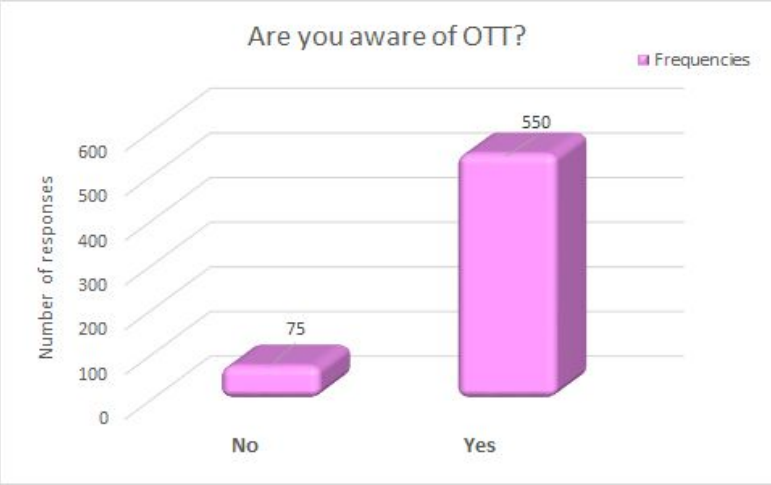
Growth of Over the top(OTT) and Factors Affecting It

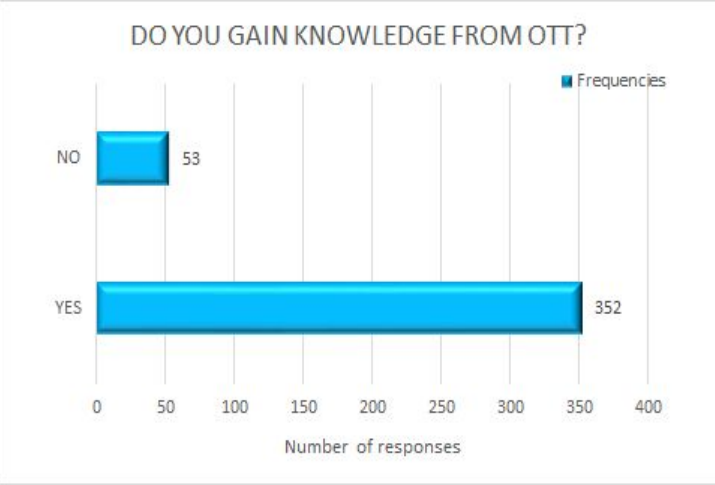
Project by-

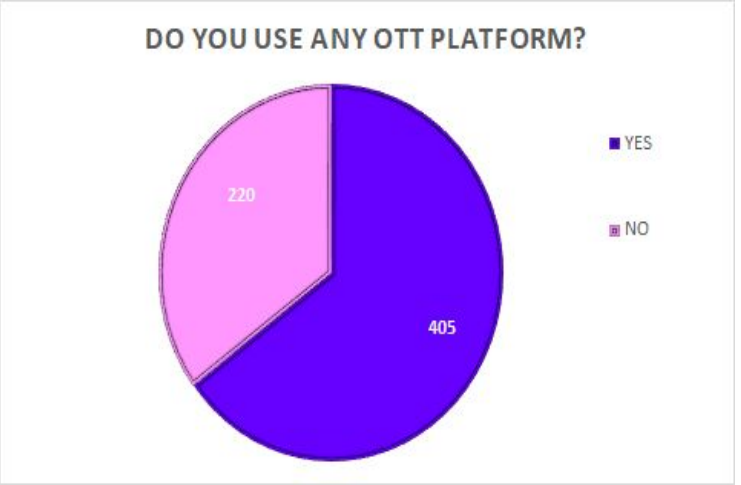
Komal Parab (12)
Krishna Barfiwala (10)
Maunika Shripati (19)
Shweta Singh (21)

EXPLORATROY DATA ANALYSIS

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



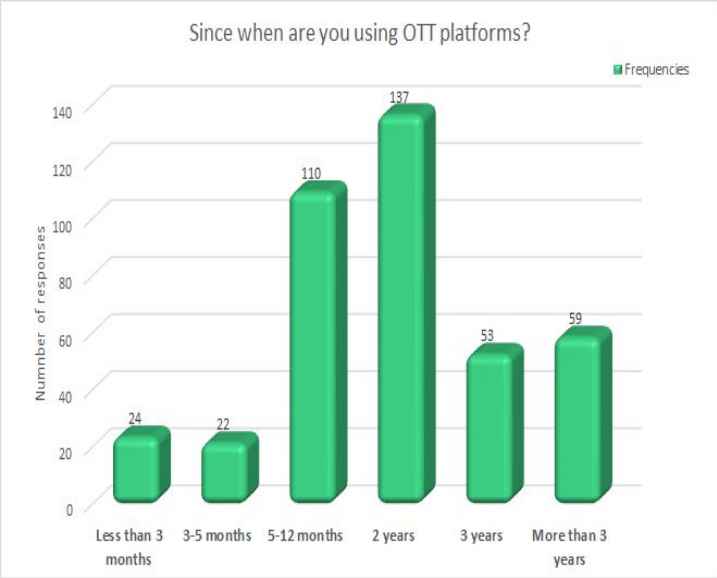




•**<u>Interpretation</u>:**

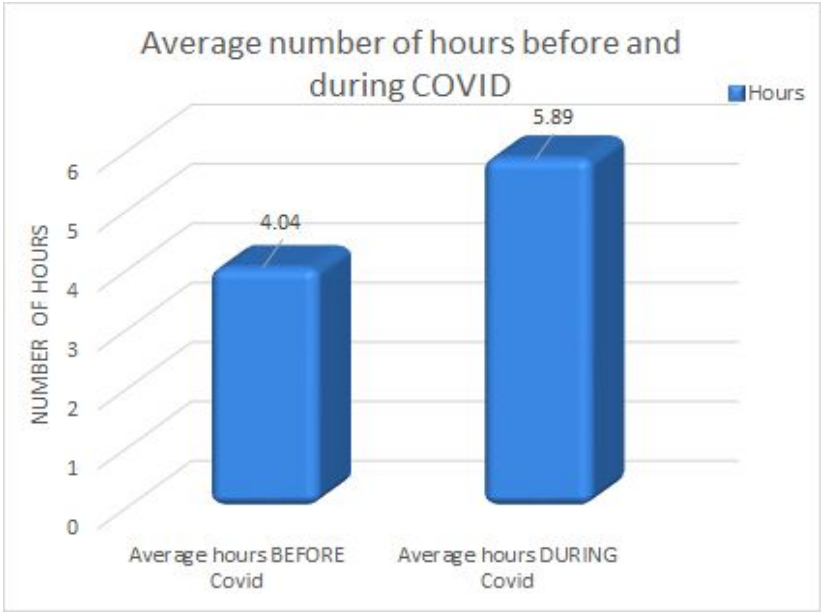
88% of the people are aware of OTT platforms of which 73.55% use them. From this it is seen that there are few people who do not use OTT platforms even though they are aware of them. 35% (220 people) of the sample do not use OTT. 87% of the viewers gain knowledge from the content they watch on OTT.

USING OTT PLATFORMS SINCE



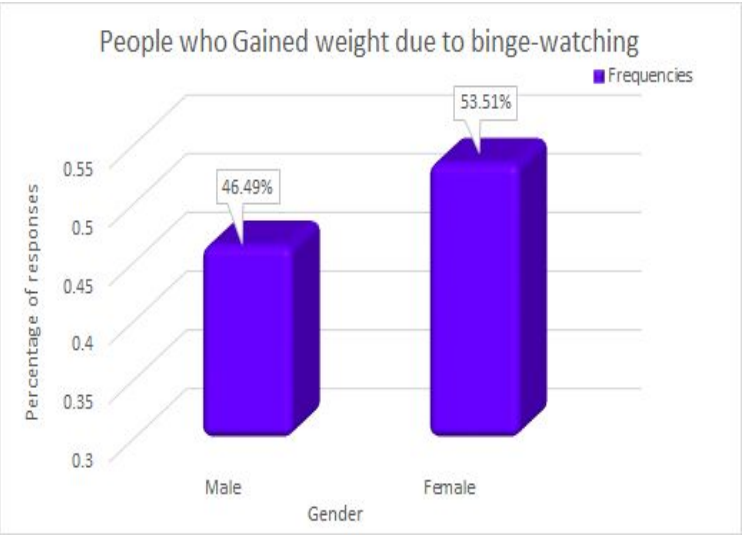
- There is an increase of 39% subscriptions during Covid.

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



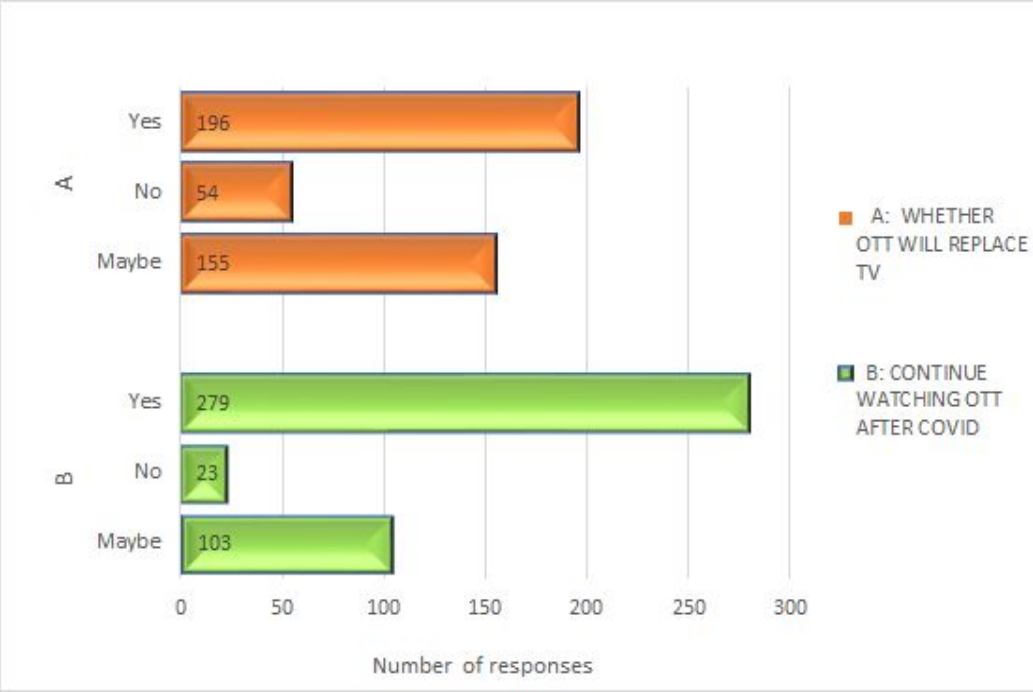
- The average number of viewing hours increased from 4.04 before Covid to 5.89 per week during Covid.

GAINED WEIGHT DUE TO BINGE WATCHING



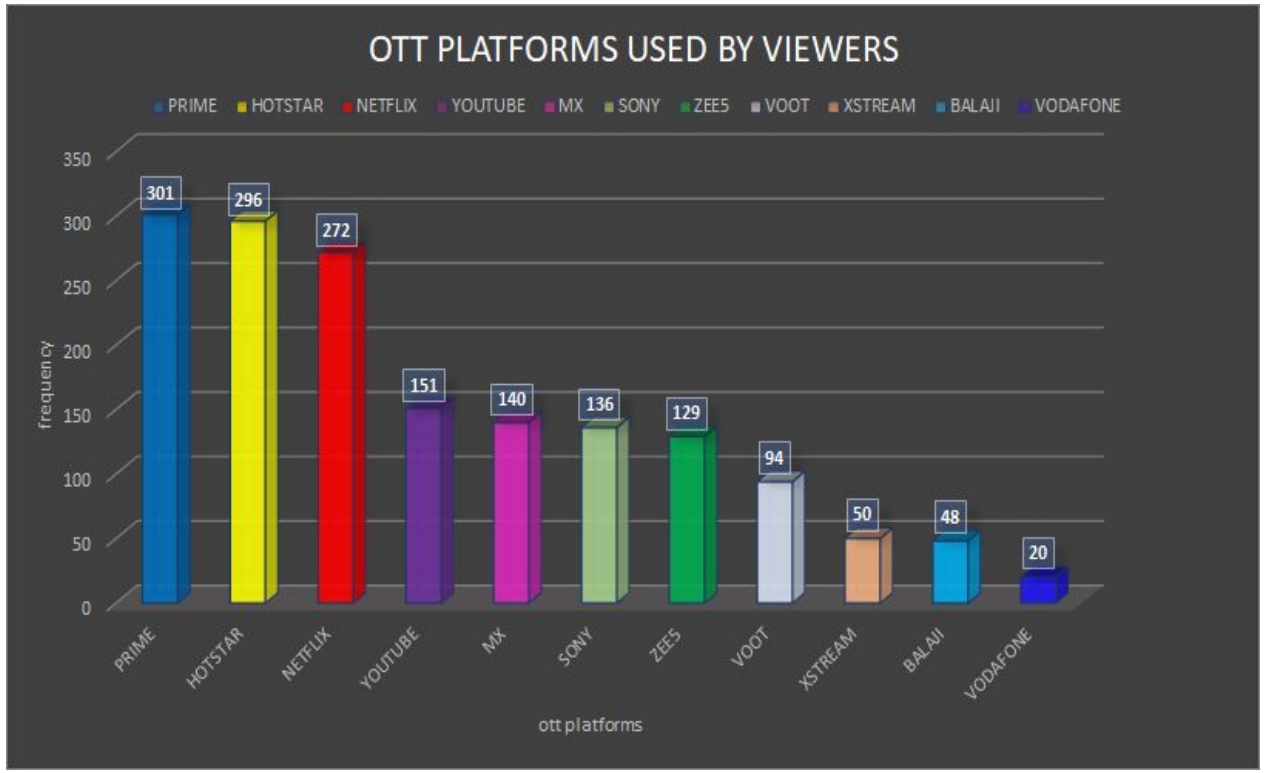
- Binge watch is an act of continuously watching for a long period of hours.
- The percentage of Females who have gained weight is more than the percentage of Males who have gained weight due to Binge watching.

FUTURE OF OTT



- 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.

OTT PLATFORMS USED BY USERS



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

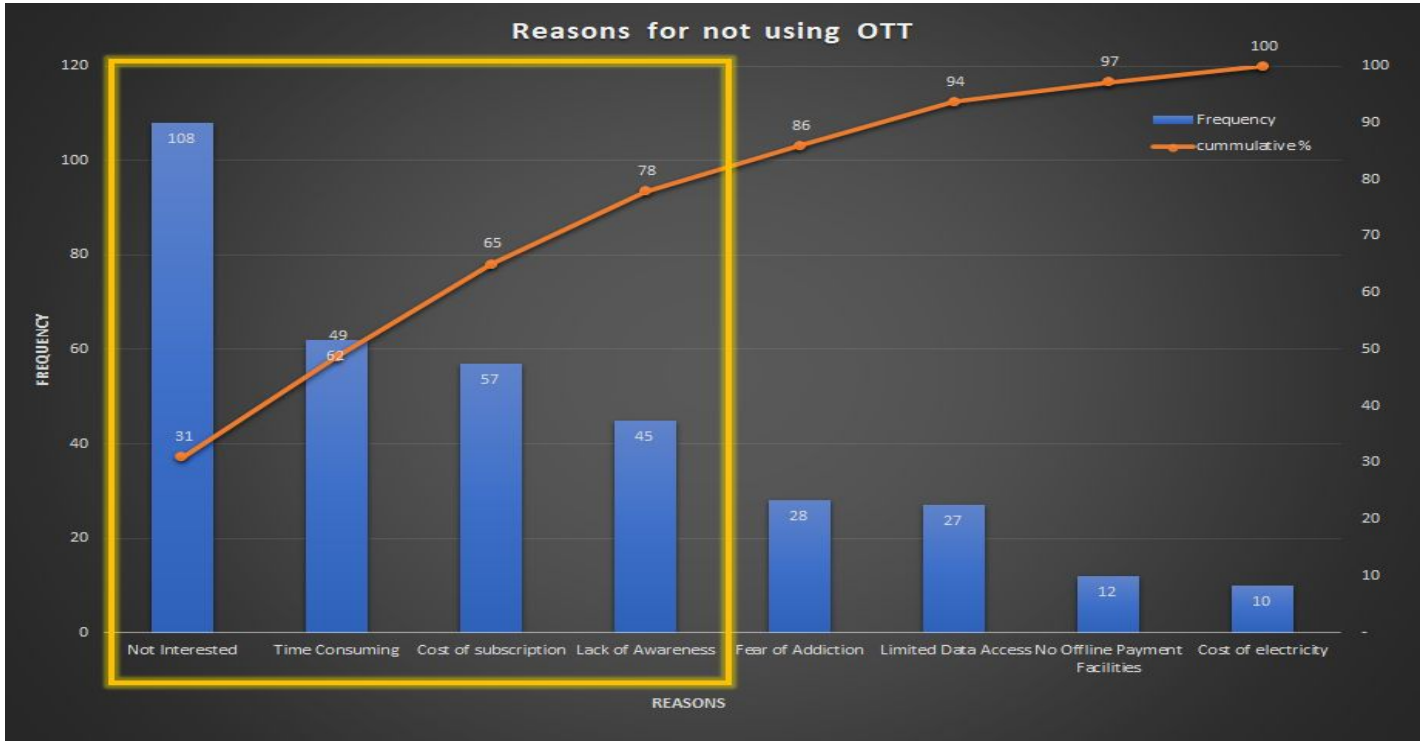
PARETO ANALYSIS:

Pareto analysis is based on the idea that 80% of a project's benefit can be achieved by doing 20% of the work or conversely 80% of problems are traced to 20% of the causes.

Objective: To Identify reasons for not using OTT

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

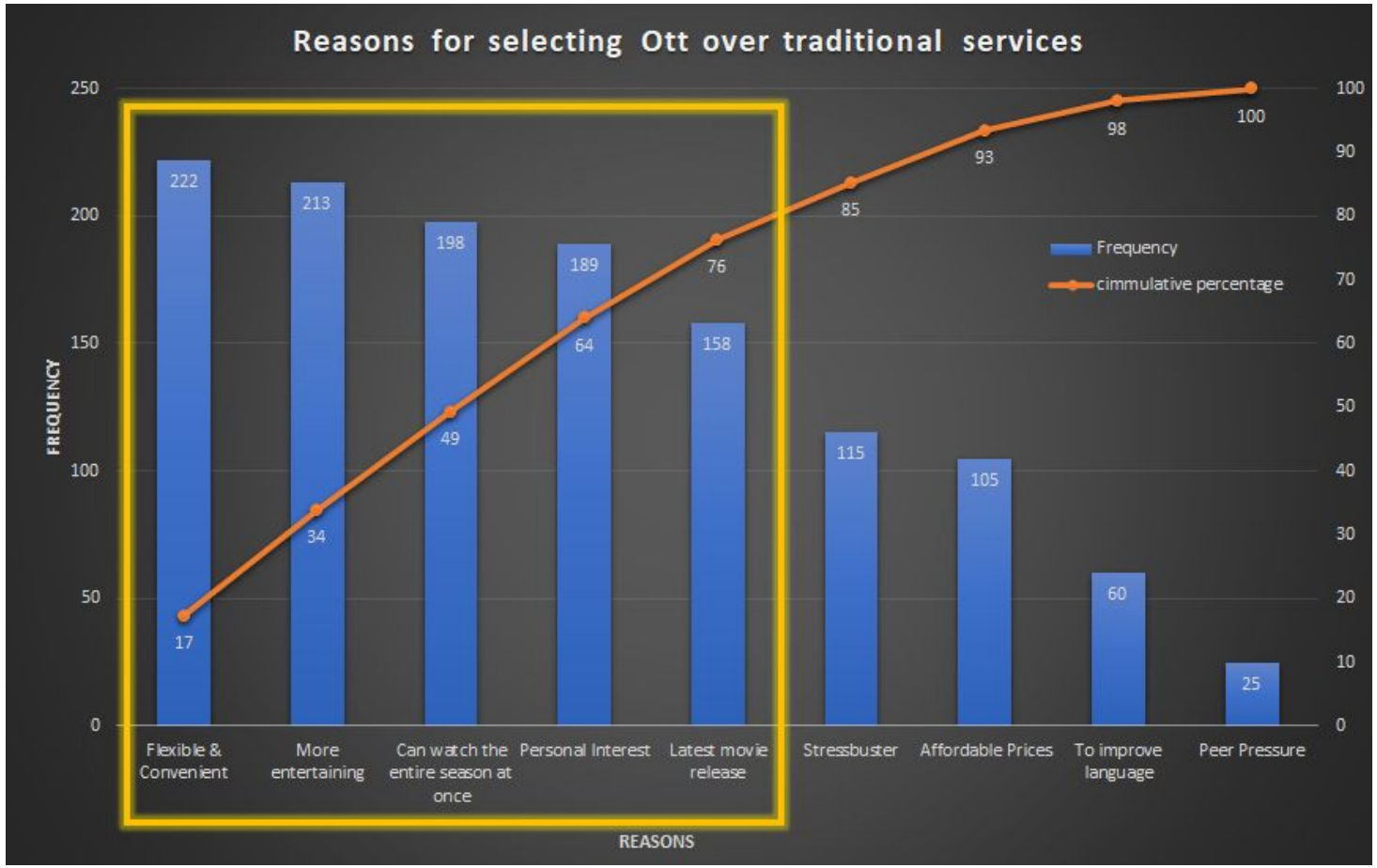


From Pareto Analysis, it is conclude that the main reasons for not using Ott are ‘Not Interested’, ‘Time consuming’, ‘Cost of subscription’, ‘Lack of Awareness’. To overcome these reasons, providers should focus on Awareness of OTT platforms, produce quality contnent and focus on providing different subsciption plans.

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

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



From Pareto Analysis, it is concluded that 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.

K-Means Cluster Analysis

- **Objective**: To compare and classify the viewers based on the user's preference.
- **K-means cluster analysis** is an algorithm that groups similar objects into groups called clusters. The endpoint of cluster analysis is a set of clusters, where each cluster is distinct from each other cluster.
- **Assumptions:** The clusters are spherical and are of resemble in size.

Variables used in analysis are:

1	Gender	15	Yearly spend on OTT
2	Age	16	Regular watching
3	Marital Status	17	Internet Source
4	Education	18	Language
5	Employment	19	Device used for streaming
6	Area	20	Ideal Time for streaming
7	family members	21	OTT platforms in use
8	Income	22	OTT Subscriptions
9	Awarness about OTT	23	Content
10	OTT user	24	Genre
11	Platform for entertainment	25	Duration of OTT use
12	How did you know about OTT	26	Time spent before lockdown
13	OTT over Traditional	27	Time spent after lockdown
14	Subscription BL		

Calculating optimum number of clusters using elbow method:

In this method, the number of clusters are varies within a certain range.For each number, within-cluster sum of square (wcss) value is calculated and stored in a list. These value are then plotted against the range of number o clusters used before. The location of bend in the plot indicates the appropriate number of clusters.

From Fig 1. we choose the number of clusters as ‘3

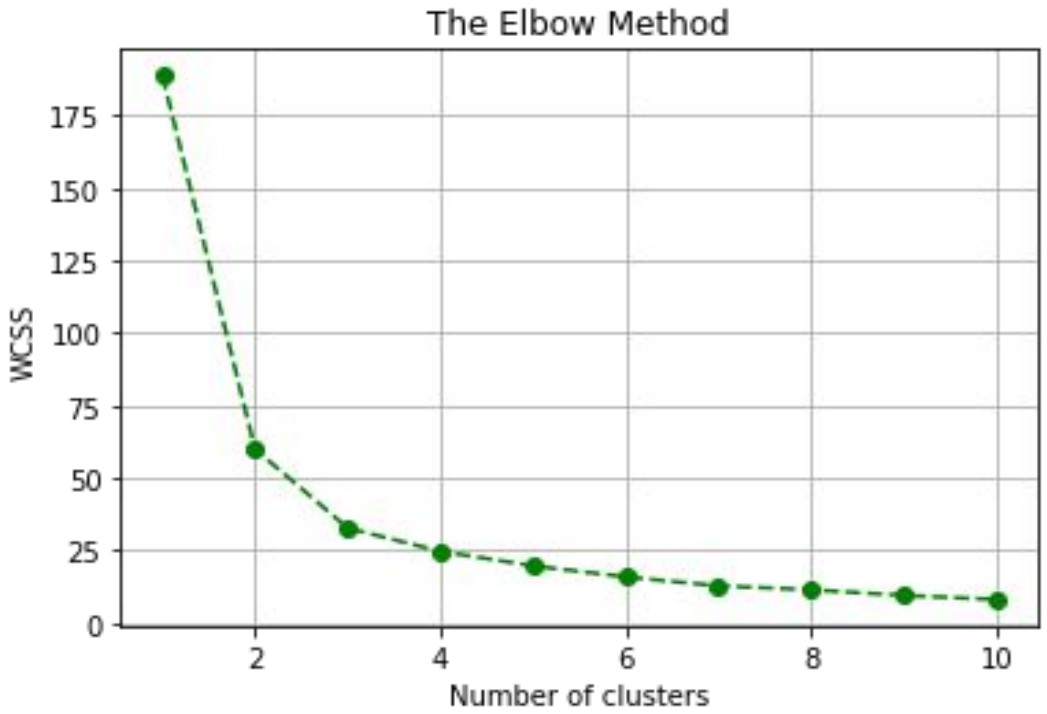


Fig 1.

In [15]:

```
from sklearn.cluster import KMeans
wcss = []
for i in range(1, 11):
    kmeans = KMeans(n_clusters = i, init = 'k-means++', random_state = 42)
    kmeans.fit(feature_scaled)
    wcss.append(kmeans.inertia_)
plt.plot(range(1, 11), wcss,'go--', color='green')
plt.title('The Elbow Method')
plt.xlabel('Number of clusters')
plt.ylabel('WCSS')
plt.grid()
plt.show()
```

Using the optimum no of clusters k-means model is trained :

In [16]:

```
km = KMeans(n_clusters=3)
y_predicted = km.fit_predict(feature_scaled)
y_predicted
```

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

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

```
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
```

Table 1

	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

The table 1 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

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 2

Table 2 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.

XG Boost Analysis

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

▪**XGBoost** is a scalable ensemble technique based on gradient boosting that has demonstrated to be a reliable and efficient machine learning challenge solver. This work proposes a practical analysis of how this novel technique works in terms of training speed, generalization performance and parameter setup.

▪The target variable is stored in ‘Y’ and features in ‘X’.

Y : User having subscription of Amazon Prime

Y = 1 ; if yes

= 0 ; otherwise.

Variables used in analysis are:

1	Gender	15	Yearly spend on OTT
2	Age	16	Regular watching
3	Marital Status	17	Internet Source
4	Education	18	Language
5	Employment	19	Device used for streaming
6	Area	20	Ideal Time for streaming
7	family members	21	OTT platforms in use
8	Income	22	OTT Subscriptions
9	Awarness about OTT	23	Content
10	OTT user	24	Genre
11	Platform for entertainment	25	Duration of OTT use
12	How did you know about OTT	26	Time spent before lockdown
13	OTT over Traditional	27	Time spent after lockdown
14	Subscription BL		

```
#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)
```

- Splitting data into training and testing data sets.
- Training the XG Boost model on train dataset.

- Calculation of Correlation matrix, VIF and Confusion Matrix.
- Removal of features which were highly correlated, and which are having high VIF then again running the model using remaining features.
- Top 15 features are selected using feature important function

```
#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)
```

```
# 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)
```

- Table 3 is the final correlation matrix table.

	Yearly spend	OTT_Subscript	Subscription	OTT_Subscript	No_of_Devic	Internet_Sou	Genre_Actic	Genre_Sci-	time spent A	Content_Mo	OTT_Subscri	OTT_Subscri
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 3

- Table 4 is indicating final VIF matrix,VIF for all the variables is under 8 hence there is little or 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 5 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 5

Table 6 is the confusion matrix for testing 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 6

- The Accuracy of the training set is 78.87% & the testing set is 75.61% .
- Hence the results of the testing 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.
- Significant features for this model are given in the table 7 along with their significance.

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 7

Binary Logistic Regression

- **Objective:** To study the socio- demographic factors that affect the usage of OTT platforms.
- **Logistic regression** is the statistical technique used to predict the relationship between predictors (independent variables) and a predicted variable (the dependent variable).
- **Assumptions**: Independent samples, Dependent variables should be binary, Little or no multicollinearity, Logistic regression assumes linearity of independent variables, Large sample size, No outliers.
- Since all the VIF values for the variables are > 1 and < 5 which suggests the absence of 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

❑ FORWARD LIKELIHOOD RATIO SELECTION:

Step Summary[a,b]

Step	Improvement Chi-square	df	Sig.	Model Chi-square	df	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.

❑ SIGNIFICANCE OF THE NEW MODEL:

- $H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$
  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 the model is fitting the data significantly better than a null model with no predictors.

❑ 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 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.

❑ HOSMER AND LEMESHOW TEST (GOODNESS OF FIT)

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
4	4.073	8	.850

- H_0: The REDUCED model is a good fit

  H_1 : The REDUCED model is not a good fit
- Here 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]

			Predicted		
			Use_OTT		Percentage Correct
	Observed		No	Yes	
Step 4	Use_OTT	No	73	107	40.6
		Yes	10	435	97.8
	Overall Percentage				81.3

a. The cut value is .500

- 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 can correctly classify 81.3% of the cases. Hence any value above 80% is called a good model.

❑ FITTED MODEL

- 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)

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.

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.

❑ Estimate (B)

- The odds of an older person using OTT platforms is low.
- A person with 4 and 5 family members are more likely to use OTT.
- Individuals with annual family income between ‘Below 1 Lakh’ are least likely to use OTT platforms.
- Respondents who were aware of OTT platforms were highly likely to use OTT platforms.

❑ Wald Statistic-

•H_0: Individual independent variables are significant

H_1: Not H_0

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

❑ Odds ratio:

- 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.
- For an additional year in age, the odds of a person using OTT platform is lowered by a factor of 0.961.
- 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.
- A person with 5 family members is more likely to use an OTT platform than a person with a More than 5 members.
- People who are aware of OTT platforms are 22 times more likely to use OTT platforms than people who are not aware of OTT.

APRIORI ALGORITHM

- Apriori is an algorithm for frequent item set mining and association rule learning over relational databases. it uses prior knowledge of frequent itemset properties.
- **Objective:** To understand the association and frequency between different genres
- The dataset looks like,

```
: df
['Comedy;Thriller;Romance;Drama;Horror;Crime'],
['Comedy;Thriller;Romance;Action;Drama;Horror;Sci-Fi;Crime;Children & Family'],
['Thriller;Romance;Drama'],
['Comedy;Thriller;Action;Crime;Children & Family'],
['Comedy;Thriller;Romance;Action;Drama;Horror;Sci-Fi;Crime;Children & Family'],
['Thriller;Action;Drama'],
['Comedy;Thriller;Sci-Fi'],
['Comedy;Thriller;Romance;Action;Drama;Horror'],
['Comedy;Thriller;Romance;Action;Drama;Crime;Children & Family'],
['Comedy;Romance;Action;Horror'],
['Comedy;Thriller;Romance;Action;Horror'],
['Comedy;Thriller;Romance;Action;Drama;Horror;Sci-Fi;Crime'],
['Comedy;Thriller;Romance;Action;Crime'],
['Comedy;Thriller;Action;Sci-Fi'],
['Comedy;Sci-Fi'],
['Comedy;Romance;Horror;Children & Family'],
['Comedy;Romance;Action;Drama;Sci-Fi'],
['Comedy;Thriller;Action;Horror;Sci-Fi;Crime'],
['Comedy;Thriller;Romance;Horror;Crime;Children & Family'],
```

Support

- 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.

```
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)

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

```
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

- 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.

CHI-SQUARE

	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	0.000000000002	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	0.000000000002	Therefore, we reject Ho and conclude that in our study out of 305 undergraduates 280 people are aware of OTT.

Sr no.	Ho	• Ha	Chi-square value	P-value	Interpretation
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.

Assumptions of Chi-Square:

- Independence of observations
  Each observed frequency is generated by a different individual
- Size of expected frequencies
  Chi-Square test should not be performed when the expected frequency of any cell is less than 5

Interpretation:

- 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.

SWOC ANALYSIS

STRENGTHS :
- Connect to Multiple devices.
- No Ads are shown while using Premium Content.
- Flexible and Convenient
- More Entertaining.
- Watch the entire season at once.

WEAKNESSES :
- Mental and Physical Health problems due to Binge Watching.
- Consumes a lot of Mobile data.
- Too many Ads in Freemium Content.
- Some OTT platforms have higher Subscription costs.
- Lack of Awareness in rural areas.

OPPORTUNITIES :
- People are more used to Movies and Web-series.
- Users can be of any age.
- Scope of innovation and digital development.
- More Live Content.
- More Regional Content can be added as per viewers response.

CHALLENGES :
- Providing Niche Content.
- Enhance content Viewing discovery.
- Cannot afford high cost subscriptions.
- Some of the viewers are Not Interested.
- Lack of Awareness.

- 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.
- 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.
- 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 web-series as viewers want a variety of entertaining content to suit every mood.
- Lack of Awareness in rural areas may be the reason of network connectivity.

Conclusion: Suggestions

- Collect user's feedback when they delete accounts to deliver a better streaming experience to other customers.
- Produce more Advertisements for better reach.
- 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.
- Focus more on the originality of the shows.
- Speed up content discovery by creating stories.

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.