)

distribution with estimate of parameter as $\hat{\theta} = 0.097169$.

Hence, we can conclude that the fitting of exponential distribution to the given data is good.

E] Test for equality of population proportions

Test to check if proportion of individuals who renew their OTT subscription regularly is same among students and other professions.

	Students	Other Professions
No of persons who renew subscription regularly	64	49
Total no of persons surveyed	185	93

Consider,

$P_1 \rightarrow$ Proportion of students who renew their OTT subscription regularly.

$P_2 \rightarrow$ Proportion of non-students who renew their OTT subscription regularly.

Here we want to test,

$H_0: P_1 = P_2$ vs $H_1: P_1 \neq P_2$

Test on R-software:

```
> x=c(64,49)
> n=c(185,93)
> prop.test(x,n,conf.level=0.95)
```

2-sample test for equality of proportions with continuity correction

data: x out of n

X-squared = 7.665, df = 1, p-value = 0.00563

alternative hypothesis: two.sided 95 percent

confidence interval:

-0.3114690 -0.0504026

sample estimates:

prop 1 prop 2

0.3459459 0.5268817

Criteria: Reject H_0 at α% level of significance if

p-value < α (l.o.s.) otherwise

accept H_0.

Decision: Here p-value = 0.00563 & level of significance, α = 0.05.

So, here, p-value < α

Hence, we may reject H_0 at 5% level of significance.

Conclusion: We can conclude that proportion of individuals who renew their OTT subscription is not same among students and people from other professions.

F] Chi Square Test for Independence of Attributes.

Test to check whether type of content viewed and OTT platform are dependent on each other. ($\alpha = 0.05$)

	Movies	Web Series	Sports	Daily Soaps	Cooking Shows	Dramas	Others
Netflix	127	133	9	3	4	29	10
Disney+ Hotstar	107	95	86	23	5	27	11
Sony Liv	23	30	23	31	4	17	12
Amazon Prime Video	103	105	8	5	3	16	8
Jio Cinema	63	33	62	15	4	20	11
Others	25	20	14	8	3	13	11

Here we want to test,

H_0: OTT Platform used and type of content streamed/viewed are independent of each other.

v/s

H_1: OTT Platform used and type of content streamed/viewed are dependent on each other.

<u>Test on R-software:</u>

```
> N=c(127,133,9,3,4,29,10)
> D=c(107,95,86,23,5,27,11)
> S=c(23,30,23,31,4,17,12)
> A=c(103,105,8,5,3,16,8)
> J=c(63,33,62,15,4,20,11)
> O=c(25,20,14,8,3,13,11)
> m=c(N,D,S,A,J,O)
> y=matrix(m,nrow=6,ncol=7,byrow=T)
```

```
> y
     [,1] [,2] [,3] [,4] [,5] [,6] [,7]
[1,]  127  133    9    3    4   29   10
[2,]  107   95   86   23    5   27   11
[3,]   23   30   23   31    4   17   12
[4,]  103  105    8    5    3   16    8
[5,]   63   33   62   15    4   20   11
[6,]   25   20   14    8    3   13   11
> chisq.test(y, correct=T)

        Pearson's Chi-squared test data:
y
X-squared = 286.57, df = 30, p-value = 2.2e-16 Warning
message:
In chisq.test(y, correct = T) : Chi-squared approximation may be incorrect
```

Criteria: Reject H_0 at α% level of significance if

p-value $< \alpha$ otherwise,

accept H_0.

Decision: Here p-value $= 2.2\times 10^{-16}$ & level of significance, $\alpha = 0.05$.

i.e. p-value $< \alpha$

Hence, we may reject H_0 at 5% level of significance.

Conclusion: It may be concluded that type of content streamed and OTT Platform used to stream the content are dependent on each other.

G] Classification using Naïve Bayes Classifier

Fitting of Naïve Bayes Classification Model to predict whether a customer is satisfied with his/her OTT subscription on the basis of :

- Use of OTT per week (in no of hours)
- Customer's view about the price point (whether the customer thinks the subscription is Cheap, Reasonable or Expensive)
- Customer's rating of the OTT platform x1 → Usage time per week x2 → Price Point x3 → Customer Rating y → Customer Satisfaction

Code using R-software:

```
>
x1=c(2,4,3,2,2,1,3,5,2,3,2,1,3,1,3,2,1,2,1,3,1,2,1,4,1,1,2,1,2,2,1,2,2,1,4,1,2,2,2,
1,1,1,4,1,1,1,1,4,5,1,1,4,1,1,1,2,2,2,3,2,5,1,2,1,4,1,3,4,2,2,3,4,5,2,3,1,1,2,1,2,3,
2,1,1,4,1,2,2,1,1,2,4,1,1,5,4,1,3,5,5,2,2,1,1,4,1,1,5,2,2,2,5,2,1,1,1,1,1,1,1,4,1,1,
1,2,3,1,1,3,5,2,1,1,1,3,4,1,1,2,2,3,2,2,4,5,4,5,2,2,3,2,2,3,1,2,1,2,2,1,2,1,1,1,1,1,
5,1,2,3,3,1,1,5,1,3,2,1,2,2,2,5,1,3,1,1,1,2,2,1,5,4,3,1,1,1,3,2,1,3,1,1,2,3,2,2,1,3,
5,1,1,1,1,1,1,2,4,2,2,3,1,1,4,2,1,2,1,2,1,1,2,2,2,1,2,2,4,2,1,1,1,2,1,3,3,2,2,2,1,3,
1,3,1,1,1,2,1,2,1,2,1,1,1,1,5,1,1,1,1,1,1,3,2,1,3,3,2,1,3,1,5,5,4,4,1,1,2,3,1,2,1,1,
1,5,1,1,1,1,1,3,5,2,1,1,2,1,1,3,3,3,3,2,1,3,1,1,1,1,4,3,1,1,1,5,1,3,1,3,2,2,1,1,2,1,
2,2,2,1,1,3,1,2,5,1,1,2,4,2,1,1,1,1,1,1,1,1,1,1,1,2,1,3,3,2,5,4,3,2,2,2,2,1,5,1,1,2,
1,4,1,1,2,4,1,2,4,2,1,2,1,1,1,1,2,1,1,3,3,3,1,1,2,1,1,1,1,1,1,1,1,3,2,3,2,2,5,2,2,1,
1,1,4,1,1,1,1,1,1,1,1,1,1,2,1,2,1,2,1,1,1,1,2,2,2,1,1,2,3,1,2,3,1,2,5,4,2,1,1,3,1,2,
2,1,1,1,2,3,2,1,1,4,1,3,1,1,5,1,1,4,2,1,1,2,2,1,2,1,1,2,2,3,1,1,1,1,3,3,1,5,3,1,2,1,
2,1,3,2,3,4,2,1,1,2,1,1,2,1,1,1,1,1,3,5,3,1,5,5,4,2,1,1,4,3,1,2,1,2,1,1,5,2,1,1,1,1,1,
1,1,2,1,1,1,1,1,2,3,3,1,1,1,1,1,2,1,1,1,3,1,1,1,2,1,1,1,1,1,1,1,1,1,2,1,1,1,2,1,2,1,
1,2,4,1,5,1,4,1,2,3,2,2,1,2,3,2,1,1,1,4,1,1,1,1,1,1,3,2,1,1,3,1,1,1,3,1,1,1,1,1,2,3,
1,1,1,1,2,1,3,4,1,2,1,1,1,2,2,1,5,1,1,1,3,1,2,1,2,1,2,1,1,3,4,1,2,1,1,2,2,1,1,2,2,1,
1,2,1,1,1,2,1,1,1,1,2,1,4,3,1,1,3,1,3,1,1,1,2,1,1,4,3,2,1,1,1,2,1,5,1,1,1,2,1,1,3,3,
1,2,5,2,1,2,1,1,1,2,1,4,1,2,1,3,5,1,1,4,1,1,1,1,3,1,5,2,4,1,1,2,1,3,1,1,5,2,3,1,2,1,
1,3,3)

>
x2=c(1,1,3,1,1,3,3,1,3,3,3,1,3,3,1,1,1,1,2,1,3,3,1,1,1,1,1,3,1,1,3,1,1,1,3,3,3,3,3,3,
3,1,3,3,1,3,1,3,1,1,3,1,3,3,2,3,1,3,1,1,1,3,3,1,3,3,1,3,1,1,1,1,1,3,1,3,1,3,3,1,3,3,
1,3,3,1,3,3,3,1,3,1,1,3,3,1,1,1,1,3,3,1,1,3,1,1,2,3,1,1,1,3,3,1,3,3,1,3,1,1,1,2,1,3,
3,3,1,3,3,1,1,3,1,1,1,3,1,3,1,3,1,1,3,3,3,3,1,3,3,3,3,1,3,3,3,1,1,2,1,3,1,1,1,1,3,3,
1,1,1,1,2,3,1,1,1,1,1,1,3,1,1,2,1,1,1,1,1,2,1,2,1,1,1,1,3,3,1,1,1,1,2,1,1,1,2,3,1,1,
2,1,1,1,1,3,2,1,1,2,3,3,1,1,3,3,1,1,3,2,1,1,2,1,1,1,1,3,1,1,3,2,1,3,1,3,1,1,1,1,1,1,
```

2,1,1,1,2,1,1,1,2,1,1,3,1,1,2,1,1,1,1,1,3,1,1,1,1,1,2,1,1,3,3,3,1,1,3,3,1,1,1,1,1,1,
3,1,1,3,1,2,1,1,3,1,1,3,1,3,1,1,1,1,1,1,3,1,1,1,2,3,1,1,1,1,1,2,3,1,3,1,3,2,3,1,2,1,
3,1,1,1,1,1,3,1,1,1,1,3,3,1,1,1,2,3,3,2,3,2,3,1,3,1,3,1,2,3,1,1,1,3,1,1,1,1,1,3,1,3,
1,2,1,1,3,1,2,3,3,3,1,1,3,3,1,3,1,3,1,3,1,1,1,1,1,3,3,1,3,1,1,1,3,1,2,1,1,1,3,2,1,3,
1,2,1,2,3,1,3,3,1,2,2,1,3,1,1,1,1,2,3,3,1,1,1,1,3,3,1,3,2,1,1,1,1,1,1,1,1,1,3,1,1,1,
1,1,1,3,1,1,1,1,1,2,3,1,1,3,1,1,1,3,3,1,2,2,1,3,1,3,3,1,1,3,3,3,1,1,1,1,1,1,1,1,1,1,
3,1,1,1,2,1,1,1,3,1,1,1,3,2,3,1,1,1,1,1,1,1,2,1,1,3,2,1,1,2,3,1,1,1,3,3,1,1,1,1,1,1,
1,1,1,3,3,3,1,1,2,3,1,2,1,2,1,1,3,2,3,3,1,1,3,2,1,1,3,1,1,1,1,1,1,3,1,2,2,2,2,1,2,1,
2,1,2,2,1,1,1,1,1,1,1,1,1,1,1,2,1,2,1,1,2,1,2,1,1,1,2,3,1,2,2,2,3,1,2,2,2,2,2,1,2,1,2,
1,3,2,1,2,1,1,1,2,2,1,2,2,3,1,1,2,1,2,1,1,2,1,1,2,1,1,3,2,1,1,1,2,1,2,2,2,1,2,1,1,2,
2,1,2,1,2,2,1,1,2,2,1,1,3,2,1,1,3,3,1,2,1,1,3,1,1,1,3,1,2,2,3,1,1,1,1,1,2,1,2,3,2,1,
2,3,1,1,1,1,3,1,1,1,3,2,1,3,1,1,1,3,1,1,1,2,1,1,2,1,3,2,3,2,3,2,1,1,2,1,3,3,2,2,1,2,
3,2,3)

>

x3=c(5,4,5,4,5,4,5,5,4,3,4,5,4,4,2,5,5,5,5,5,5,5,4,5,3,5,5,5,5,5,4,5,5,5,4,3,5,5,4,
4,5,5,5,5,4,5,4,5,2,4,5,5,3,5,3,4,2,4,5,5,4,5,5,3,4,4,4,5,5,5,4,5,4,5,5,4,5,3,5,5,5,
5,5,4,5,4,4,5,4,5,5,5,4,5,5,5,5,5,4,4,5,5,3,5,5,5,3,4,5,4,4,5,4,5,3,5,5,4,3,4,5,5,5,5,
3,5,4,4,5,5,4,5,5,4,5,5,4,5,1,5,4,4,2,3,3,1,5,2,5,5,5,4,3,4,5,5,5,5,5,5,4,5,5,5,3,5,
5,3,5,3,4,5,2,3,2,3,4,5,2,5,4,4,4,5,3,5,5,5,4,5,5,5,5,5,5,2,4,3,4,4,3,5,4,5,3,4,4,4,4,
5,4,3,2,4,4,4,3,4,5,5,5,5,4,5,3,1,5,1,4,4,4,5,3,5,4,5,3,2,4,5,3,5,3,4,4,3,4,3,5,2,5,
4,5,2,5,4,4,5,3,3,4,4,3,3,5,5,3,3,4,5,4,4,4,4,3,5,5,5,5,4,4,4,5,4,4,5,5,3,4,4,5,3,3,
5,3,5,5,4,3,5,4,3,4,5,5,5,4,5,5,5,4,4,3,5,4,3,4,2,4,5,4,3,4,2,5,5,5,5,4,3,4,4,4,5,4,
4,5,3,4,4,5,3,4,5,5,1,4,4,4,5,5,5,2,4,5,2,2,5,3,2,3,3,5,5,2,2,5,3,3,4,3,3,3,4,2,3,3,
3,5,2,2,5,3,3,4,5,3,2,5,4,3,4,5,4,5,3,3,5,5,4,2,4,3,3,3,3,2,5,4,3,4,3,4,5,2,5,4,3,3,
4,3,3,4,2,5,3,3,3,4,4,5,3,2,5,5,4,4,1,5,5,4,5,4,5,5,3,4,2,4,5,5,5,5,5,5,5,3,4,3,3,4,
5,4,5,3,5,3,4,4,4,5,4,4,4,4,5,4,5,5,3,5,3,5,3,5,4,3,1,5,5,3,5,4,5,3,3,3,3,4,4,5,3,4,
5,5,3,4,5,4,5,4,5,4,3,4,5,4,4,3,3,4,5,5,5,3,3,5,3,3,3,5,5,4,2,3,4,4,4,5,5,5,5,3,4,5,
3,4,5,3,4,4,4,5,5,4,4,4,3,3,4,4,5,5,3,4,4,4,4,5,5,5,5,5,5,5,5,5,4,3,5,2,2,4,4,3,4,4,
5,5,5,4,5,3,5,5,4,4,5,3,4,4,2,5,4,3,3,5,5,4,5,1,3,3,3,5,2,3,3,5,5,1,5,2,4,4,3,5,3,3,
3,5,2,2,2,5,4,5,3,3,4,3,3,3,4,3,5,4,3,4,4,4,4,3,5,4,3,3,5,4,1,5,3,4,3,2,4,3,3,3,4,2,5,
3,4,3,4,5,4,4,4,2,5,4,3,2,4,5,4,3,3,3,3,3,2,1,4,5,4,3,5,3,3,2,5,5,3,5,3,4,5,3,5,4,3,
4,4,5,4,3,4,5,4,4,2,5,5,5,2,3,4,4,4,1,5,5,4,3,4,5,5,4,4,4,2,4,5,4,5,5,4,4,5,3,5,5,2,
4,4,5)

>

y=c(1,1,1,1,0,1,1,1,1,1,0,1,0,0,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,0,1,0,1,0,0,1,1,1,
0,1,1,1,1,1,1,0,1,0,0,1,0,0,1,0,1,0,1,1,1,1,0,1,0,1,1,1,1,1,1,1,1,0,1,1,1,1,0,1,0,1,

1,1,1,1,1,1,0,1,0,0,1,1,1,1,1,1,1,0,0,1,1,1,1,1,1,1,1,1,1,1,0,1,1,1,1,1,1,1,1,1,1
,
1,1,1,0,1,1,1,1,1,1,1,0,1,0,1,1,1,1,1,0,1,1,0,0,0,0,0,0,1,0,0,1,1,1,1,1,1,1,1,1,0,0
,
1,1,1,1,1,1,0,0,0,0,1,1,0,1,1,1,1,1,1,0,1,1,1,1,1,1,1,1,1,0,0,1,0,1,1,1,1,1,1,1,0,1,1
,
1,1,1,1,1,1,0,1,1,0,1,1,0,1,1,1,1,0,0,1,0,0,1,1,1,1,1,1,0,1,1,1,0,0,1,0,0,1,1,0,1,1,1,1,1
,
1,1,1,1,1,1,1,1,1,1,0,1,0,1,1,1,1,1,1,1,1,0,1,0,1,1,1,1,1,1,1,1,1,0,1,1,1,1,0,1,0,1,1,1,1,1
,
0,0,1,1,1,0,1,1,0,1,1,1,1,1,1,1,1,1,1,1,0,1,1,1,1,1,1,0,1,1,1,1,1,1,1,1,1,1,1,0,1,0,0,0,1
,
1,1,0,1,1,1,1,1,1,1,1,1,0,0,1,1,1,1,1,0,1,0,0,0,1,0,1,1,0,1,0,0,1,1,0,0,1,1,0,1,1,1,1,1,1,
0,1,1,1,1,1,1,0,0,0,0,0,1,1,1,1,1,0,1,0,1,1,1,1,1,1,0,1,0,0,0,0,0,0,1,0,1,1,0,1,1,0,0,1,0,
1,1,0,0,1,1,0,0,0,1,1,1,1,0,0,1,1,1,1,1,0,0,0,1,1,1,1,1,1,0,1,0,1,1,1,1,1,1,1,1,1,0,0,0,1,1,
0,1,1,0,1,1,0,1,1,1,1,1,1,1,1,1,1,1,1,0,0,1,1,1,1,1,1,1,0,0,1,1,1,1,0,1,1,1,1,1,1,1,1,1,1,1,
1,1,0,0,1,0,1,1,1,0,1,1,0,1,1,1,1,1,1,1,1,1,1,0,1,1,0,1,1,1,1,1,0,0,0,1,1,1,1,1,1,1,1,1,1,
1,1,1,1,1,1,0,1,1,1,0,1,1,0,1,1,0,1,0,1,1,1,1,0,0,0,1,0,1,1,1,1,1,1,1,1,1,1,0,1,1,0,1,1,1,1,1,
1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,0,1,0,1,1,1,1,1,1,1,1,1,1,1,1,
1,0,1,0,0,1,1,1,0,1,1,1,1,1,1,1,1,1,1,0,1,1,1,0,0,1,1,1,1,1,0,0,0,0,0,1,1,1,1,1,1,0,1,
1,1,0,0,1,1,1,1,1,1,1,0,1,0,1,1,1,1,1,0,1,1,1,0,0,0,1,1,0,1,1,0,1,0,1,1,0,1,0,1,1,1,1,1,0
,
1,0,1,1,1,1,1,0,1,0,0,0,1,1,1,1,1,1,1,1,1,1,0,1,1,0,1,0,1,1,1,1,1,1,0,1,1,1,1,1,0,1,0,0,1,1,0
, 1,0,1)

> data=data.frame(x1,x2,x3,y)

> sam=sample(2,nrow(data),prob=c(0.8,0.2),replace=TRUE)

> data_train=data[sam==1,]

> data_test=data[sam==2,]

> library('e1071')

> model_nb=naiveBayes(y~.,data_train)

> pred_nb=predict(model_nb,data_test)

> cm=table(pred_nb,data_test[,4])

> cm

pred_nb 0 1

```
0  19  18
1  14 101
> a=19
> b=18
> c=14
> d=101
> accuracy=((a+d)/(a+b+c+d))*100
> accuracy
[1] 78.94737
> error=100-accuracy
> error
[1] 21.05263
> sensitivity=(a/(a+c))*100
> sensitivity
[1] 57.57576
> specificity=(d/(b+d))*100
> specificity
[1] 84.87395
```

Conclusion: We can conclude that the fitted Naïve Bayes classification model fits well since it has an accuracy more than 75% with sensitivity around 57% and specificity almost 85%.

H] Clustering using k-means algorithm

Clustering on the basis of –

- Employment Status (Profession)
- Number of OTT Platforms subscribed to.

x1 → Profession x2 → Number of OTT Platforms

Code on R-software:

```
>
x1=c(4,4,4,4,4,4,1,4,1,3,1,1,4,4,3,4,1,1,2,1,1,4,1,1,1,4,4,4,2,2,1,1,1,4,4,4,4,1,2,
2,2,1,1,1,1,4,4,1,1,4,1,2,6,4,1,2,5,1,4,4,2,4,4,1,1,4,2,1,4,4,4,4,4,1,1,2,4,2,1,4,4,
1,3,1,4,4,6,4,1,1,4,4,4,4,4,4,4,4,1,4,4,4,4,4,2,4,4,4,4,4,4,4,4,2,4,4,4,4,4,1,1,1,1,
3,1,4,1,1,2,1,1,2,1,1,4,4,1,2,4,4,4,1,4,4,4,4,3,4,4,4,1,4,4,4,4,4,4,2,4,4,3,4,4,4,4,
4,4,4,4,5,4,4,4,1,4,4,4,3,4,4,1,4,1,1,5,1,4,4,1,4,4,4,2,1,1,4,4,4,2,4,4,4,4,4,4,4,4,
4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,2,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,
4,4,4,4,4,4,4,4,4,4,4,2,4,4,4,4,4,4,4,1,4,4,4,4,4,4,1,4,4,4,4)

>
x2=c(2,5,5,2,2,5,3,2,5,6,1,1,1,3,2,1,1,2,1,2,4,6,3,2,2,4,3,2,2,5,1,1,1,5,1,3,5,3,3,
2,2,2,2,3,5,5,3,1,5,2,6,3,6,1,3,3,5,4,1,2,6,4,5,1,5,2,4,4,1,2,2,1,3,1,1,3,1,5,3,5,1,
2,6,1,5,2,4,4,2,1,4,2,1,3,5,4,2,6,4,4,3,5,5,3,2,2,3,1,2,4,1,8,2,4,4,4,1,3,4,2,4,5,2,
6,3,5,4,5,3,2,5,5,4,13,1,5,2,1,1,1,7,1,3,2,4,3,2,2,1,1,8,7,2,2,1,2,2,1,1,3,4,4,1,4,2
,3,3,4,1,4,1,3,2,3,5,3,2,3,4,1,2,4,2,3,5,4,1,4,2,1,2,1,2,4,3,2,1,4,3,3,2,2,3,2,5,6,5,
1,4,2,2,2,2,3,2,1,1,2,1,1,2,3,1,1,1,1,3,1,1,3,1,3,3,3,1,1,1,12,2,2,1,2,2,5,3,1,1,1,1
,2,3,10,3,3,2,1,1,3,1,3,2,1,3,3,3,1,4,1,4,5,3,1,1,4,3,3,2,1)


> data=data.frame(x1,x2)

> scaled_d=scale(data)

> library('cluster')

> library('factoextra')

> model1=kmeans(scaled_d,5,nstart=15)

> model1
```

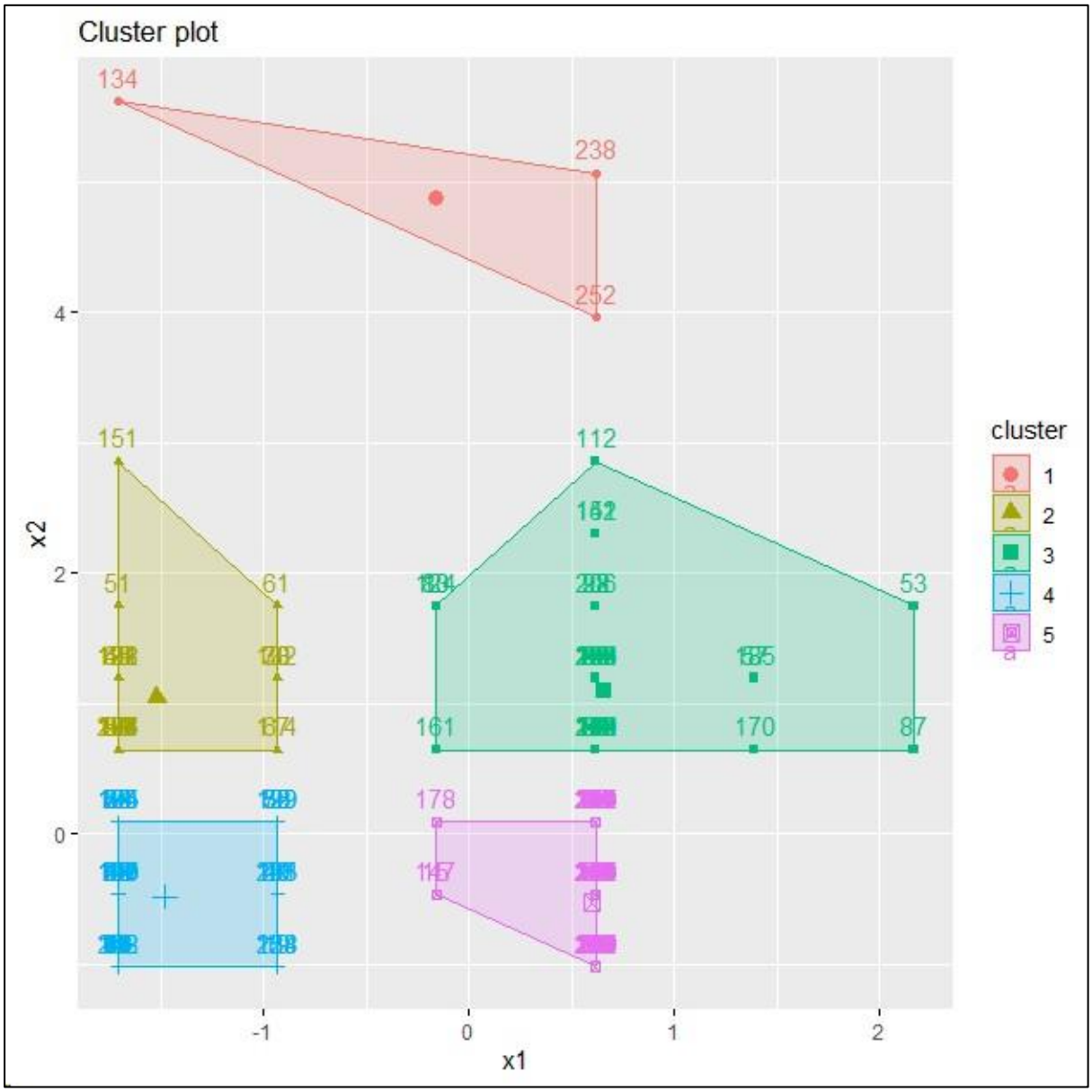
K-means clustering with 5 clusters of sizes 55, 55, 25, 3, 140 Cluster means:

```
       x1        x2
1  0.6582794  1.0934374
2  -1.4832446 -0.4857873
3  -1.5226938    1.0391202  4  -0.1588811    4.8822355  5    0.5994077  -
   0.5288962

Clustering vector:
 [1] 5 1 1 5 5 1 2 5 3 1 2 2 5 5 5 5 2 2 2 2 3 1 2 2 2 1 5 5 2 3 2 2 2 1 5 5 1
 [38] 2 2 2 2 2 2 2 3 1 5 2 3 5 3 2 1 5 2 2 1 3 5 5 3 1 1 2 3 5 3 3 5 5 5 5 5 2  [75]
2 2 5 3 2 1 5 2 1 2 1 5 1 1 2 2 1 5 5 5 1 1 5 1 3 1 5 1 1 5 2 5 5 5 5 1 5
[112] 1 5 3 1 1 5 5 1 2 3 3 2 1 2 1 3 3 2 2 3 3 3 4 5 1 2 2 5 5 1 2 5 5 1 5 5 5
[149] 5 5 3 1 5 5 5 5 5 2 5 5 1 1 5 1 5 5 5 1 5 1 5 5 5 2 1 5 5 5 1 5 2 1 2 2 1
[186] 3 5 1 2 5 5 5 2 3 2 5 5 1 2 5 5 5 5 5 1 1 1 5 1 5 5 5 5 5 5 5 5 5 5 5 5 5
[223] 5 2 5 5 5 5 5 5 5 5 5 5 5 5 5 4 5 5 5 5 5 1 5 5 5 5 5 5 5 4 5 5 5 5 5 5 5
[260] 5 2 5 5 5 5 5 1 2 1 1 5 5 5 3 5 5 5 5

Within cluster sum of squares by cluster:
[1] 23.043290 17.196449  9.177709  5.031065 29.973244
 (between_SS / total_SS =  84.8 %)

Available components:
[1] "cluster"     "centers"     "totss"       "withinss"    "tot.withinss"
[6] "betweenss"   "size"        "iter"        "ifault"
> fviz_cluster(model1,scaled_d)
```



Conclusion: We have got 5 clusters using K-means algorithm for profession and number of OTT platforms used. We can observe that the 170th observation is included in the third cluster.

Conclusions & Observations

Through this project we have studied the use of OTT among citizens. Some of the observations made from the above all the analysis are listed below:

- Maximum users have subscribed to Disney+ Hotstar, followed by Netflix, Amazon Prime video and Jio Cinema respectively.
- From the pie-chart we can observe that only about 40% of the users renew their subscription regularly.
- The proportion of users sharing their subscription is almost similar (45%-60%) in all platform except in among Jio-Cinema where its only almost 70%.
- Mobile is the most used device to stream OTT content. TV, Laptop and PC are also by considerable number of people. Tablet is the least used device with less than 10% users on all platforms.
- Through the Shapiro-Wilk test we can conclude that the collected data is not Normally distributed.
- According to the Chi-Square test for goodness of fit, the Exponential Distribution fits well to the average weekly OTT usage times.
- The results of the test for equality of proportions suggest that the proportion of individuals who renew their subscriptions regularly is not same among students and individuals from other professions.
- In the Chi-Square test for independence of attributes, the null hypothesis is rejected. Hence, we can conclude that the type of content viewed and OTT platform used are independent of each other
- The Naïve Bayes classification model is fitted on the basis of
  - Use of OTT per week
  - Customer's view about the price point
  - Customer's rating of the OTT platform

  The model fits well and predicts customer satisfaction with an accuracy of almost 80%.
  - The clustering using k-means algorithm results in 5 distinct clusters.

Limitations

- The sample in the survey may not be a proper representation of all the users in the city.
- A sample of size 278 is relatively small and it may be inappropriate to draw conclusion regarding all users.
- As the variation in the data increases the results may differ from the significantly.
- We rely on respondents to provide accurate and correct information. Arbitrary and inaccurate information may affect the precision of the analysis.
- We have used basic software such as MS-Excel, R-software and Python for the analysis. The use of advanced software may help in a more precise observation.

References

1. Nirali Prakashan B.Sc. Textbooks –
    - ST – 111 – 'Descriptive Statistics – I' by Dr. P. G. Dixit and Dr. (Mrs.) V. R. Prayag.
    - ST – 232 – 'Continuous Probability Distributions' by Dr. P. G. Dixit and Prof. P. S. Karpe.
    - ST – 241 – 'Test of Significance and Statistical Methods' by Dr. P. G. Dixit and Prof. P. S. Karpe.
    - ST – 242 – 'Sampling Distributions and Exact Tests' by Dr. P. G. Dixit and Prof. P. S. Karpe.

2. Reference book 'Statistical Methods' by W Snedecor and G Cochran.

3. Website - https://real-statistics.com/.