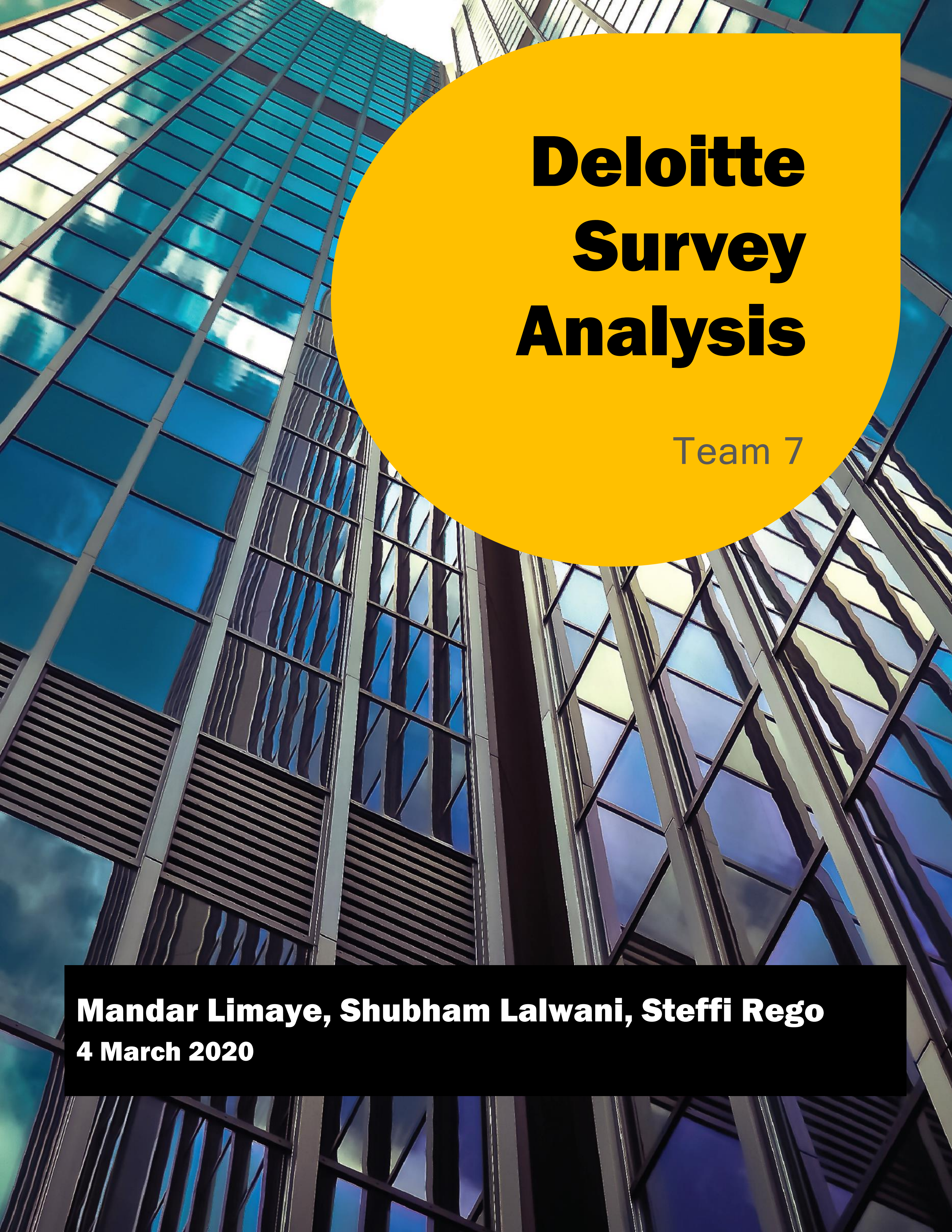


Deloitte Survey Analysis

Team 7

Mandar Limaye, Shubham Lalwani, Steffi Rego

4 March 2020

Contents

1. Executive Summary

The project aims to decipher the hidden insights inside a survey conducted by Deloitte. Deloitte, a multinational professional services network firm, conducted a survey across the United States of America in the year 2009,2010 and 2011. The survey was focused on media consumption habits and was utilized for Deloitte's Digital Democracy report.

We, as Deloitte Consultants, were able to derive critical business insights from the analysis of survey data. The analysis helped generate larger revenue for our clients by bundling products and services. Additionally, predictive analytics helped us understand consumer behavior and willingness to pay for an Ad-supported video streaming platform. These inferences will help clients make better utilization of their marketing budgets.

2. Objectives

The objective of the project is to find useful insights for our clients using supervised and unsupervised machine learning models. Results related to a product owned and tolerance towards advertisements will help us understand consumer behavior. These results will help our clients make better decisions for targeted advertisement and product placements on their online and offline platforms.

3. Beneficiaries

The biggest beneficiaries of our analysis will be the Smartwatch and Fitness Band Manufacturers. The bundling recommendations will help subscription service providers find stronger business collaborations to boost their customer base. The assessments for low-cost Ad-based streaming services will help advertising agencies such as Media.net, Epsilon, Wunderman Thompson to get better outcomes for their clients with the help of targeted advertisement recommendations.

4. Dataset

The dataset used was made available as part of the coursework for 95-851: Data Science for Product Management.

Data Characteristics:

- Years: 2009, 2010, 2011
- Rows (Y-o-Y): 2076, 2205, 2131 (indicates individual survey responses)
- Columns (Y-o-Y): 190,196, 197 (indicates questionnaires)

5. Pre-processing Data

5.1.1. Initial Shape

The data was loaded into pandas dataframe after importing relevant libraries in the Jupyter notebook environment. The data from 3 different years were merged.

5.1.2. Curtailment of excessive columns

The dataframes relating to different years had an inconsistent number of features. Hence, to maintain consistency certain columns from the dataframe were dropped.

5.1.3. Imputation of null values

The survey results had multiple columns having null values. According to best practices, columns having more null values than a given threshold are dropped from the dataframe. Although, since the survey had dependent questions, null values were not an indication of ill responses. Hence, to preserve the data points null values were imputed with -9999.

5.1.4. Handling Outliers

In statistics, an outlier is a data point that differs significantly from other observations. An outlier may be due to variability in the measurement or it may indicate the experimental error; the latter are sometimes excluded from the data set. An outlier can cause serious problems in statistical analyses. In our case, data points having age of respondents greater than 100 and less than 10 were considered as outliers and were removed from the dataframe. Respondents who did not disclose their incomes were also removed from the dataframe.

6. Exploratory Data Analysis

The initial exploratory analysis gave us insights regarding the following:

6.1.1. Gender Balance

We could see that there is an equal distribution of male and female respondents.

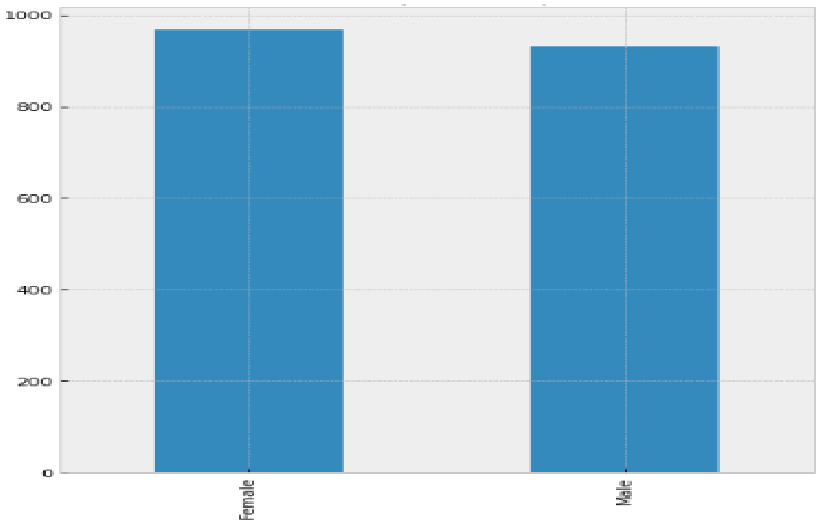


Figure: Frequency distribution of Male Vs Female Respondents

6.1.2. Employment-related information

We found that the majority of respondents had a Job (Full-time or Part-time) or were retired. Additionally, students were a sizeable portion of the total response population.

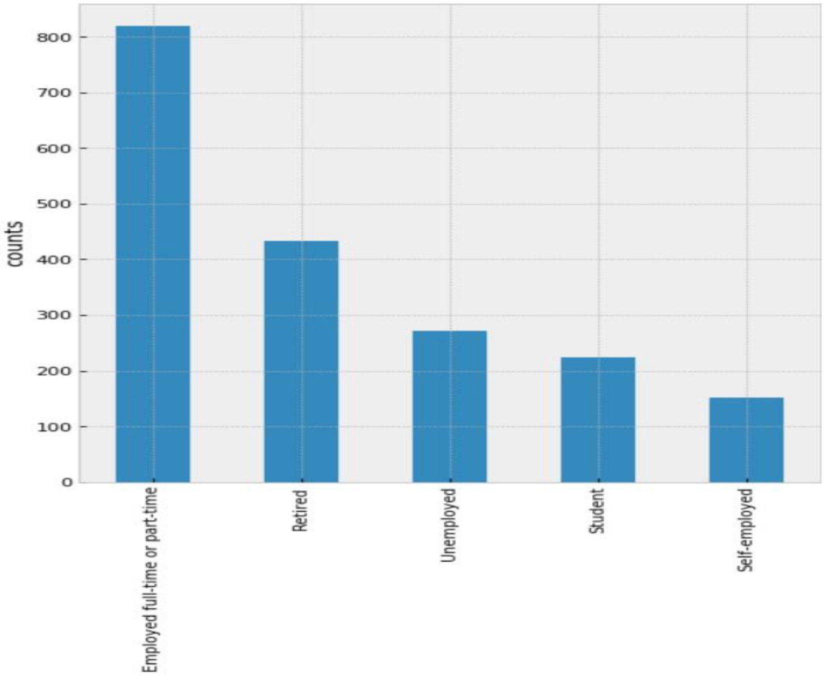


Figure: Distribution of employment-related information

6.1.3. Geographical distribution of respondents

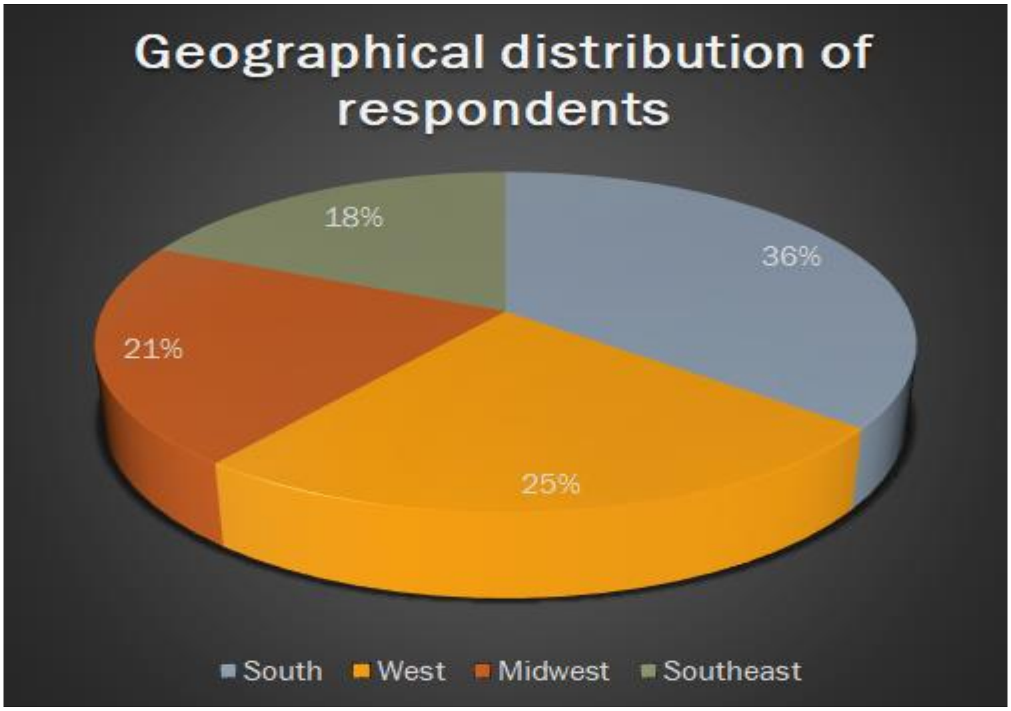


Figure: % distribution of geographical states for respondents

As we could see from the above figure, Southern and Western states were the majority states of origin for people who took the survey.

6.1.4. Demographic distribution of respondents

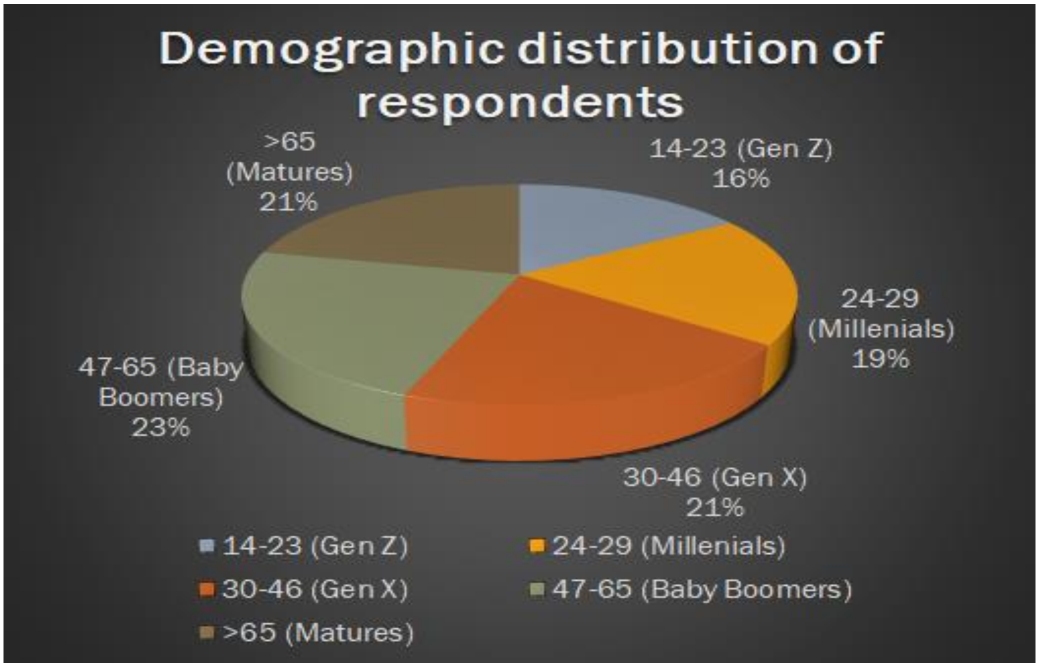


Figure: % distribution of ages across the respondents

From the above figure, there is a good amount of distribution in the age categories of respondents.

7. Business Case 1

7.1.1. Summary

As part of the retail industry, bundling of products and services together is a common practice in order to enhance sales of any business. Depending on the consumer trends of the market at the time, products like laptops, smartphones, Television are bundled with various services like Home Internet , Mobile Data plan , Pay TV in order to target a wider audience who to tend to buy at least one of these two services being provided by a business. Bundling not only helps boost sales, but it also improves customer confidence as shopper feel invested in the decision and increases the likelihood of a sale, since the customer is likely to return to a company that understands his or her needs.

7.1.2. Data preparation

In order to prepare data for this task, firstly, the columns that corresponded to questions related to consumer features, was converted into consumer feature variables like Age , Gender , Employment Status . Then, the respondent answers in the survey that corresponded to Yes / No were converted to Binary 1/0 values.

For this task, we particularly concentrated on 2 questions of the survey:

Q 8 – Which of the following media/ home entertainment equipment does your household own?
Q 26 – Which of the following subscriptions does your household purchase?
Q 10 - Of those products you indicated you do not currently own, which of the following do you plan to purchase in the next 12 months. [only serve up those products that consumers indicated they do not own in Q8

And created a dataframe consisting of consumer features and columns corresponding to answers related to these 3 questions.

7.1.3. Data Analysis

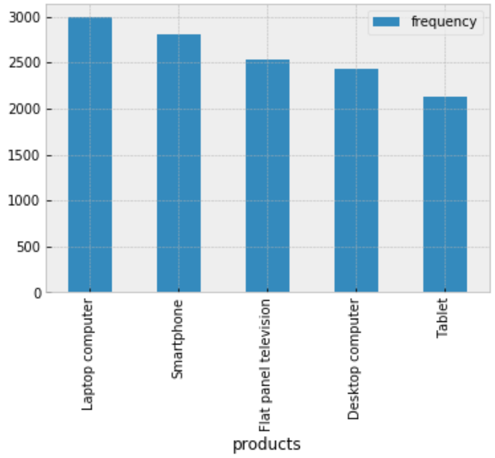
In order to understand which products and subscriptions should be bundled together, we performed the following steps:

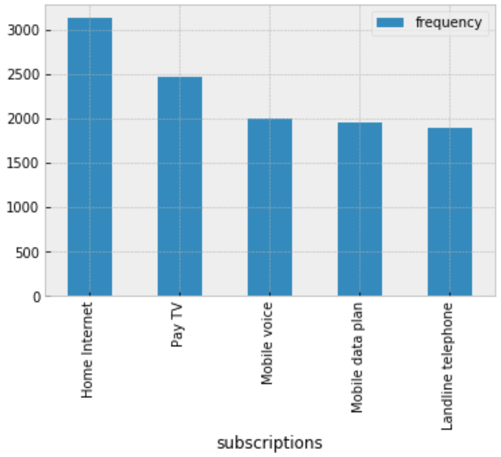
Step 1: Created a dataframe consisting only of Q8 related columns and determined all the products that each consumer currently owns by traversing through each row that corresponds to all the survey answer for that user.
Similarly created dataframes consisting of Q26 to determine all the subscriptions that the user currently purchases and dataframe consisting of Q10 to determine all the products that a user will buy in the future.

Step 2: Combine / merge the three dataframe such that the final dataframe consists of columns that corresponds to the products, subscriptions and future products for each Consumer.

Now to determine the top 5 products and top 5 subscriptions that the users currently own , we determine the number of users that currently own each of the products by calculating the sum of the answers to each question (each column of the dataframe) , and then capture the top 5 products that have the maximum number of users .

Similarly, to compute the top 5 subscriptions , we calculate the sum of answers corresponding to the questions related to subscriptions purchased and then capture the top 5 subscriptions that have the maximum number of users .





7.1.4. Top 5

PRODUCTS	SUBSCRIPTIONS
1. Laptop Computer	1. Home Internet
2. Smartphone	2. Pay TV
3. Flat panel television	3. Mobile voice
4. Desktop computer	4. Mobile data plan
5. Tablet	5. Landline telephone

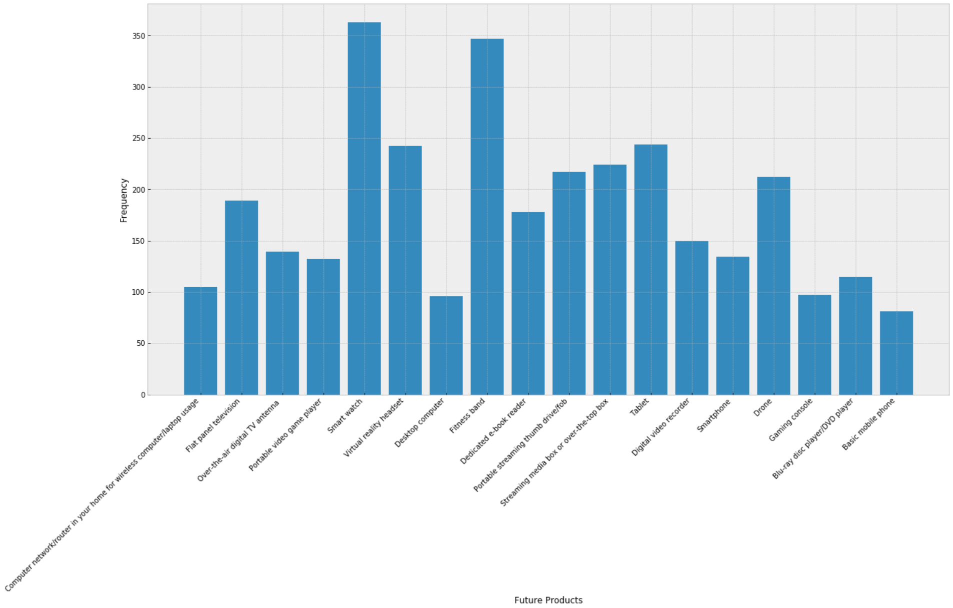
Step 3: Determining most likely future products that a user will buy, given he/she currently owns a product and subscriptions

In order to do this, we first eliminated all rows that had null values in either products, subscriptions or future products. We then filtered the dataframe by any combination of product and subscription from the top 5.

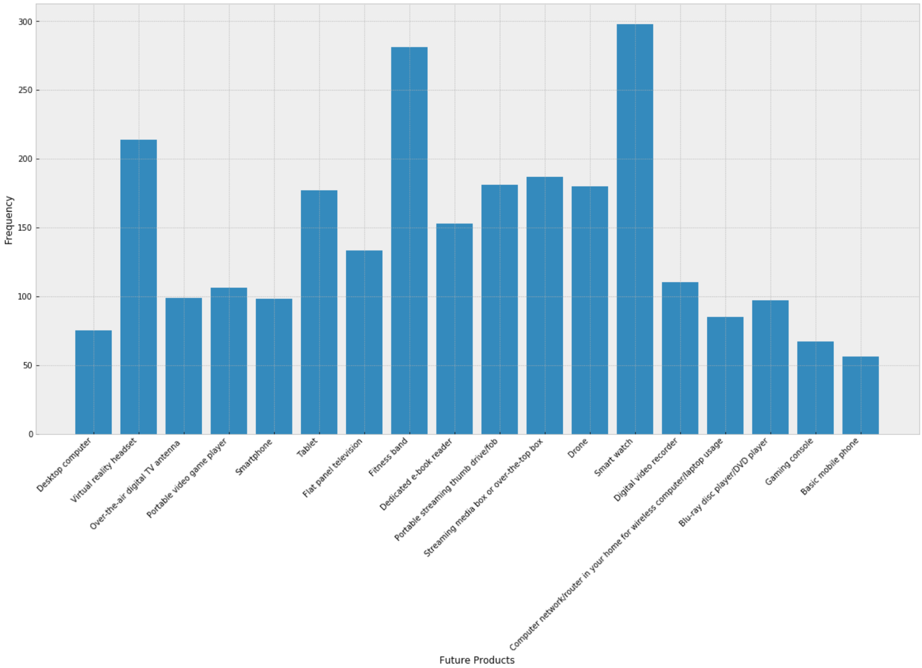
After, we determined all the future products that each of these users will buy. These future products were then ranked in order of their frequency using a dictionary and the most likely future product was determined.

Analyzing various combinations of product and subscriptions:

1. Laptop Computer and Home Internet



2. Laptop Computer + Pay TV



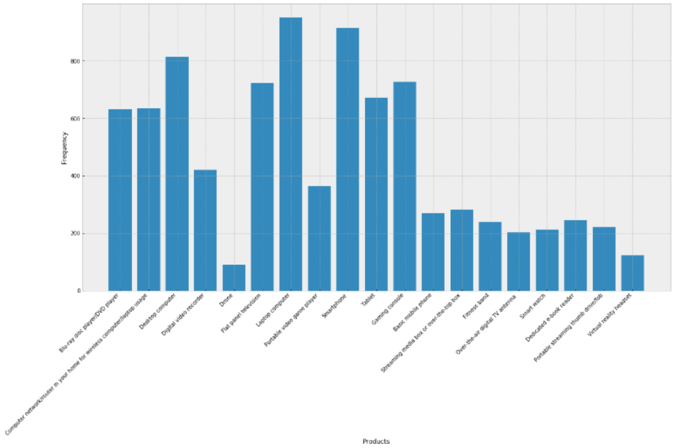
Similarly, we can determine the future products that a user will buy from other combinations of products and subscriptions of the top 5.

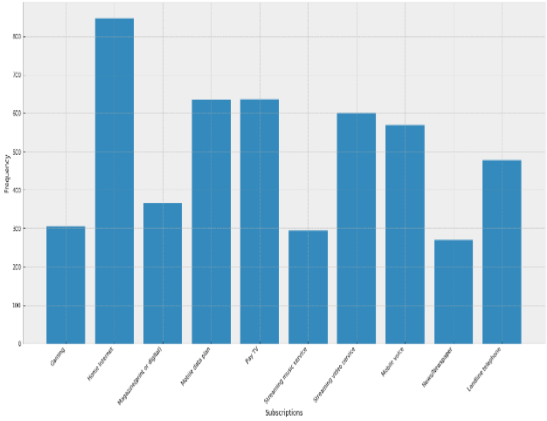
Keenly observing the results of these graphs, we concluded that **Fitness band** and **Smartwatches** are the most likely product that a user will buy in the future.

7.1.5. Gender Based Analysis

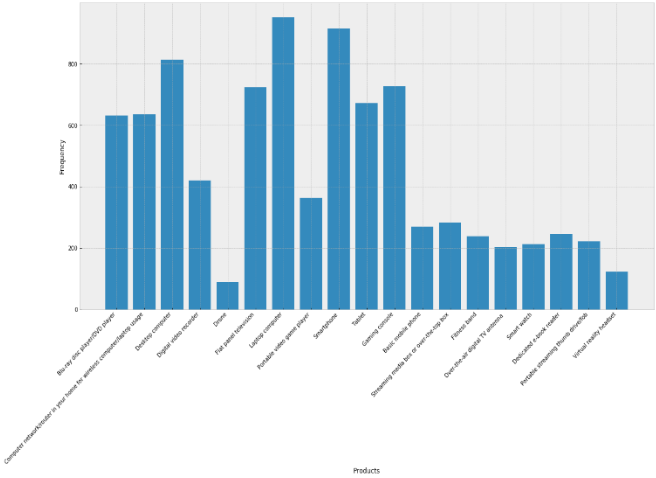
To understand which products and subscriptions can be bundled together in order to target a particular gender of the users, we found out the top 3 products and top 3 subscriptions that the users of the respective genders currently owns. Based on these results we based our recommendations for the respective gender.

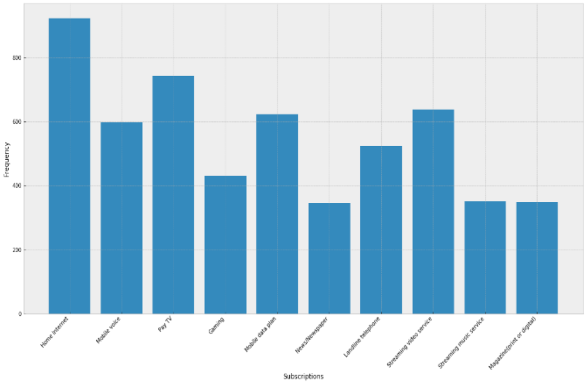
Female:





Male:





Top 3:

Products		Subscriptions	
Male	Female	Male	Female
1.Laptop Computer	1.Laptop Computer	1. Home Internet	1. Home Internet
2.Smartphone	2. Smartphone	2. Pay TV	2. Pay TV
3.Desktop	3. Flat panel TV	3. Mobile data plan	3.Streaming video service

7.1.6. Business Recommendations

1. For Bundling of future products, we would recommend Fitness bands and Smart watch companies to partner with Laptop Computer or Smartphone selling companies or also sell their products with a bundled subscription like Home Internet or Pay TV.
   Bundling of the Fitness bands and Smart watch with any of the top 5 products or subscriptions based on user data will help improve sales and also boost consumer confidence the company.

2. For gender-based bundling:

   To target everyone irrespective of gender, Laptop computer and Smartphones should be bundled with Home Internet or Pay TV

   For only Male Consumers, Desktop Computers should be bundled with Home Internet or Pay TV or Mobile data plan

   For only female Consumers, Flat panel television with Home Internet or Pay TV or streaming video service

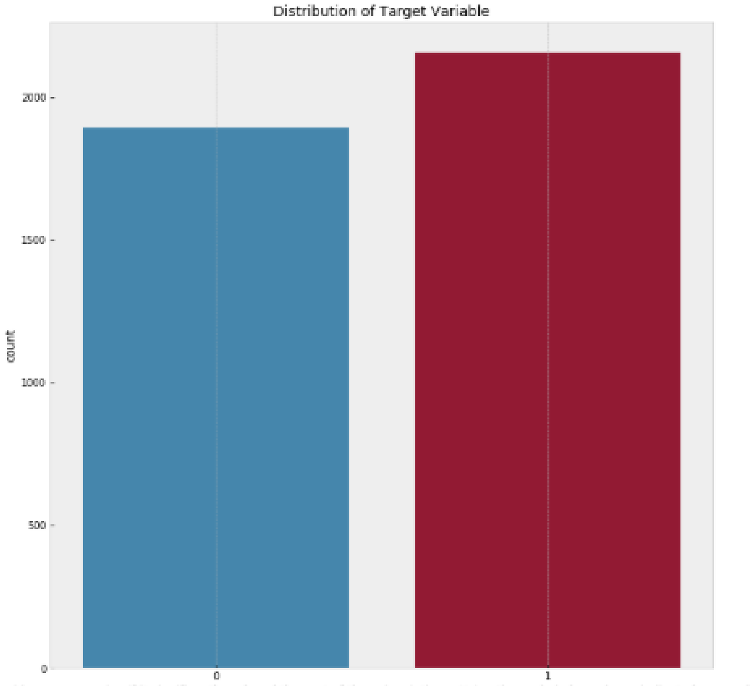
8. Business Case 2

8.1.1. Summary

Traditionally, subscription videos on-demand services have been popular in the industry. However, there is another business model in the industry called ‘Advertising Video on Demand’ that is gaining popularity these days. In this part, we built a classification model to predict the customer's willingness to pay for an ad-supported video streaming platform at a lower cost.

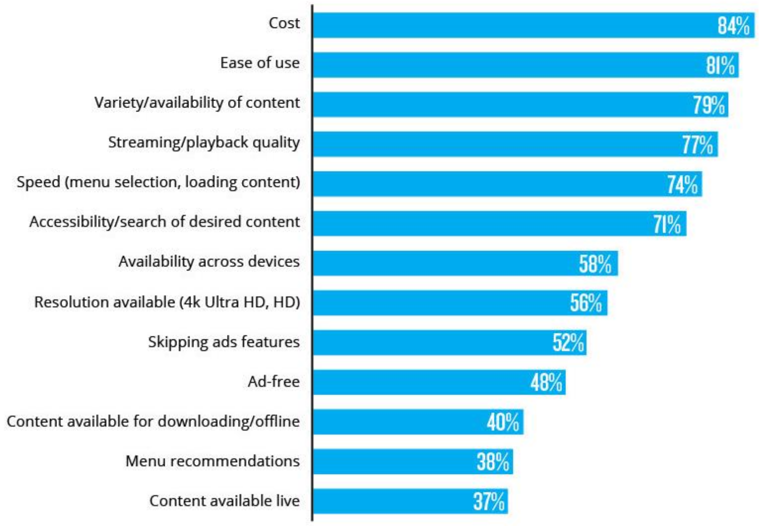
8.1.2. Data Preparation

For this task, we have considered the two data frames (task2_10 and task2_11) that consist of consumer survey data from 2010 and 2011. We also used label encoder to convert the categorical feature data into a numerical data. The survey data with responses ‘No’ and ‘Yes’ was converted into a binary variable 0 and 1 to perform crisp classification. This conversion also ensured that the target variable in the classification would have equal distribution to avoid bias and overfitting. The following graph explains the distribution of the target variable after converting into binary format.



8.1.3. Market Analysis

We also did some research to identify the market significance of our implementation. In the United States, nearly half of the households subscribe to video streaming services. The industry generates a revenue of 2 billion from user subscriptions. The following data from the audience report collected by Nielsen explains major factors considered by users while buying a video subscription service. Cost is the primary factor considered by all the consumers while only 48% take the factor of advertising into consideration.



Source: February 2020 Nielsen Total Audience Report

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. Hence, we wanted to identify which consumer would be willing to move to an ad-based video streaming platform. One very important point that we want to mention is that the data is collected in the year 2010 and 2011. Therefore, the perspective of the analysis is different as during those years the video streaming industry, as well as ad-tech industry, was gradually growing.

8.1.4. Data Modeling: Classification

We used multiple survey questions as the input features for the classification model. The questions comprised of the following attributes:

1. Customer Detail (Age, Region, Employment, Salary, Ethnicity, Number of Children)
2. Video Streaming Habits (Using someone else’s account)
3. Willingness to pay in exchange of zero ads (Multiple columns for each media channel)
4. Willingness to provide more information to get targeted advertising

The following question was used as the target variable which basically comprised of willingness to have an ad-based video streaming subscription.

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

The training data and testing data were divided using a 75-25 split percentage. In order to classify the users, we have used two supervised machine learning models:

1. Random Forest Classifier

We used a forest consisting of 100 decision trees to classify the users.

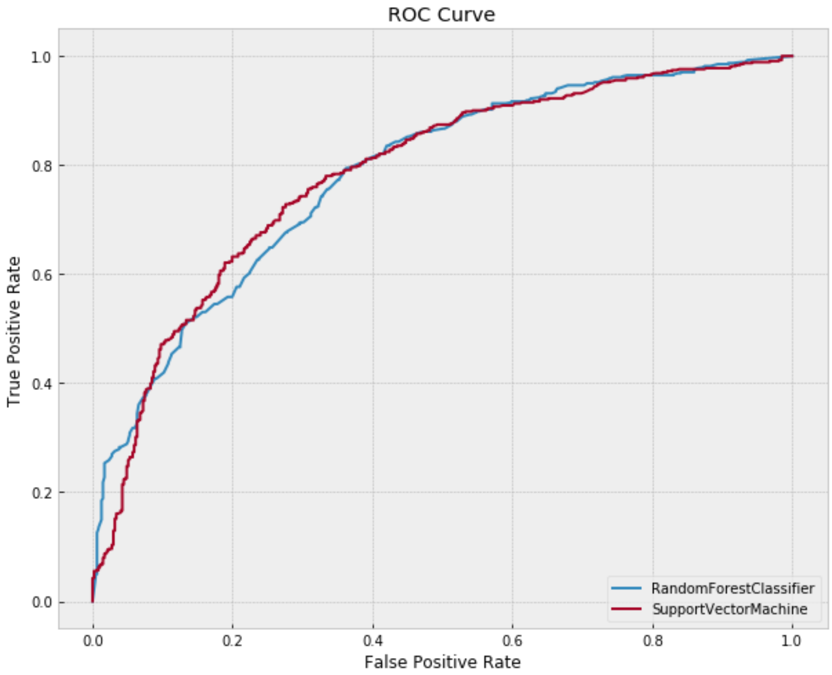
Accuracy: 71.25%.

2. Support Vector Machine

Accuracy of 72.23%

(Please refer to python notebook to find a detailed confusion matrix to understand model performances)

In order to evaluate model performance, we also plotted the ROC curve given below. Using the accuracy and AUC in the ROC curve we determined that the Support Vector Machine model works best for our task.



8.1.5. Business Recommendation

We believe that this data is very useful for our media client for two reasons.

- **User Growth**
  As cost is a critical factor while making the decision, such a pricing strategy can be very useful to grow the user base by launching targeted marketing campaigns.
- **Customer Acquisition**
  Companies entering video-streaming space can utilize this insight to make strategic decisions and introduce ad-based subscriptions at a calculated price point.
- **Subscriber Retention**
  Video-streaming platforms lose a lot of users due to changes in subscription prices. Classifying the user's willingness can help these services to provide targeted packages.

9. Business Case 3

9.1.1. Summary

Over the years, advertisers have identified traditionally successful channels to market a specific type of product. For example, electronics have been successfully marketed through online channels. However, with the growth of programmatic ads and machine learning, it is possible to target individual consumers using their preferred medium. This increases the operational cost as requirement for each user might differ. In this scenario, segmenting the customers based on their willingness to avoid advertisements on different channels would give us a reverse insight on the channels that could be utilized to target that customer segment.

9.1.2. Data Preparation

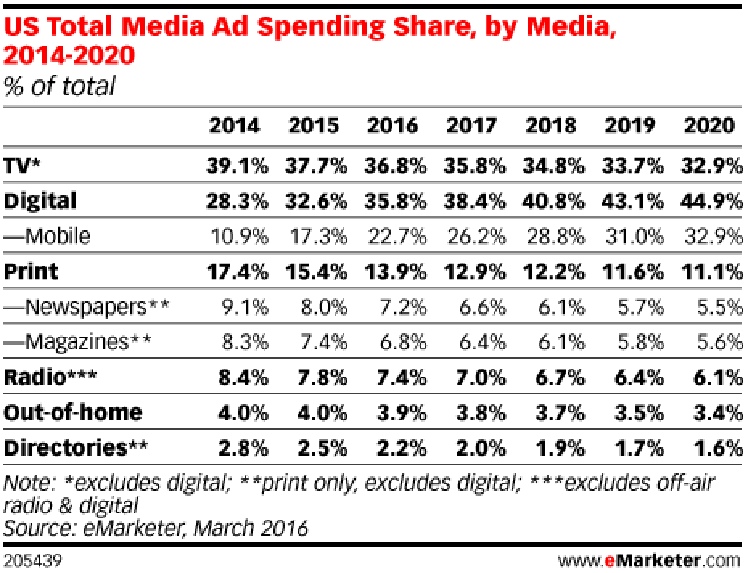
For this task, we have combined two data frames (task2_10 and task2_11) that consist of consumer survey data from 2010 and 2011. We also used label encoder to convert the categorical feature data into a numerical data. The survey data with responses 'No' and 'Yes' was converted into a binary variable 0 and 1 to form distinct clusters.

9.1.3. Market Analysis

In digital ad-tech industry customer are segmented using following different approaches:

- Demographic Segmentation
- Geographic Segmentation
- Behavioral Segmentation
- Device Segmentation
- Channel Segmentation

The following data shows the total US total media ad spending share by different media.



US Total Media Ad Spending Share, by Media, 2014-2020
% of total

	2014	2015	2016	2017	2018	2019	2020
TV*	**39.1%**	**37.7%**	**36.8%**	**35.8%**	**34.8%**	**33.7%**	**32.9%**
Digital	**28.3%**	**32.6%**	**35.8%**	**38.4%**	**40.8%**	**43.1%**	**44.9%**
—Mobile	10.9%	17.3%	22.7%	26.2%	28.8%	31.0%	32.9%
Print	**17.4%**	**15.4%**	**13.9%**	**12.9%**	**12.2%**	**11.6%**	**11.1%**
—Newspapers**	9.1%	8.0%	7.2%	6.6%	6.1%	5.7%	5.5%
—Magazines**	8.3%	7.4%	6.8%	6.4%	6.1%	5.8%	5.6%
Radio***	**8.4%**	**7.8%**	**7.4%**	**7.0%**	**6.7%**	**6.4%**	**6.1%**
Out-of-home	**4.0%**	**4.0%**	**3.9%**	**3.8%**	**3.7%**	**3.5%**	**3.4%**
Directories**	**2.8%**	**2.5%**	**2.2%**	**2.0%**	**1.9%**	**1.7%**	**1.6%**

*Note: *excludes digital; **print only, excludes digital; ***excludes off-air radio & digital*
Source: eMarketer, March 2016

205439 www.eMarketer.com

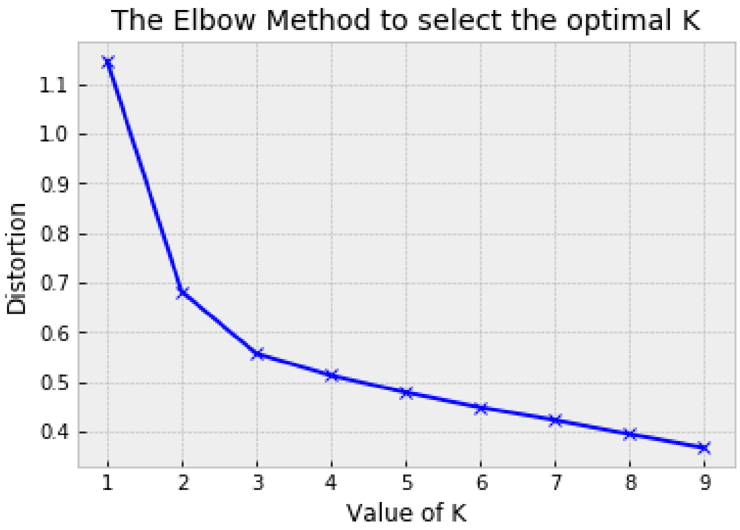
With the growth of digital ad spending, it is important to limit unnecessary advertising spend. The survey gives us data about user demography and location. It also gives us data regarding user behavior through different questions about their willingness to pay a fee in exchange for advertisements. This data would be helpful to target segments of customers and increase the ROI.

9.1.4. Data Modeling: Clustering

We used multiple questions from the survey (Q39r1, Q39NEW1, Q39NEW2, Q39NEW3, Q39NEW4, Q39NEW5). These questions described the consumer willingness to pay for not receiving any advertisements across the following media channels:

- News
- Sports
- Online Games
- Music
- TV Shows
- Movies

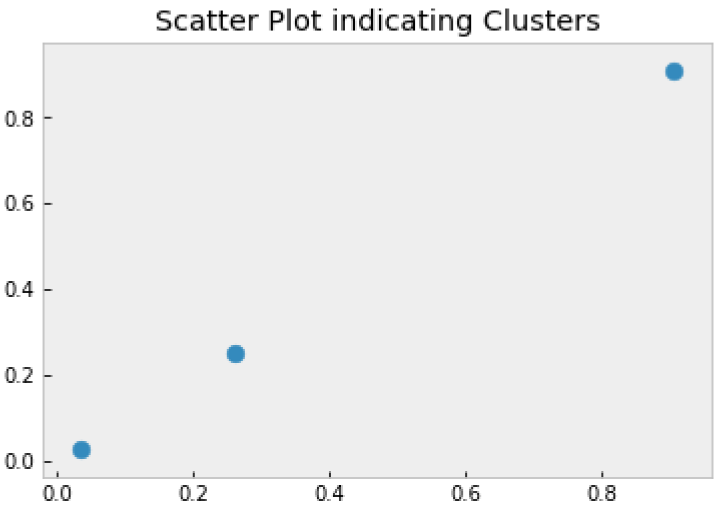
We clustered the consumers using these survey questions. Each output customer segment gave an idea about the gender, age bracket, employment status, and salary bracket of consumers. In order to determine an ideal number of clusters, we used the elbow method. Please see the following snapshot gives a plot of elbow method-



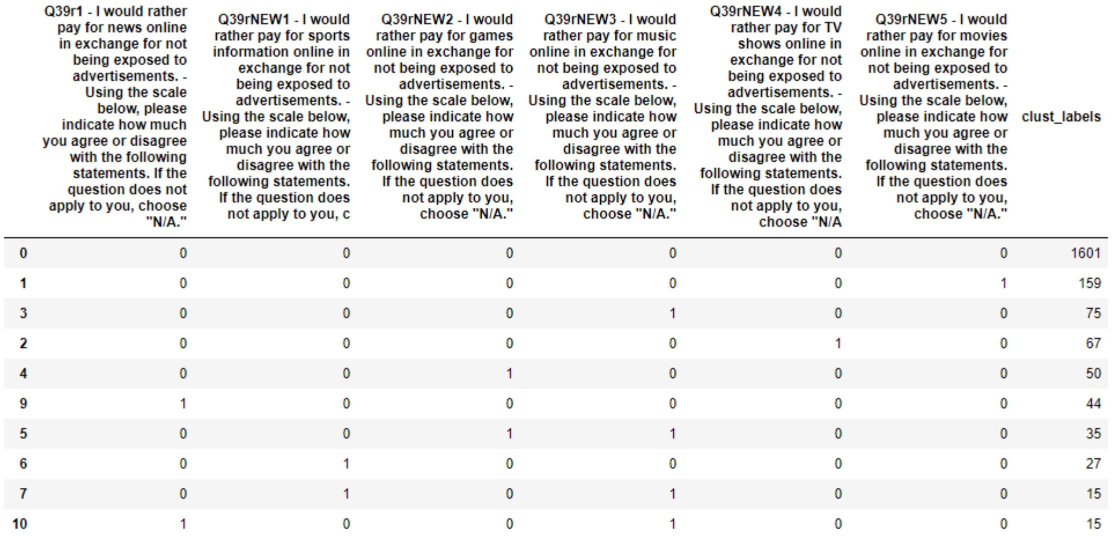
The elbow plot clearly shows an ideal number of clusters is 3. The reason for selecting ideal clusters as 3 is that the elbow plot shows the incremental value of slope of at K = 3 is maximum and the gradient of slope does not increase moving ahead. We performed analysis using two unsupervised machine learning models to cluster the data

1. KModes Clustering
2. KMeans Clustering

The scatter plot on the next page shows the cluster centroid of three distinct clusters.

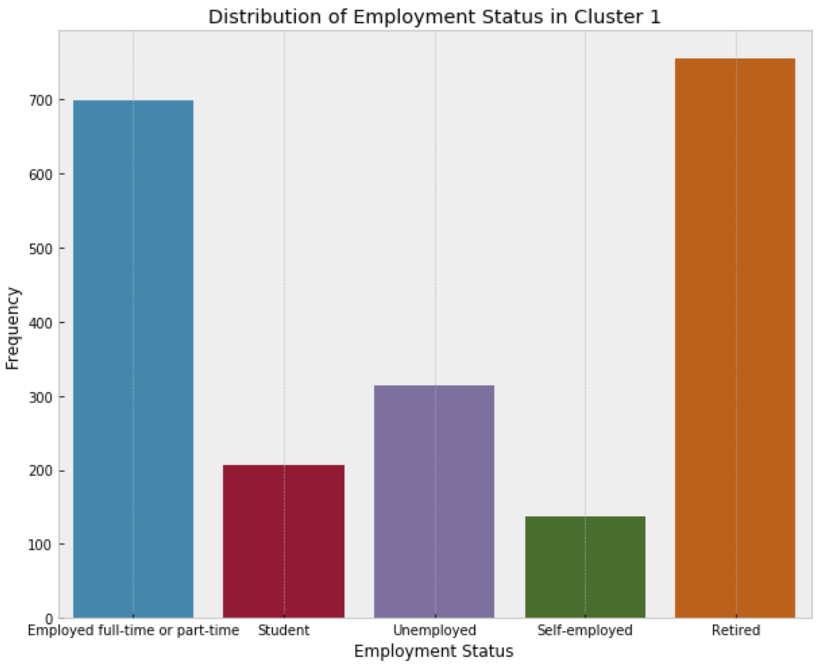


In order to identify the frequency count of groups of unique behavior across different clusters, we grouped these behaviors to identify most occurring behavioral trends. Following is the snapshot of the frequency count of unique behavioral groups in the first cluster.



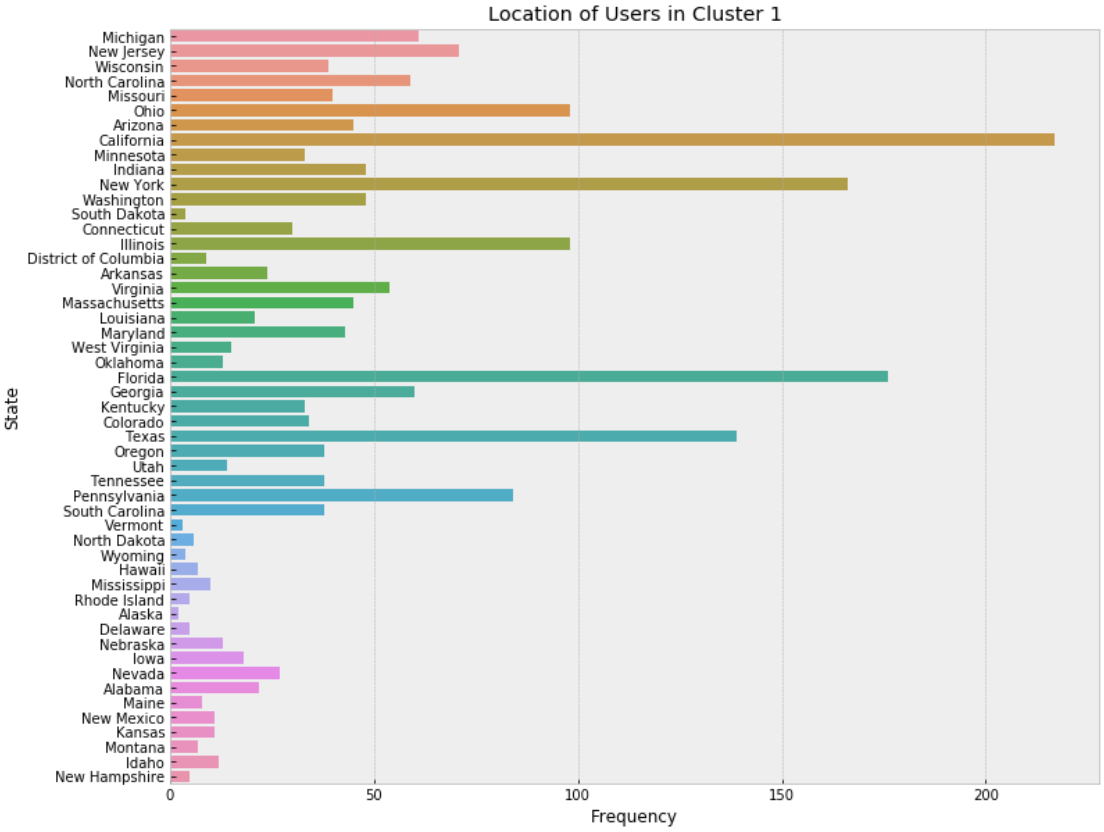
	Q39r1 - I would rather pay for news online in exchange for not being exposed to advertisements. - Using the scale below, please indicate how much you agree or disagree with the following statements. If the question does not apply to you, choose "N/A."	Q39rNEW1 - I would rather pay for sports information online in exchange for not being exposed to advertisements. - Using the scale below, please indicate how much you agree or disagree with the following statements. If the question does not apply to you, c	Q39rNEW2 - I would rather pay for games online in exchange for not being exposed to advertisements. - Using the scale below, please indicate how much you agree or disagree with the following statements. If the question does not apply to you, choose "N/A."	Q39rNEW3 - I would rather pay for music online in exchange for not being exposed to advertisements. - Using the scale below, please indicate how much you agree or disagree with the following statements. If the question does not apply to you, choose "N/A."	Q39rNEW4 - I would rather pay for TV shows online in exchange for not being exposed to advertisements. - Using the scale below, please indicate how much you agree or disagree with the following statements. If the question does not apply to you, choose "N/A	Q39rNEW5 - I would rather pay for movies online in exchange for not being exposed to advertisements. - Using the scale below, please indicate how much you agree or disagree with the following statements. If the question does not apply to you, choose "N/A."	clust_labels
0	0	0	0	0	0	0	1601
1	0	0	0	0	0	1	159
3	0	0	0	1	0	0	75
2	0	0	0	0	1	0	67
4	0	0	1	0	0	0	50
9	1	0	0	0	0	0	44
5	0	0	1	1	0	0	35
6	0	1	0	0	0	0	27
7	0	1	0	1	0	0	15
10	1	0	0	1	0	0	15

The above table shows that the consumers in this cluster did not want to pay fees for any media channel to get rid of ads. This means this set of users can be targeted across all the media channels. The second group wanted to restrict advertisements on movies by paying some fee. Using these insights, we can hyper-target these clusters/segments of consumers on specific media channels that are most occurring. We can also identify different characteristics of users in each cluster. The graph below shows the distribution of employment status in the first cluster.



This distribution compliments our finding as a higher number of retired people ideally do not care about the ad-free experience. The majority consumers in this survey can be targeted across all platforms with few platforms being exception.

The below graph gives location-wise data of the consumers in the first clusters.



9.1.5. Business Recommendations:

The Deloitte survey describes key findings regarding millennials and generation z. On the other hand, such analysis can be used to form different segments and hyper-target these consumers.

The location-based segment data can be used to launch regional campaigns that target specific media channels to grow the business. Other characteristics of consumers in the segment can also be used to perform targeting and retargeting on different platforms.