Composition ID Survey Data Analysis

There are in 310 observations in the data set. The first step that I took was to identify the input variables for the algorithm. I took the responses from Question 39 and copied the responses to a new text csv file(comma separated). I added a ID column to every row, which is unique for every response. Apart from the responses for question 39, I added Gym and Dietary supplement columns.

Next step was to assess data quality, I found out that many values were missing. Generally, the rows with missing values are removed as correct values cannot be imputed. But here data set had very less number of rows, so removing rows would have resulted in very less number of rows. But imputing the rows with zero won't impact clustering, as these rows are likely to be in same cluster. So, I filled the missing values with 0. I assumed that the users didn't fill these survey questions as they were of less importance.

I didn't use K-Means, as for it the initial cluster centers are chosen randomly. So, sometimes it has a tendency to collapse on local minima instead of global minima due to poor initialization most likely because of the initialization of the K number of cluster centers to their respective points. I used Bisecting K-Means, which is a top down approach for clustering. It starts with all the points in a single cluster and divide it with into multiple clusters with every iteration. Bisecting K-Means algo has been shown to result in better cluster assignment for data points, converging to global minima as than that of getting stuck in local minima as K-Means does.

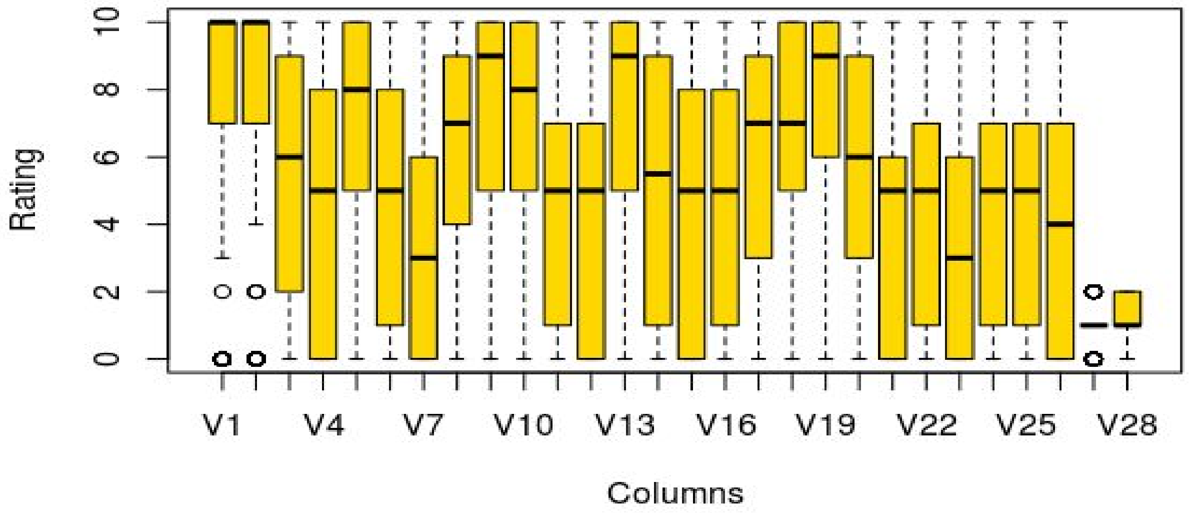
I developed my code in Apache Spark using scala. I used these variables:-

1)Good Value for the Price
2)Frequent DEXA scanning and other testing
3)Convenient, online access to your PHR
4)Comparison to the averages of other people's historical data
5)Data analytics of your historical testing data
6)Virtual reality viewing of your health data
7)Ability to post online about your Composition ID membership activities (goals, events, progress)
8)Being able to forecast your improvement using predictive analytics
9)Cybersecurity and privacy/security of your data
10)HIPAA compliant storage of your data
11)High quality online content (videos, blogs, etc)
12)Email newsletter updates to introduce you to new content and events
13)Discounts and Promotions
14)A one-on-one relationship with a Composition ID nutritional advisor
15)Access to nutrition challenge programs (example: 30-day challenge)
16)Anytime access to a Composition ID nutritional advisors by phone, email, or in-person
17)Ability to use Composition ID services at your regular gym
18)Ability to use Composition ID services at a dedicated Composition ID location
19)Convenient location to me
20)A Composition ID mobile app
21)Events
22)Educational talks about nutrition and health
23)Ability to buy high quality Composition ID- approved health supplements and fitness products
24)Assistance in creating your PHR
25)Assistance in maintaining your PHR
26)Family plan - Combined Composition ID membership for a family group

27)Gym
28)Dietary Supplement.

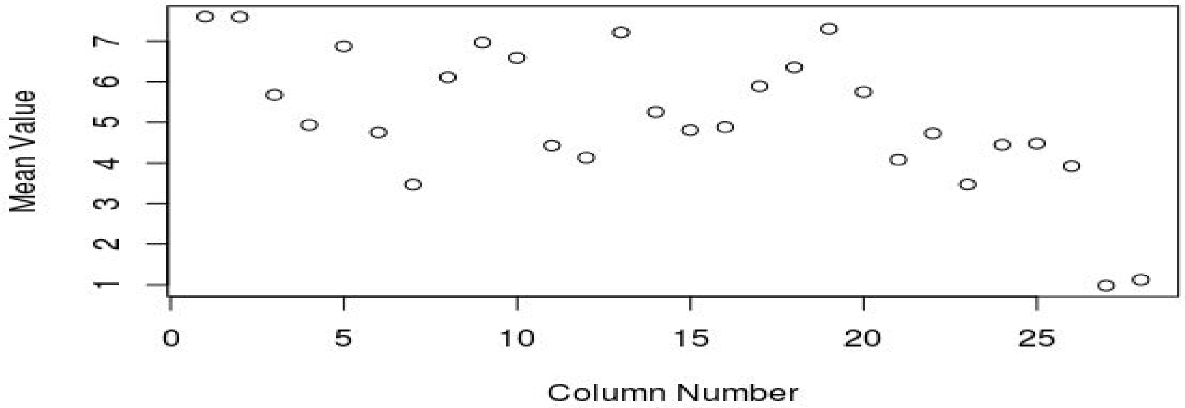
Exploratory Data Analysis:

BoxPlot for variables:-



Here V1 to V28 are variables used for analysis and are listed above in same order. From the above plot, I observed that most of the variables had higher ratings and only some had lower ratings .Columns V27 and V28 are variables which denote whether the participant goes to the Gym and whether they take Dietary Supplement, these have values only in yes or no(1 or 2).

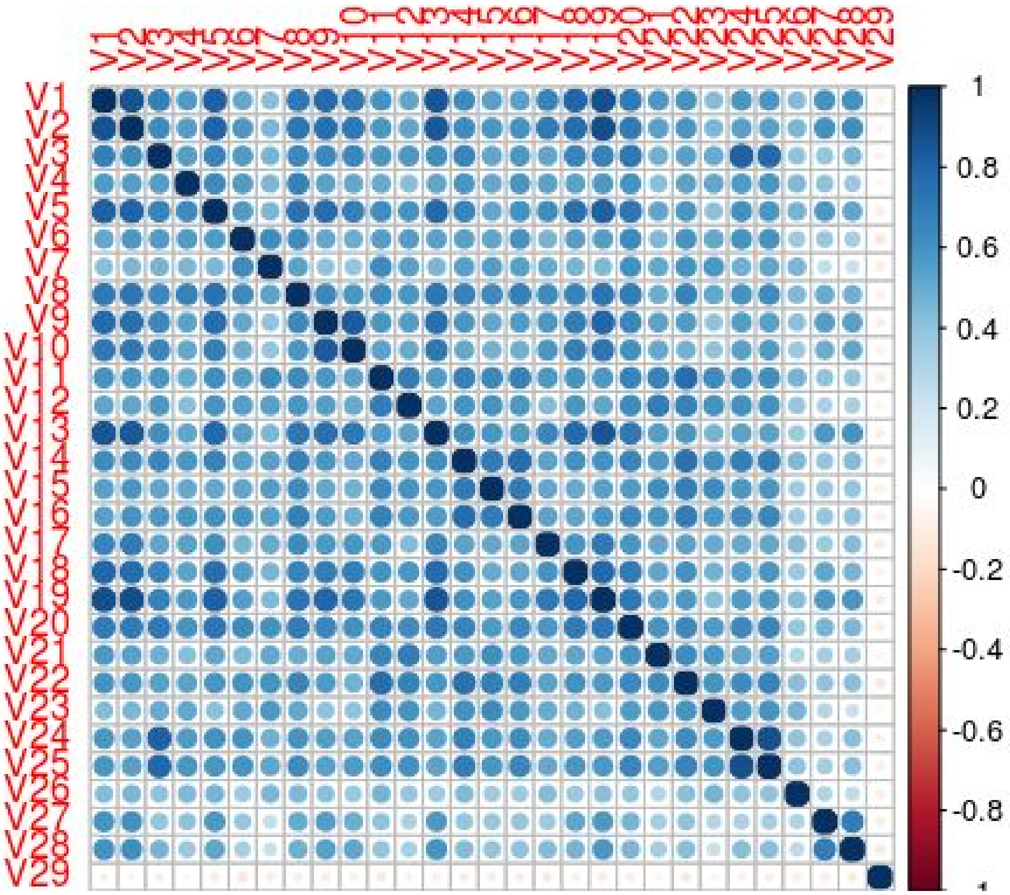
Plot for mean rating for variables:-



Here column number are columns(variables) used for analysis and are listed above in same order. Above graph shows the mean value for 28 columns(variables). Mean value is high for most of the variables.

Correlation plot for variables:-

I made a correlation plot for all the variables, to identify column which are less correlated.(labels are V1-V28 for columns listed above, in same sequence as above.V29 is ID):-



Here V1 to V28 are variables that I used for analysis and are listed above in same order. V29 is the ID column.

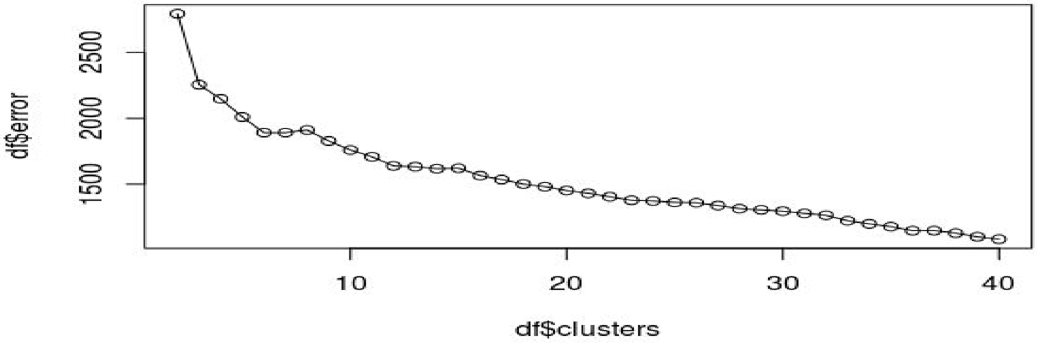
I found out that the following columns were less correlated
Family plan - Combined Composition ID membership for a family group", "Gym", "Dietary supplement", "Events", "Virtual reality viewing of your health data", "Ability to post Composition activities", "Access to nutrition challenge programs (example: 30-day challenge)", "Educational talks about nutrition and health", "Ability to buy high quality Composition ID -approved health supplements and fitness products"

I used PCA to reduce the dimensionality, by combining the columns which are less correlated and then took first three components which accounted for most of the variance.

Selecting appropriate cluster size:

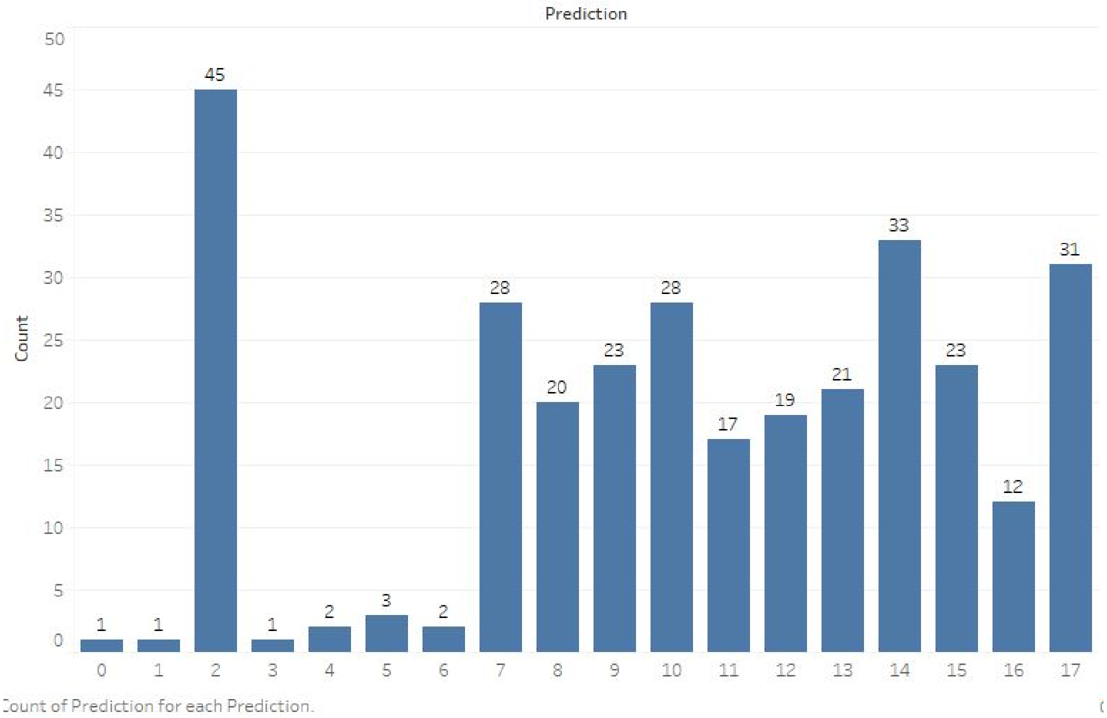
K is cluster size. To select K, I wrote code to use K from 2 to 40 and cluster iteratively for every value of K and record errors.

Plot error vs K(clusters):-



So I selected k=18 for clustering.

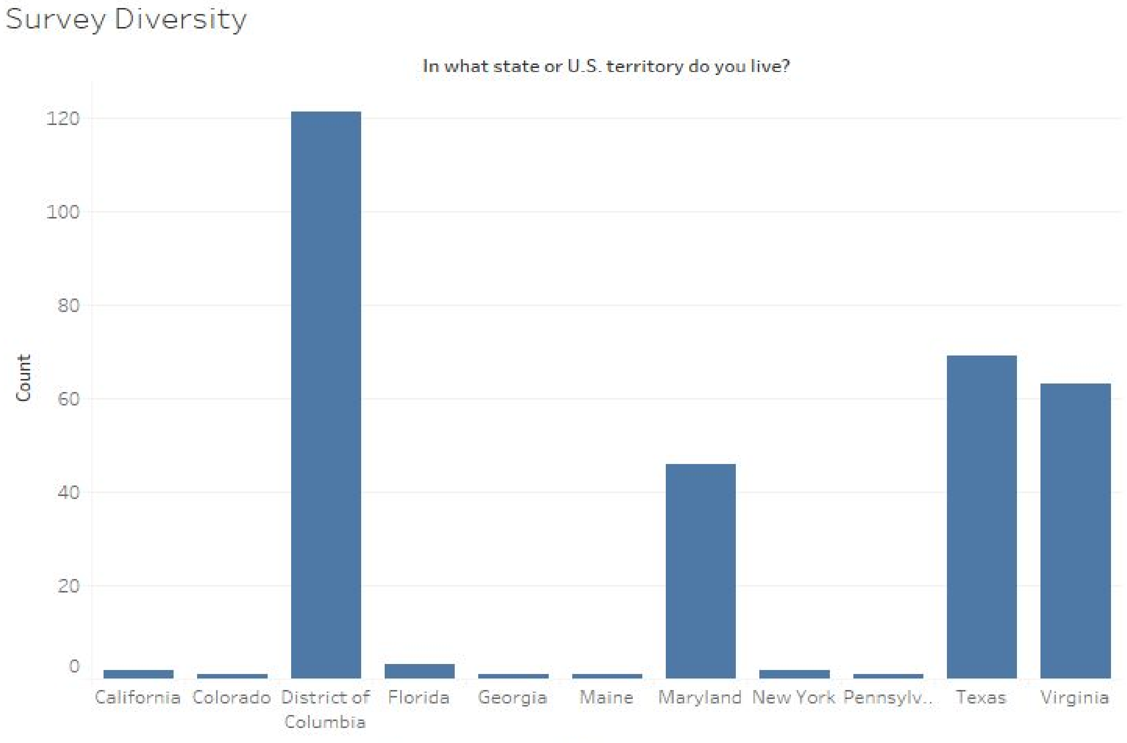
Plot for count of customers in segment(cluster):-



Above plot shows the count of participants in every cluster. Cluster 2 has forty five participants, but all the participants are those from whom the values were blank.

Participants regional diversity:-

Plot shows the regional diversity among the participants in the survey. From the plot I saw that majority of participants are from texas and DMV. This is because Composition ID has offices only in Houston and Washington DC.

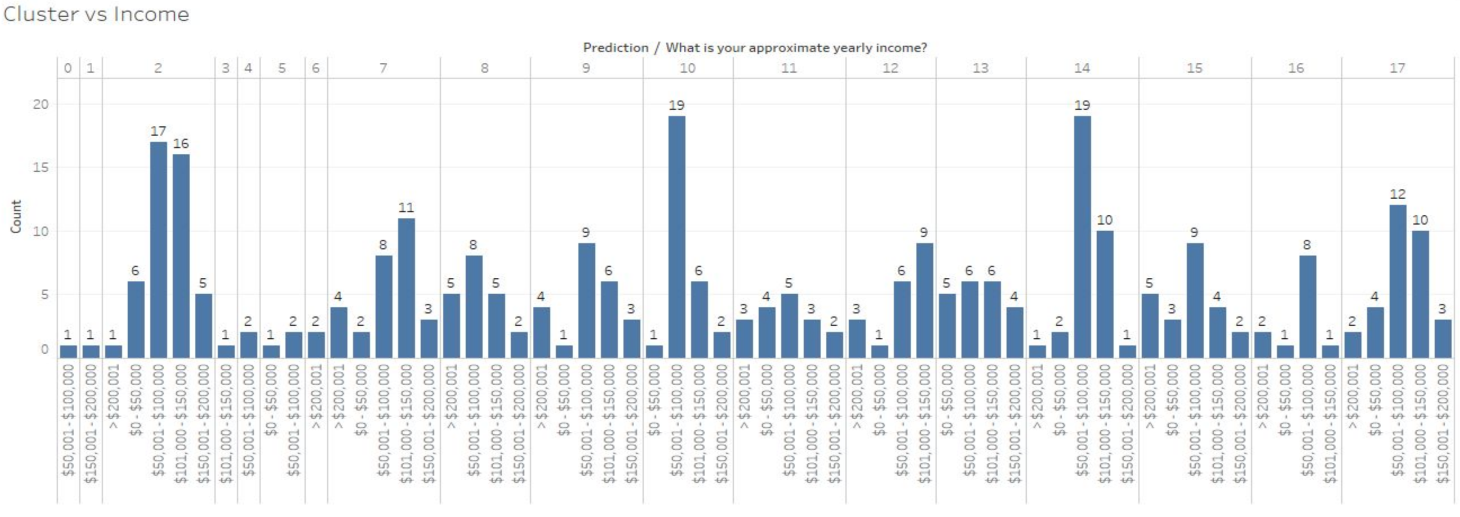


Count of Prediction for each In what state or U.S. territory do you live?.

Customer Segments on the basis of clusters:

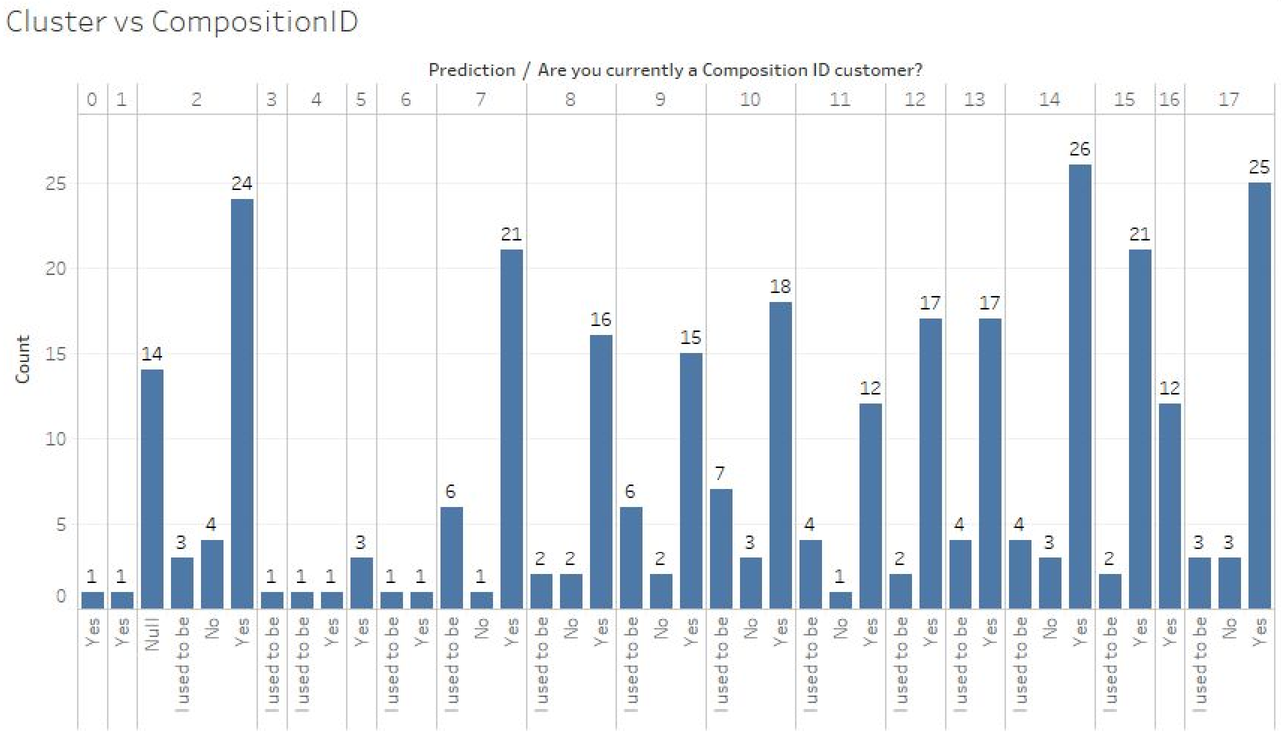
Cluster Vs Yearly Income:-

Following graph shows the spread of income in every customer segment(cluster). I observed that the clusters 0,4,5,10,14,16 have customers whose yearly income is low as compared to clusters 1,2,3,6,7,8,9,11,12,13,15,17. Here by low income I mean that around fifty percent of the customers have income below 100k and for high income around fifty percent of the customers have income above 100k.



Count of Prediction for each What is your approximate yearly income? broken down by Prediction.

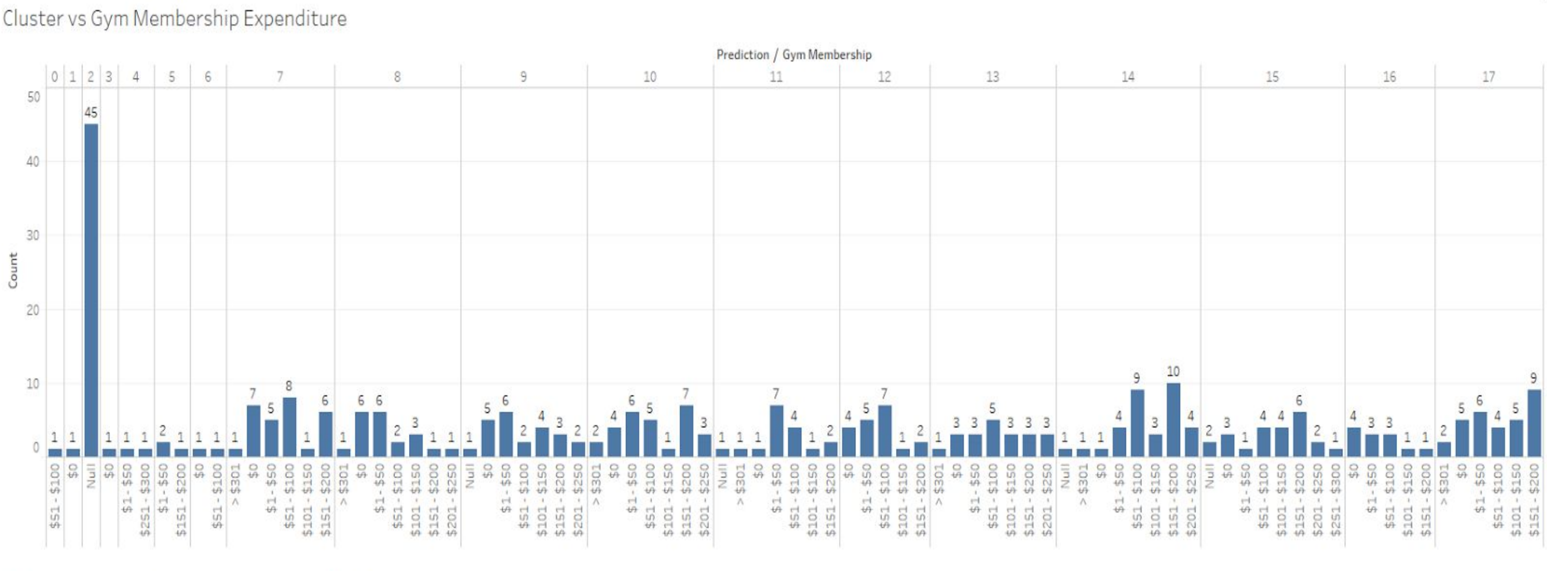
Cluster Vs Composition ID Customer:-



Count of Prediction for each Are you currently a Composition ID customer? broken down by Prediction.

Above graph shows count of customers of Composition ID in every customer segment(cluster). I observed that in the clusters 2,7,8,9,10,11,12,13,14,15,16,17 more than fifty percent of the participants are Composition ID customers. So, it can be said that they will take subscription in future also.

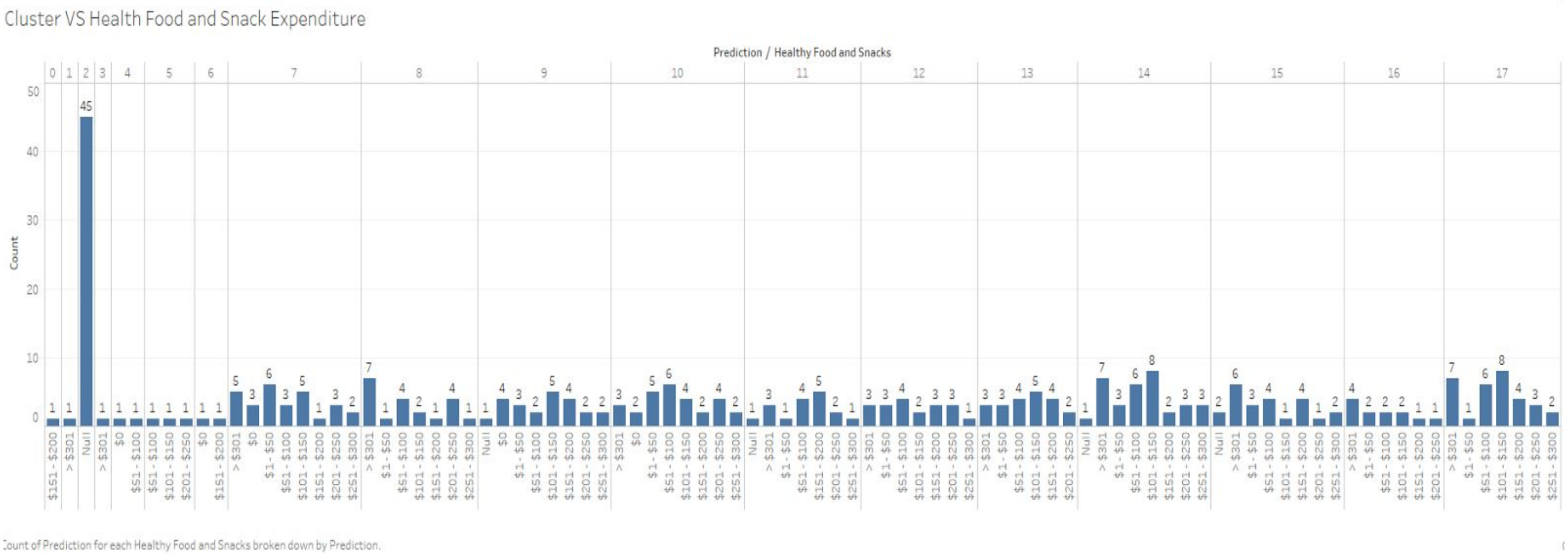
Cluster Vs Gym Membership Expenditure:-



Count of Prediction for each Gym Membership broken down by Prediction.

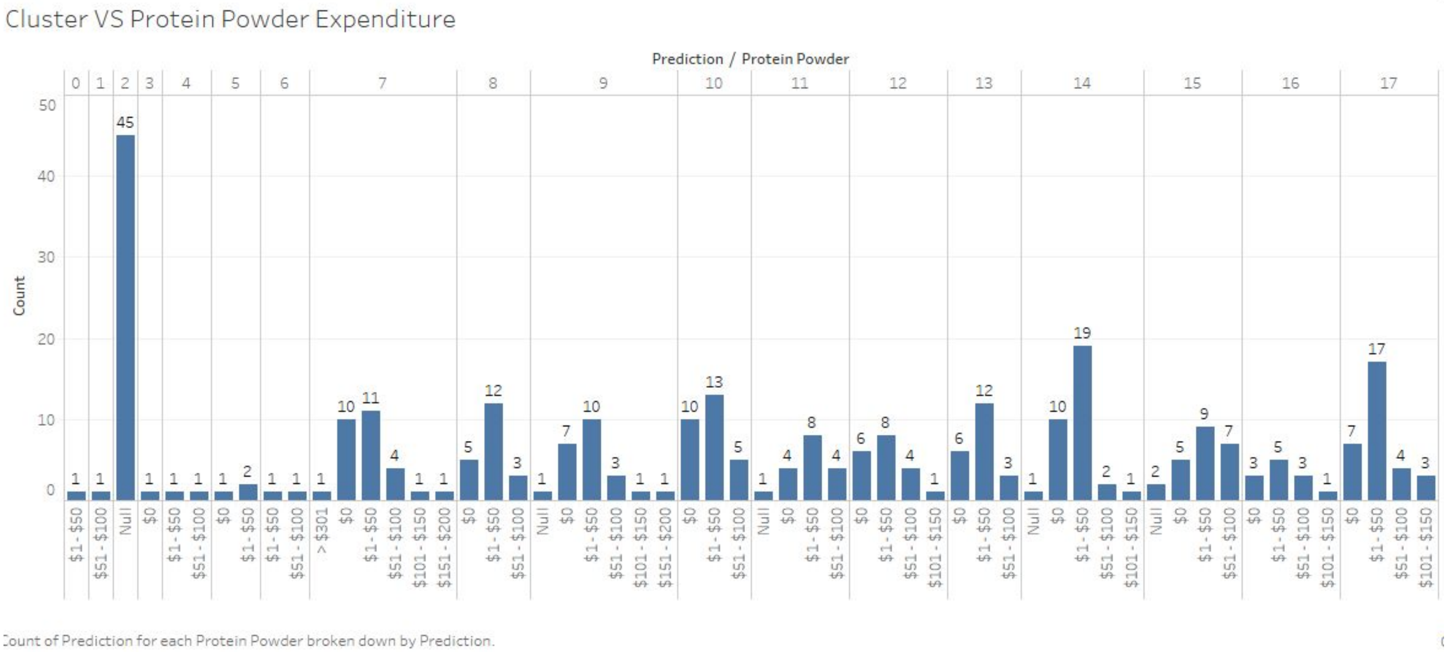
Above graph shows expenditure on gym membership in customers segments(clusters). I noticed that in the clusters 17,15,14,13,10 more than fifty percent participants spend more than hundred dollars on gym membership.

Cluster Vs Expenditure on Healthy Food and Snack:-



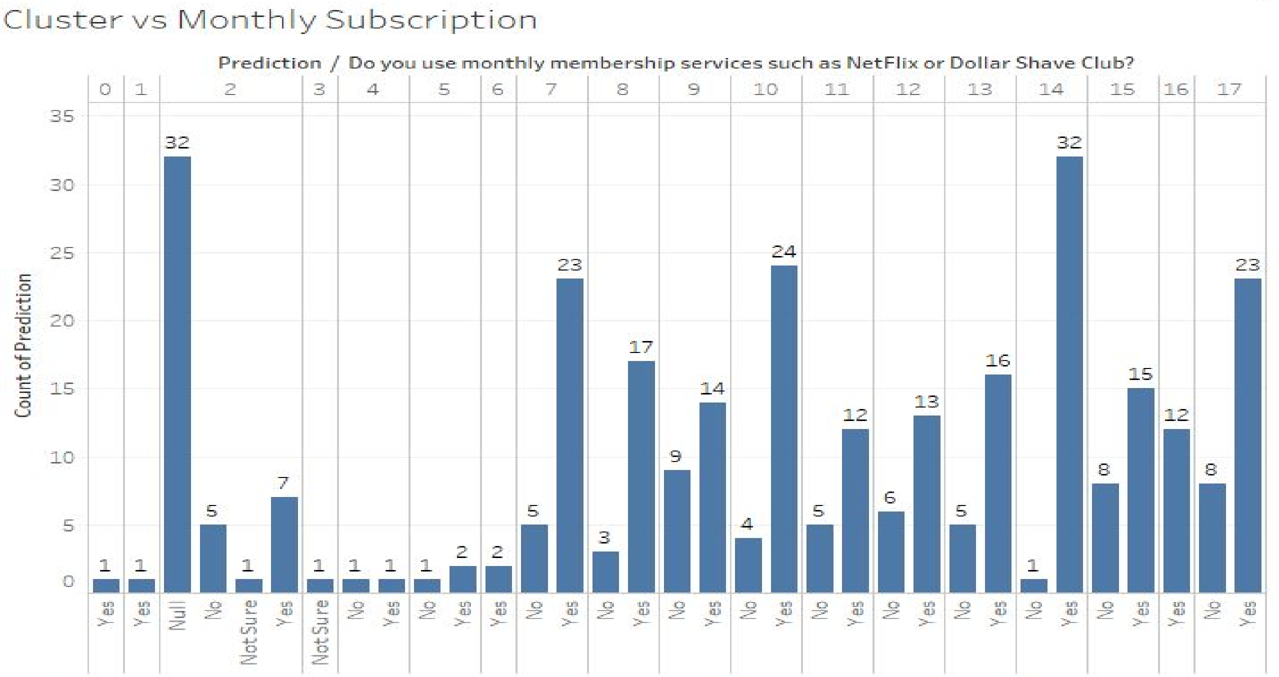
Above graph shows expenditure on health food in cluster segments(clusters). Except for cluster 2,4 every cluster has high expenditure on healthy food i.e more than fifty percent of the participants spend more than hundred dollars on healthy food and snack.

Cluster Vs Protein Powder Expenditure:-



Above graph shows the protein powder expenditure for various customer segments(clusters). I noticed that though most people in each cluster spend zero, but a considerable amount of participants spend money on protein powder. The clusters in which participants spend money on protein powder are 7,8,9,10,11,12,3,13,15,16,17. Though in clusters 0,1,4,5 participants do spend money but number of participants are very less.

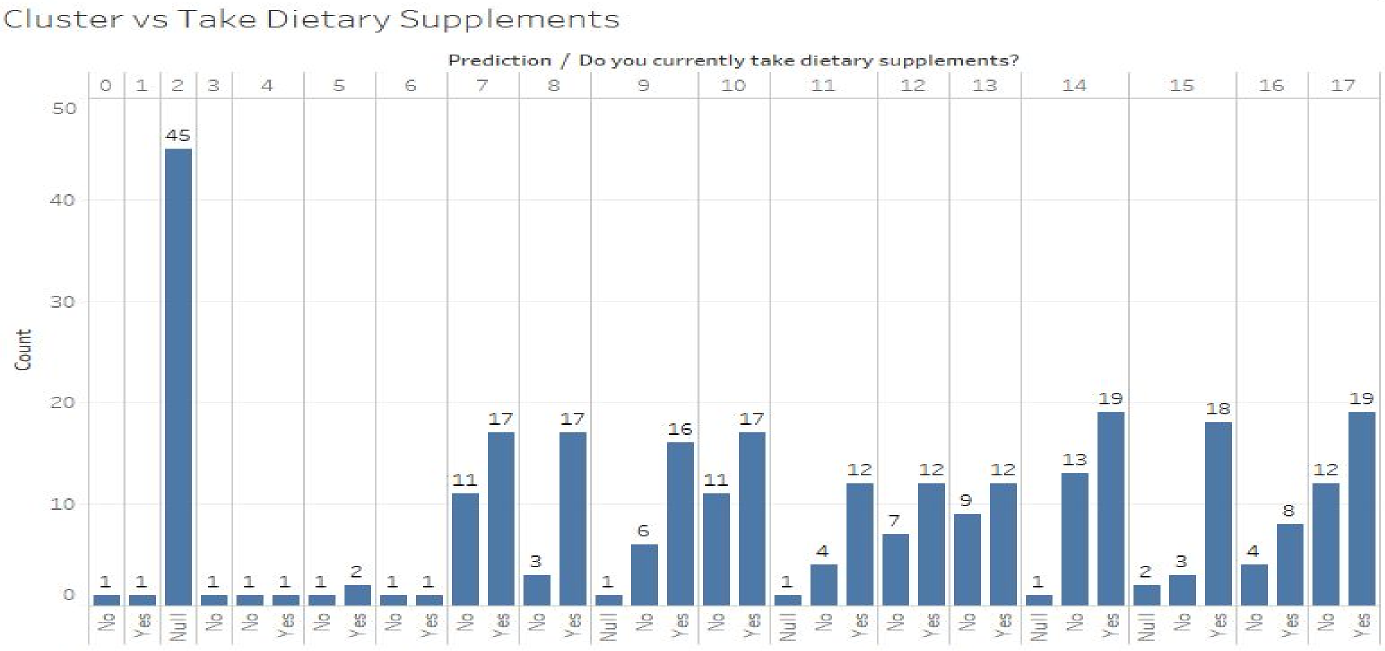
Cluster Vs Using Monthly Subscription:-



Count of Prediction for each Do you use monthly membership services such as NetFlix or Dollar Shave Club? broken down by Prediction.

Above graph shows the how many participants are using monthly services such as NetFlix or Dollar Shave Club in customer segments(clusters). Here, I noticed that in clusters 7,8,9,10,11,12,13,14,15,16,17 more than sixty percent of the participants are using monthly subscription, so they might take monthly subscription of Composition ID also.

Cluster Vs Taking Dietary Supplements:-



Count of Prediction for each Do you currently take dietary supplements? broken down by Prediction.

From the above plot, I noticed in the clusters 7,8,9,10,11,12,13,14,15,16,17 more than seventy percent of the participants take dietary supplements.

<u>Reason for low participants in some cluster:</u>

Cluster analysis shows that the clusters 0,1,3,4,5,6 have less participants(<10). This is because of the reason that most of the values in these clusters are either 0 or have very less values in different variables. Cluster 2 has 45 participants, these participants are together because they didn't answer the questions and I replaced missing values with 0, as a result they are in same cluster.

Identifying high and low profit segments:

From the above analysis, I identified that customer segments 7,10,12,13,14,15,16,17 are high profit segments and 0,1,2,3,4,5,6,8,9,11 are low profit segments.

Customer segment 14 is most profitable segment as their expenditure is high in gym membership, high income, almost everyone has a monthly membership, almost everyone takes dietary supplements and most participants are already member of Composition ID. So this is highly profitable segment.

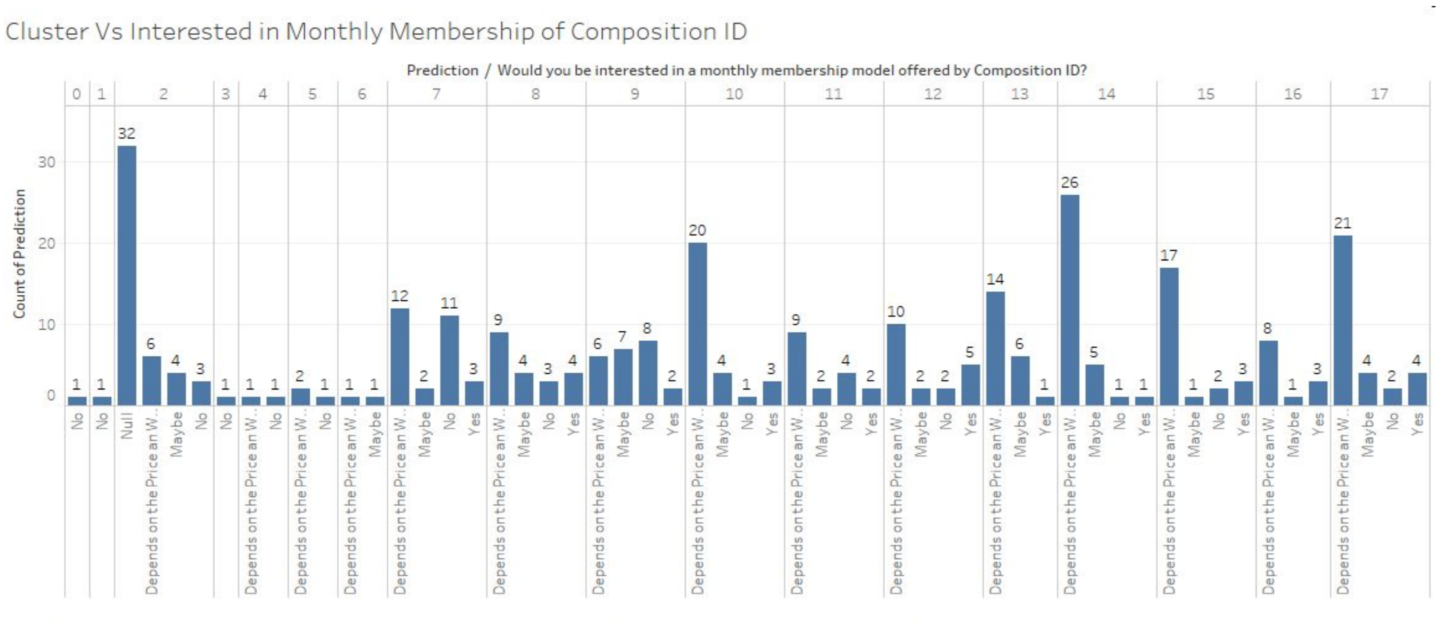
Customer segment 14 predicted spending in various categories:-

gym membership	> 100 dollars
health food and snack	> 150 dollars
personal training	> 50 dollars
nutritional advice	< 50 dollars
health diagnostic testing	< 50 dollars
protein powder	< 50 dollars
Dietary expenditure	> 50 dollars
Exercise clothes	> 50 dollars

Differences between High Profit segment Vs Low Profit Segment:

High Profit Segment	**Low Profit Segment**
Generally high expenditure in gym membership.	Generally low expenditure in gym membership.
Generally high expenditure on healthy food.	Generally low expenditure on healthy food.
Generally high income.	Generally low income.
High number of participants Composition ID customer.	Low number of participants Composition ID customer.
More Likely to take monthly subscription.	Less likely to take monthly subscription.
Generally take dietary supplements.	Most participants don't take dietary supplements.

Comparing Results of “Would you be interested in a monthly membership model offered by Composition ID?” in customer segments:



Count of Prediction for each Would you be interested in a monthly membership model offered by Composition ID? broken down by Prediction.

Above graph shows the result of question which asked about whether the participants will be interested in taking up the monthly membership model by Composition ID. For every segment most participants are skeptical about the taking the membership as their decision is based on the price offered by the Composition ID for the membership. Data analysis showed that the segment 14 is most profitable but the graph above shows that the customers are skeptical about taking the monthly membership. So with the right balance in price and services provided in membership, customers in segments 10,12,13,14,15,16,17 will take monthly membership.

Moreover they provide four services(DEXA, VO2 Max, Nutritional Counselling, RMR). DEXA is mostly done once in two years, RMR for non athlete but people who train or exercise can consider taking RMR test in a span for 4-5 months and for athletes depends on their training. Nutritional counselling and VO2 Max are the only services whose frequency can be more than once in a month. So, monthly memberships plans can be made for customers who are athletes, non-athletes but they do some sort of training and for customers who do not train or exercise but are health conscious. For athletes the plan should include nutritional counselling, VO2 Max, RMR and for non-athletes but they do some sort of training the plan should include nutritional counselling,RMR as it is necessary when diet changes. For customers who do not train or exercise but are health conscious, plan should include nutritional counselling.