Deloitte Digital Democracy Survey Analysis

Akshay Munde

Contents

Introduction

For decades before 2000 media consumption remained dominant. People used television, radio as the main devices for entertainment. But with the internet powered **tech boom**, a simple click replaced the need to go to the rental stores. People shifted to streaming music, video, and movies from their preferred devices anywhere. Consumers are ditching cable TV for content streaming to pave the way for the digital era.

This shift in media consumption has obviously affected digital marketing. For instance, digital marketing strategies used to include text messages, TVs, as their platform. However, with changing media consumption habits, our analysis will help rethink content creation strategies to include platforms like social media and devices such as smartphones, tablets, and laptops.

Objective

The objective of this project is to analyze the tide of change in media consumption which is driving the revolution in digital marketing.

Unsupervised learning will be used to explore:

1. **To own or to Stream?**
   In the past 5 years or so, consumers have largely shifted from renting movies to streaming platforms like Netflix. This analysis will help clients to come up with new marketing strategies to ensure content reaches viewers on these streaming platforms like Netflix, Spotify.
2. **The growth of Social Media**
   With billions of active users every month, social media has contributed immensely to shift in media consumption. This will help the digital marketing department to focus in designing engaging content that connects with the audience on social media.

Supervised learning will be used to predict if a consumer is willing to watch more advertisements for a lower subscription cost to streaming services.

Data Description

In the 11th edition of Deloitte Survey data, we have a total of **2131 rows and 198 columns**, giving us a view of what consumers are doing. It explains the demographics of a consumer along with the digital questions asked on the devices and services used by them. This survey would help them understand how the interaction and engagement are taking place and how Deloitte and others can use it for advertising and other means for their benefit. This would also help them to understand where they stand in the market and where they headed.

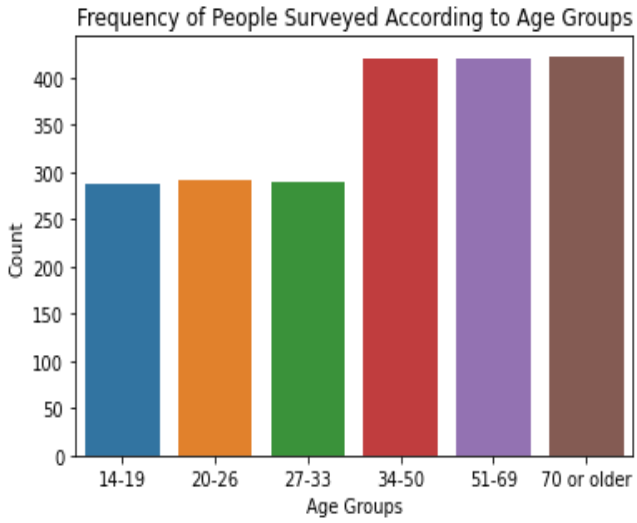
Some of the most common fields available for consumers here are: Age, Age-Group, Gender, Region of stay, Employment Status, Race, Household Income, Devices and Services they use.

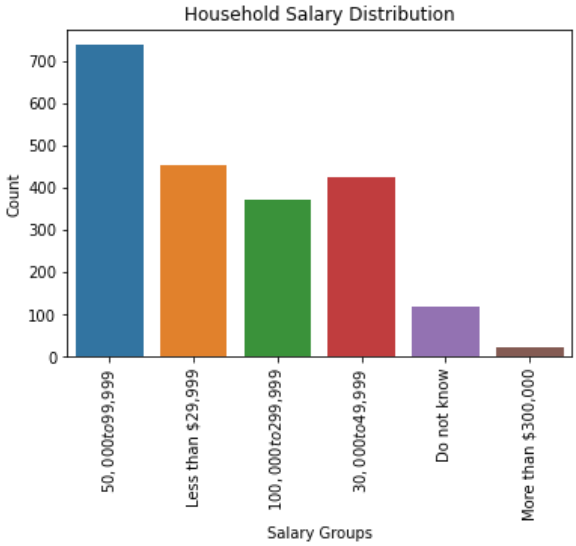
Data Pre-processing

We started preparing our dataset by handling the names of the columns in the dataset. As it is a survey, it had questions asked as the column headings which would be difficult to handle while working on the data. So, we trimmed the questions to include only the question number. For an example, Q4 represents the gender, QNEW3 the employment status, Q5 the race of the consumer.
Also, to perform exploratory data analysis on this dataset and have better understanding, we changed the null values to No for the questions where the consumer didn't answer. This was done on those columns in the dataset where null and "no" were assumed to be equivalent.

Exploring Data Analysis

After performing cleaning operations and replacing null values, we explored the most important attributes in the dataset and created visualizations for better interpretability.

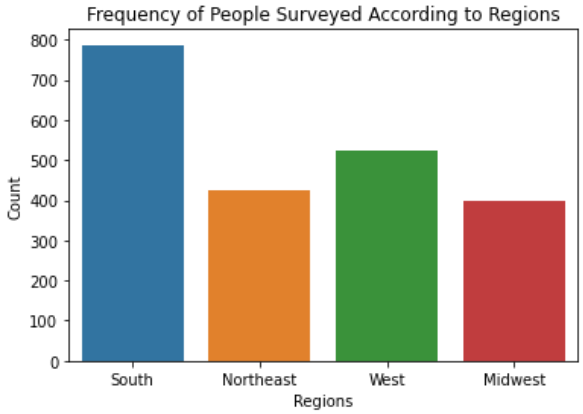


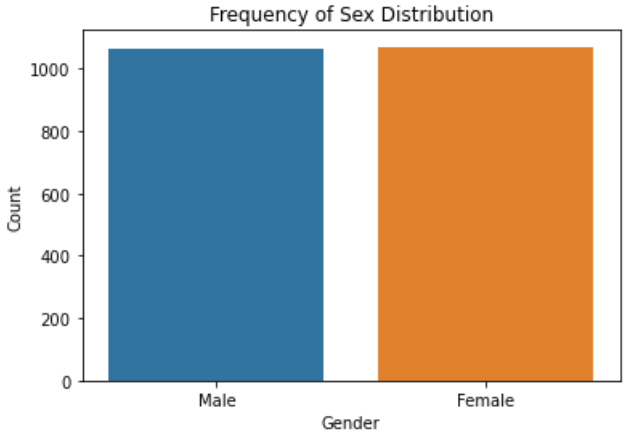


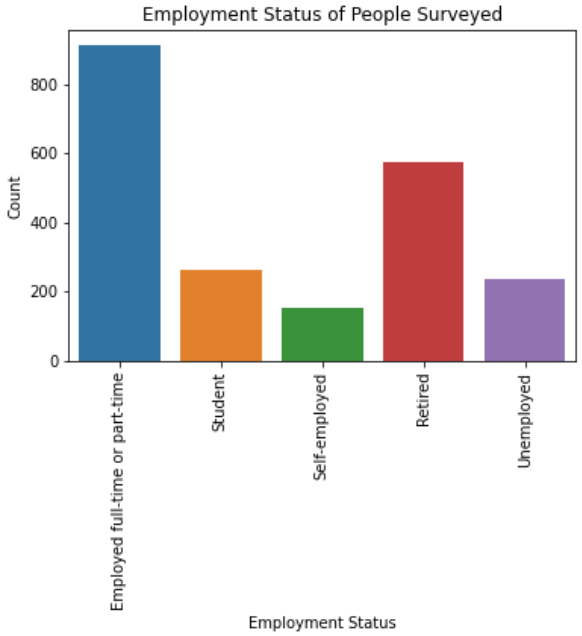
From the histogram we can see that, more than half of the consumers (1250+) were above the age of 34. Also, we have a good number of consumers from the age 14 to 33 i.e. almost 900 and hence there is no skewness in the age distribution.

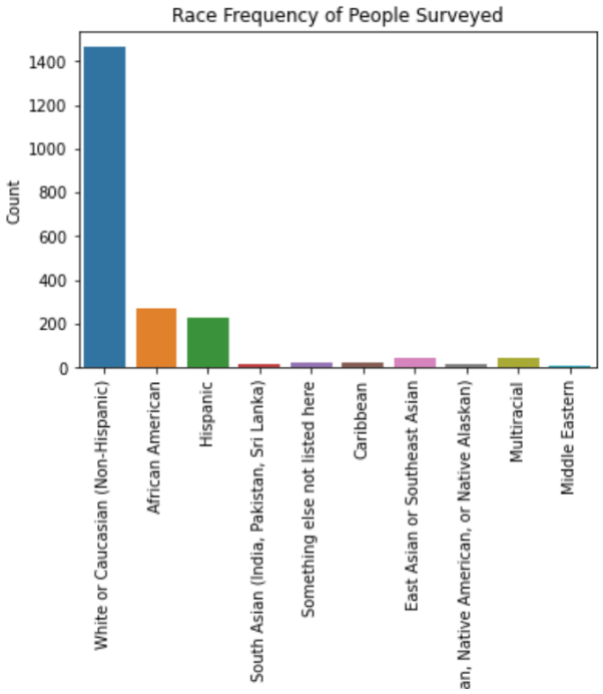
Moreover, more than half of the consumers household income is between $50,000 to $300,000. As the median household income in the US falls around ~$65,000 we can see the expected count (730+) for range $50,000 to $99,000. This is a good target audience for the media companies to advertise their products.

Similarly, we explored and created histogram distribution for regions, gender, employment status, race, and household salary of the consumers.
From the region distribution we found out that most of the people (~800) were surveyed from the South region.





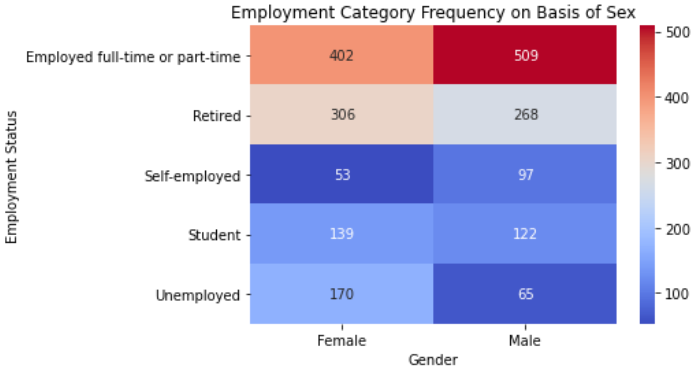




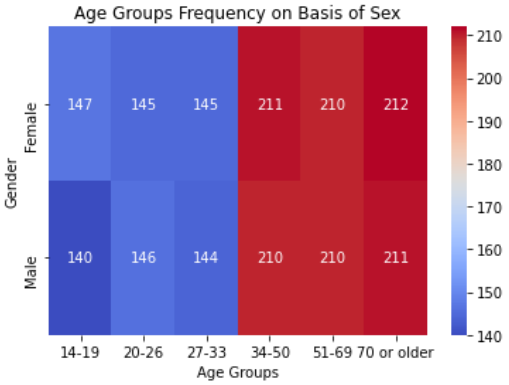
From the above distributions, we found out that almost 70% of people surveyed were of white race and only 25% of people were of African, American, and Hispanic race. This could be because of the sampling bias. We saw an equal distribution of males and females surveyed.

Crosstab Distribution:
Moving forward we explored the crosstab between these attributes to get a more detailed and deeper understanding of our dataset.
From these crosstabs, we got the relationship between 2 variables and a better view for our modeling questions.

Employment Vs Gender:

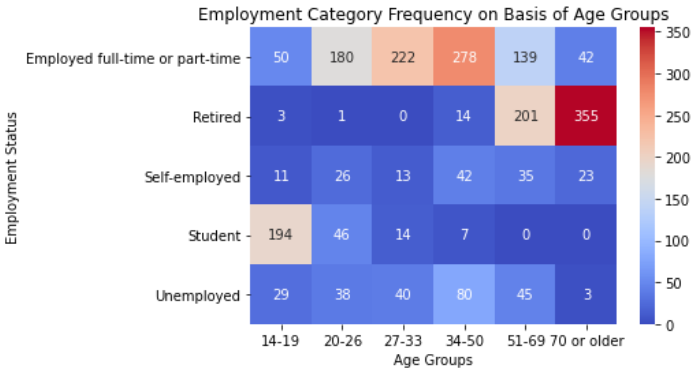


Age Groups Vs Gender:

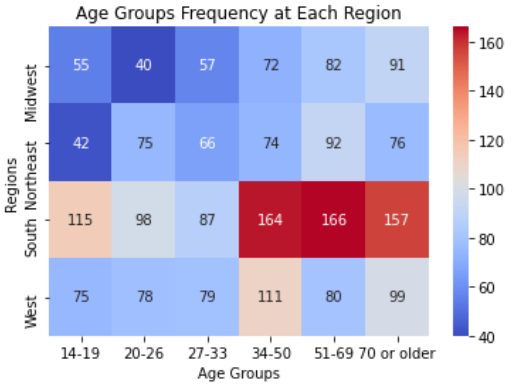


From the employment vs gender tab, we found out that more men are employed compared to women. Also, the unemployment numbers are almost 3 times for females. Studying this we can design separate marketing strategies keeping in mind the employment status of a consumer. Moreover, we have 60% of our audience above age 34 and there is equal distribution of males and females surveyed. Seeing these numbers, we can sense that online advertising won't be the only means for our advertising. But studying their social media usage would help us understand the norms of their likeness.

Employment Vs Age Groups:



Region Vs Age Groups:

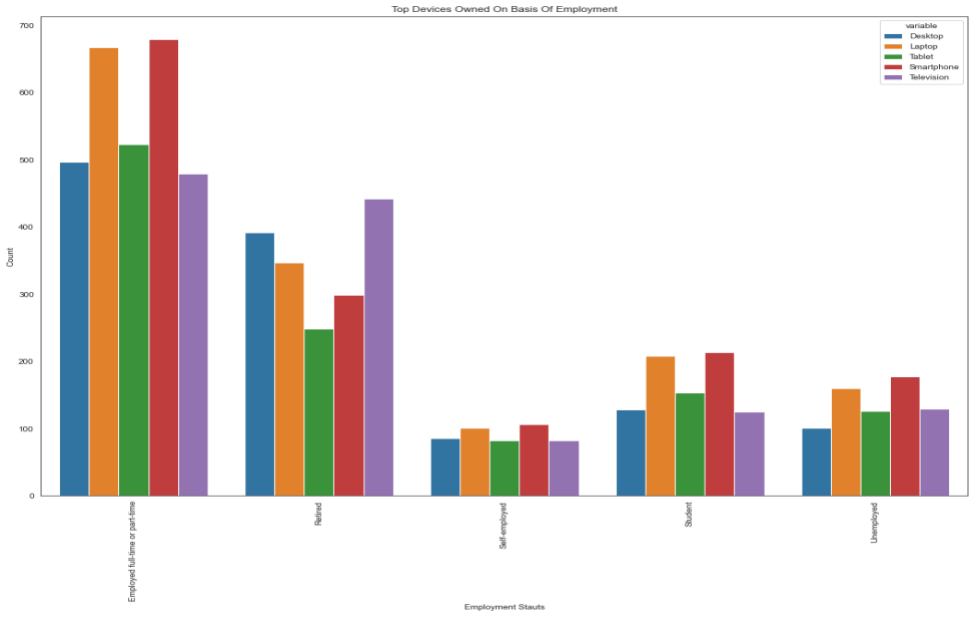


Employment and Age Group relationships of consumers surveyed behaves as expected. Majority of retired people are over 70 and almost 900 people who are employed fall between the ages 20-69. As mentioned earlier, companies can use these demographics to design their marketing strategies. However, region and age groups crosstab help us see some biases as most of the people are surveyed from the South region and their age is more than 34+. This is not an equal distribution of consumers as it exists in all other regions.

Moving forward, as our dataset contains the details about devices used, services used and applications favored by consumers, we explored them over multiple aspects i.e. age groups, gender and among employment status to study the behaviors of these consumers and to study which is the best means to target them.

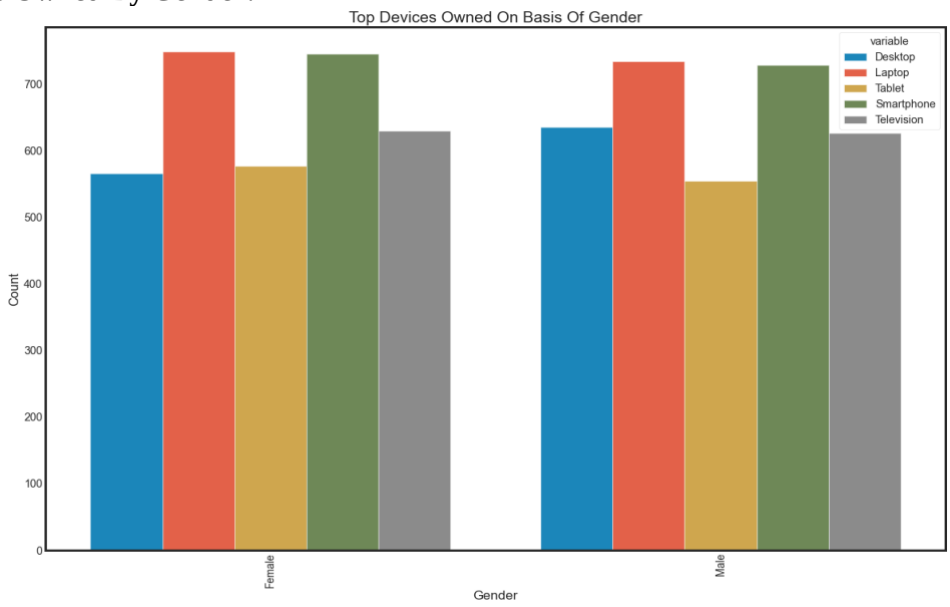
While exploring the devices used by these consumers, we found out that **Top Devices owned by them are**: Desktop, Laptop, Tablet, Smartphone, and Television in that order. As in the last two decades digital devices have taken over the market we can differentiate between our audience and the means to get their unvaried attention.

Top Devices Owned on Basis of Employment:



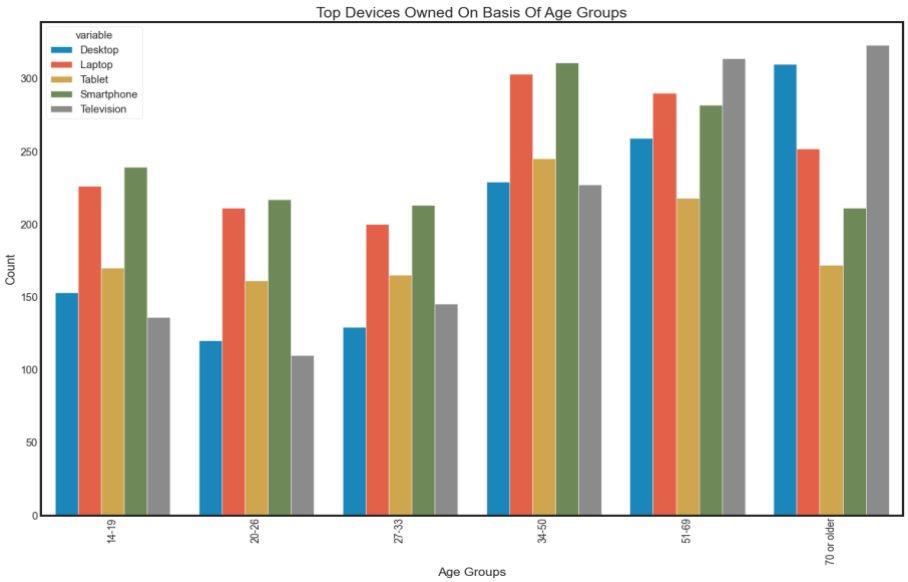
From our dataset we can see that smartphones and laptops are the most used devices for employed peoples. Hence, we should target them through online promotions of our products. Whereas for people who are retired, the most owned devices are television. So, we can promote our products there through tv-advertisements.

Top Devices Owned By Gender:



From the above graph, we found out that there is equal distribution of devices owned by both males and females. Hence, we do not need specific strategies for each one of them.

Top Devices Owned by Age Groups:



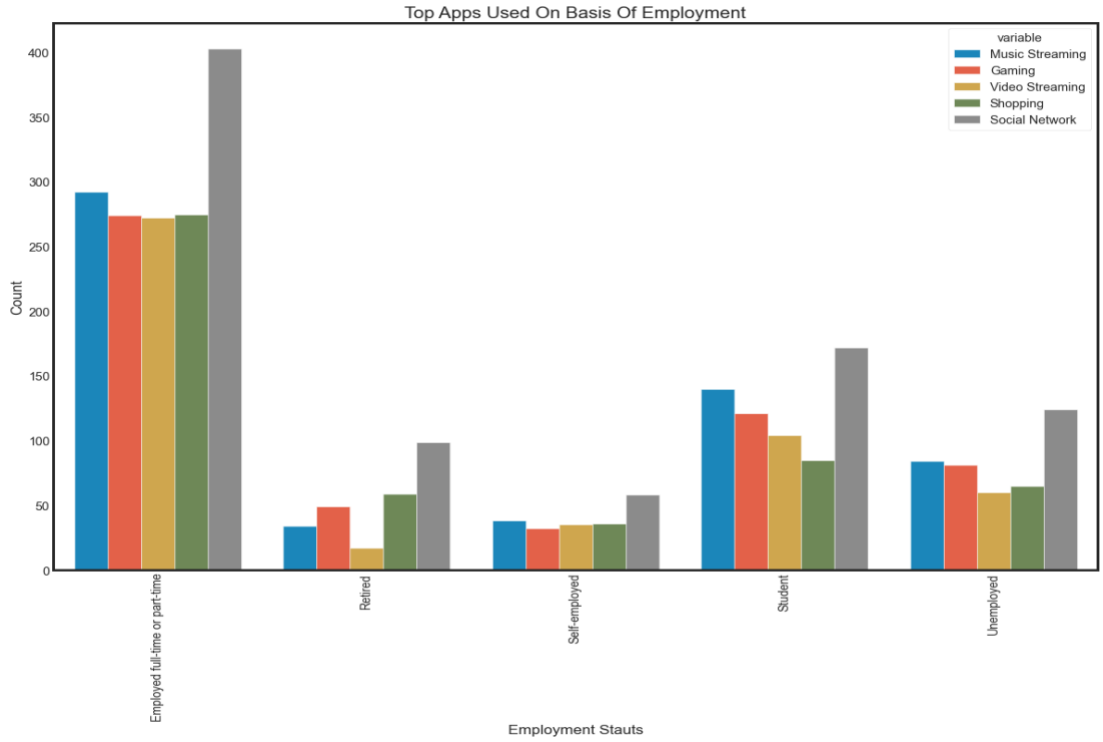
For people above age 50, they own Television the most which seems correct as Laptop and Smartphones came into the picture in the last 20 years only. We can also see teenagers below 19 have smartphones as the most owned device. Hence as mentioned above targeted marketing would be more effective here. We need to lure older people through television marketing. Go local and target the audience of that specific region through the tv ads. To make the product visible, TV ads can go a long way and people can search for the products online later. In today's generation, how much visible the product is, more effective it is. Hence make use of all the above-mentioned devices and get the attention of the consumer.

Top Apps Usage:

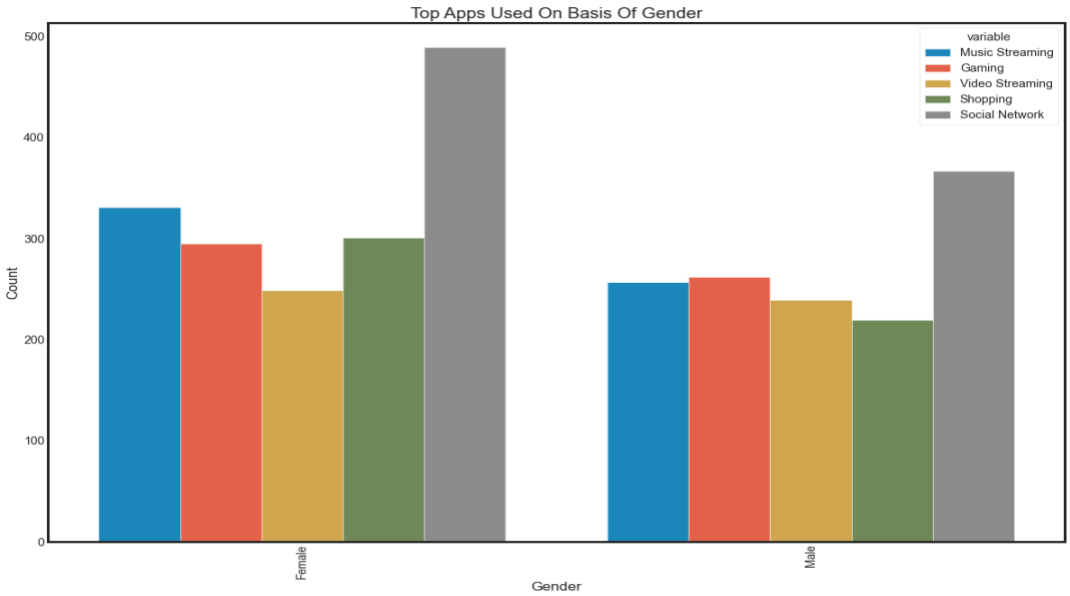
In Order to understand the app usage/favored by consumers, we found the top apps and studied the distribution among multiple demographics.

Top Apps are: Music Streaming, Gaming, Video Streaming, Shopping, Social Network

Top Apps Used by Consumers on Basis of Employment:

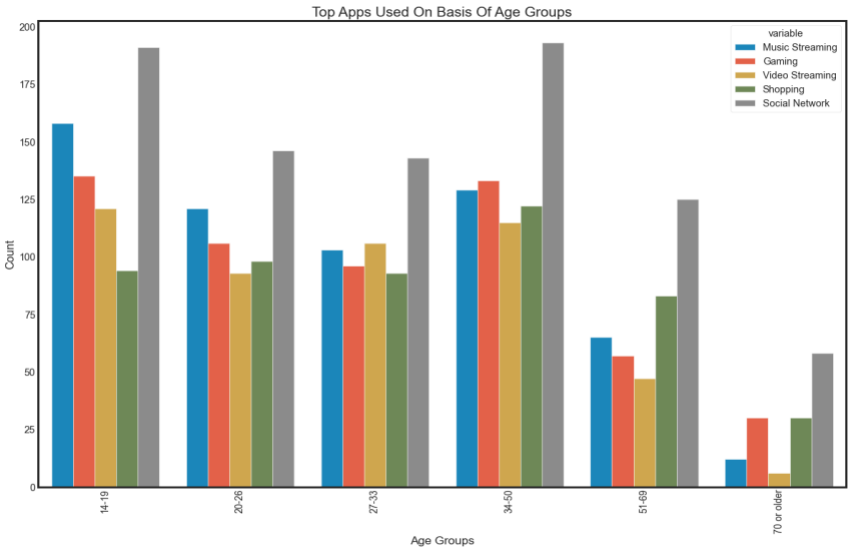


Top Apps Used by Gender:



Interestingly, females use these top apps more than males. Most interesting fact is that they use more Gaming apps than males by a significant margin. So, to target a broad audience companies can make use of Social Networking apps. In recent time, Instagram and Facebook have lots of influencers and product owners can advertise their products using these influencers services. They have a broad reach and with their huge list of followers, and one product campaign can reach thousands of people in a few minutes. With these apps, promoters can market and sell their products for free.

Top Apps Used by Age Groups:



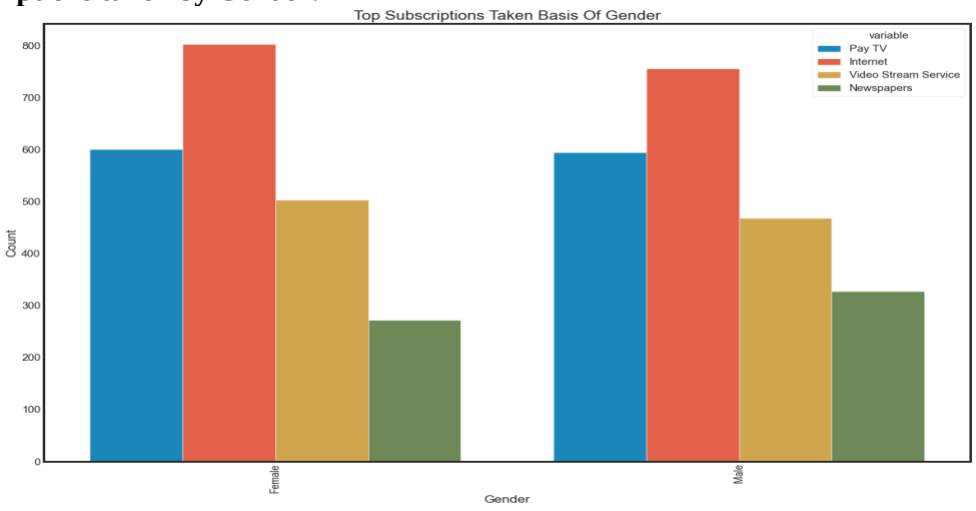
As expected, teenagers are the one who uses Social Networking apps the most. Interestingly following them are the people in their late 30's and early 50's. Gen Z is moving from Facebook to Instagram and YouTube. With the trend of everyone being an influencer or a Youtuber, companies can target the influencers first rather than traditional marketing strategies. As the ads on these apps are friendly, the word of mouth and good recommendations can do wonders for the product owners. Teenagers do not consider influencers posts as ads and hence these are cheaper, scalable, and more effective compared to regular advertisements.

Top Subscriptions:

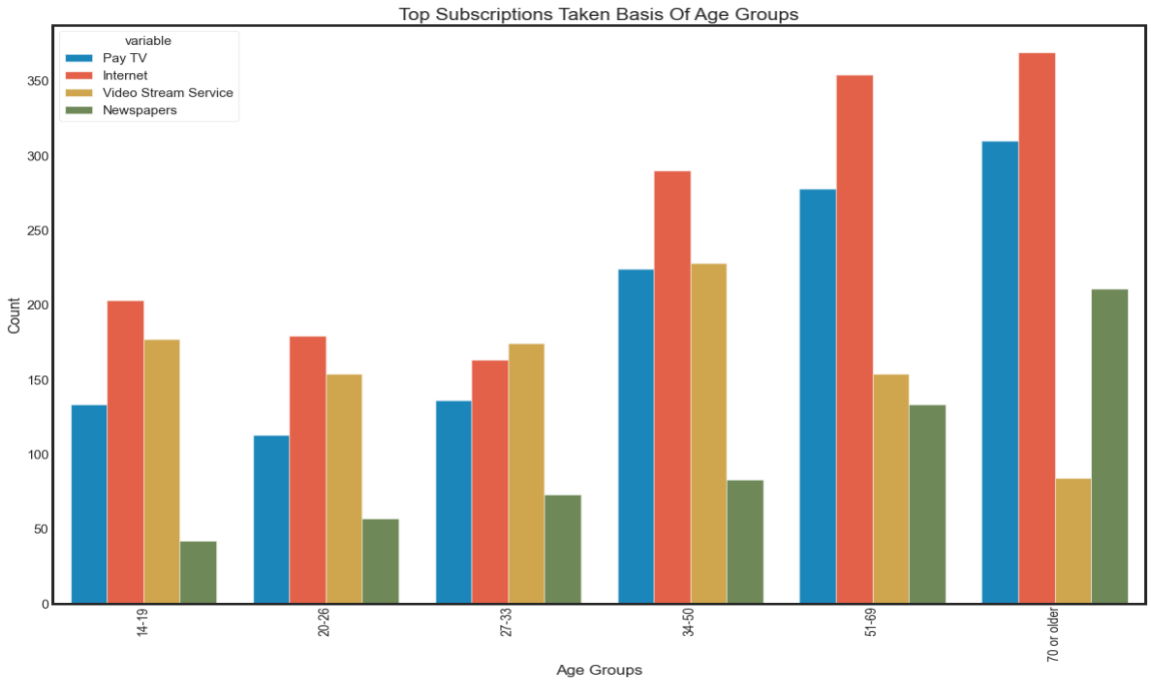
After studying the devices and apps usage by consumers. We moved on to the subscriptions side. The **Top subscriptions are:** Pay Tv, Internet, Video Streaming Service and Newspapers for the people surveyed.

Here below we will find the top subscriptions taken by people based on the gender and the age groups. Following the above trend, the distribution of these top subscriptions is almost similar between males and females. But we can see, the usage of the internet is highest among both. As today's public spend almost 2 hours of the day on Social networking, the internet becomes the most valuable subscription for them. In order to stream music and movies, they require undisturbed connections. As the printed newspapers are on the decline, the promoters can still promote their product via digital newspaper copies. Regulars have shifted from hard copy to digital but still there are a handful of people who spend a good amount of time studying them.

Top Subscriptions taken by Gender:



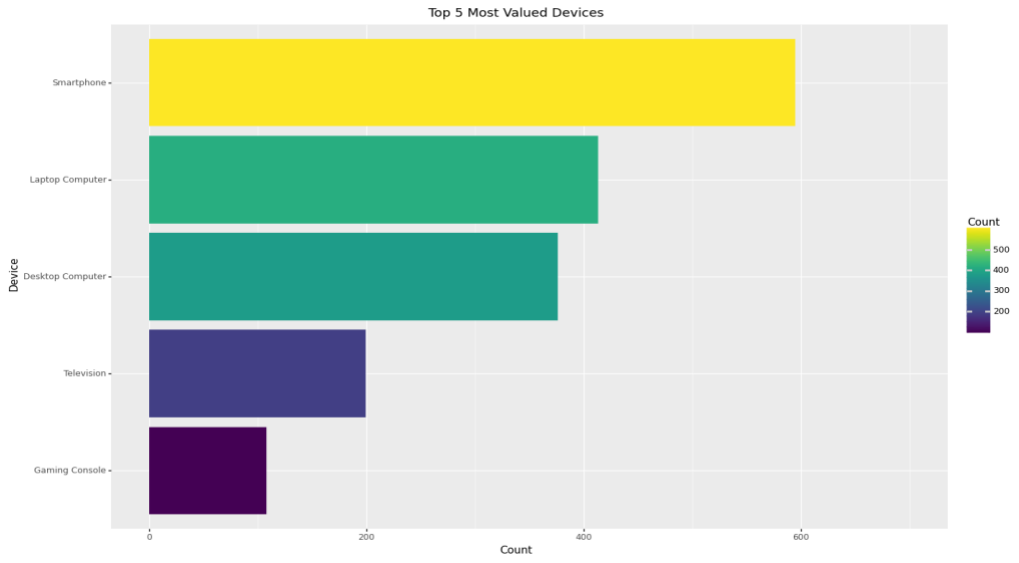
Top Subscriptions taken by Age Groups:



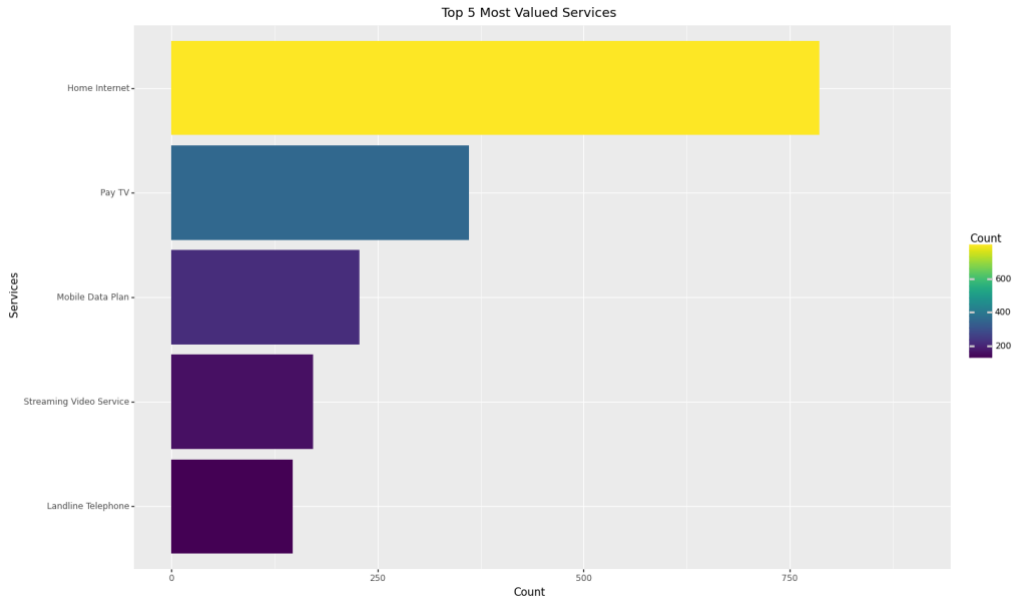
Gen Z and people in their 30s spend most of their time on social media and video streaming platforms. From our surveyed people only, we found out that more than 700 people binge watch regularly, i.e, they spend 3-4 hrs continuously watching a tv show or movies. To make this more interesting, people in recent time spend 15hrs per week on these streaming platforms. Availability of multiple platforms gives the option for the product owners to personalize the ads according to consumers based on the platform. They can perform A/B testing to assess the effectiveness of their marketing campaign.

Most Valued Devices by Consumers:

Of the devices owned by consumers, we found that their most favored devices are the following:



Most Valued Services by Consumers:
Similarly, of the device's consumers use, their most preferred services are:



The internet of things has been growing steadily over a couple of years and an increasing number of gadgets are connected to home internet. Hence, we can see that the most valued service for consumers currently is Home Internet. This presents a new digital marketer new way to engage with customers. For example, advertisements can be pushed through smart refrigerators, smart watches and opportunities are endless in this era. The availability of smartphones, video platforms and the internet make the perfect opportunity for marketers to be consistent with their proven ads and implement multiple campaigns in short times.

Data Issues/Limitations:

Though we have a good amount of data, there are still some limitations and issues with this dataset. As we saw above 70% of surveyed consumers are White, we do not get a proper distribution of surveyors. We cannot expect to have similar results for the people of other races for our data modeling objectives. Similarly, we found out though the distribution of males and females is quite equal, still most of the people were surveyed from the south region and that can lead to some biases. Also, we had many columns with null values i.e. for columns where people do not own the device/service and hence they cannot tell their review about it. That may lead to some inconsistencies as we neither have a positive or negative take on that thing. Though we took care of most of the columns containing null values, it changes the understanding of the data on the whole level. Also, the number of surveyed people is at a small scale and hence to be more robust and thorough with our findings we need more dataset. To make better strategies for people who are unemployed, people in their teens and people from other regions than south, we need more detailed surveys.

Business Case I

Summary
A survey conducted in 2016 by the Pew research center found that out of the billions of active users, 62% of adults in America consume news via social media. The survey also found that 18% of adults got news from social media on a regular basis. Thus, social media is well attended by all age-groups to discover new content, news and resolve customer-service issues.

In the past five years, consumers largely shifted from renting movies from rental stores to streaming online on Netflix, Hulu, and Amazon Prime. In fact, 60% of households with broadband connection in the U.S. have at least one subscription to streaming service.

This analysis will help devise new marketing strategies to ensure content reaches the consumers through streaming platforms and social media.

Data Preprocessing

We replaced all columns containing categorical values:
{'Yes': 1, 'No': 0},
{‘Male’: 1, ‘Female’: 0},

We converted following columns with numeric values to categorical with value >50 to 1 else 0
Q15r1, Q15r2, Q15r3, Q15r4, Q16r1, Q16r2, Q16r3, Q16r4, Q17r1, Q17r2, Q17r3, Q17r4

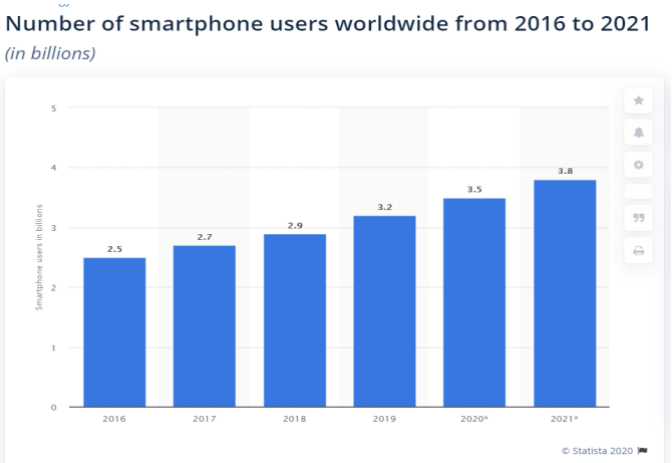
We converted columns with categorical values to dummy indicator variables for clustering and analysis.

For instance, Q11r1 Television-Of the products you indicated you own, which [total count] do you value the most? Is converted to Q11r1_1_1.0, Q11r1_1_2.0, Q11r1_1_3.0 to illustrate if Television is ranked 1 or ranked 2 or ranked 3.

Data Analysis
With boom in technology and shift of people from traditional Television and radio to Laptop and smartphones, we will analyze following in different age-groups, employment status and salary:

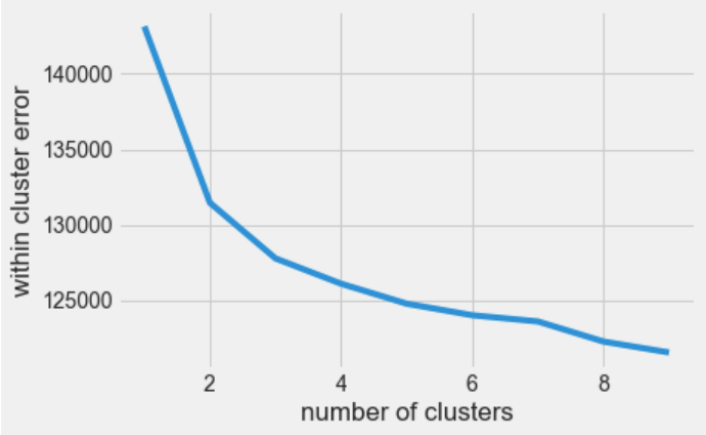
1. **Most valued device**
2. **Most preferred applications for consumers**
3. **Most common activity while watching television**
4. **Most preferred entertainment activity**
5. **Most preferred activity even while using other people subscription**

This analysis will help understand consumer behavior and limit unnecessary advertising spend across all platforms by selecting most used devices, applications, and mode of entertainment.



Above image shows how worldwide consumption of smartphones is. We will start our analysis by finding out the most valued devices.

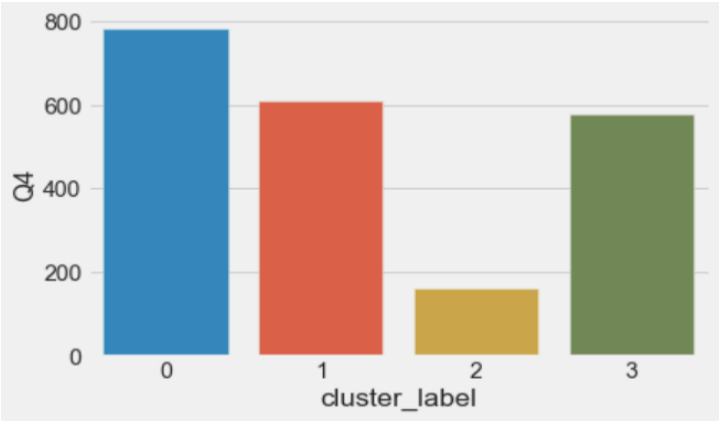
Data Modeling: Clustering
We clustered the consumers based on survey questions using K-Mode clustering. We performed an elbow method to find out the number of clusters.



We choose 4 clusters as the curve has an elbow at 4.

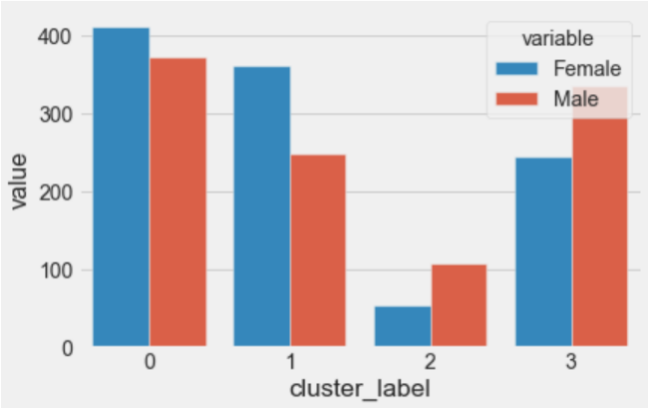
Exploring 4 clusters

Number of People in each cluster



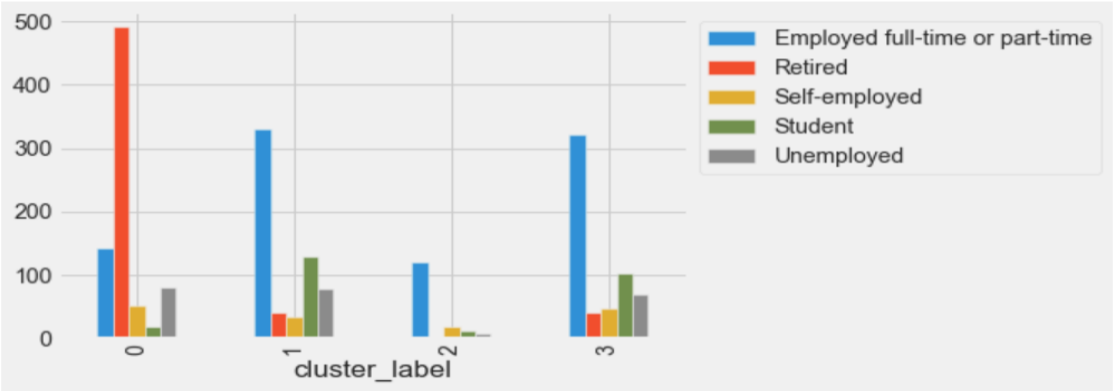
Majority of people cluster 0, 1 and 3. We will focus our analysis using these 3 clusters.

Male and Female in each cluster



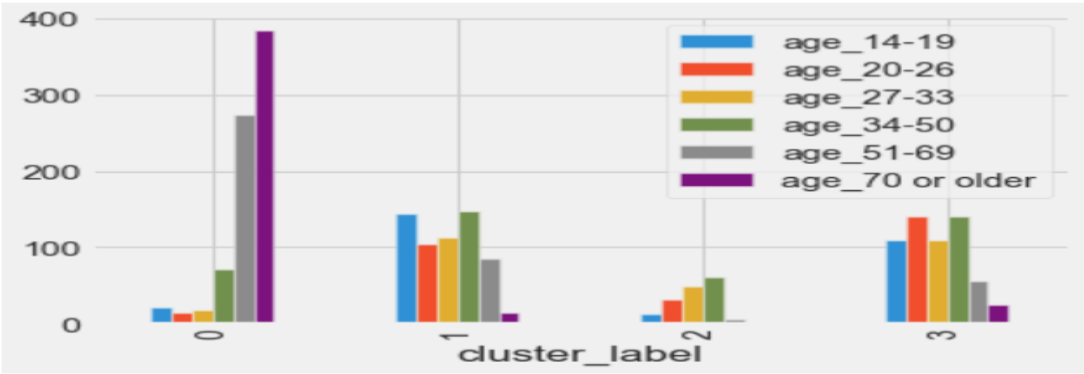
Females are dominant in cluster 0 and 1 and male are dominant in cluster 2 and 3

Employment status across Clusters



Cluster 0 has mainly retired people, Cluster1 has employed full-time or part time and students
Cluster 3 has employed full-time or part time and students

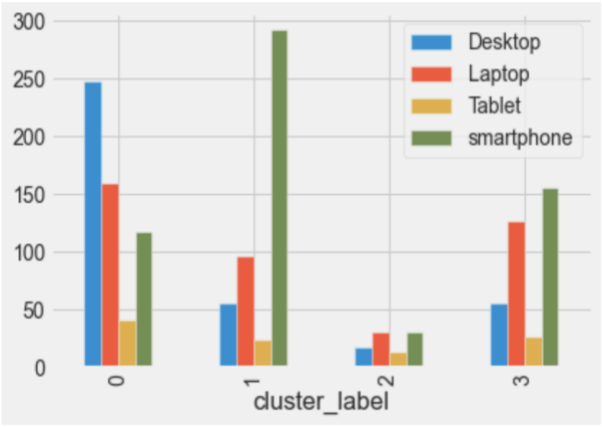
Age-Groups vs clusters

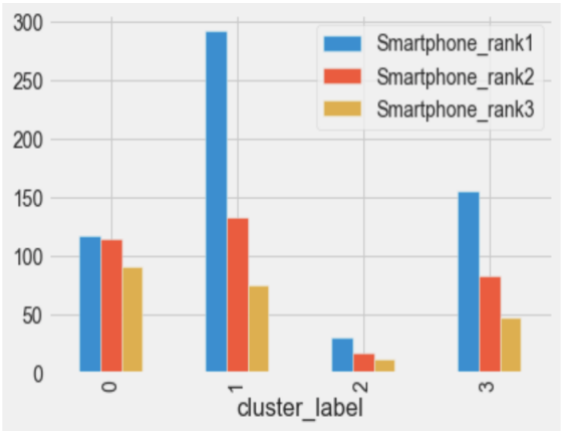


Summary on Clusters

Cluster 0: Age group 70 older, retired,
Cluster 1: Age group 14-19, 34-60, employed and students
Cluster 2: Age group 27-50, but number of people are less in this cluster
Cluster 3: Age group 20-50, employed and students

Most valued devices

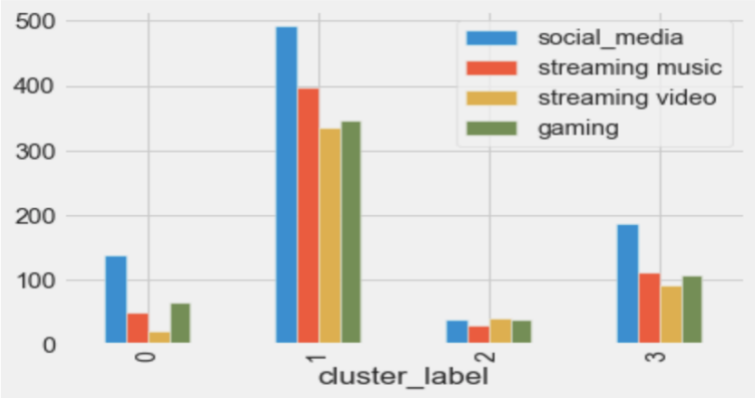




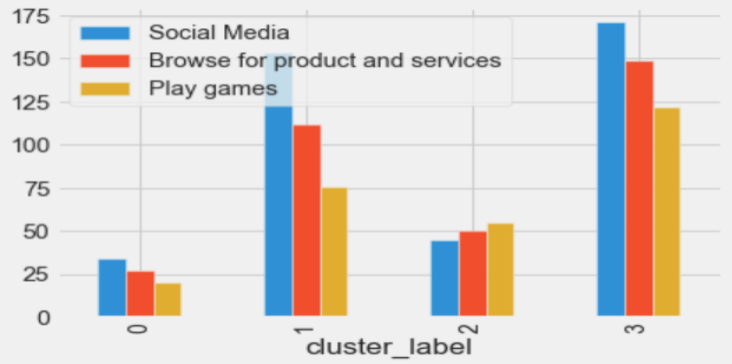
In cluster 0 laptop is the most valued device. Retired old age people use laptops preferable because of screen size.

In cluster 1, 2, 3 smartphones are the most valued device as it is easy to use while multitasking. For example, traveling to work in the metro and using a phone for Facebook.

Most preferred application on Smartphones

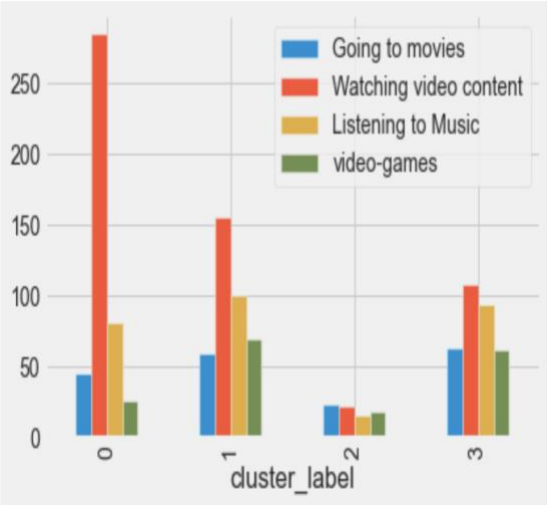


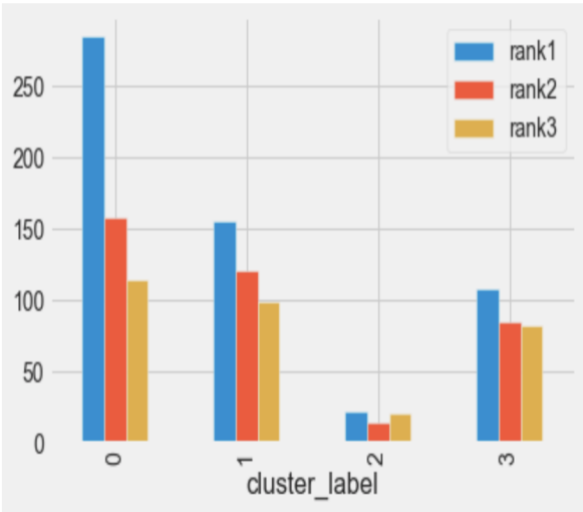
Social Media is the most preferred application by all 4 clusters including people from all age groups and all employment types on smartphones.



In fact, social media is the top ranked activity for people while watching television. Grabbing attention with TV is the most challenging because people simultaneously multitask while watching television. Thus, TV ads are no longer the most effective avenue for getting through an audience. Creating social buzz will matter most.

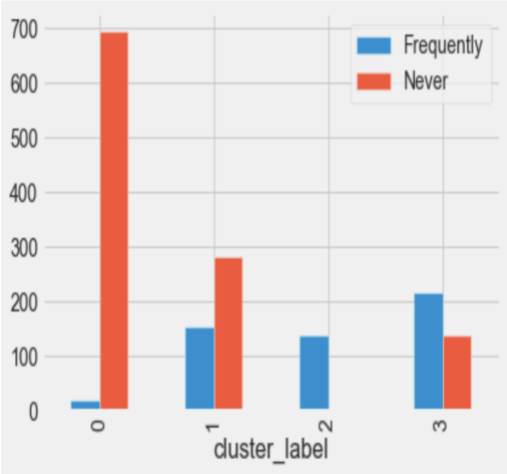
Top ranked activity for entertainment

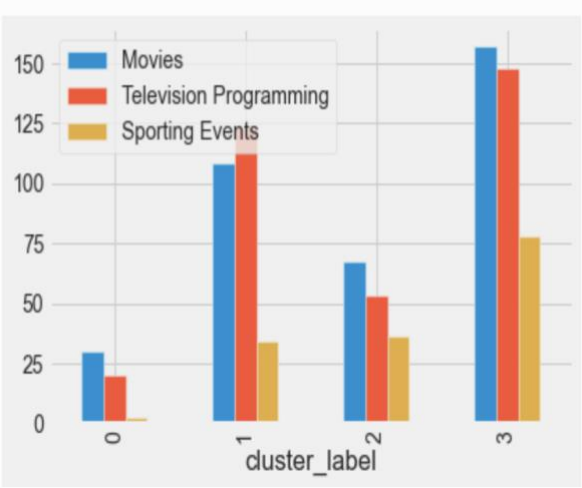




Watching video content on online streaming on any device is top ranked in all 4 clusters. Second image shows in all 4 clusters most people rank video streaming as the most preferable activity for entertainment. Therefore, video streaming platforms are a perfect opportunity for digital marketers to reach out to consumers in a timely, unscripted, and authentic manner.

Using other's Subscription and content watched





Cluster 0 consists of retired old age people who do not use other's subscriptions. Even cluster 1 consisting of people from 34-50 majority follow the same pattern. But here the target audience is people from cluster 2 and 3 who use other people's subscriptions. Referring to the second plot, we can say that people mainly use subscriptions for watching movies online.
For target audiences in cluster 2 and 3, advertisement videos can be creatively embedded in Netflix movie streaming.

Recommendation

1. **Growth of Social Media-** Our analysis shows that all age-groups are adopting social media as the main medium of media consumption.
2. **Live video and video streaming-** Watching video online and movies are top ranked activities for entertainment in all age-groups and employment types.

- Platforms such as Facebook and Instagram can be dug with live video and used as a perfect opportunity by the digital marketing team to reach out to all types of consumers across different demographics.
- As smartphones are the top ranked devices, videos can be creatively repurposed into many things, including embedding videos in emails as part of your email marketing strategy.
- We recommend digital marketing teams to target video streaming platforms like Netflix, Amazon prime to reach out to people shifting from traditional ad platforms like cable and print media to online video streaming.

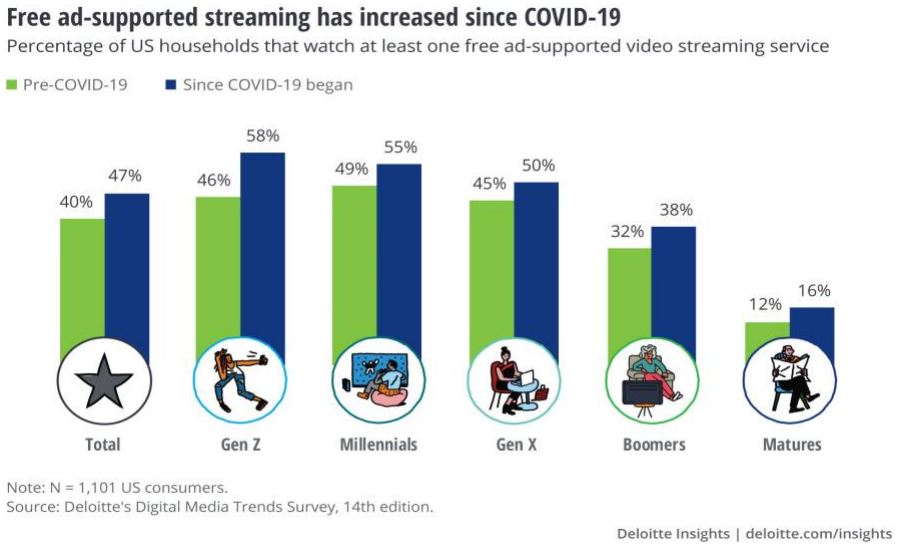
Business Case II

Summary

Subscription to video streaming services have grown in popularity in recent years. It gives brands predictable, recurring revenue streams. However, many video streaming platforms are now partnering up with advertisers to generate more revenue as the ad-based video subscription model is gaining traction.

These streaming platforms have a vast amount of data that can be used to deliver targeted advertising based on demographic and geographic user data.

Based on the 14th edition of Deloitte's Digital Media Trends Survey, 47 percent of American consumers are watching free ad-supported video services such as Pluto TV, Tubi, and the Roku Channel, up from 40 percent before the start of the COVID-19 pandemic. That is 18 percent growth in a matter of weeks. As budgets tighten, consumers could rely more on free video streaming. Thirty-nine percent of consumers—including 43 percent of free ad-supported viewers—say their household has lost income since the pandemic hit the United States.

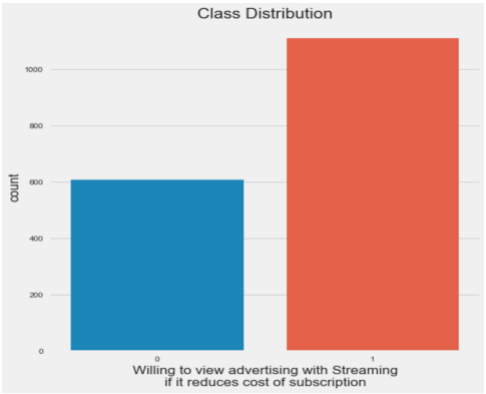


Corporations can leverage the customer survey data to model customers that are likely to respond well to advertisements on streaming platforms. Therefore, in this part, we have developed machine learning classifiers that will predict if a customer is willing to use ad-supported video streaming platforms for a reduced or no cost.

Data Preprocessing

In this task we have again considered the latest survey data that was available to us which consisted of user preferences in 2011. We used to one-hot-encode to convert categorical into numerical features so that are made suitable for machine learning models. The target variable consists of 5 possible responses that have been binarized to improve interpretability and to steer away from a multi class problem.

Orginal	Mapped
Disagree strongly	0
Disagree somewhat	0
Agree somewhat	1
Agree strongly	1



As our input features, we used multiple survey questions and used stepwise elimination until we obtained the desired model performance. We removed the data points of users that stated that they "do not have a basis to answer the question".
Other features that we dropped were:

- Question 39 - related to the output
- Columns that took numerical inputs such as percentages
- Age - Has already been incorporated as age range
- Final Weights
- Record Numbers

The target variable used was chosen based on the business requirements and consisted of the following:

Q39r4 - I would be willing to view advertising with my streaming video programming if it significantly reduced the cost of the subscription.

Model Comparison

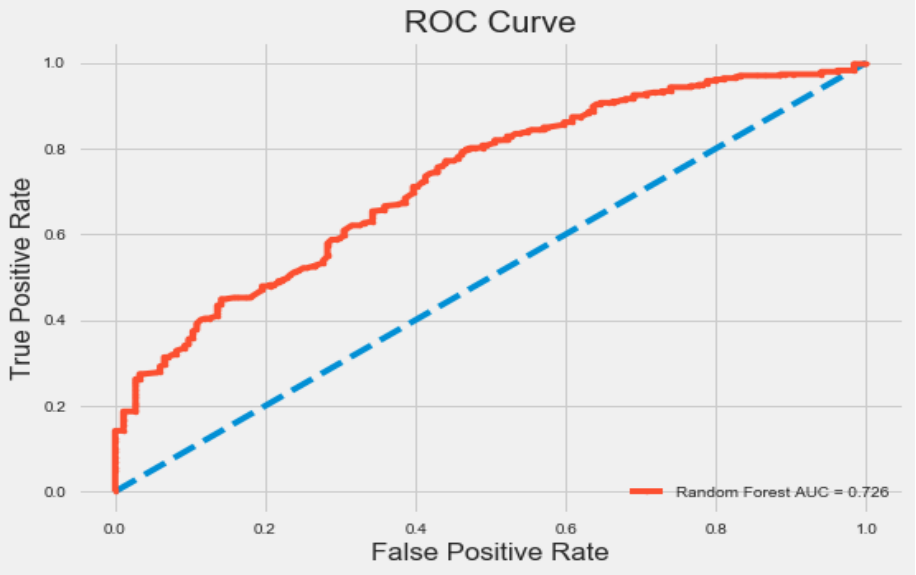
The data was randomly split into training and testing sets using a 70-30 split. To classify the users, we have used a tree-based and as well as a linear classifier (Logistic Regression):

1. **Random Forest Classifier**
We used a forest consisting of 1000 decision trees and default hyper-parameters to classify the users. Please refer to the attached Python Notebook to understand the model optimization process used.

Test Accuracy of the model: 69.9 %
OOB accuracy: 69.3%
R-squared: 70%
AUC ROC Score: 72.6%

Predicted / Actual	0	1	All
0	57	127	184
1	28	304	332
All	85	431	516

Test Set Confusion Matrix

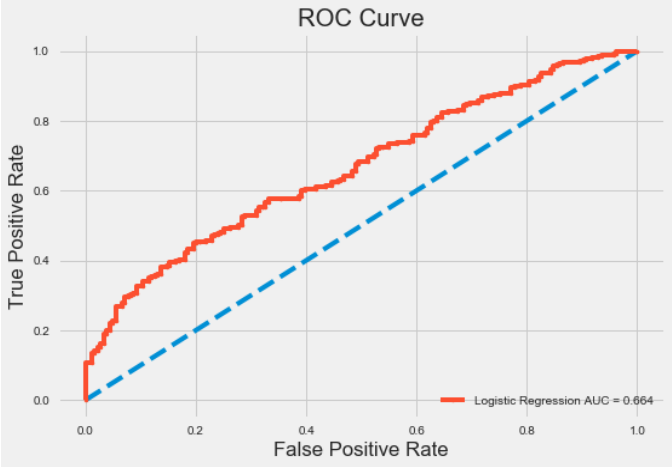


2. Logistic Regression

We achieved a similar performance using logistic regression (LR). However, this model assumes a linear relationship between variables and can't learn nonlinear decision boundaries. Logistic Regression has high bias as compared to RF which has low bias and is flexible enough to learn highly nonlinear decision boundaries. The variance in RF is also reduced due to bootstrapping and voting. In LR, features need to be scaled and normalized unlike RF which is unaffected by it.

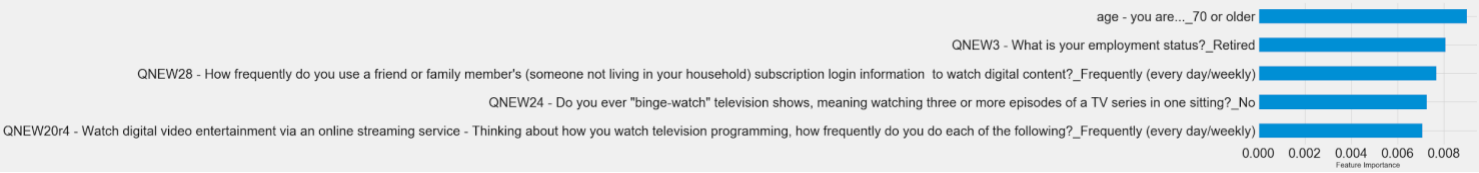
Test Accuracy of the model: 62.6 %
R-squared: 62.6%
AUC ROC Score: 66.4%

Predicted / Actual	0	1	All
0	75	109	184
1	84	248	332
All	159	357	516



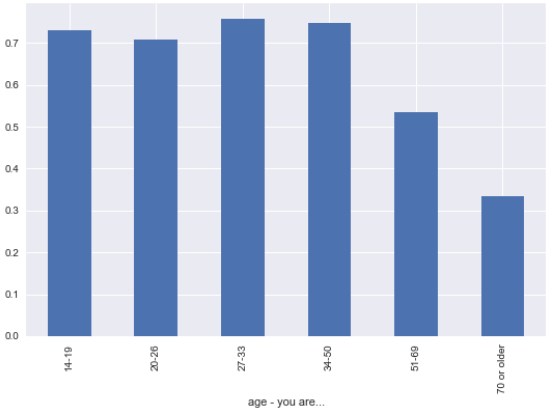
Inference and Recommendations

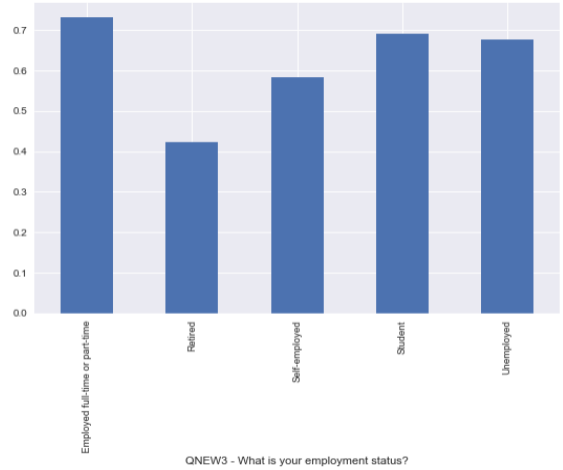
Based on the Random Forest model, we derived the following features that were most important while prediction the output variable. Looking at these features we are able to determine the questions that provide the maximum information gain in predicting the output. We can then use partial dependence plots to isolate the effect of individual questions.



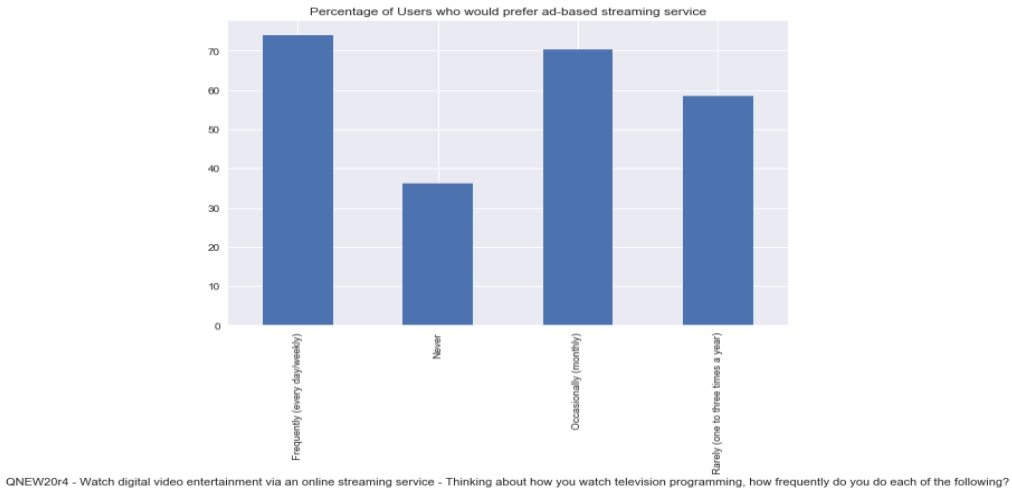
In cases when there are a very large number of features, random forest is able to identify the features most important for making predictions about the output. For example, based on the above results, we can observe the following:

- Users that are older and retired have a lower preference for advertising.





- Users that tend to use a friend's or family member's login information to view online content are also users that would prefer an ad-based streaming service at a lower cost.



Other interesting features to look at are whether a user likes to binge-watch and whether they watch tv-programmed frequently via streaming services.

We believe that our model and the underlying data can be used by client for the following:

1. Targeted Advertising
   Online advertising can get very costly if proper targeting is not used. Choosing the right user demographics is a critical factor while running ad campaigns. Our model can be used to inform a targeted marketing strategy that can be very useful to grow the user base while keeping the cost per customer acquisition in check.
2. Pricing and Marketing Strategy
   Based on preferences of users on ad-based streaming platforms, our client can create strategies such as bundling and price discrimination such that they can maximize their ROI. For example, healthcare services can be bundled with subscription to streaming services to market to users.

Future Steps and Recommendations

Moving forward we would like to gather a more consistent dataset which is more specific to the digital content and that would help us in better understanding the target audience segmentation. Also, we would like to have a dataset available at a more granular level along with other demographics with equal set of frequencies. That would help us remove any biases and get more specific results. I believe more dataset would require thorough research and similar granular level for the interoperability.

Project Challenges

There are tons of challenges when it comes to adoption of new media consumption platforms by the digital marketing community. Tackling issues like tech support, data privacy, regulatory compliance and security will continue to be challenging for the marketing industry.
But as technology improves and more media consumers become part of the digital movement, digital marketers should have an easier time in this space of technology moving forward.