



INDEX

1. Introduction 2
2. Data cleaning and Exploration 3
3. Data Analysis and Model 17
4. Conclusion 24

**1. Introduction**

Sterling-Rice Group has been a flourishing marketing services company for more than 30 years, offering brand positioning, creation, architectures, mergers and acquisitions. The company serves customers ranging from startups to Fortune 500 companies. The company has worked with several food and beverages giants including Starbucks, Kraft, and Coors.

The data provided is SRG’s quantitative research study conducted in 2015 to better understand health and wellness, and help their clients in any of the numerous ways – brand positioning, expanding market share of food and beverage companies, launching market strategies, etc.

We aim to help their clients maximize their profits by increasing product sales. For this, we are studying the given data to find the target audience would be the best way to maximize profits. Through clustering analysis, we determined a target audience and established a target profile. The target audience for the particular company is people who are “Aspiring Wellness” and “Epitomizing Wellness” and have BMI in the range 18.5 to 25. We termed the target as Healthy Eater. We also recorded the shopping preferences of the Healthy Eater like the store that they buy their groceries. This way the product can be featured or displayed in such stores to maximize profits. We also have to find out various factors about these people like the average age, average salary and education of each of these groups as that will help us take better and informed business decisions.

Our goal here is to provide a complete plan on how the Company can launch its new products into the market and how they can maximize their sales and market cap by targeting the correct people by not wasting resources on people who won’t buy wellness or health related products.

**2. Data Cleaning and Exploration**

The dataset initially consisted of 1152 columns and 1004 rows. The columns were reduced to 461 since most of the deleted columns were missing more than 50% of data. All the categorical variables were converted to quantitative variables. A new column called BMI was added by calculating the BMI from the height and weight of the audience.

We are using BMI as the criteria for segmenting people as healthy and not healthy, where BMI in the range of 18.5 to 25 is healthy and anything outside that is unhealthy.

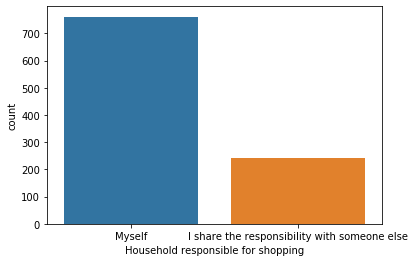


Figure 2.1

The above graph displays the count of the person responsible for shopping in a household. We can notice that most of the audience are themselves responsible for shopping. So, if we were to target our product to such an audience then it would guarantee an increase in profits. Since a person is more likely to purchase a product if they themselves are responsible for shopping in their household.



Figure 2.2

The above graph displays the count of gender. We can notice that there are equal numbers of males and females in our dataset.

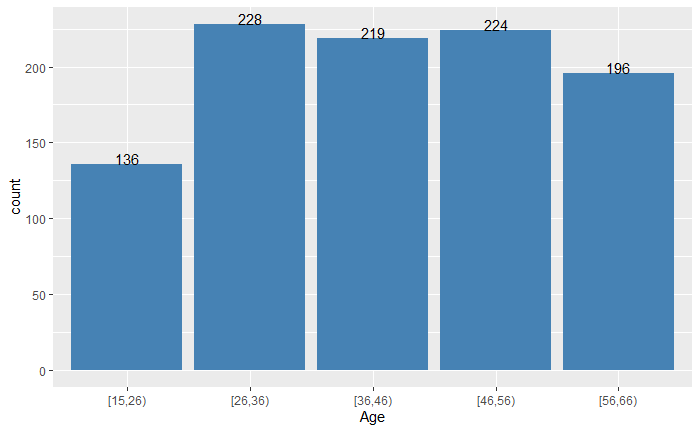


Figure 2.3

The above graph displays the count of each age group in the dataset. As we can notice, the count for all age groups is almost the same. The age group 15 – 26 has the lowest count at 136.

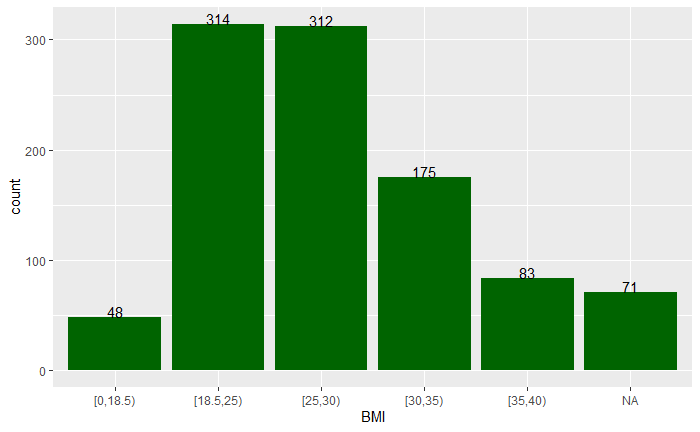


Figure 2.4

The above graph displays the count of BMI of the audience. A person falling under the BMI range of 18.5 – 25 is healthy. We can notice that in our data set, the range 18.25 – 25 has the highest count at 314, which means the audience is mostly healthy. Therefore, we can gain more information about people who are healthy and the reasons why certain people are more healthy compared to the other group.

We can see that the ideal BMI and slightly overweight groups are quite close to one another; this shows that there as many people who are slightly overweight as there are fit people in the dataset.

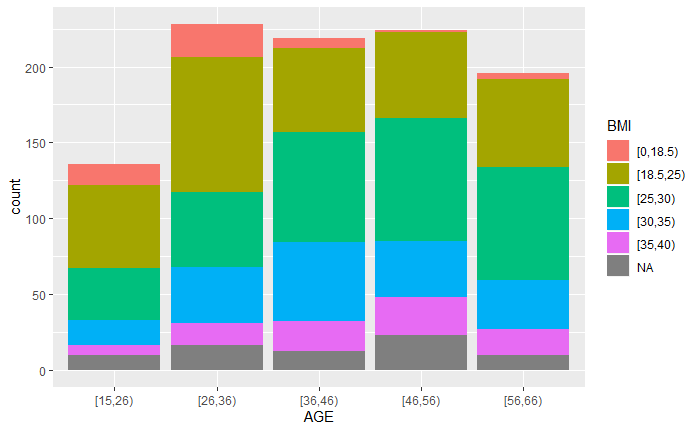


Figure 2.5

The above graph displays the count of age groups with respect to the BMI. We can notice that the age group 26 – 36 is the healthiest because they have normal BMI in the range 18.25 – 25. So, for most of our target audience, the healthiest group of people belong to the age group 26 – 36.

The reason for this most people in their late twenties and early thirties try to keep fir for various reasons, maybe to look good as most of them are still unmarried or even because they get more free time than all the other people in different age groups. As people grow older and older, they have less time for themselves as they have to work as well as spend time with the family, which can be really difficult and stressful, Thus leading to an increase in weight in that particular group.

But as you grow older we realize that health is really important in order to survive and hence we start healthy and take walks and do other such activities and thus we can see there is a significant amount of people in the healthy section in the oldest age group.

People in the youngest age group are bound to be healthy because they get enough time to play sports as well as participate in various activities be it in school or in college and also because they haven’t been exposed to the corporate lifestyle where we get less time to worry about our physical health.

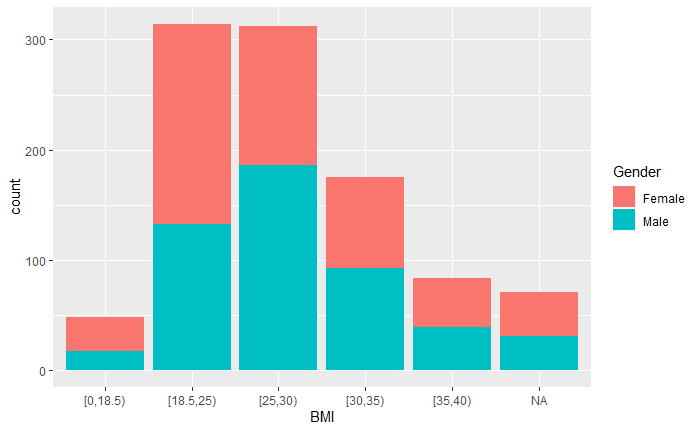


Figure 2.6

The above graph displays the count of different groups of BMIs with respect to the Gender of the audience. We can notice that for our target BMI 18.5 – 25, the number of males and females is almost equal. So, we may assume that gender does not have much effect on the BMI.

From this we can say that when we target our customers the campaigns should be in such a way that it attracts both male and female audiences at the same time. We also see that in almost all the BMI ranges the male and female proportion of the people are the same, so all our campaigns should target male and female customers equally.

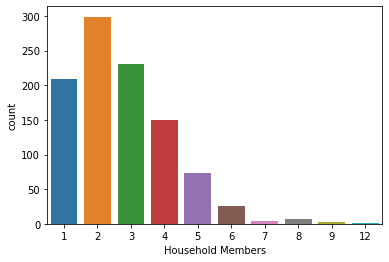
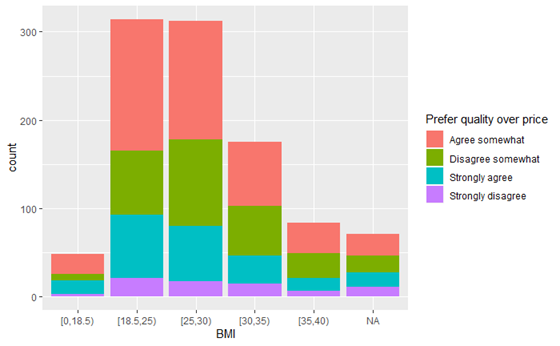
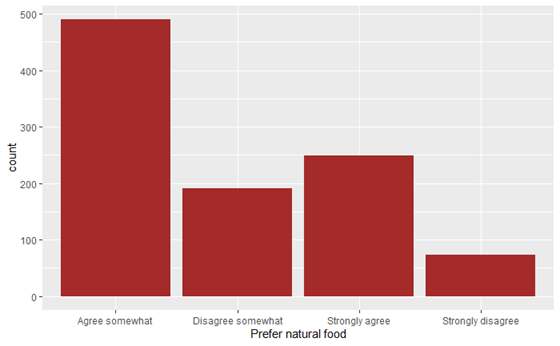


Figure 2.7

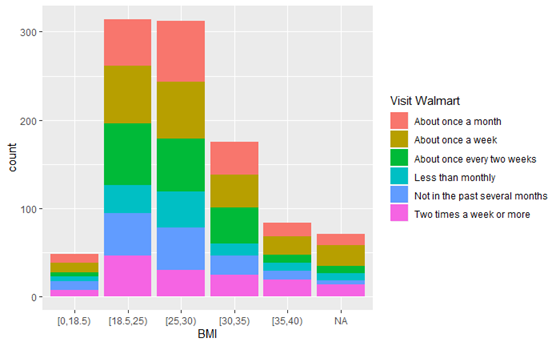
The above graph displays how many people live in a household. Majority of the audience have 2 people living in their household. That means that most of the time they cook any food it will be for all the people in the family. So, if we can market our products with success to the head of the family or the one who is responsible for all the shopping in the family we might end up selling more products now since everyone in the family will also use the same products. We see that majority of the people in the dataset have family members living with them and we can use that point to our advantage about how nothing is more important for us than the health of the customers and their family members



The above graph displays the relationship between the BMI of a person and their priority of quality over price. We can notice that people who agree on prioritizing quality over price generally have an ideal BMI. From this we can say that people who prefer their health do not mind spending money if they are getting good and healthy products.



The above graph displays the preference of the audience for natural foods. Majority of the audience either somewhat agree or strongly agree that natural foods are better for them and the environment .We can also suggest to the company to sell healthy and natural foods, they could market the product as being good for an individual and good for the environment as well. This could well be used in Marketing campaigns as well where we can tell how the company not only cares for the health of the people but also about the environment. The company can achieve this by using Natural Foods and Ingredients and using environment friendly packaging. This will only help us improve the brand image of the company on a global scale as the need for being environment friendly is really important at this age of time because of problems such as global warming and plastic pollution around the world.



We notice that from the dataset most of the people visit Walmart and hence we try to find how many times people go to Walmart and what their BMI ranges. This is important as we can decide how much money the company can spend for various activities in Walmart.

The above graph shows the frequency of visit to Walmart with respect to the BMI. We can notice that Walmart has a high number of visits by the audience with an ideal BMI. We can also see that most people who are in the healthy range go to Walmart at least once every two weeks or more than that. This can be used to our advantage where we use this data to have various Displays and Promotional events in Walmart regularly to improve the brand image of the company.

We can also ask the company to tie with Walmart as to have certain rows in the Store where products are more visible to the people walking inside the store as this has a significant impact on how many people buy the products.

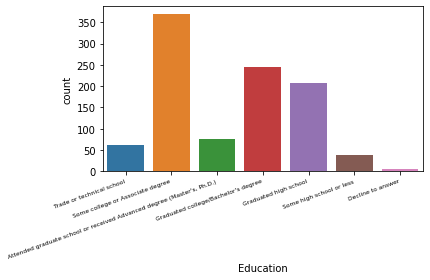
****

Figure 2.8

The above graph displays the count of the different educational backgrounds of the audience. We can notice that Some college or Associate degrees have the highest count. If most of the dataset have some sort of college degree or education, it becomes easier to market our products as people can understand our message of health and wellness quite well.

This is because people who have completed some level of education know the dangers of becoming overweight and the health complications that arise with it. We can use this point to carefully market our product.

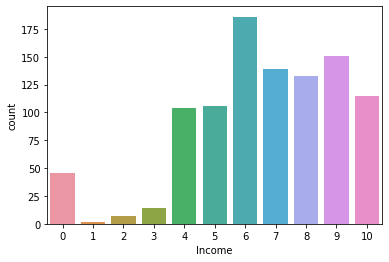
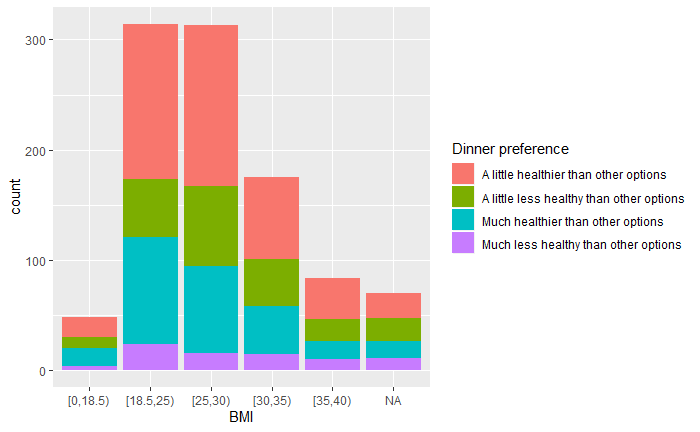


Figure 2.9

The above graph displays the count of the income groups. "Less than $15,000" is 10, "$15,000 but less than $25,000" is 9, "$25,000 but less than $35,000" is 8, "$35,000 but less than $50,000" is 7, "$50,000 but less than $75,000" is 6, "$75,000 but less than $100,000" is 5, "$100,000 but less than $200,000" is 4, "$200,000 but less than $300,000" is 3, "$300,000 but less than $500,000" is 2, "$500,000 or over" is 1, "Decline to answer" is 0. We can notice that most of the audience have a salary greater than $50,000 and less than $75,000. This means that our sample represents the population very closely, because the average salary in the US is $56,516.

Since we know that most of the responders here are the average American people, we can further use the same evidence to run our campaigns across the country with similar success. There are also people in the dataset who have salaries greater than $100,000. We can further focus on this group to see their eating habits and wellness, as they would be willing to spend money if they know the benefits of leading a healthy lifestyle.



From the above graph we can see most of the people who are in the ideal range of the BMI(18.5-25) prefer eating healthy even when they option to eat something else and also the group that is slightly overweight also eats healthy this shows they are putting in efforts to become healthier.

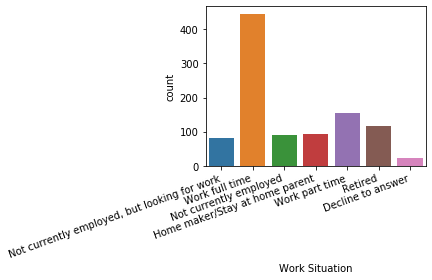


Figure 2.10

The above graph displays the work situation of the audience in the dataset. We can notice that the majority of the audience work full time.

After sufficient exploration, we determined our area of focus to be segmentation into healthy and unhealthy and how health is affected by the following- age, income, exercising, did people choose healthy options for their last meal and overall attitude towards wellness.

For this purpose, we chose the following variables –

* Age – numerical variable (S3) describing age is renamed to Age
* Income – categorical variable (D6) with multiple levels changed into a binary categorical variable with two levels – less than $50,000 and $50,000 and above
* Exercise – multi-level categorical variable (Q6) about how often people exercise, changed to a binary categorical variable with two levels – two to three times a week and more than two to three times a week
* HealthyDinner – multi-level categorical variable (Q18), changed to binary categorical variable with two levels as 1 for healthy and 0 for not healthy
* Points – numerical variable (Points) that is a cumulative response of different wellness based questions, with highest points being for epitomizing wellness.

Table 2.1: Frequency table for Exercise variable

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Exercise** | | | | |
| **Exercise** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| **0** | 532 | 53.15 | 532 | 53.15 |
| **1** | 469 | 46.85 | 1001 | 100.00 |

Table 2.2: Frequency table for Income variable

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Income** | | | | |
| **Income** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| **0** | 724 | 72.33 | 724 | 72.33 |
| **1** | 277 | 27.67 | 1001 | 100.00 |

Table 2.3: Frequency table for HealthyDinner variable

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Healthydinner** | | | | |
| **Healthydinner** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| **0** | 294 | 29.37 | 294 | 29.37 |
| **1** | 707 | 70.63 | 1001 | 100.00 |

Table 2.4: Frequency table for Points variable

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Points** | | | | |
| **Points** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| **0** | 651 | 65.03 | 651 | 65.03 |
| **1** | 350 | 34.97 | 1001 | 100.00 |

Table 2.5: Frequency table for NormalBMIOrNot variable

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **NormalBMIOrNot** | | | | |
| **NormalBMIOrNot** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| **0** | 687 | 68.63 | 687 | 68.63 |
| **1** | 314 | 31.37 | 1001 | 100.00 |

**3. Data Analysis**

From the above tables 2.1-2.5, we observed that the majority of the survey takers earn less than $50,000. There are more people with abnormal BMI than normal, suggesting that unhealthy people can be studied to expand markets for various health and food products.

As stated previously, we are using BMI as the criteria for segmenting people as healthy and unhealthy.

Our focus is on the following variables after segmentation: - Age, income, whether they exercise, eat healthy dinner, the points for aspiring wellness, epitomizing wellness etc.

**Note:** Analysis is based on 10% significance level

Table 3.1 Correlation between age, exercise, healthyDinner, income, points, normalBMIOrNot

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Pearson Correlation Coefficients, N = 1001 Prob > |r| under H0: Rho=0** | | | | | | |
|  | **Age** | **Exercise** | **Healthydinner** | **Income** | **Points** | **NormalBMIOrNot** |
| **Age** Age | 1.00000 | 0.00110 0.9724 | 0.00025 0.9936 | -0.03480 0.2713 | 0.02382 0.4515 | -0.09524 0.0026 |
| **Exercise** Exercise | 0.00110 0.9724 | 1.00000 | 0.16592 <.0001 | 0.13074 <.0001 | 0.30653 <.0001 | 0.10304 0.0011 |
| **Healthydinner** Healthydinner | 0.00025 0.9936 | 0.16592 <.0001 | 1.00000 | 0.02626 0.4065 | 0.16925 <.0001 | 0.08142 0.0100 |
| **Income** Income | -0.03480 0.2713 | 0.13074 <.0001 | 0.02626 0.4065 | 1.00000 | 0.09903 0.0017 | 0.07753 0.0141 |
| **Points** Points | 0.02382 0.4515 | 0.30653 <.0001 | 0.16925 <.0001 | 0.09903 0.0017 | 1.00000 | 0.20413 <.0001 |
| **NormalBMIOrNot** NormalBMIOrNot | -0.09524 0.0026 | 0.10304 0.0011 | 0.08142 0.0100 | 0.07753 0.0141 | 0.20413 <.0001 | 1.00000 |

From the last column in the above table, we observed that there is a strong correlation between NormalBMIOrNot and remaining variables Age, Exercise, HealthyDinner, Income, Points.

Thus, on performing Logit for NormalBMIOrNot against all other variables, we find the following result:

Table 3.2: AIC for the Logit

| **Model Fit Statistics** | | |
| --- | --- | --- |
| **Criterion** | **Intercept Only** | **Intercept and Covariates** |
| **AIC** | 886.330 | 853.526 |
| **SC** | 890.883 | 880.841 |
| **-2 Log L** | 884.330 | 841.526 |

Table 3.3: Parameter estimates for Logit

| **Analysis of Maximum Likelihood Estimates** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| **Parameter** |  | **DF** | **Estimate** | **Standard Error** | **Wald Chi-Square** | **Pr > ChiSq** |
| **Intercept** |  | 1 | -0.2991 | 0.2765 | 1.1704 | 0.2793 |
| **Age** |  | 1 | -0.0175 | 0.00625 | 7.8127 | 0.0052 |
| **Income** | **1** | 1 | 0.1719 | 0.0911 | 3.5615 | 0.0591 |
| **Exercise** | **1** | 1 | 0.0513 | 0.0886 | 0.3350 | 0.5627 |
| **Healthydinner** | **1** | 1 | 0.1175 | 0.0964 | 1.4855 | 0.2229 |
| **Points** |  | 1 | 0.8217 | 0.1780 | 21.3054 | <.0001 |

We see that age, income and points are significant.

Age has a negative effect on NormalBMIOrNot, implying that for a person, every year the chance of being healthy decreases by 0.0175, meaning younger people seem to have better chances to have normal BMI, or in other words, are healthier, than older people.

On the other hand, with every one unit increase in income of a person, the chances of being healthy increases by 0.1719, implying that higher income people are healthier.

And clearly, with every one unit increase in points, chances of being healthy increases by 0.8217 for a person suggesting they are pursuing wellness indeed.

Using the assumption that with increasing age, people might have more health problems and might start eating healthy and thus attain normal BMI, or in other words, have increased chances of becoming healthy, we ran Logit with Age\*Age.

Table 3.4: AIC for the Logit with age2

| **Model Fit Statistics** | | |
| --- | --- | --- |
| **Criterion** | **Intercept Only** | **Intercept and Covariates** |
| **AIC** | 886.330 | 853.065 |
| **SC** | 890.883 | 884.933 |
| **-2 Log L** | 884.330 | 839.065 |

Table 3.5: Parameter estimates for logit with age2 variable

| **Analysis of Maximum Likelihood Estimates** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| **Parameter** |  | **DF** | **Estimate** | **Standard Error** | **Wald Chi-Square** | **Pr > ChiSq** |
| **Intercept** |  | 1 | 0.9544 | 0.8431 | 1.2815 | 0.2576 |
| **Age** |  | 1 | -0.0843 | 0.0430 | 3.8427 | 0.0500 |
| **Income** | **1** | 1 | 0.1781 | 0.0914 | 3.7991 | 0.0513 |
| **Exercise** | **1** | 1 | 0.0598 | 0.0890 | 0.4519 | 0.5015 |
| **Healthydinner** | **1** | 1 | 0.1128 | 0.0967 | 1.3594 | 0.2436 |
| **Points** |  | 1 | 0.8203 | 0.1783 | 21.1572 | <.0001 |
| **age\_2** |  | 1 | 0.000800 | 0.000509 | 2.4722 | 0.1159 |

We observed a slight improvement in AIC and that the variables age, income and points remain significant.

The above implies that every year a person’s chances of being healthy decreases by 0.0843. Thus, young people are healthier and could serve as the target audience for the healthy food products market such as organic food for millennials.

On the other hand, with each one unit increase in income of a person, the chances of being healthy increases by 0.1781, implying that higher income people are healthier.

And clearly, with each one unit increase in points, chances of being healthy increases by 0.8203 for a person suggesting they are pursuing wellness indeed.

Furthermore, we ran Logit with interaction effect between age and income to understand whether increase in age and income group of people (high or low) influences their health (BMI-wise).

Table 3.6: AIC for the Logit with interaction effect of age and income

| **Model Fit Statistics** | | |
| --- | --- | --- |
| **Criterion** | **Intercept Only** | **Intercept and Covariates** |
| **AIC** | 886.330 | 855.223 |
| **SC** | 890.883 | 887.090 |
| **-2 Log L** | 884.330 | 841.223 |

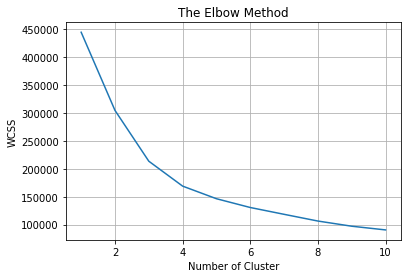
Table 3.7: Parameter estimates for logit with interaction effect of age and income

| **Analysis of Maximum Likelihood Estimates** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| **Parameter** |  | **DF** | **Estimate** | **Standard Error** | **Wald Chi-Square** | **Pr > ChiSq** |
| **Intercept** |  | 1 | -0.3610 | 0.2983 | 1.4645 | 0.2262 |
| **Age** |  | 1 | -0.0159 | 0.00686 | 5.3773 | 0.0204 |
| **Income** | **1** | 1 | 0.0195 | 0.2912 | 0.0045 | 0.9465 |
| **Exercise** | **1** | 1 | 0.0517 | 0.0887 | 0.3397 | 0.5600 |
| **Healthydinner** | **1** | 1 | 0.1174 | 0.0964 | 1.4821 | 0.2234 |
| **Points** |  | 1 | 0.8223 | 0.1780 | 21.3326 | <.0001 |
| **Age\*Income** | **1** | 1 | 0.00377 | 0.00684 | 0.3037 | 0.5816 |

We found the income as well as its interaction effect with age to be insignificant, implying there could be other factors involved with income – such as the state in which people reside, meaning people living in regions with low state taxes could still afford to eat healthy and thus are already healthy.

Since the AIC for the classification models are not good, we further ran clustering models to see if we can find any patterns in the data and if there is any particular cluster that is more healthy than any other cluster so we can target those particular clusters in order to market our products.

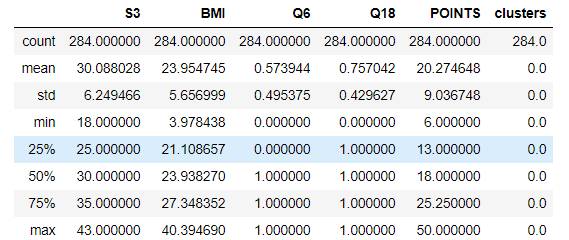
We use Age, Income, Exercise, Healthy Dinner, BMI and Points to cluster people as these are the factors that decide which people are healthy.



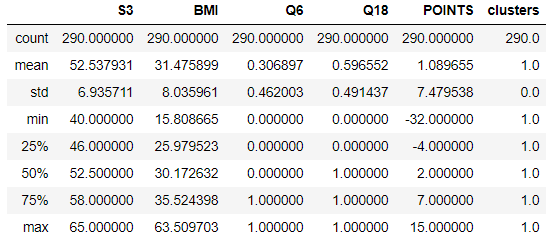
Using the elbow method where we use the WCSS(within cluster sum of squares) vs the number of clusters we can say that there is a small elbow at around the value of 4 thus we choose n=4 to be the ideal number of clusters for our model.

We run the K-means algorithm with K=4 and then find patterns in these clusters and find more information about those clusters.

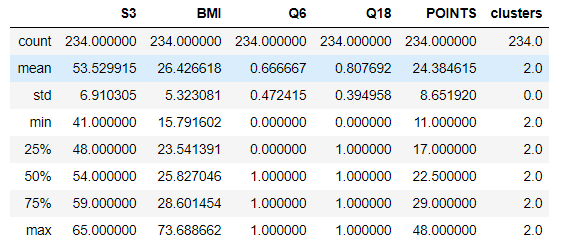
For Cluster 0:



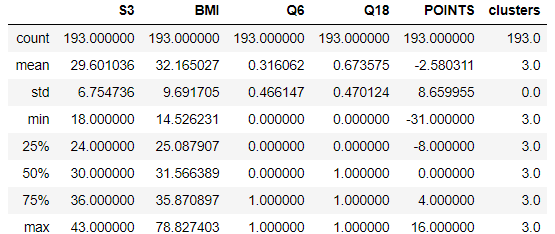
For Cluster 1:



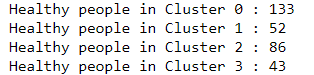
For Cluster 2:



For Cluster3:



We can see that the Clusters 0 has the mean and median BMI around the ideal range. We must further find more about the people in this Cluster so that we can get more information about this demographic.



We see that in Cluster 0 we have the most percentage of healthy people, we try to find out more information about this demographic.

The average age of all the people in the cluster 0 is



This again is as we expected that younger people have more reasons and time to be fit than people who are in their late thirties and early forties.

The income in this cluster is like that on the entire dataset it is less than $75,000 for most of the people. We can say that income is not as important as Exercise and eating healthy also plays an important role in whether a person is healthy or not.



We then see at what education level the people in this cluster we have seen that most of them have a degree which shows that education plays a really important role whether a person is healthy or not.



We also try to find any more patterns which can help us see if that behavior aids in the person being healthier or not.

We see that even though the majority of the people in this cluster don’t exercise regularly most of those who do exercise regularly have a BMI in the ideal range, thus showing exercise is really important even if we eat healthy.

**4. Conclusion**

From the above analysis, we found out the demographics of people who lead a healthy lifestyle and also eat healthy food. This segment of people is the most important for our clients as these are the people who would most likely be interested in their products compared to any other demographic of people.

We see that across the dataset and also across the clusters that most of the BMI in the ideal BMI range are the people between the ages of 25 and 35, which is reasonable considering this age group has relatively more time for their health, but as they grow older there are more issues such as work and other stresses that affects the time they spend on themselves thus becoming unhealthy. This is also evident from the logistic regression models we ran where we can see that the ‘age’ factor plays a significant role in whether the person is healthy or not.

We also observe that even though income plays a significant role if the person is healthy or not, it is not important as shown in the clusters that most of the people even with lower incomes have a healthy lifestyle. The reasons for this might be plenty as people who earn more money need to work longer hours and have more stressful jobs. Also, people who earn large sums of money travel very frequently and thus do not have a control on their diet as they must eat outside every time they travel or go out for business meetings.

We also saw the role of education in the health of a person. It can be noted that most people with qualification of a degree or above are unhealthy compared to people who only have gone to a Trade School or lower. This is because usually jobs for people from Trade School or lower qualifications involve a lot of physical effort and thus help them maintain their fitness compared to the former where most of the jobs don’t take any physical effort.

We also see that a lot of customers who are healthy go to Walmart to buy their products, we can have various displays around important areas to improve the visibility of our clients’ products and to market their products. We also see that most people do not mind spending more money if the product is good. This can help us fix a competitive market price for our clients’ wellness products.

We also see that people care if the product is all natural or not, hence our clients’ can also release an all whole food-based product. The packaging of which will also be environment friendly. This helps us get customers on both sides of the aisle.

To conclude, we would suggest our clients to target a younger audience as they are the people who are most likely to buy the product. We can also ask them to target people who buy for their whole family by creating campaigns or advertisements showcasing a family's collective health.

The price of the product is not of great significance as people are willing to spend more but it should not be exorbitant as most healthy people earn very little and have families to take care of. Hence, our clients can have two kinds of products: one that is healthy and one that is healthy based on whole foods (natural ingredients).