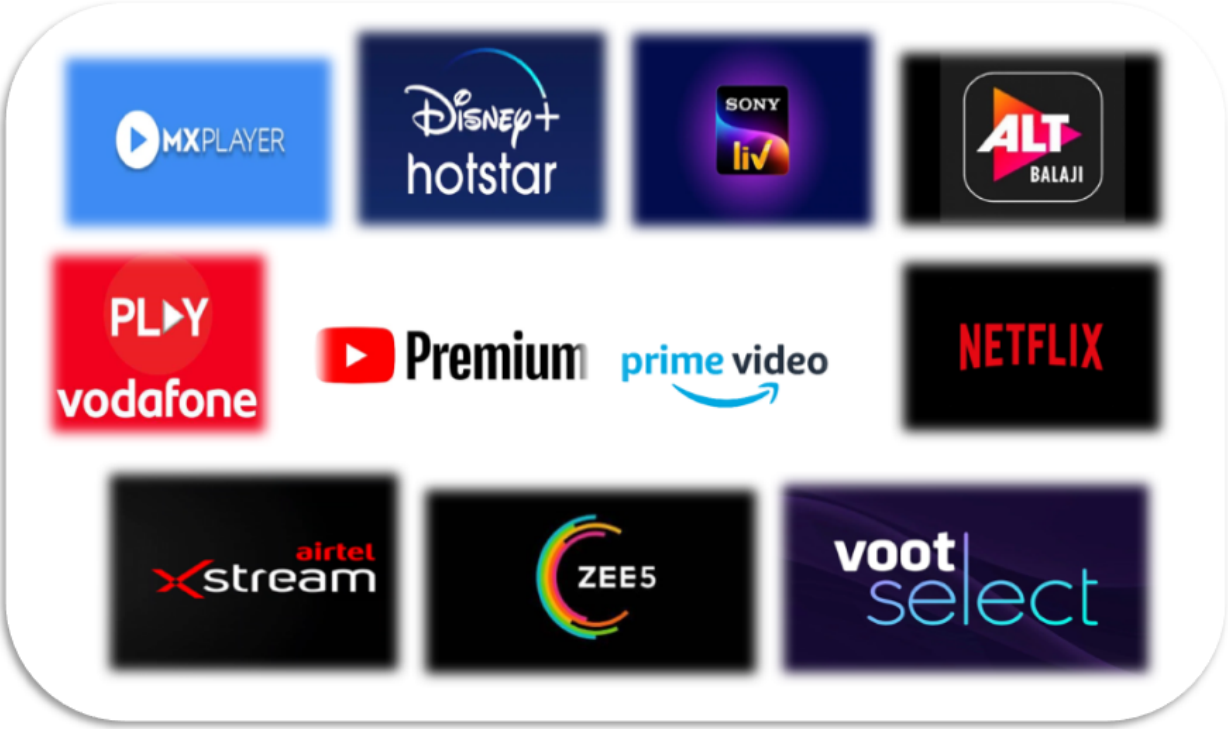


Growth of Over The Top(OTT) Video Services and Factors Affecting It

Acknowledgement

We would like to thank our Principal, Vice – Principal, all the faculty members of M.Sc. Statistics Department of KC College and the KC College office staff for providing us with the necessary infrastructure every time we needed one.

We express the deepest gratitude to our guide and co-ordinator **Mrs. Shailaja Rane** for guiding and helping us whenever required.

And a huge thankyou to each and every one who has helped us in every possible way, right from the start of our project by means of pitching in their ideas, collection of data to the completion of our project.

Last but not the least, we would like to give our special thanks to our respondents who diligently and honestly filled our survey purely for academic purposes and for taking the time out of their schedule to answer our questions genuinely.

TABLE OF CONTENTS

---

1. Abstract

OTT- Over The Top media is an online platform for users to view content like movies, news, etc. OTT dodges other content-watching platforms like cable, satellite television. Some of the OTT platforms include Netflix, Disney+ Hotstar, Amazon Prime, Zee5, Hulu, etc. Throughout the world, the number of people watching content on OTT platforms has increased over the time.

The study aims to determine awareness of online streaming services, growth, factors affecting it and finally determining a prediction model.

A questionnaire was designed for data collection and the data was collected through google forms and direct personal investigation. Statistical techniques like chi-square test, pareto analysis, binary logistics regression, K-Means cluster analysis, xg-boost analysis, SWOC analysis, Apriori algorithm & graphs were applied for analysis. Python, R, SPSS, MS Excel softwares are used for analysis. The study is based on 625 subjects out of which 405 are OTT users & 220 are non-OTT users. The results of the study revealed that the main reasons for not using OTT are that most viewers are not interested, it is time-consuming, due to the cost of subscription and lack of awareness. Amazon Prime, Netflix & Disney+ Hotstar are leading OTT platforms. According to the sample data, the main reasons behind the growth of OTT over traditional services are its flexibility, convenience, more entertaining, the entire season can be watched at once, more interest in OTT content and availability of the latest movies. The selection of online services is not associated with the Gender of a person. The study shows that there is an association between marital status, education and income with the awareness of OTT. The study helps to classify viewers based on their preference and predict which type of subscribers will prefer the leading OTT platform. From Logistic regression, the important factors affecting the subscription of OTT platforms are Age, Family Income and Awareness about OTT platforms. From the Apriori algorithm, viewers like to watch Comedy, Thriller, Crime with Action and Horror. They also like to watch Drama and Horror with Romance or Crime.

Keywords: OTT platform, Television, SWOC analysis, Pareto Analysis, Apriori analysis, XG-Boost, Traditional Services, Internet, SPSS.

2. Introduction

Over the years the importance of Television is fading. These days it is not necessary to have a Television to watch your favorite show as almost all the shows can be watched using Internet connection. Dozens of live streaming platforms across the globe provide different types of content such as movies, TV series, news, sports events, etc. Streaming services are an addition to the digital download contribution and DVD with a trickle of secondary movies and TV shows. Streaming services are now second-run movies and TV shows. With an increase in speedier internet connection, the usage of video streaming devices have accelerated resulting in the decline of traditional cable networks. More and more viewers are shifting to OTT from traditional services because of the availability of more alternatives.

Video Streaming is a technology that has completely changed the entertainment industry as well as consumption models among audience members. Tons has changed since that very first Real Player transmission in 1995. Since then, technology has been continuously upgrading, creating content and delivering with an easier approach.

Video content used to be supplied to the consumers by cable distributors with a box connected to a television. As technology evolved, the internet became a big part of our lives and with it, mobile connectivity also developed. This resulted in content providers such as Disney+ Hotstar, ZEE5, Netflix inaugurating a different type of television model. With these models, users could stream TV shows on any device connected to the internet. This includes smartphones, laptops, televisions and tablets. Even though the subscription of OTT platforms is more costly, still they are more popular than traditional services.

As OTT platforms come with different payment models they are more preferred than traditional closed TV infrastructure. Hence the term OTT or over the top. Netflix is one of the best examples of the Business model for Video Streaming. With millions subscribing to the service all over the world, the company has found a way to capitalize its services using its title stock and outsourced infrastructure.

We have seen some repercussions of Video Streaming. Film critics feel the viewers may turn “platform agnostic” consuming content, irrespective of the size of the screen or the quality of the image. But, viewers have proven the opposite. It is observed that they are willing to return to cinema halls if the movie is worth its price. However, content abundance has made audiences “socially autistic” by continuously connecting themselves to a device and getting isolated from others. Apparently, audiences are willing to sacrifice their social experiences “offline” for the sake of personalized content.

Historical concept

The term OTT (Over The Top) was first termed as Telco-OTT, by telecoms industry analyst Dean Bubley in February 2012, in which a telecommunications service provider supplies one or more services across an IP network. The IP network may be a public network which runs through the corporation's existing IP-VPN from another provider, as against the carrier’s own access. This supports an eclectic range of telco services including communications, content (TV and music) and cloud- based services.
In 2008, BIGFlix was the first dependent Indian OTT platform which was launched by Reliance Entertainment. In 2010, nexGTv was India’s first OTT mobile app which was launched by Digivive, which provides access to both live TV and on–demand content. In 2013 and 2014, nexGTV was the first app to live-stream Indian Premier League matches on smartphones. After the launch of Ditto TV (Zee) and Sony Liv in 2013 OTT actually gained momentum in Indian market. The main aggregator was DittoTV, a platform containing shows across all media channels including Star, Sony, Viacom, Zee, etc. One of the highest watched OTT platforms in India is Hotstar, now named Disney+ Hotstar, which was launched in 2015. As of April 2021, Hotstar has 34.5 million paid users whereas it has more than 300 million active users.Well known global OTT platform Netflix began its operations in India in the year 2015. With popular players like Amazon Prime, Disney+Hotstar,etc, Netflix faces stiff competition in India with other popular OTT platforms.
Amazon Prime and Voot were launched in 2016. Eros Now awarded as the ‘Best OTT Platform of the Year 2019’. MX Player is the youngest OTT platform active from 2018. There was significant growth in the consumption of OTT which could be menacing to the Television industry. Due to the national lockdown, people were confined to their homes and were unable to go to theatres. Because of no new content available for streaming,

past content is shown repeatedly on the television due to this viewers are opting to watch foriegn shows on OTT platforms.

Reason Behind Choosing This Topic :

The whole world is affected by Coronavirus and social distancing and quarantine have become a new norm are the most correlated terms with the present circumstances. Staying inside the house would've been extremely boring had it not been for the endless trove of content that's provided on streaming services like Netflix, Amazon Prime, or Disney+ Hotstar, etc.
Many people around the world started the consumption of internet and technical devices for streaming OTT platforms. OTT platforms have a variety of genres that can be viewed by any age group. With the increased usage of OTT platforms amongst kids, youngsters, and adults there was a thought to identify the reasons for the rise in the popularity of OTT platforms.
The motive behind the topic of this survey was to address a larger group of people and not be confined to a particular group of people.

3. Literature Review

There are no gender differences when it comes to having access over the OTT apps. Both girls and boys are equally drawn into the world of streaming culture. 50% of teenagers use the OTT apps for entertainment purposes, around 25% for their academic purposes whereas the remaining 25% of them use it for both. 50% of college going students prefer watching foreign shows, only 20% of students prefer watching Indian shows or movies while the remaining 30% prefer both (Reshma and Chaitra)[1].

During the lockdown 66.3% viewers say they spend 0 to 2 hours watching the OTT applications, 24.7% viewers spend 2 to 4 hours and the remaining 9% respondents say they spend more than 4 hours watching OTT applications in a day. 65.2% of viewers are satisfied with subscribing 0 to 2 OTT channels whereas 20.2 % viewers subscribed to 2 to 4 channels and remaining 14.6% subscribed for more than 4 channels. 82% of viewers stated that they watch less than 2 movies in cinema halls in a month whereas 13.5% viewers said that they watch 2 to 4 movies in a month in a cinema hall and the remaining 4.5% watch more than 4 movies. 41.6% of viewers say they still prefer watching movies in the cinema hall as before. (Manoj Kumar Patel[1], et al.)[2]

.

Approximately 44 percent of viewers rate Netflix as the most used OTT channel for watching online video content.The most preferred device among users for watching OTT content is smartphone. Approximately 56 percent of the users use smartphones,26 percent of the respondents feel that the content openness on OTT platforms prevents many from using OTT services. Lack of technical knowledge and cost are other important reasons for customers not using OTT services (Tripti Kumari)[3].

More than 50% preferred Hotstar over other apps and Netflix could garner only 41% and the remaining users chose other apps over Hotstar and Netflix.The reason for the popularity of Netflix is TV series and Movies. Hotstar is more popular for its Sports Section and Live Streaming facilities offered for International Cricket and IPL(Rachita Ota,et al.)[4].

There was a significant positive relationship between the online media adoption and how easy it was to use. There is no statistical relationship between cost and online streaming. Customer service is the main driver to customer satisfaction while social trends persuade the adoption of online streaming(C. Christopher Lee, et al.)[5].

4. Objectives

Sr no.	Objectives	Techniques Used
1.	Identification of leading OTT apps.	Graphical Representation

2.	To identify the reasons for not using OTT.	Pareto Analysis
3.	Reasons behind switching over to Online Video Streaming (OTT) from Traditional services.	Pareto Analysis
4.	To compare and classify the viewers based on their preference.	K-means Cluster Analysis
5.	To predict which type of subscribers will prefer the leading OTT platforms.	XG-Boost Analysis
6.	To study the socio- demographic factors that affect the usage of OTT platforms.	Binary Logistic Regression
7.	To understand the association and frequency between different genres.	Apriori Algorithm
8.	To identify whether people are aware of streaming services or not based on factors	Chi-Square Test
9.	To understand Strengths, Weaknesses, Opportunities & Challenges for OTT platforms.	SWOC Analysis
10.	To prepare a diagrammatic representation for easy identification of choice of OTT platforms and suggestions.	Word Cloud

5. Material and Methods

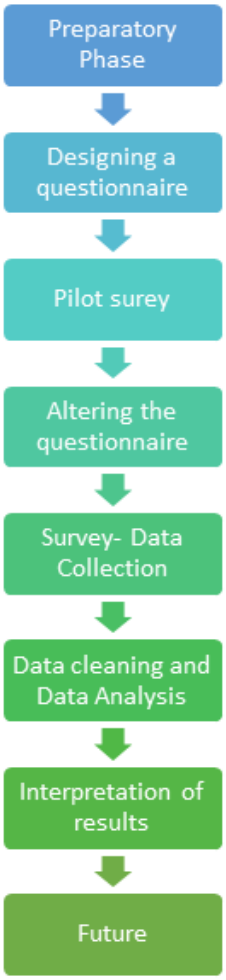
Keeping objectives in mind, the questionnaire was prepared. The types of questions included in the Questionnaire were:

- **Dichotomous Questions**
- **Multiple Choice Questions**
- **Rating Questions**
- **Likert Scale Questions**
- **Categorical Questions**

Due to the pandemic a convenience sampling method was used. Survey for collection of information was conducted online using WhatsApp, Facebook, LinkedIn, and mail. Also the survey was conducted offline from housing societies, family groups, etc.
A pilot study on 100 sampling units was conducted to check the reliability of the questionnaire. After a slight modification, the questionnaire was improved before using it for the actual survey. A sample size of 625 sampling units was selected consisting of 405 OTT users & 220 non-OTT users. 29% of the people have filled the questionnaire on Tuesday, while a very less number of people prefer to fill it on weekends.This may be because the users may prefer to relax on the weekend.
Outliers were replaced using percentiles. Columns which contain multiple options were converted into single separate variables using ‘if else’ condition. Categorical data was converted into integers using label encoder.
The statistical analysis was conducted using statistical techniques such as:

- **Graphical Representation**
- **Pareto Analysis**
- **K-means cluster Analysis**
- **XG-Boost Analysis**
- **Binary Logistics Regression**
- **Apriori Learning**
- **Chi-Square Test**
- **SWOC Analysis**
- **Word Cloud**

STEPS INVOLVED IN CONDUCTING THE SURVEY:


Preparatory Phase
Designing a questionnaire
Pilot surey
Altering the questionnaire
Survey- Data Collection
Data cleaning and Data Analysis
Interpretation of results
Future

6. Analysis and Results

6.1 Exploratory Data Analysis :

The main purpose of graphical representation is to readily give some idea about the entire data and draw instant conclusions.

- **DEMOGRAPHIC DETAILS**
  - The sample consists of 76% Unmarried individuals.
  - The data contains almost equal numbers of Males and Females.
  - Majority of the respondents are from age 22-31.The average age of the sample is approximately 27 years (27 ± 11.3).
  - The maximum number of responses was from 4 family members.
  - 51% of the respondents are Graduates (bachelors) while 39% are Post Graduates or Above (included Masters)
  - Average annual family Income is 5.5 lakhs.
  - Majority of the OTT users are from the Urban area, while only 6% of the viewers are from the Rural area.

- **EMPLOYMENT STATUS AND LIVING AREA OF TOTAL POPULATION:**

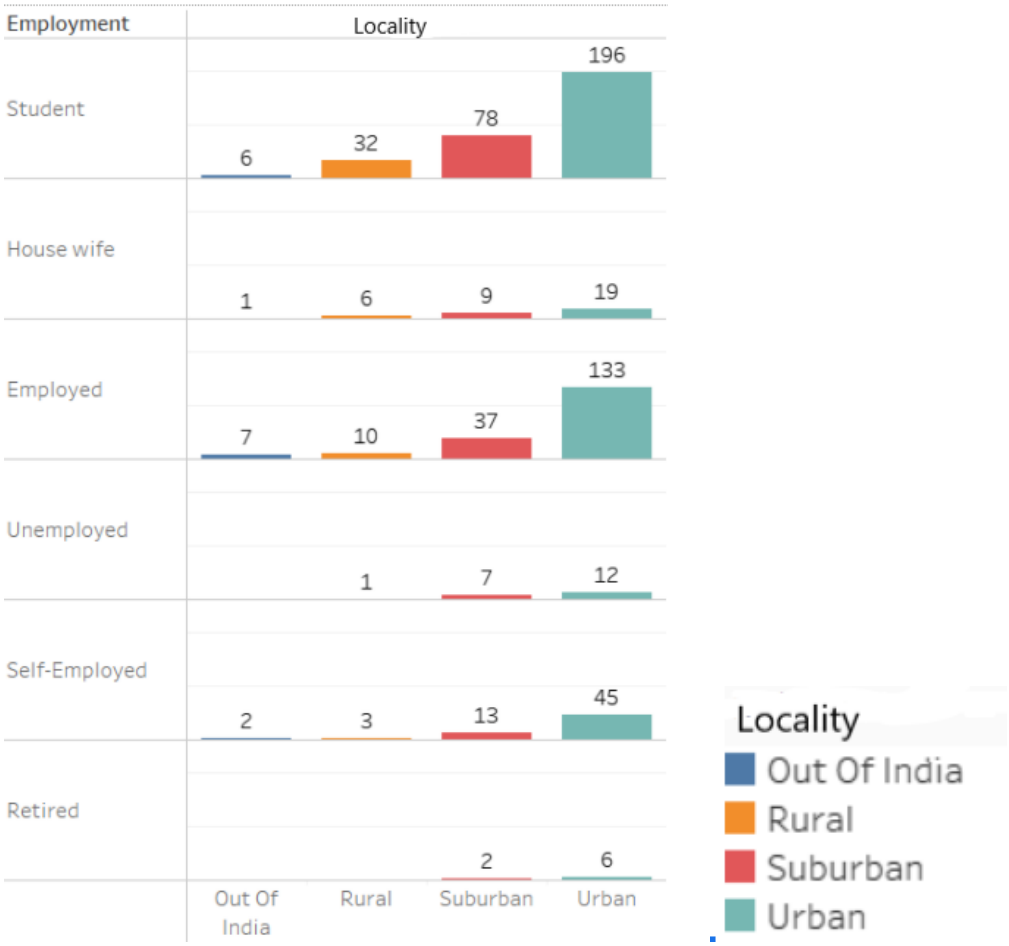


Fig 6.1.1

- **AWARENESS ABOUT OTT PLATFORMS AND USAGE OF OTT PLATFORMS**

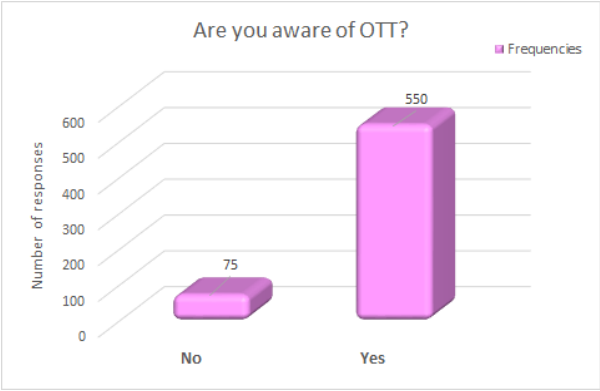


Fig 6.1.2

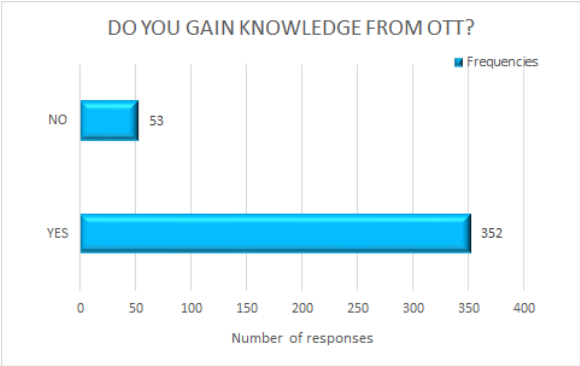


Fig 6.1.3

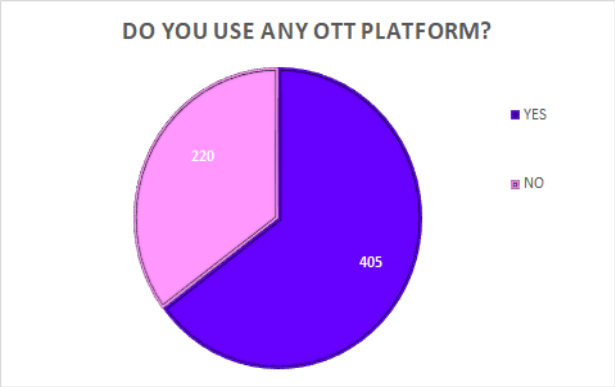


Fig 6.1.4

Interpretation:

88% of the people are aware of OTT platforms of which 73.55% use them.From this it is seen that there are few people who do not use OTT platforms even though they are aware of them. 35% (220 people) of the sample do not use OTT. 87% of the viewers gain knowledge from the content they watch on OTT.

- **REASONS FOR NOT USING OTT (For those who do not use OTT)**

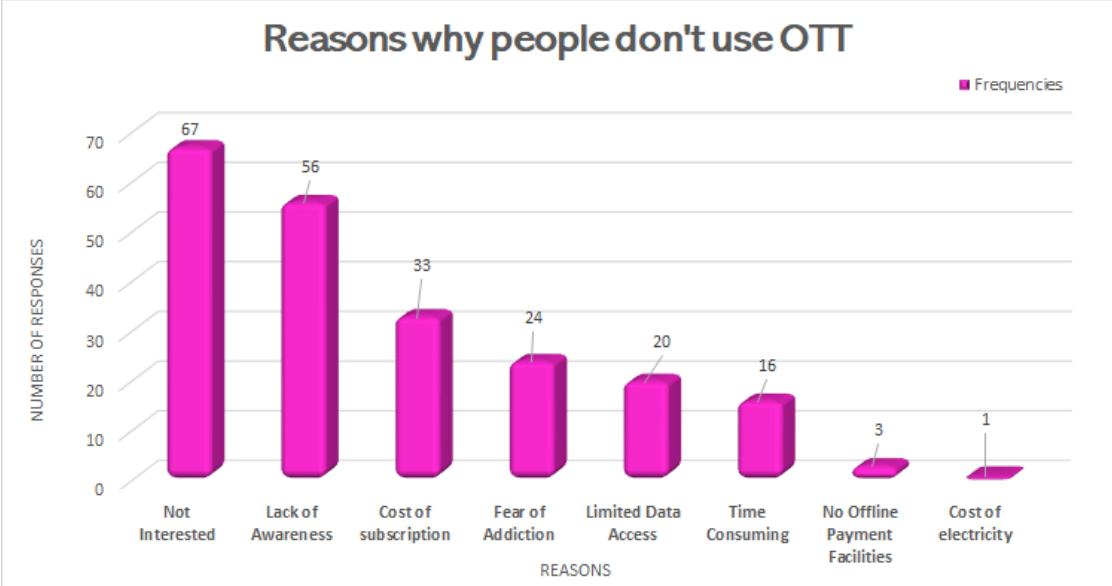


Fig 6.1.5

Interpretation:

Majority of people are Not Interested (30%) in watching OTT platforms. Whereas 25% of the people have said that they are not Aware of OTT and around 26% have said that the Cost of subscription and Fear of Addiction are the reasons for not choosing OTT.

- **FIRST GAINED KNOWLEDGE ABOUT OTT THROUGH**

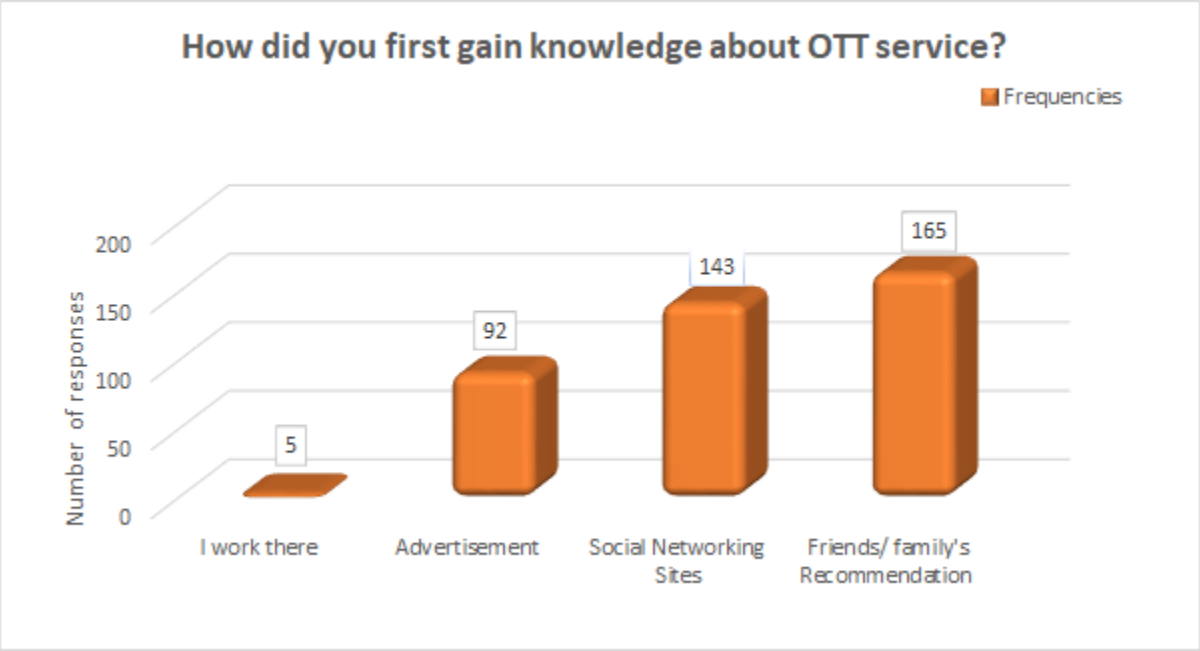


Fig 6.1.6

Interpretation:

Out of 405 people who use OTT, 40.7% first got to know about OTT from Friends/ family. 35% of people first got to know about OTT from Social Networking Sites and 23% came to know through Advertisement.

- **TRADITIONAL SERVICES OR OTT PLATFORMS**

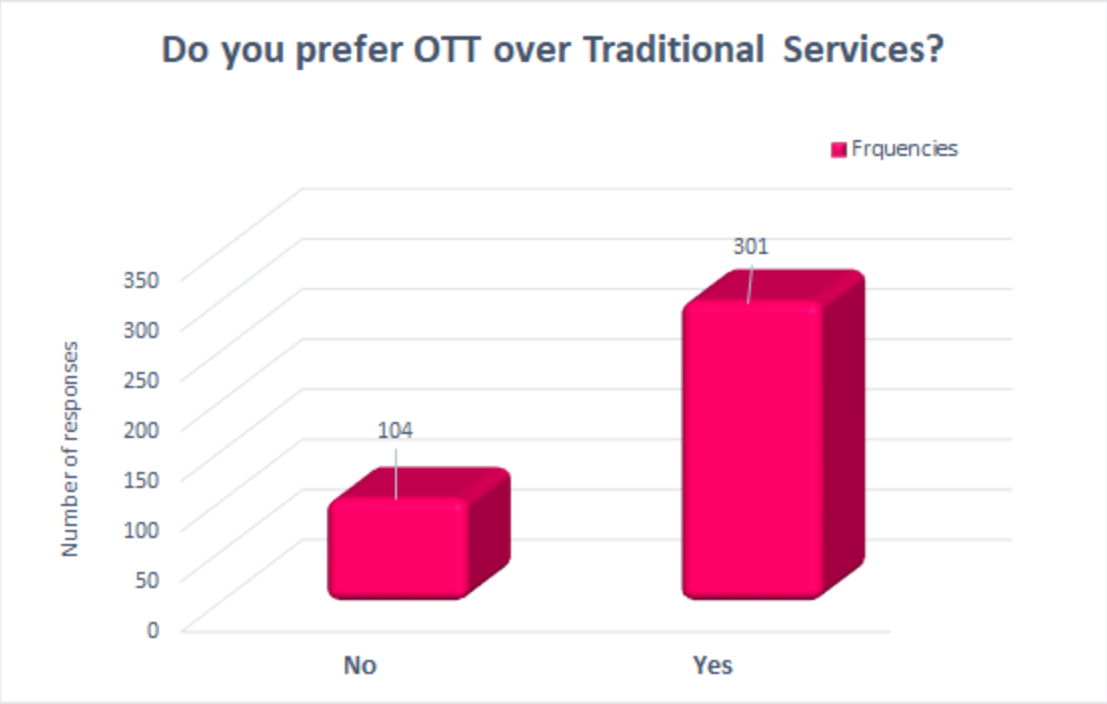


Fig 6.1.7

Interpretation:

About 74% of the respondents opt for OTT services via WiFi or Mobile data. But 26% of the present users may switch over to Traditional services.

- **USING OTT PLATFORMS SINCE**

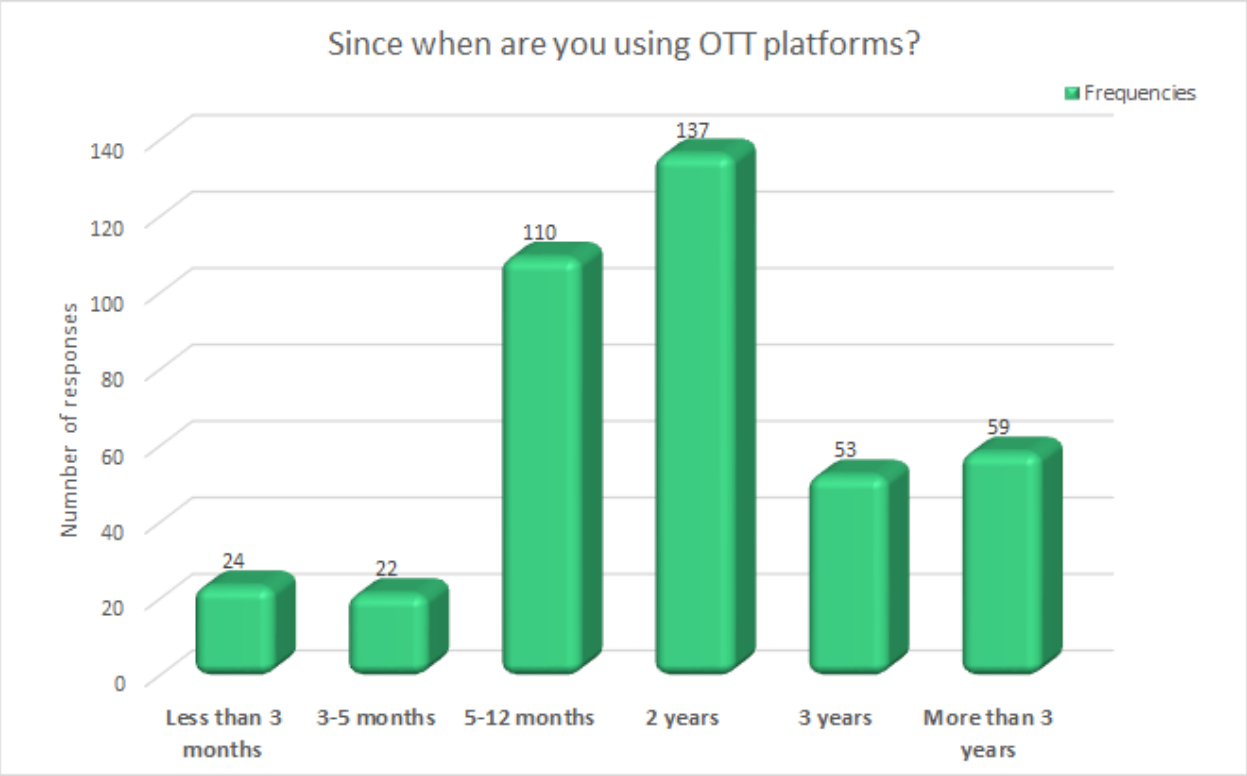


Fig 6.1.8

Interpretation:

There is an increase of 39% subscriptions during Covid.

- **VIEWING HOURS BEFORE LOCKDOWN AND DURING LOCKDOWN DUE TO COVID-19**

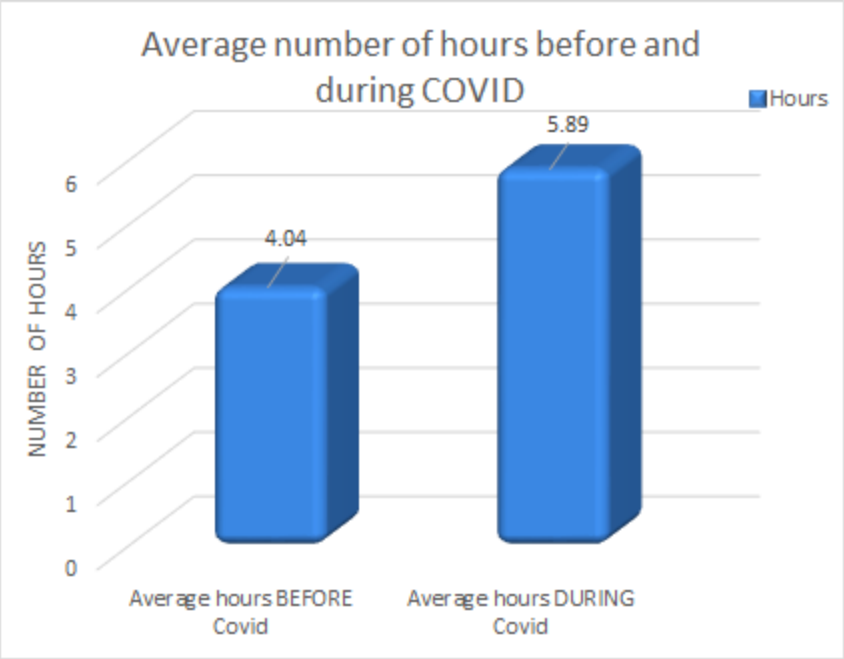


Fig 6.1.9

Interpretation:

The average number of viewing hours increased from 4.04 before Covid to 5.89 per week during Covid.

- **GAINED WEIGHT DUE TO BINGE WATCHING**

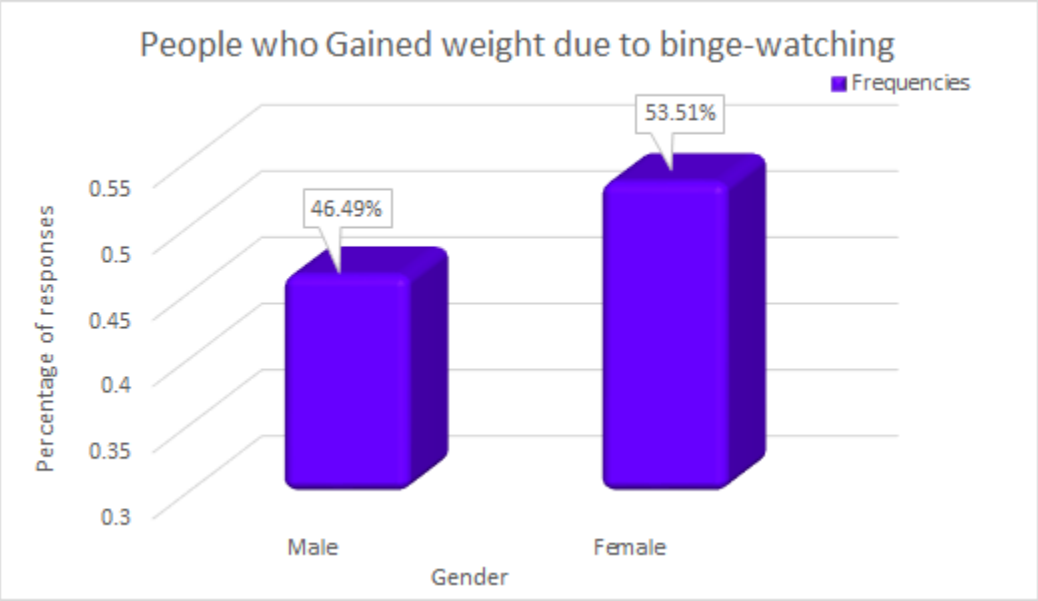


Fig 6.1.10

Interpretation:

Binge watch is an act of continuously watching for a long period of hours.

The percentage of Females who have gained weight is more than the percentage of Males who have gained weight due to Binge watching.

- **REGULAR VIEWERS**

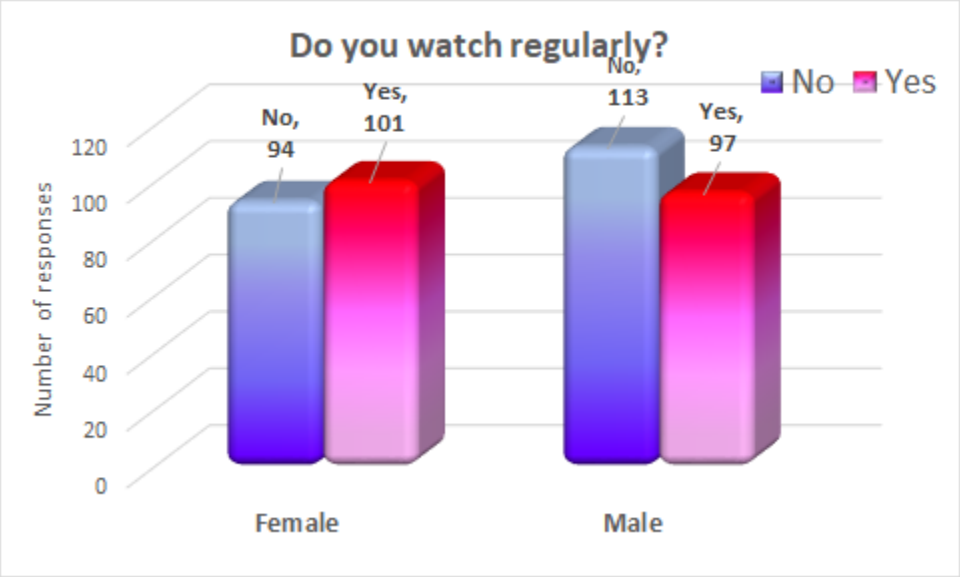


Fig 6.1.11

Interpretation:

The percentage of regular viewers among Males and Females is almost the same.

- **FUTURE OF OTT**

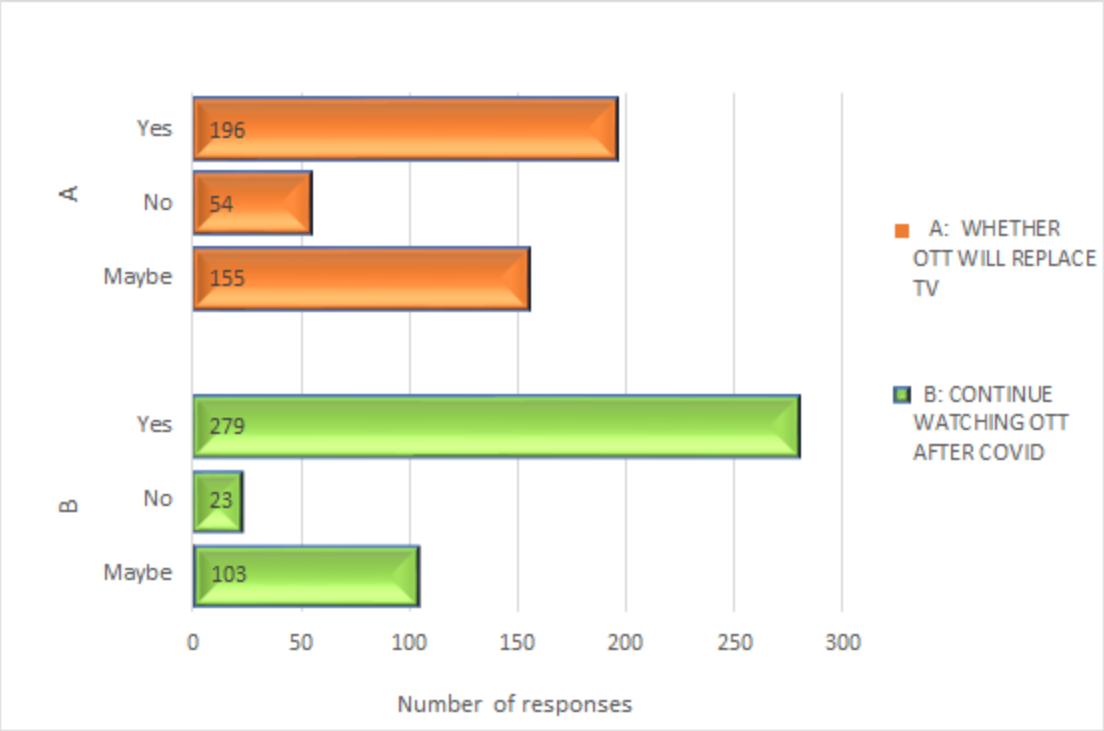


Fig 6.1.12

Interpretation: Around 69% of the viewers are sure about continuing with OTT but only 48% feel that OTT will replace Television.
31% of the people are not sure about using OTT after Covid.

- **SOURCE OF INTERNET PREFERRED BY VIEWERS**

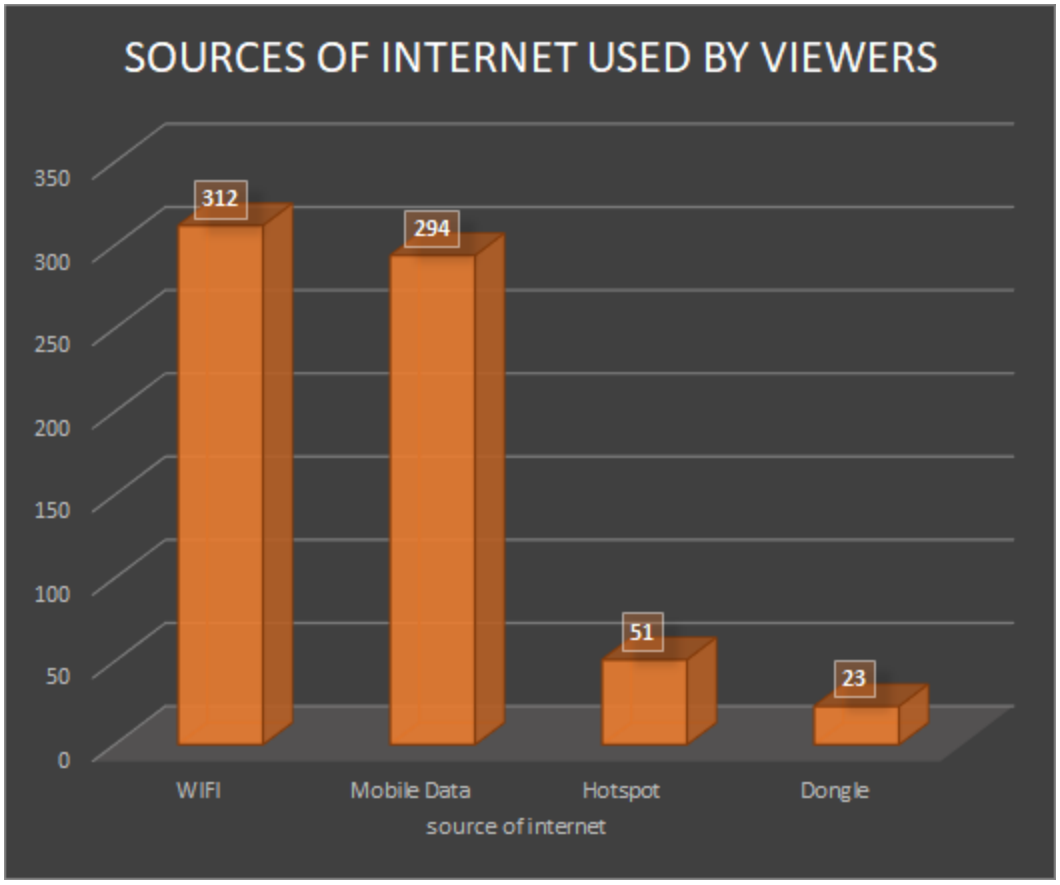


Fig 6.1.13

Interpretation:

Majority of the viewers use WiFi or Mobile data.

- **PREFERABLE TIME FOR STREAMING**

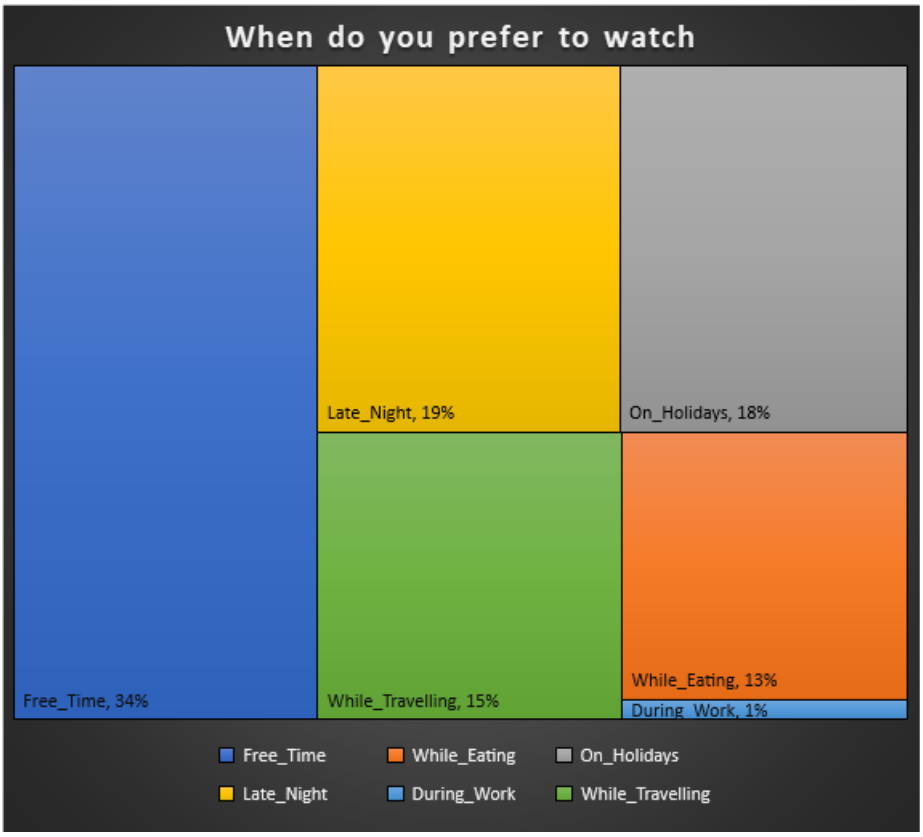


Fig 6.1.14

Interpretation:

Maximum users stream during leisure hours, also 19% users stream during late night.

- **GENRES PREFERRED BY VIEWERS**

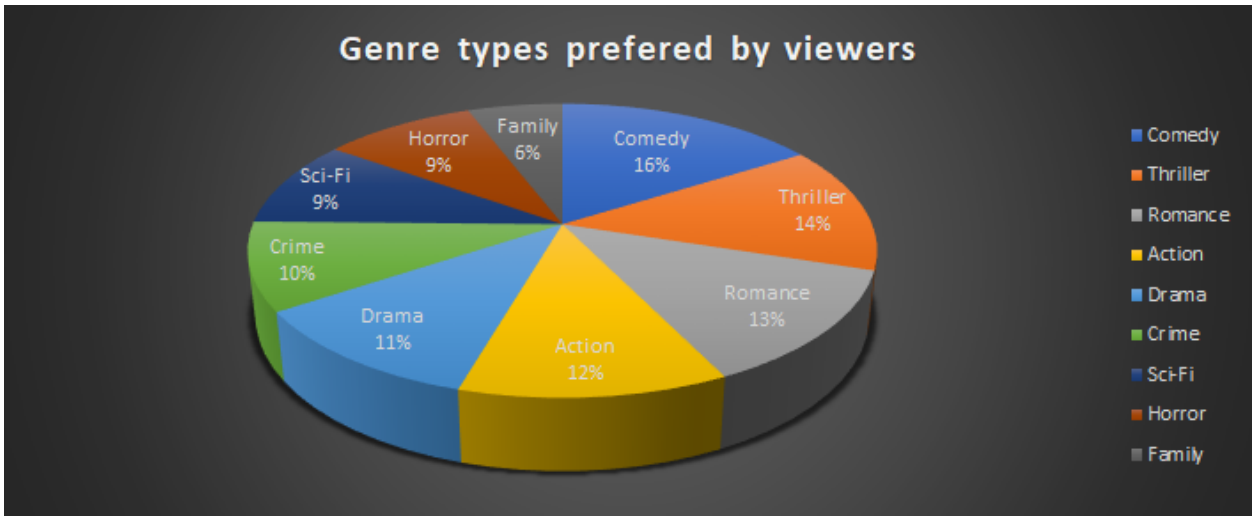


Fig 6.1.15

Interpretation:
According to the user's preference, the top 5 genres are Comedy, Thriller, Romance, Action, Drama.

- **TYPE OF CONTENT PREFERRED BY VIEWERS**

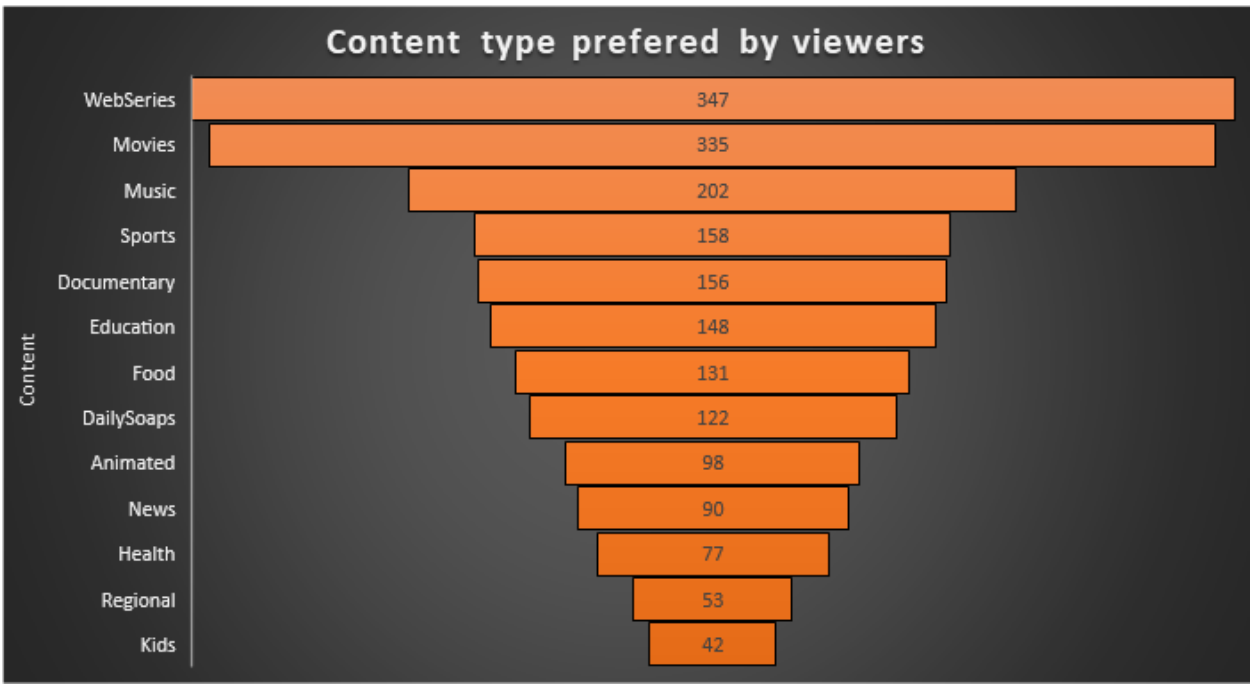


Fig 6.1.16

Interpretation:
Viewers prefer Web series & movie content more as compared to other types of content.

- **OTT PLATFORMS USED BY USERS**

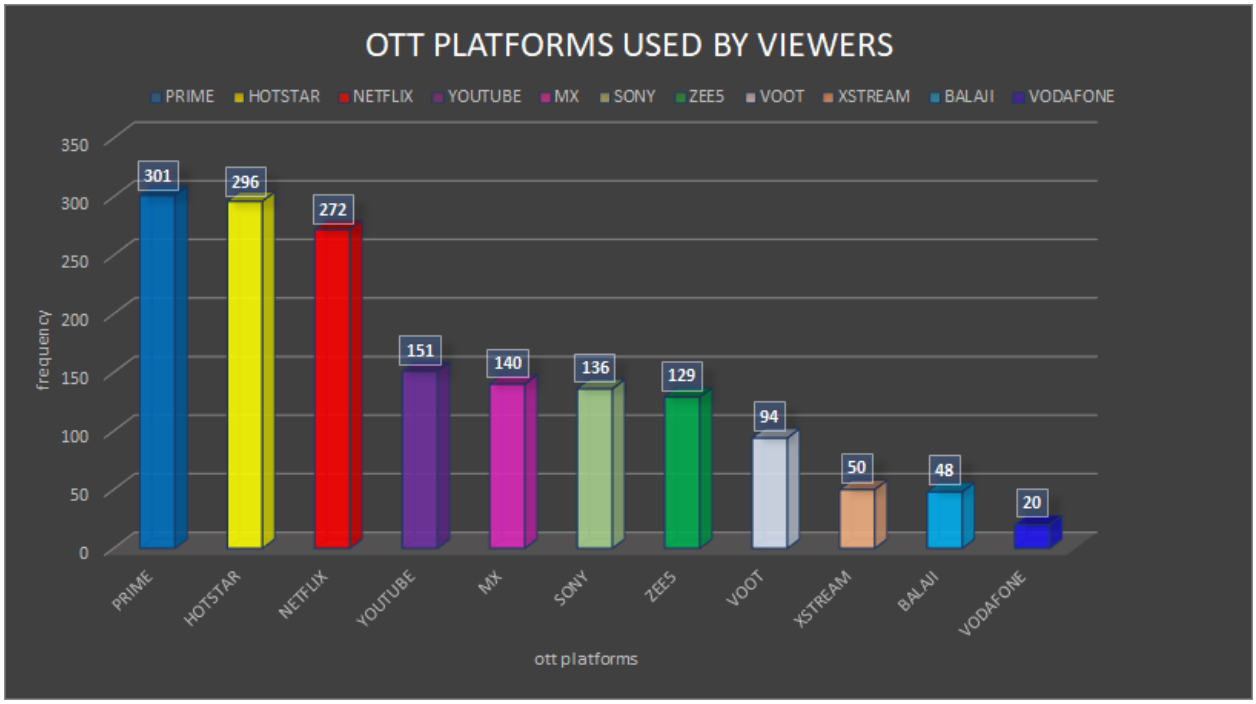


Fig 6.1.17

Interpretation:

Amazon Prime, Disney+ Hotstar, Netflix,Youtube,MX Player are top 5 OTT platforms used by users.

- **OTT SUBSCRIPTIONS TAKEN BY VIEWERS**

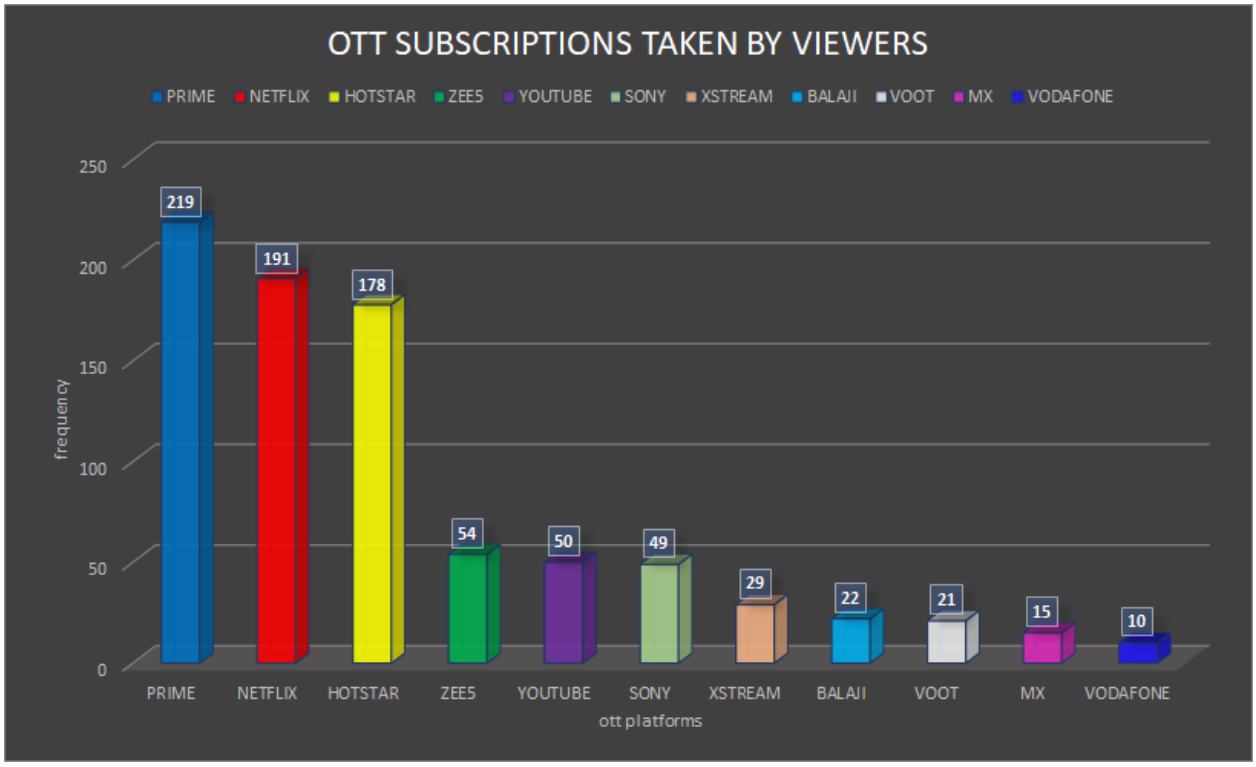


Fig 6.1.18

Interpretation:

Maximum user's take a subscription to Amazon Prime, Netflix, Disney+ Hotstar, ZEE5, and Youtube. From above both graphs it concludes that Amazon prime is the leading OTT platform.

6.2 OBJECTIVE:

To Identify reasons for not using OTT.

Technique used:Pareto Analysis

Tool used: MS Excel

The variables used in the analysis are

Not Interested
Time Consuming
Cost of subscription
Lack of Awareness
Fear of Addiction
Limited Data Access
No Offline Payment Facilities
Cost of electricity

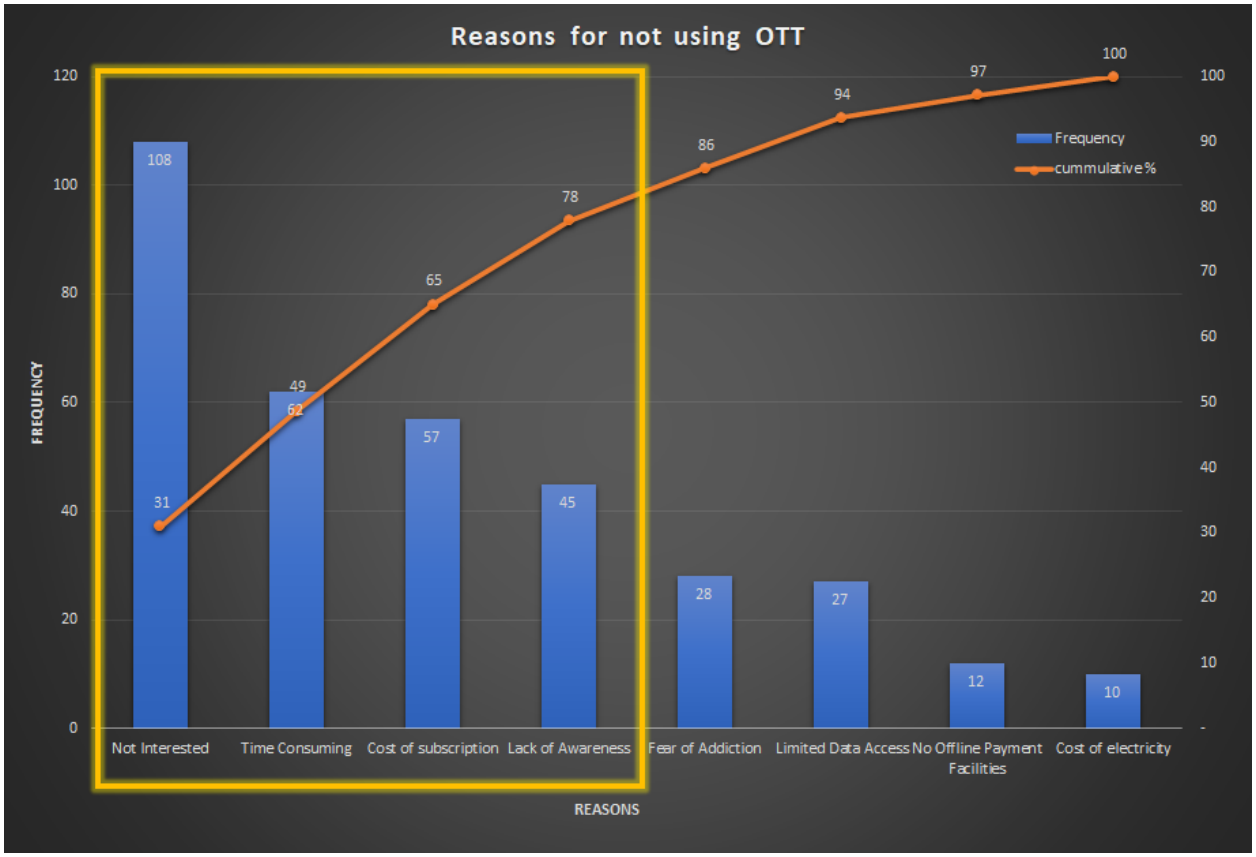


Fig 6.2.1

<u>Interpretation</u>:

From Pareto Analysis, it is conclude that the main reasons for not using Ott are 'Not Interested', 'Time consuming', 'Cost of subscription', 'Lack of Awareness'

6.3 OBJECTIVE:

Reasons behind switching over to Online Video Streaming(OTT) from Traditional services.

Technique used: Pareto Analysis

Tool used: MS Excel

The variables used in the analysis are,

Flexible & Convenient
More entertaining
Can watch the entire season at once
Personal Interest
Latest movie release
Stressbuster
Affordable Prices
To improve language
Peer Pressure

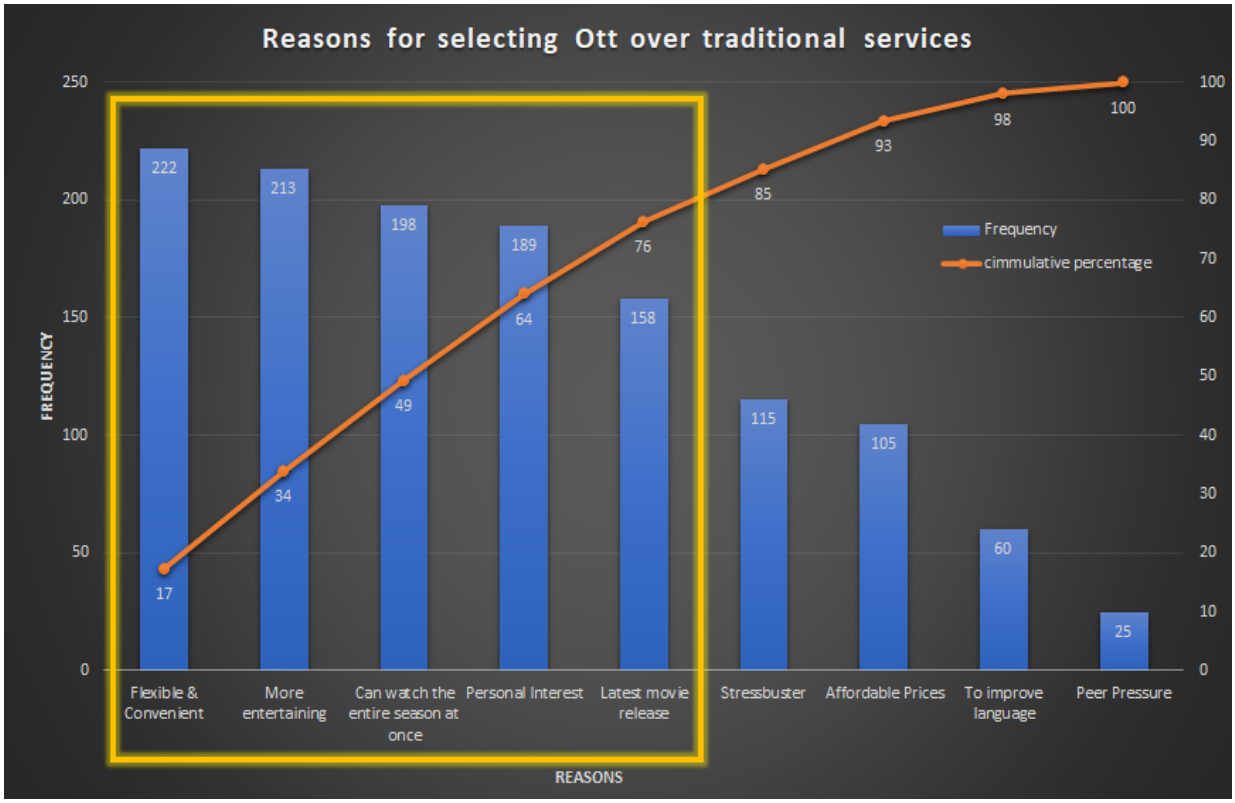


Fig 6.3.1

<u>Interpretation</u>:

From Pareto Analysis, it is concluded that the main reasons for selecting OTT over traditional services are 'Flexible & convenient', 'more entertaining', 'can watch the entire season at once', 'personal interest', 'Latest movie release'.

6.4 OBJECTIVE:

To compare and classify the viewers based on the user's preference.

Technique used: K-means Cluster Analysis
Tool used: Python software.

Variables used in analysis are:

1	Gender	36	Device_Mob	71	OTT_Subscription_SONY
2	Age	37	Device_stick	72	OTT_Subscription_MX
3	Marital Status	38	Device_Other	73	OTT_Subscription_VOOT
4	Education	39	No_of_Device	74	OTT_Subscription_BALAJI
5	Employment	40	Purchased_TV	75	OTT_Subscription_XSTREAM
6	Area	41	Purchased_Stand	76	OTT_Subscription_VODAFONE
7	family members	42	Purchased_Headphones	77	Total_Subscription
8	Income	43	Purchased_internet	78	Content_Daily_Soaps
9	Awarness about OTT	44	Purchased_data_plans	79	Content_Web_Series
10	OTT user	45	Purchased_Nothing	80	Content_Music
11	Platform for entertainment	46	No_of_Purchased	81	Content_News
12	How did you know about OTT	47	Ideal_time_FreeTime	82	Content_Movies
13	OTT over Traditional	48	Ideal_time_Travelling	83	Content_Sports
14	Subscription BL	49	Ideal_time_Holidays	84	Content_Animated
15	Duration of OTT use	50	Ideal_time_Eating	85	Content_Health
16	Yearly spend on OTT	51	Ideal_time_Night	86	Content_Food
17	Regular watching	52	Ideal_time_Work	87	Content_Kids
18	time spent BL	53	Total_Time	88	Content_Regional
19	time spent AL	54	OTT_Platform_YOUTUBE	89	Content_Documentary
20	OTT vs tv	55	OTT_Platform_NETFLIX	90	Content_Education
21	Internet_Source_Wi-Fi	56	OTT_Platform_ZEE5	91	No_of_content
22	Internet_Source_Mobile data	57	OTT_Platform_PRIME	92	Genre_Comedy
23	Internet_Source_Hotspot	58	OTT_Platform_HOTSTAR	93	Genre_Thriller
24	Internet_Source_Dongle	59	OTT_Platform_SONY	94	Genre_Drama
25	No_of_Internet_Source	60	OTT_Platform_MX	95	Genre_Romance
26	Language_English	61	OTT_Platform_VOOT	96	Genre_Action
27	Language_Hindi	62	OTT_Platform_BALAJI	97	Genre_Crime
28	Language_Marathi	63	OTT_Platform_XSTREAM	98	Genre_Family
29	Language_South_Indian	64	OTT_Platform_VODAFONE	99	Genre_Sci-Fi
30	Language_Foreign	65	Total_Platform	100	Genre_Horror
31	Language_Gujurati	66	OTT_Subscription_YOUTUBE	101	Total_Genre
32	Language_Other	67	OTT_Subscription_NETFLIX		
33	No_of_languages	68	OTT_Subscription_ZEE5		
34	Device_TV	69	OTT_Subscription_PRIME		
35	Device_PC	70	OTT_Subscription_HOTSTAR		

The optimum number of clusters using elbow method:

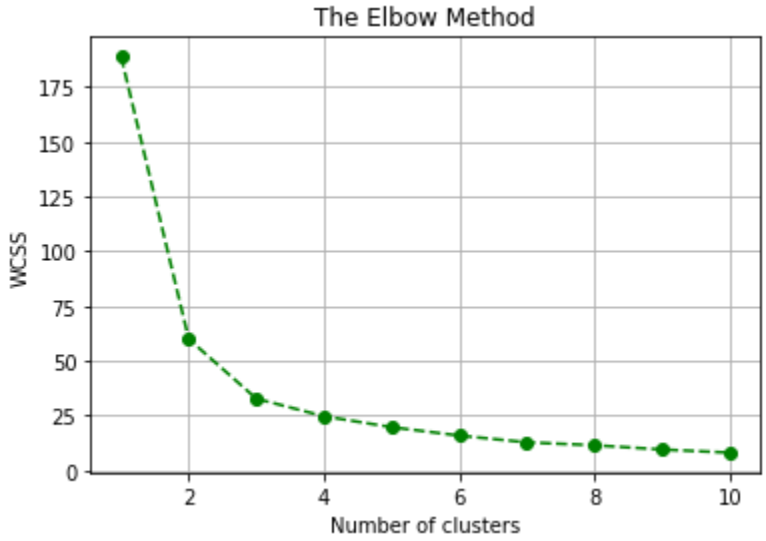


Fig 6.4.1

From this we choose the number of clusters as '3'

Using the optimum no of clusters k-means model of the data:

```
In [15]: km = KMeans(n_clusters=3)
         y_predicted = km.fit_predict(feature_scaled)
         y_predicted

Out[15]: array([2, 1, 1, 0, 2, 2, 1, 0, 0, 1, 1, 1, 1, 1, 0, 0, 2, 2, 0, 2, 1, 0,
                1, 1, 1, 1, 1, 2, 1, 1, 2, 2, 0, 0, 2, 1, 2, 0, 0, 2, 0, 1, 0, 1,
                2, 0, 0, 2, 2, 1, 0, 1, 2, 2, 0, 2, 2, 0, 2, 0, 2, 0, 1, 2, 2, 0,
                0, 2, 0, 0, 1, 1, 0, 2, 1, 1, 1, 0, 1, 2, 1, 0, 1, 1, 0, 0, 1, 2,
                0, 2, 2, 2, 0, 1, 0, 0, 2, 0, 2, 1, 2, 0, 0, 0, 2, 2, 2, 0, 1, 1,
                0, 2, 2, 2, 2, 1, 1, 0, 0, 2, 1, 1, 1, 0, 2, 0, 2, 1, 1, 2, 1, 2,
                2, 2, 1, 0, 0, 0, 0, 1, 2, 1, 1, 1, 2, 0, 1, 2, 1, 2, 2, 2, 0, 1,
                1, 2, 0, 1, 0, 1, 0, 2, 1, 2, 1, 2, 2, 1, 1, 1, 2, 0, 0, 0, 0, 0,
                2, 1, 2, 2, 2, 2, 2, 0, 0, 1, 2, 1, 0, 1, 0, 0, 1, 1, 2, 0, 1, 0,
                1, 2, 1, 2, 2, 0, 1, 0, 0, 2, 2, 1, 2, 0, 0, 1, 2, 0, 2, 1, 2, 1,
                1, 2, 1, 2, 2, 2, 0, 2, 1, 2, 1, 1, 1, 0, 0, 1, 2, 2, 2, 1, 0, 1,
                1, 0, 1, 1, 0, 1, 1, 2, 1, 0, 2, 1, 2, 2, 0, 0, 0, 0, 2, 2, 1, 0,
                0, 2, 2, 2, 2, 2, 1, 1, 0, 2, 0, 0, 0, 2, 2, 2, 1, 0, 0, 0, 2, 2,
                1, 1, 2, 2, 2, 0, 2, 2, 0, 1, 2, 2, 2, 0, 0, 2, 2, 1, 1, 1, 2, 2,
                2, 2, 1, 2, 2, 0, 2, 0, 1, 0, 2, 0, 0, 0, 1, 1, 0, 2, 1, 1, 0, 0,
                1, 1, 1, 1, 1, 0, 2, 0, 2, 1, 1, 1, 0, 2, 0, 1, 0, 2, 2, 1, 2, 1,
                2, 1, 1, 2, 0, 2, 2, 2, 0, 2, 2, 2, 1, 2, 1, 1, 0, 0, 0, 0, 1, 1,
                2, 1, 1, 1, 0, 1, 0, 1, 2, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 2, 2, 2,
                1, 2, 1, 2, 2, 1, 2, 1, 1, 1, 1])
```

Calculation of cluster centroids and addition of cluster columns to original data;

```
In [24]: kmeans.cluster_centers_
```

```
Out[24]: array([[0.79487179, 0.30246914, 0.69230769, ..., 0.48717949, 0.1025641 ,
                 0.25641026],
                [0.375     , 0.20216049, 0.95833333, ..., 0.70833333, 0.75      ,
                 0.84895833],
                [0.43478261, 0.1763285 , 0.86956522, ..., 0.93478261, 0.93478261,
                 0.91576087],
                ...,
                [0.46808511, 0.19858156, 0.87234043, ..., 0.29787234, 0.44680851,
                 0.43085106],
                [0.6097561 , 0.22583559, 0.82926829, ..., 0.12195122, 0.56097561,
                 0.42682927],
                [0.93333333, 0.21440329, 0.84444444, ..., 0.8       , 0.46666667,
                 0.50555556]])
```

```
In [25]: #adding cluster column to data
         feature_scaled['cluster']=y_predicted
```

```
In [17]: #to check no of observations in each cluster.
         feature_scaled['cluster'].value_counts()
```

```
Out[17]: 2    145
         1    141
         0    121
         Name: cluster, dtype: int64
```

The entire data is divided into 3 clusters and the clusters are named as **Minimal viewer, Normal viewer, Extreme viewer**

The distribution of each cluster is as follows:

```
In [5]: pd.pivot_table(data=data,index="cluster",values=["Age",
        "Yearly spend on OTT",
        "No_of_Internet_Source",
        "No_of_languages",
        "No_of_Device",
        "No_of_Purchased",
        "Total_Platform",
        "Total_Subscription",
        "No_of_content",
        "Total_Genre",
        ],aggfunc="mean").T
```

	Minimal viewer	Normal viewer	Extreme viewer
No_of_Device	2	2	3
No_of_Purchased	1	1	2
No_of_content	4	4	7
Total_Genre	4	5	7
Total_Platform	3	4	5
Total_Subscription	1	2	3
Yearly spend on OTT	514	2210	2758

Table 1

The above table shows that extreme viewers spend more money on the subscriptions of different types of OTT platforms, also they prefer to watch a variety of content, genre and they use more devices.

```
In [11]: Income=pd.pivot_table(data=data,index="Income",columns="cluster",values="key_0",aggfunc="count")

         Duration of OTT use=pd.pivot_table(data=data,index="OTT user",columns="cluster",values="key_0",aggfunc="count")

         Regular watching=pd.pivot_table(data=data,index="Regular watching ",columns="cluster",values="key_0",aggfunc="count")
```

Duration of OTT use	Minimal viewer	Normal viewer	Extreme viewer
Less than 3 months	67%	21%	13%
3-5 months	59%	23%	18%
5-12 months	42%	33%	24%
2 years	31%	43%	26%
3 years	28%	35%	37%
more than 3 years	5%	39%	56%
Regular watching			
no	50%	29%	20%
yes	16%	44%	40%
Income			
below 1L	53%	15%	33%
1-4L	43%	34%	24%
4-8L	32%	32%	37%
8-12L	33%	28%	39%
above 12L	10%	39%	51%

Table 6.4.2

Above table shows that minimal viewers have been using OTT platforms for less than three months and Extreme viewers have been using it for the past three years.Number of Minimal viewers are more in the below 1 lakh income category whereas Extreme viewers are more in 8-12 lakhs income category. Also Extreme viewers prefer to watch regularly.

6.5 OBJECTIVE:

To predict which type of subscribers will prefer the leading OTT platform(Amazon Prime).

Technique used: XG-Boost
Tool used: Python software.

Variables used in analysis are:

1	Gender	36	Device_Mob	71	OTT_Subscription_SONY
2	Age	37	Device_stick	72	OTT_Subscription_MX
3	Marital Status	38	Device_Other	73	OTT_Subscription_VOOT
4	Education	39	No_of_Device	74	OTT_Subscription_BALAJI
5	Employment	40	Purchased_TV	75	OTT_Subscription_XSTREAM
6	Area	41	Purchased_Stand	76	OTT_Subscription_VODAFONE
7	family members	42	Purchased_Headphones	77	Total_Subscription
8	Income	43	Purchased_internet	78	Content_Daily_Soaps
9	Awarness about OTT	44	Purchased_data_plans	79	Content_Web_Series
10	OTT user	45	Purchased_Nothing	80	Content_Music
11	Platform for entertainment	46	No_of_Purchased	81	Content_News
12	How did you know about OTT	47	Ideal_time_FreeTime	82	Content_Movies
13	OTT over Traditional	48	Ideal_time_Travelling	83	Content_Sports
14	Subscription BL	49	Ideal_time_Holidays	84	Content_Animated
15	Duration of OTT use	50	Ideal_time_Eating	85	Content_Health
16	Yearly spend on OTT	51	Ideal_time_Night	86	Content_Food
17	Regular watching	52	Ideal_time_Work	87	Content_Kids
18	time spent BL	53	Total_Time	88	Content_Regional
19	time spent AL	54	OTT_Platform_YOUTUBE	89	Content_Documentary
20	OTT vs tv	55	OTT_Platform_NETFLIX	90	Content_Education
21	Internet_Source_Wi-Fi	56	OTT_Platform_ZEE5	91	No_of_content
22	Internet_Source_Mobile data	57	OTT_Platform_PRIME	92	Genre_Comedy
23	Internet_Source_Hotspot	58	OTT_Platform_HOTSTAR	93	Genre_Thriller
24	Internet_Source_Dongle	59	OTT_Platform_SONY	94	Genre_Drama
25	No_of_Internet_Source	60	OTT_Platform_MX	95	Genre_Romance
26	Language_English	61	OTT_Platform_VOOT	96	Genre_Action
27	Language_Hindi	62	OTT_Platform_BALAJI	97	Genre_Crime
28	Language_Marathi	63	OTT_Platform_XSTREAM	98	Genre_Family
29	Language_South_Indian	64	OTT_Platform_VODAFONE	99	Genre_Sci-Fi
30	Language_Foreign	65	Total_Platform	100	Genre_Horror
31	Language_Gujurati	66	OTT_Subscription_YOUTUBE	101	Total_Genre
32	Language_Other	67	OTT_Subscription_NETFLIX		
33	No_of_languages	68	OTT_Subscription_ZEE5		
34	Device_TV	69	OTT_Subscription_PRIME		
35	Device_PC	70	OTT_Subscription_HOTSTAR		

The target variable is stored in ‘Y’ and features in ‘X’.
Y : User having subscription of Amazon Prime
Y = 1 ; if yes
= 0 ; otherwise

Splitting data into training and testing data sets.

In [5]:

```
if __name__ == "__main__":

    target_col = "OTT_Subscription_PRIME"

    print(model_base[target_col].value_counts())


    X_train, X_test, y_train, y_test = train_test_split(model_base, model_base[target_col],
                                                        test_size=0.3,
                                                        random_state=1, stratify=model_base[target_col])

    X_train.to_csv(r"C:\Users\HP\Desktop\KOMAL\OUTPUT_n\train_df.csv", index=False)
    print(X_train[target_col].value_counts())
    print(X_train.shape)
    X_test.to_csv(r"C:\Users\HP\Desktop\KOMAL\OUTPUT_n\test_df.csv", index=False)
    print(X_test[target_col].value_counts())

    print(X_test.shape)
```

Training the XG-boost model on train dataset,

```
#training xg-boost model
xg_clf = xgb.XGBClassifier(colsample_bytree=0.6, gamma=1, learning_rate=0.01, max_depth=8,
                           n_estimators=250, objective='binary:logistic',
                           subsample=0.7, reg_alpha=0.1, reg_lambda=2,
                           seed=123, max_delta_step=0, verbosity=3, n_jobs=3,
                           scale_pos_weight=0.5*np.sum(training_df[target_col]==0)/np.sum(training_df[target_col]==1))

    xg_clf.fit(training_df[feature_list], training_df[target_col])

    preds = xg_clf.predict(training_df[feature_list])


# training_df["pred_prob"] = xg_clf.predict_proba(training_df[feature_list])[:,1]
    pred_prob = xg_clf.predict_proba(training_df[feature_list])

    prob_name_list = ["pred_prob_"+str(xg_clf.classes_[0]), "pred_prob_" + str(xg_clf.classes_[1])]

    training_df["pred_prob_"+str(xg_clf.classes_[0])] = pred_prob[:,0]
    training_df["pred_prob_" + str(xg_clf.classes_[1])] = pred_prob[:, 1]

    training_df["y_pred"] = preds

    # training_df.ix[0, "max_prob"] =  training_df["pred_prob"].max()
    # training_df.ix[0, "min_prob"] =  training_df["pred_prob"].min()
    # training_df.ix[0, "avg_prob"] =  training_df["pred_prob"].mean()
    # , "max_prob", "min_prob", "avg_prob"
    training_df[["LeadID", target_col,"y_pred"]+prob_name_list]\
        .round(2).to_csv(r"training_data_pred.csv", index=False)
```

Top 15 features predicted using feature importance function.

```
# feature importance df
    feat_imp_df = pd.DataFrame()
    feat_imp_df["Feature_Name"] = feature_list
    feat_imp_df["Imp"] = xg_clf.feature_importances_

    feat_imp_df.sort_values(["Imp"], ascending=False, inplace=True)
    feat_imp_df.reset_index(drop=True, inplace=True)
    # feat_imp_df.round(2).to_csv("D:/Analytics/Wayne/PMS_Redemption/xgboost_model_building/new_y
    #         index=False)

    feat_imp_df.to_csv(r"training_data_feat_imp.csv",
            index=False)
```

Calculation of Correlation matrix,VIF and Confusion matrix:
Removal of features which were highly correlated and/or which are having high VIF then again running the model using remaining features.

```
#calculating confusion matrix
def confusion_matrix(df, target_col, top_per=0.0):
    op_df = deepcopy(df)

    if top_per !=0:
        op_df.sort_values(["pred_prob_1"], ascending=False, inplace=True)
        op_df.reset_index(drop=True, inplace=True)
        top_len = int(top_per * op_df.shape[0])
        op_df["y_pred"] = 0
        op_df.loc[0:top_len, "y_pred"]=1

    pivot_df = pd.pivot_table(op_df, index=target_col, columns="y_pred", values="LeadID", aggfunc="count")
    pivot_df["Total"] = pivot_df.sum(axis=1)
    pivot_df.loc["Total", :] = pivot_df.sum(axis=0)
    pivot_df.loc["Precision", :]=[np.nan, np.nan, pivot_df.loc[1, 1]*100/pivot_df.loc["Total", 1]]
    pivot_df.loc["Recall", :]=[np.nan, np.nan, pivot_df.loc[1, 1]*100/pivot_df.loc[1, "Total"]]
    pivot_df.loc["Per_Base_Pred", :]=[np.nan, np.nan, pivot_df.loc["Total", 1]*100/pivot_df.loc["Total", "Total"]]

    pivot_df.reset_index(inplace=True)
    pivot_df["OTT_Subscription_PRIME"] = ["Actual_0", "Actual_1" ,"Total", "Precision", "Recall", "Per_Base_Pred"]
    pivot_df.rename(columns={0:"Pred_0", 1:"Pred_1"}, inplace=True)
    return pivot_df
print("number of features", len(feature_list))
```

```
#calculating correlation matrix

    t_corr_df = training_df[feature_list].corr()
    t_corr_df.round(2).to_csv(r"t_corr.csv")
#calculating VIF
    t_vif = pd.DataFrame()
    t_vif["VIF Factor"] = [variance_inflation_factor(training_df[feature_list].values, i) for i in range(training_df[feature_lis
    t_vif["features"] = feature_list
    t_vif.to_csv(r"training_data_vif.csv", index=False)
```

	Yearly spend on	OTT_Subscription	Subscription BL	OTT_Subscription_	No_of_Device	Internet_Source	Genre_Action	Genre_Sci-Fi	time spent AL	Content_Movies	OTT_Subscription_	OTT_Subscription_XSTREAM
Yearly spend on OTT	1	0.38	0.28	0.25	0.13	0.17	0	0.06	0.18	0.13	0.32	0.09
OTT_Subscription_NETFLIX	0.38	1	0.38	0.26	0.26	0.25	0.04	0.2	0.09	0.11	0.33	0.07
Subscription BL	0.28	0.38	1	0.26	0.24	0.23	0.15	0.19	0.14	0.21	0.23	0.11
OTT_Subscription_HOTSTAR	0.25	0.26	0.26	1	0.23	0.13	0.02	0.02	0.13	0.06	0.39	0.23
No_of_Device	0.13	0.26	0.24	0.23	1	0.28	0.06	0.19	0.11	0.09	0.23	0.09
Internet_Source_Wi-Fi	0.17	0.25	0.23	0.13	0.28	1	-0.04	0.11	0.12	0.05	0.06	-0.03
Genre_Action	0	0.04	0.15	0.02	0.06	-0.04	1	0.16	0.05	0.13	0.07	0.09
Genre_Sci-Fi	0.06	0.2	0.19	0.02	0.19	0.11	0.16	1	0.1	0.16	0.02	0.16
time spent AL	0.18	0.09	0.14	0.13	0.11	0.12	0.05	0.1	1	0.14	0.18	0.12
Content_Movies	0.13	0.11	0.21	0.06	0.09	0.05	0.13	0.16	0.14	1	0.11	0.06
OTT_Subscription_ZEE5	0.32	0.33	0.23	0.39	0.23	0.06	0.07	0.02	0.18	0.11	1	0.24
OTT_Subscription_XSTREAM	0.09	0.07	0.11	0.23	0.09	-0.03	0.09	0.16	0.12	0.06	0.24	1

Table 6.5.1

Above table is the final correlation matrix where all variables are perfectly correlated. Below table is indicating final VIF matrix,VIF for all the variables is under 8 hence there is no multicollinearity,

VIF Factor	features
1.631915698	Yearly spend on OTT
2.801459977	OTT_Subscription_NETFLIX
4.104388923	Subscription BL
2.208447032	OTT_Subscription_HOTSTAR
6.100228036	No_of_Device
4.405732579	Internet_Source_Wi-Fi
2.538700975	Genre_Action
2.250108685	Genre_Sci-Fi
2.76209053	time spent AL
4.778500154	Content_Movies
1.58695512	OTT_Subscription_ZEE5
1.217599688	OTT_Subscription_XSTREAM

Table 6.5.2

Below table is the confusion matrix for training data set:

OTT_Subscription_PRIME	Pred_0	Pred_1	Total
Actual_0	116	15	131
Actual_1	45	108	153
Total	161	123	284
Precision			87.8
Recall			70.59
Accuracy of model			0.788732

Table 1

The Accuracy of the training set is 78.87%

Below table is the confusion matrix for validation data set:

OTT_Subscription_PRIME	Pred_0	Pred_1	Total
Actual_0	51	6	57
Actual_1	24	42	66
Total	75	48	123
Precision			87.5
Recall			63.64
Accuracy of model			0.756098

Table 2

The Accuracy of the testing set is 75.61%
Hence the results of the validation set are very close to the results of the training set hence we can say that our model is a good fit and ready for future prediction.

Feature_Name	Imp
Yearly spend on OTT	0.229117
OTT_Subscription_NETFLIX	0.19697
Subscription BL	0.120353
time spent AL	0.072684
OTT_Subscription_HOTSTAR	0.07069
No_of_Device	0.067124
OTT_Subscription_ZEE5	0.059562
Internet_Source_Wi-Fi	0.055217
Genre_Action	0.045168
Content_Movies	0.042665
Genre_Sci-Fi	0.040451

Table 3

Significant features for this model are given in the above table along with their significance.

6.6 OBJECTIVE:

To study the socio- demographic factors that affect the usage of OTT platforms.

Technique used: Binary Logistic Regression
Software: SPSS

➔ Design matrix:

Categorical Variables Codings

		Frequency	Parameter coding				
			(1)	(2)	(3)	(4)	(5)
Employment	Employed	187	.000	.000	.000	.000	.000
	House wi	35	1.000	.000	.000	.000	.000
	Retired	8	.000	1.000	.000	.000	.000
	Self-Emp	63	.000	.000	1.000	.000	.000
	Student	312	.000	.000	.000	1.000	.000
	Unemploy	20	.000	.000	.000	.000	1.000
Members_family	1	2	1.000	.000	.000	.000	.000
	2	26	.000	1.000	.000	.000	.000
	3	111	.000	.000	1.000	.000	.000
	4	262	.000	.000	.000	1.000	.000
	5	140	.000	.000	.000	.000	1.000
	More tha	84	.000	.000	.000	.000	.000
Family_Income	1-4 Lakh	220	.000	.000	.000	.000	
	4-8 Lakh	153	1.000	.000	.000	.000	
	8-12 Lak	72	.000	1.000	.000	.000	
	Above 12	87	.000	.000	1.000	.000	
	Below 1	93	.000	.000	.000	1.000	
Living_area	Out Of I	16	1.000	.000	.000		
	Rural	52	.000	1.000	.000		
	Suburban	146	.000	.000	1.000		
	Urban	411	.000	.000	.000		
Education	Graduati	304	.000	.000			
	Post Gra	221	1.000	.000			
	Undergra	100	.000	1.000			
Marital_status	Married	147	.000				
	Unmarrie	478	1.000				
Aware_of_OTT	No	75	.000				
	Yes	550	1.000				
Gender	Female	306	.000				
	Male	319	1.000				

➔ Assumptions:

- Observations should come from Independent samples
- The dependent variable should be binary
- Logistic Regression requires little or no multicollinearity
- Large sample size
- Logistic regression assumes linearity of independent variables and log odds
- No outliers

➔ **Checking for multicollinearity:**

VARIABLE	VIF
Gender	1.08
Age	2.2
Marital_status	1.94
Education	1.19
Employment	1.42
Members_family	1.03
Living_area	1.03
Family_Income	1.13
Aware_of_OTT	1.25

Table 6.6.1

Interpretation:

Since all the VIF values for the variables are > 1 and < 5 which suggests the absence of multicollinearity.

➔ Dependent Variable Encoding

Do you use OTT platforms?	Internal Value	Frequency
No	0	220
Yes	1	405

Table 6.6.2

➔ USED FORWARD LIKELIHOOD RATIO SELECTION: This selection method with entry testing based on the significance of the score statistic, and removal testing is based on the probability of a likelihood-ratio statistic given by the maximum partial likelihood estimates.

Step Summary[a,b]

Step	Improvement Chi-square	Improvement df	Improvement Sig.	Model Chi-square	Model df	Model Sig.	Correct Class %	Variable
1	140.857	1	<.001	140.857	1	<.001	80.6%	IN: Aware_of_OTT
2	11.011	1	<.001	151.868	2	<.001	80.8%	IN: Age
3	20.792	4	<.001	172.660	6	<.001	81.0%	IN: Family_Income
4	11.445	5	.043	184.104	11	<.001	81.3%	IN: Members_family

a. No more variables can be deleted from or added to the current model.

b. End block: 1

There happens to be 4 steps as described in the above table. Awareness about OTT, Age and Family Income are highly significant whereas Members in a family are significant but not as high as the above variables.

➔ Checking for significance of the new model

H_0: All the variables in the model are statistically insignificant i.e

H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$

H_1: At least one of the independent variable is statistically significant i.e

H_1: at least one $\beta_i \neq 0$ (i=1,2,3,4)

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 4	Step	11.445	5	.043
	Block	184.104	11	<.001
	Model	184.104	11	<.001

Since P-value< 0.05, we reject H_0 and conclude that our model is fitting the data significantly better than a null model with no predictors.

Model row- The final reduced model; Step- Model previous models

➔ VARIATION EXPLAINED

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
4	566.335[a]	.255	.365

a. Estimation terminated at iteration number 20 because maximum iterations has been reached. Final solution cannot be found.

The -2 Log likelihood The **Cox & Snell R Square** and **Nagelkerke R Square** values indicate the variation explained. These values are also referred to as pseudo R2. The total variation in the model explained ranges from 25.5% to 36.5%. The independent variables explain roughly 36.5% of the variation in the dependent variables.

The **Cox & Snell R Square value** cannot reach one, hence **Nagelkerke R Square** is preferred.

For regression models with a categorical dependent variable, it is not possible to compute a single R^2 statistic that has all of the characteristics of R^2 in the linear regression model, so these approximations are computed instead.

➔ Hosmer and Lemeshow Test (Goodness of fit)

H_0: The REDUCED model is a good fit

H1: The REDUCED model is not a good fit

Hosmer and Lemeshow Test			
Step	Chi-square	df	Sig.
4	4.073	8	.850

We have a p-value: 0.850 > 0.05. Hence not rejecting H_0 and concluding that the reduced model is a good fit.

➔ Confusion Matrix

Classification Table[a]

Observed			Predicted		
			Use_OTT		Percentage Correct
			No	Yes	
Step 1	Use_OTT	No	67	113	37.2
		Yes	8	437	98.2
	Overall Percentage				80.6
Step 2	Use_OTT	No	68	112	37.8
		Yes	8	437	98.2
	Overall Percentage				80.8
Step 3	Use_OTT	No	73	107	40.6
		Yes	12	433	97.3
	Overall Percentage				81.0
Step 4	Use_OTT	No	73	107	40.6
		Yes	10	435	97.8
	Overall Percentage				81.3

a. The cut value is .500

Table 6.6.3

Sensitivity- Also called True Positives, 435 in our data. The Sensitivity rate is 98% .

Specificity- Also called True Negatives, 73 in our data.

As the model adds significant variables in the model, the prediction percentage increased from 80.6% to 81.3% in 4 steps. The model fitted is able to correctly classify 81.3% of the cases. Hence any value above 80% is called a good model.

➔ The fitted model

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 4[a]	Age	-.040	.010	14.431	1	.000	.961	.941	.981
	Members_family			6.621	5	.250			
	Members_family(1)	-21.575	28280.573	.000	1	.999	.000	.000	.
	Members_family(2)	.895	.620	2.082	1	.149	2.448	.726	8.256
	Members_family(3)	.561	.360	2.428	1	.119	1.753	.865	3.550
	Members_family(4)	.704	.305	5.315	1	.021	2.022	1.111	3.680
	Members_family(5)	.771	.352	4.808	1	.028	2.162	1.085	4.308
	Family_Income			17.551	4	.002			
	Family_Income(1)	.535	.279	3.677	1	.055	1.708	.988	2.953
	Family_Income(2)	.737	.382	3.716	1	.054	2.090	.988	4.422
	Family_Income(3)	1.025	.386	7.065	1	.008	2.787	1.309	5.935
	Family_Income(4)	-.428	.306	1.957	1	.162	.652	.358	1.187
	Aware_of_OTT(1)	3.086	.405	58.008	1	.000	21.897	9.896	48.452
	Constant	-1.527	.564	7.343	1	.007	.217		

a. Variable(s) entered on step 4: Members_family.

Table 6.6.4

f(x) = -1.527 (constant) - 0.04 (x1) -21.575(x21) + 0.895(x22) + 0.561(x23) + 0.704(x24) + 0.71(x25) + 0.535(x31) + 0.737(x32) + 1.025(x33) -0.428(x34) + 3.086(x4)

Where,

x1 = Gender,
x21= 1 Member Family
x22 = 2 Members Family
x23 = 3 Members Family
x24= 4 Members Family
x25= 5 Members Family

x31= Annual Family Income (4-8 Lakhs)
x32= Annual Family Income (8-12 Lakhs)
x33= Annual Family Income (More than 12 Lakhs)
x34= Annual Family Income (Below 1 Lakh)

x4= Aware about OTT (Yes)

Note:

1. The Members in the family variable are compared to ‘More than 5’ Family Members.
2. The Annual Family Income is compared to Annual income of Rs. 1-4 Lakhs.
3. Awareness about OTT is compared to people who are not aware of OTT platforms.

- Estimates:

  The first column of B values indicates proportionality. Positive value indicates that the variable is having a positive likelihood of people using OTT platforms. SE is the standard error of the variables.

<u>Interpretation</u>:

1. Age has a negative sign indicating as age increases, the likelihood of a person using OTT platform decreases. The odds of an older person using OTT platforms is low.
2. A person with 4 and 5 family members are more likely to use OTT.
3. Individuals with annual family income between 'Below 1 Lakh' are least likely to use OTT platforms.
4. Respondents who were aware of OTT platforms were highly likely to use OTT platforms.

- Wald Statistic-

  H_0: Individual independent variables are significant

  H_0: Not H_0

  The statistical significance of each B is tested by the Wald Chi-Square-testing the null hypothesis that the B coefficient = 0 (the alternate hypothesis is that it does not = 0). The Wald test ("**Wald**" column) is used to determine statistical significance for each of the independent variables.

  For variables whose p value<0.05 indicate they did not add to the model significantly.

<u>Interpretation</u>:

Age, Members in family(4), Members in family(5), Family Income above 1 Lakhs, Awareness about OTT platforms are all significant as their p value is <0.05. This indicates that these variables have a significant impact on predicting if a person uses OTT or not.

- Odds ratio:

  The Exp(B) column is the Odds Ratio. Exp(B) (taking the B value by calculating the inverse natural log of B) indicates odds ratio: the probability of an event occurring, divided by the probability of the event not occurring. An Exp(B) value

over 1.0 signifies that the independent variable increases the odds of the dependent variable occurring. An Exp(B) under 1.0 signifies that the independent variable decreases the odds of the dependent variable occurring, depending on the decoding that is mentioned on the variables details before.

Interpretation:

1. For an additional year in age, the odds of a person using OTT platform is lowered by a factor of 0.961.
2. Odds Ratio is a measure of association representing the odds that a person with annual family income '4-8 Lakhs' is 2 times more likely to use OTT platforms than people with '1-4 Lakhs' annual family income.
3. A person with 5 family members is more likely to use an OTT platform than a person with a More than 5 members.
4. People who are aware of OTT platforms are 22 times more likely to use OTT platforms than people who are not aware of OTT.

- Confidence Interval :

  The last column of the Variation in the equations gives the 95% confidence Interval for the ODDS ratio. There is a probability of 0.05 that the ODDS ratio will lie outside this Confidence Interval.

6.7 OBJECTIVE :

To understand the association and frequency between different genres.

Technique : Apriori Algorithm
Software : Python

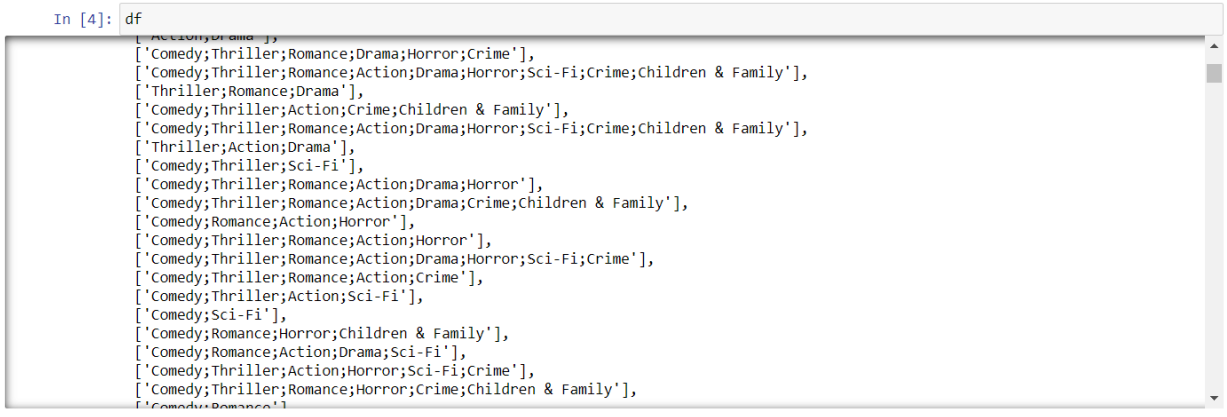
Association rules analysis is a technique to discover how items are associated with each other. There are few ways to measure association.

1. Support (freq(A, B)/N) : It says frequency and combination of frequency. This says how frequent any genre or item is. Filters out items are used less frequently. The table contains 3 different support metrics: The 'antecedent support' computes

the proportion of transactions that contain the antecedent A, and 'consequent support' calculates the support for the itemset of the consequent B. The 'support' metric then calculates the support of the combined itemset A U B. The 'support' depends on 'antecedent support' and 'consequent support' via min ('antecedent support', 'consequent support'). The range of Support is from [0,1].

2. Confidence (freq(A, B)/ freq(A)) : It says how often A and B occur together given the number of times A occurs. That is how likely is an item B purchased when A is purchased. The range of Confidence is from [0,1].
3. Lift: It is the strength of any rule. This tells how probable item B's purchase is when item A is purchased, however monitoring the popularity of item B. If A and B are independent, the Lift score will be exactly 1. Lift score ranges from 0 to infinity.
4. Leverage: Leverage computes the difference between the observed frequency of A and B appearing together and the frequency that would be expected if A and B were independent. A leverage value of 0 indicates independence. The range for Leverage is from -1 to +1.
5. Conviction: A high conviction value means that the consequents is highly dependent on the antecedent. Similar to lift, if items are independent, the conviction is 1. The range for Conviction is from 0 to infinity.

The dataset looks like



Shows the first 15 items where support of the respective items/ pairs.

```
In [100]: frequent_itemsets = fpgrowth(df, min_support=0.2, use_colnames=True)
          items= frequent_itemsets.sort_values('support',ascending= False)
          items.head(15)
```

Out[100]:

	support	itemsets
2	0.827160	(Comedy)
0	0.720988	(Thriller)
3	0.661728	(Romance)
4	0.607407	(Action)
9	0.602469	(Thriller, Comedy)
25	0.592593	(Romance, Comedy)
1	0.575309	(Drama)
28	0.525926	(Comedy, Action)
5	0.498765	(Crime)
26	0.498765	(Thriller, Romance)
6	0.496296	(Sci-Fi)
11	0.496296	(Drama, Comedy)
29	0.493827	(Thriller, Action)
7	0.481481	(Horror)
12	0.464198	(Drama, Romance)

The support for comedy is the highest indicating most viewers want to watch Comedy genre followed by Thriller and Romance. Viewers also like to watch Thriller and Comedy together and Romance and Comedy together.

Using the association rule to find the pairs with minimum confidence 0.1.

```
In [9]: from mlxtend.frequent_patterns import association_rules

        ans= association_rules(frequent_itemsets, metric="confidence", min_threshold=0.1)
        ans
```

Out[9]:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(Thriller)	(Comedy)	0.720988	0.827160	0.602469	0.835616	1.010223	0.006097	1.051440
1	(Comedy)	(Thriller)	0.827160	0.720988	0.602469	0.728358	1.010223	0.006097	1.027133
2	(Drama)	(Thriller)	0.575309	0.720988	0.422222	0.733906	1.017917	0.007432	1.048546
3	(Thriller)	(Drama)	0.720988	0.575309	0.422222	0.585616	1.017917	0.007432	1.024875
4	(Drama)	(Comedy)	0.575309	0.827160	0.496296	0.862661	1.042918	0.020424	1.258488
...	...	...	...	...	...	...	...	...	...
279	(Thriller, Action)	(Horror)	0.493827	0.481481	0.303704	0.615000	1.277308	0.065935	1.346801
280	(Horror, Action)	(Thriller)	0.323457	0.720988	0.303704	0.938931	1.302285	0.070495	4.568827
281	(Thriller)	(Horror, Action)	0.720988	0.323457	0.303704	0.421233	1.302285	0.070495	1.168939
282	(Horror)	(Thriller, Action)	0.481481	0.493827	0.303704	0.630769	1.277308	0.065935	1.370885
283	(Action)	(Thriller, Horror)	0.607407	0.412346	0.303704	0.500000	1.212575	0.053242	1.175309

284 rows × 9 columns

This table also gives the Genre - Thriller and Comedy are preferred more. Drama is the next Genre highly preferred.

Viewers who like to watch Comedy, Thriller, Crime also watch Action and Horror. Viewers who like to watch Drama and Horror will preferably watch Romance and Crime.

6.8 OBJECTIVE

To identify whether people are aware of streaming services or not based on different factors:

Technique used: Chi-square test
Tool used: Microsoft Excel

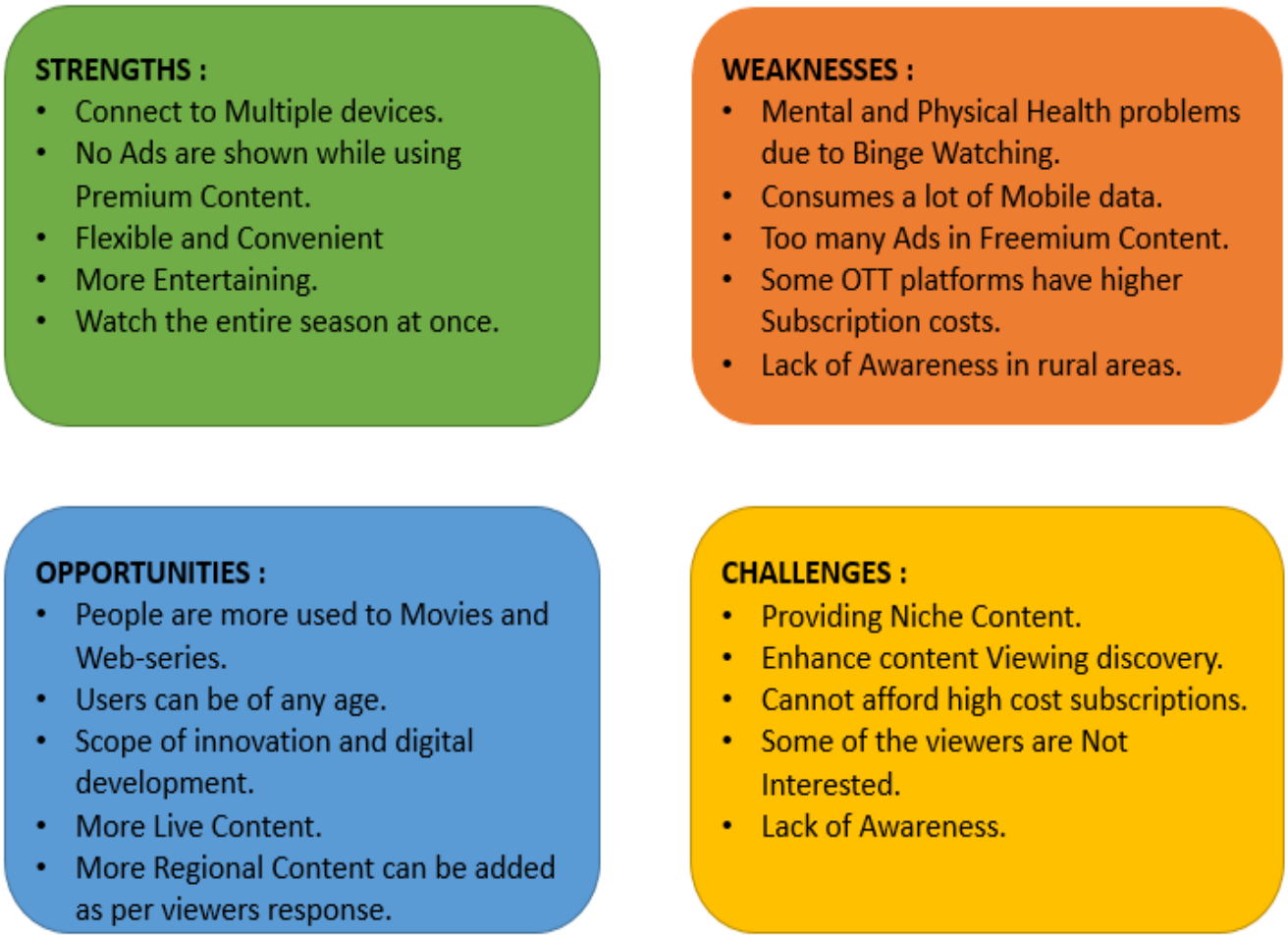
If P-value <0.05 then we Reject Ho

	Null Hypothesis	Alternative Hypothesis			
Sr.No	**Ho**	**Ha**	**Chi-square value**	**P-value**	**Interpretation**
Case 1	There is no association between Gender and Awareness of online streaming (OTT)	There is an association between Gender and Awareness of online streaming (OTT)	2.48	0.12	Therefore we do not reject Ho and conclude that there is no association between Gender and Awareness of online streaming.
Case 2	There is no association between Marital Status and Awareness of online streaming (OTT)	There is an association between Marital Status and Awareness of online streaming (OTT)	49.84	0.00000 0000002	Therefore we reject Ho and conclude that Unmarried people are more Aware of OTT than Married.
Case 3	There is no association between Education Status and Awareness of OTT.	There is an association between Education Status and Awareness of OTT.	71.90	0.00000 0000002	Therefore we reject Ho and conclude that in our study out of 305 undergraduates 280

					people are aware of OTT.
Case 4	There is no significant difference between Employment and Awareness of OTT.	There is a significant difference between Employment and Awareness of OTT.	0.52	0.47	Therefore we do not reject Ho and conclude that there is no association between Employment and Awareness of online streaming.
Case 5	There is no significant difference between Income and Awareness of OTT.	There is a significant difference between Income and Awareness of OTT.	28.34	0.00001	Therefore we reject Ho and conclude that people who have income between 1-4 lakhs and 4-8 lakhs are more aware of OTT.

OBJECTIVE:

To understand Strengths, Weaknesses, Opportunities & Challenges for OTT platforms.
Technique used: SWOC Analysis



Interpretation:

From the analysis it is observed that OTT has allowed people to watch their favorite shows on a wide range of multiple devices. It is also very flexible and convenient to use. These are the platforms that can work on Smart TVs, Mobile and Laptops. Despite its flexibility, entertainment and other advantages, still there are few weaknesses and challenges faced by these platforms.

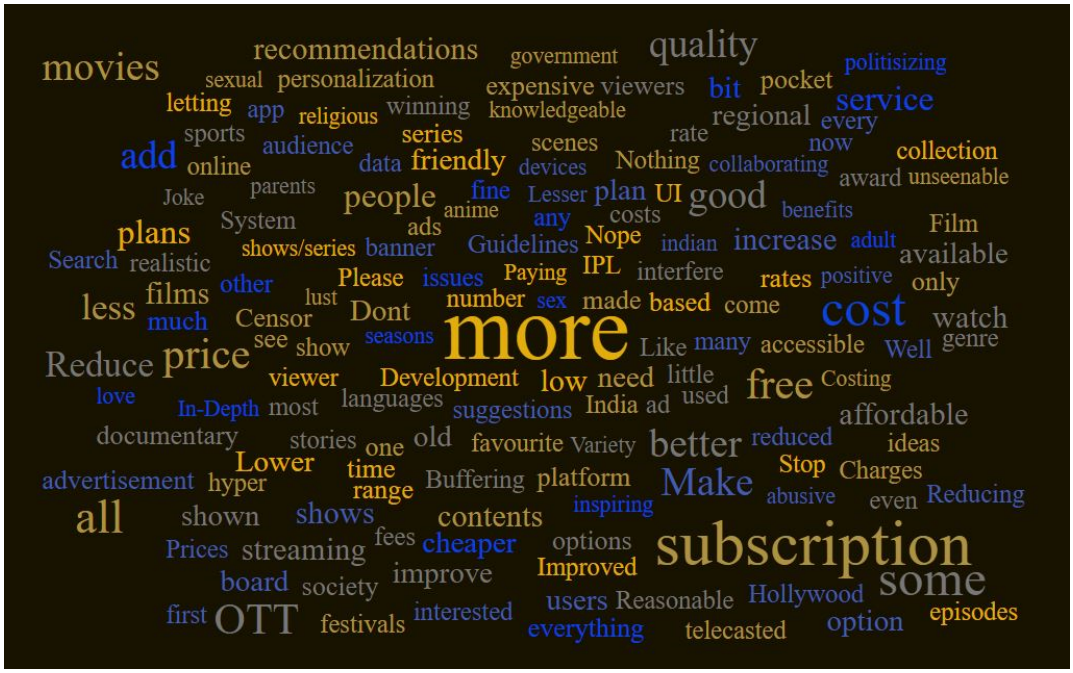
Higher subscription cost is the major reason for the users for not using OTT and due to binge watching by people, there is a high risk of mental and physical health issues. Apart from this, content viewing is the most important part of OTT, if these services are unable to create and provide the content desired by viewers, they cannot be successful in the long term. Viewers may no longer have the patience to browse through hundreds of channels to find something they might like watching, therefore, these platforms should leverage applications that can identify the viewer's preferences and habits. Creating more interesting content along with keeping it original is very much essential for staying in the

race for each platform. By taking the challenges as opportunities there should be more movies and webseries as viewers want a variety of entertaining content to suit every mood. More regional content can be added as per viewer's response based on analysis.

6.9 OBJECTIVE:

Diagrammatic representation for easy identification of choice of OTT platforms and suggestions.

Technique used: Word Cloud



Interpretation:

The above word cloud highlights:

1. More quality content
2. Less subscription cost
3. Lesser advertisement
4. More religious and family content

This shows that these larger-sized words have been expressed by most people in their responses. Some sort of improvement is needed by them in this area of Online Streaming Services.

7. Overall Conclusion

1. From the study, it is seen that Amazon Prime, Netflix & Disney+ Hotstar are leading OTT platforms ranking 1,2,3 respectively.
2. Awareness of OTT is not associated with Gender and the Employment status of an individual. Whereas Awareness of OTT is associated with Marital status, Education and Income.
3. From Pareto analysis, it is concluded that OTT providers should focus on awareness of OTT and also different subscription plans.
   Also the main reasons behind the growth of OTT are its flexibility, convenience, more entertaining, the entire season can be watched at once, more interest in OTT content and availability of the latest movies.
4. From Cluster Analysis, viewers from data are classified into three different categories which are Minimal viewers,Normal viewers, Extreme viewers & their distribution is as follows:

Minimal viewers	Normal viewers	Extreme Viewers
OTT users from less than 3 months	OTT users from 5-12 months	OTT users from last 3 years
spend less money on subscriptions	spend average amount on subscriptions	spend more money on subscriptions
Total platforms in use: 2	Total platforms in use: 4	Total platforms in use: 5
Total subscriptions: 1	Total subscriptions: 2	Total subscriptions: 3
No of genere used: 4	No of genere used: 5	No of genere used: 7
Income Category : below 1 lakh	Income Category : 1-4lakh and above	Income Category : 8 to12 lakh and above

5. From the Apriori algorithm, viewers who altogether like to watch Comedy, Thriller, Crime with Action and Horror. Viewers who like to watch Drama and Horror will preferably watch Romance and Crime.
6. From Binary Logistic Regression, factors like Awareness about OTT, Age, Family Income, Members in the family play a significant role in predicting if a person is an OTT user or not.

8. Scope

- ➢ There are so many streaming services like music streaming, video streaming, live video streaming, etc.This research study is limited to Video Streaming platforms only, so this study can be explored further to other streaming services. Other topics including freemium can also be included. App distribution by freemium and subscription based can be done.
- ➢ Large mobile penetration in developing countries and cheap data prices is one of the reasons for streaming being more popular these days.The relationship between data consumption and the use of streaming services can be studied.
- ➢ Future research should examine the impact of how consumers watch entertainment and explore different ways that Cable/DTH or streaming services are used. One should do the survey on why people are preferring a particular subscription.
- ➢ Attributes like UI design, Application Designing (graphics, animations, etc) can be taken into consideration for further study.
- ➢ This study can be done using samples of equal proportions of all age groups. Because this project was done during Covid, the responses collected were not in equal proportions of all age groups due to which the data was skewed.

9. SUGGESTION

- Provide affordable subscription plans.
- Focus more on the originality of the shows.
- Speed up content discovery by creating stories.
- Collect user's feedback when they delete accounts to deliver a better streaming experience to other customers.
- Produce more Advertisements for better reach.
- The number of viewerships can be increased by uploading content according to preference on Holidays.
- More awareness in rural areas.

- Prices for mobile data can be lowered so that more people can take subscriptions and can watch videos.
- Add more incentives for OTT subscribers in the Rural area.
- During the Covid period, everything is being escalated to online. OTT platforms (like Youtube) owners can provide subscriptions for a lower rate in the Rural area for educational purposes. This will create awareness among the areas in the Rural areas. Eventually more people will opt for OTT.
- OTT platform owners can give additional perks to such customers who are not sure about continuing using OTT by giving them an extra month of usage or recommending better videos as per the customers interest.

10. Bibliography

10.1 Reference :

[1] (Reshma , Chaithra), Proliferation of OTT apps in India: an empirical study of OTT apps and its impact on college students, February 2020
https://www.ijrar.org/papers/IJRAR2001475.pdf
[2] (Manoj Kumar Patel[1], et al.), A Study: OTT Viewership in "Lockdown" and Viewer's Dynamic Watching Experience, 2020
https://www.researchgate.net/profile/Manoj-Patel-9/publication/343444529_A_Study_OTT_Viewership_in_Lockdown_and_Viewer%27s_Dynamic_Watching_A-Study-OTT-Viewership-in-Lockdown-and-Viewers-Dynamic-Watching-Experience.pdf
[3] (Tripti Kumari), A Study on Growth of Over the Top (OTT) Video Services in India, 2020 http://www.ijlrhss.com/paper/volume-3-issue-9/11-HSS-747.pdf
[4] (Rachita Ota[1], et al.), An analysis of customer preference towards OTT Platform during a pandemic: A special reference to Jamshedpur Market,Aug 2020
http://www.journalstd.com/gallery/24-aug2020.pdf
[5] ((C. Christopher Lee, et al.)), Factors Affecting Online Streaming Subscriptions ,2018
https://scholarworks.lib.csusb.edu/cgi/viewcontent.cgi?article=1394&context=ciima
[6]
https://towardsdatascience.com/feature-engineering-for-machine-learning-3a5e293a5114
[7] https://www.econstor.eu/bitstream/10419/205203/1/Park-Kwon.pdf
[8] https://www.lexjansen.com/wuss/2018/130_Final_Paper_PDF.pdf

[9] https://stats.idre.ucla.edu/spss/dae/logit-regression/
[10]http://core.ecu.edu/psyc/wuenschk/MV/Multreg/Logistic-SPSS.PDF
[11]https://www.techsciresearch.com/report/india-ott-video-services-market/3164.html
[12]http://core.ecu.edu/psyc/wuenschk/MV/Multreg/Logistic-SPSS.PDF
[13]https://corp.phando.com/what-are-the-challenges-faced-by-an-ott-platform/
[14] https://blog.shortfundly.com/platform/ott-advantages-and-disadvantages-in-india/

10.2 Sites Used:

- YouTube
- Analytics Vidhya
- Towards Data Science
- Statisticshowto
- Wikipedia

10.3 Statistical Software Used:

- Python (jupyter notebook)
- R
- SPSS
- MS-Excel

11. Questionnaire
