Understanding and Influencing Health and Wellness

BUAN 6337.001 Predictive Analytics Using SAS

Group 2
Jian Li /JXL190021
Yuru Li /YXL180114
Xin Liao /XXL180003
Xuechen He / XXH190007
Yuzhou Huang /YXH190011
Jindong Yu /JXY180004

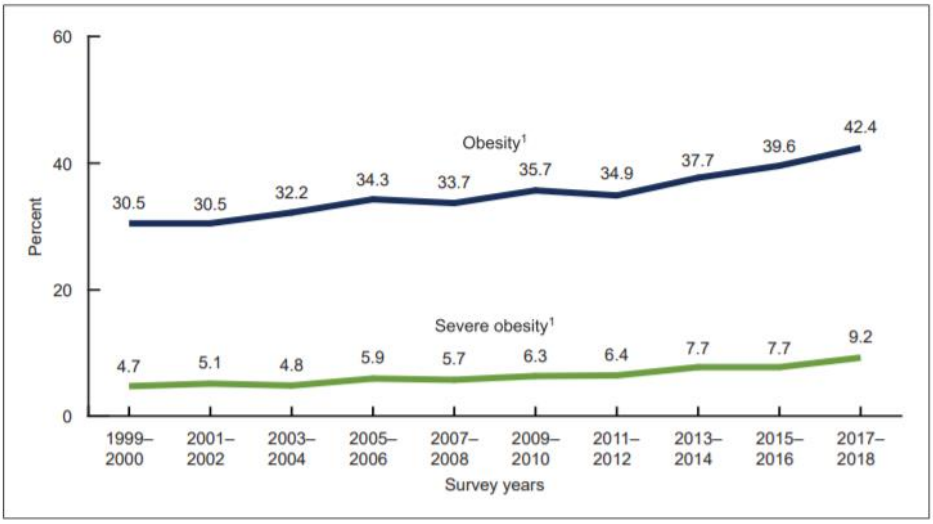
Executive Summary

Recently, with the economic prosperity and lifestyle changes, the overweight and obesity problem become one of the biggest threats to health and wellness. The market related to solving overweight and obesity problem catches huge attention of business sectors. In this study, we use data science techniques such as principal component analysis, k-means clustering, regression etc., to analyze a comprehensive customer survey data, and find that although some people have good health mindset, still suffer from overweight and obesity problem. One of the main reasons is busy and overwork. For people who have good health mindset, high BMI, and busy, food choices and structure have significant effect on BMI (overweight), especially unhealthy and high calories foods such as frozen and canned food, candy and sugar, high carbohydrate, and beer. Based on these insights and our target customer segmentation characteristics, we developed specific strategies and recommendations for businesses.

Introduction

Health and wellness are always the most important topics for every human being. Recently, the overweight and obesity problem have become one of the biggest threats to health and wellness with the economic prosperity and lifestyle changes. According to a study conducted by National Center for Health Statistics, from 1999–2000 through 2017–2018, the age-adjusted prevalence of obesity increased from 30.5% to 42.4%, and the prevalence of severe obesity increased from 4.7% to 9.2% (Figure 1). So, what are the causes of overweight and obesity, and how to help people keep away from obesity and obtain a great health level are important questions that all consumer product companies care about the most. It is a booming market with great potential.

Figure 1: Trends in age-adjusted prevalence of obesity



[1]Significant linear trend.
NOTES: Estimates were age adjusted by the direct method to the 2000 U.S. Census population using the age groups 20–39, 40–59, and 60 and over. Access data table for Figure 4 at: https://www.cdc.gov/nchs/data/databriefs/db360_tables-508.pdf#4.
SOURCE: NCHS, National Health and Nutrition Examination Survey, 1999–2018.

In this study, with the SRG Proprietary Shopper Insights Study survey dataset, we will try to uncover business insights about overweight and obesity, specifically, why people become overweight and obese, how the foods choices / structure impact health status for different groups of people.

Data Pre-processing

There are more than 30 survey questions and 1003 observations in the original dataset. With our research question in mind, we screened the original data table and selected part of the variables we need, and code them for our use.

Overweight / Obesity

We choose BMI as the measure of overweight and obesity level, which is widely used in research papers. We calculate BMI using participants' height and weight data: survey question D9, D10ra, D10rb.

D9:	Please tell us your current weight in pounds
D10ra:	ENTER FEET: - Please tell us your current height
D10rb:	ENTER INCHES: - Please tell us your current height

Health mindset

We think one of the key factors to health conditions and overweight problem is

people's lifestyle and attitudes, especially people's mindset and awareness of health. In the available dataset, we are interested in the following survey questions related to the health mindset:

Variables	How much do you agree or disagree with the following statements?
Q30rbm:	I go out of my way to buy products that are all natural
Q30raf:	I believe natural foods are both better for me and better for the environment
Q30rah:	I actively seek out information about nutrition and health
Q30rak:	I give up good taste for health benefits
Q30ral:	I give up convenience for health benefits
Q30ran:	I regularly eat organic foods
Q30rao:	I don't allow junk food in my home
Q30raq:	I prefer cooking with fresh food rather than canned or frozen
Q30rat:	I am willing to change stores in order to eat healthier
Q30rbm:	I eat for taste enjoyment more than for health purposes

The chosen questions are all about health mindset and the correlation analysis shows they are correlated with each other (Figure 2). So, we can use Principal Components Analysis to extract one component from these questions as a measure for health mindset (Figure 3).

Figure 2: Correlation matrix for questions related to health mindset

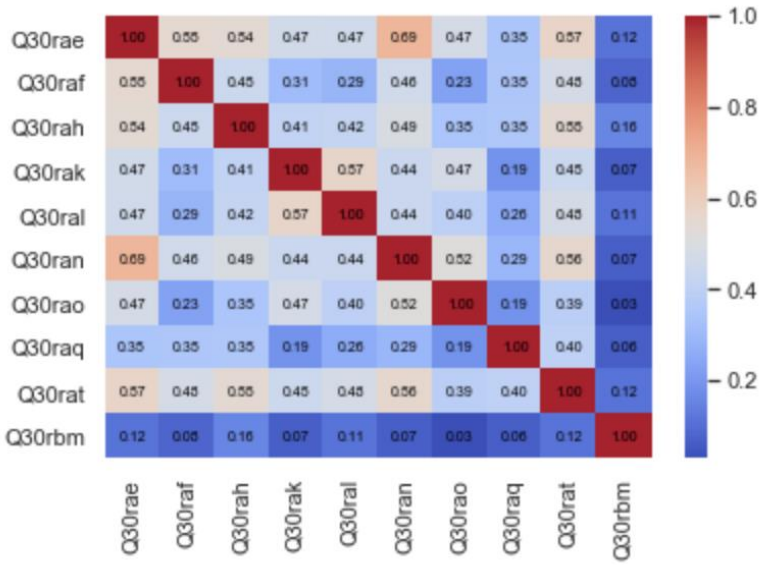


Figure 3: PCA for health mindset

Principal components/correlation — Number of obs = 897; Number of comp. = 10; Trace = 10; Rho = 1.0000
Rotation: (unrotated = principal)

Component	Eigenvalue	Difference	Proportion	Cumulative
Comp1	4.4884	3.42521	0.4488	0.4488
Comp2	1.06319	.0911617	0.1063	0.5552
Comp3	.972031	.241139	0.0972	0.6524
Comp4	.730891	.0737349	0.0731	0.7255
Comp5	.657156	.144276	0.0657	0.7912
Comp6	.51288	.0408215	0.0513	0.8425
Comp7	.472059	.0516775	0.0472	0.8897
Comp8	.420381	.0336714	0.0420	0.9317
Comp9	.38671	.0904128	0.0387	0.9704
Comp10	.296297	.	0.0296	1.0000

Principal components (eigenvectors)

Variable	Comp1	Comp2	Comp3	Comp4	Comp5	Comp6	Comp7	Comp8	Comp9	Comp10	Unexplained
d30ae	0.3903	0.0229	-0.0945	-0.2751	0.0445	0.1284	-0.2275	0.2637	-0.2013	-0.7630	0
d30af	0.3093	0.3110	-0.2550	-0.3758	-0.3507	0.4358	0.4004	-0.0847	0.3106	0.1569	0
d30ah	0.3456	0.2005	0.0121	-0.0061	-0.1647	-0.7852	0.2996	0.3191	0.0490	0.0498	0
d30ak	0.3228	-0.3681	0.2483	0.2165	-0.3202	0.1715	0.3538	-0.1162	-0.6133	0.0716	0
d30al	0.3236	-0.2754	0.2455	0.3693	-0.3586	0.1303	-0.3745	0.2187	0.5362	0.0405	0
d30an	0.3696	-0.0898	-0.1189	-0.3362	0.2521	0.0368	-0.4233	0.2160	-0.2535	0.6109	0
d30ao	0.2954	-0.4211	0.0800	-0.0252	0.6576	0.0296	0.4114	-0.0763	0.3416	-0.0582	0
d30aq	0.2340	0.4571	-0.3057	0.6872	0.2956	0.2152	0.0452	0.1356	-0.1212	0.0424	0
d30at	0.3704	0.1087	-0.0793	0.0456	-0.0144	-0.2566	-0.2864	-0.8299	0.0335	-0.0681	0
d30bm	0.0885	0.4951	0.8269	-0.1093	0.1826	0.1278	-0.0134	-0.0217	-0.0141	0.0295	0

From the results of the PCA, we can see that component 1 accounts for 44.88% of the effects of total 10 variables. We will be using component 1 as a measure for health mindset.

Busy – Leisure level

Joelle Abramowitz (2016) studied the relationship between working hours and BMI index in the U.S and concluded that the longer the working hours, the higher the BMI. According to this study, we think the level of busy - leisure may be a key factor to health conditions and overweight problem. We measure busy – leisure level using the following variables:

Variables	How much do you agree or disagree with the following statements?
Q1ra	Stress keeps me from being the type of person I really want to be - Please tell us how much you agree or disagree with the following statements.
Q1rb	I work much more than I'd like - Please tell us how much you agree or disagree with the following statements.
Q1re	Most nights I don't get enough sleep - Please tell us how much you agree or disagree with the following statements.

Q1rp	I often wish I had more energy - Please tell us how much you agree or disagree with the following statements.
Q1rq	I am so busy; I often can't finish everything I need to in a day - Please tell us how much you agree or disagree with the following statements.
Q3ra	Busy - And which of the following images best represents your day-to-day life right now?
Q3rb	Time for me - And which of the following images best represents your day-to-day life right now?

The chosen questions are all about busy – leisure level and the correlation analysis show they are correlated with each other (Figure 4). So, we can use Principal Components Analysis to extract one component from these questions as a measure for busy – leisure level (Figure 5).

Figure 4: Correlation matrix for questions related to busy – leisure level

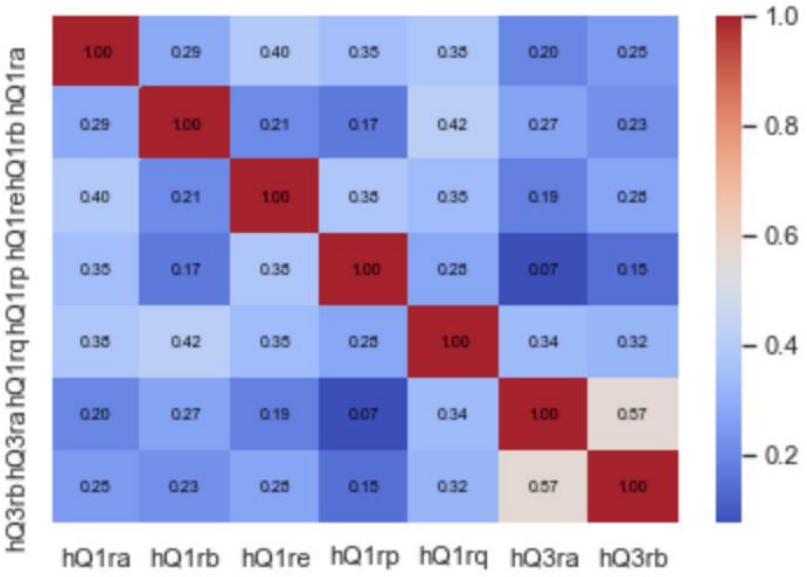


Figure 5: PCA for busy – leisure level

Principal components/correlation — Number of obs = 897
Number of comp. = 7
Trace = 7
Rotation: (unrotated = principal) — Rho = 1.0000

Component	Eigenvalue	Difference	Proportion	Cumulative
Comp1	2.83236	1.65829	0.4046	0.4046
Comp2	1.17408	.336499	0.1677	0.5723
Comp3	.837579	.220629	0.1197	0.6920
Comp4	.61695	.0451445	0.0881	0.7801
Comp5	.571806	.0322558	0.0817	0.8618
Comp6	.53955	.111876	0.0771	0.9389
Comp7	.427674	.	0.0611	1.0000

Principal components (eigenvectors)

Variable	Comp1	Comp2	Comp3	Comp4	Comp5	Comp6	Comp7	Unexplained
d1a	0.4049	-0.2969	-0.0110	-0.4224	-0.7289	0.1286	0.1468	0
d1b	0.3452	-0.0144	-0.7849	0.2077	0.1517	0.4392	-0.0739	0
d1e	0.3938	-0.2748	0.3273	-0.4170	0.6371	0.2630	0.1156	0
d1p	0.3315	-0.4580	0.3239	0.7544	-0.0700	-0.0246	0.0281	0
d1q	0.4392	-0.0304	-0.2535	-0.1099	0.1448	-0.8254	-0.1665	0
d3a	0.3445	0.6111	0.0960	0.1505	-0.0128	-0.0300	0.6891	0
d3b	0.3747	0.5019	0.3134	0.0262	-0.1176	0.1968	-0.6754	0

From the results of the PCA, we can see that component 1 accounts for 40.66% of the effects of total 7 variables. As for all the survey questions, the higher the rating, the more leisure time the participants have. We will be using this component as a measure for leisure.

Food choices / Structure

Food choices and structure is a very direct predictor for overweight and obesity level. We will take into accounts all variables from Q7, which are questions indicating how often the participants typically eat or drink each of the specific foods or beverages.

Demographic and exercise

We will also consider demographic factors and frequency of exercise which may be factors to overweight and obesity level. We will control these factors in our models.

S2	What is your gender?
S3	Please tell us your age?
D5	What is your educational background?
D6	What is your household's annual income before taxes?
D11	In which state do you live?
S6	About how often do you do exercise?

BMI - health mindset clustering

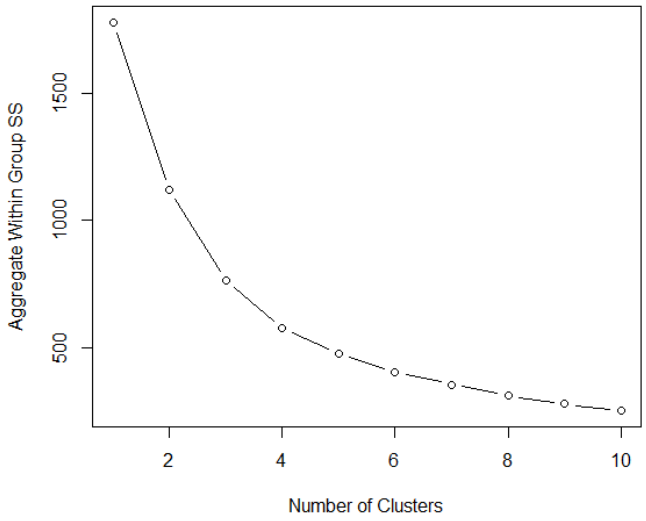
With all dataset well coded, we proceed to analyze the participants by clustering them based on different dimensions. From there, we will future explore why people become overweight and obese , how the foods choices / structure impact BMI index for different clusters of people.

We use k-means algorithm to explore the clusters and clusters distribution on BMI and health mindset dimensions. To optimize the results of k-means clustering, we need to drop outliers and standardize BMI and health mindset score before we apply clustering.

For BMI, we have identified some unreasonable data points, such as someone with 200 pounds body weight but only 3 inches body height. We scaled the data using 21.5 (the middle score of healthy BMI range) as center and drop all the outliers with more than 2 standard deviation shifts around center, and finally we have 897 observations left.

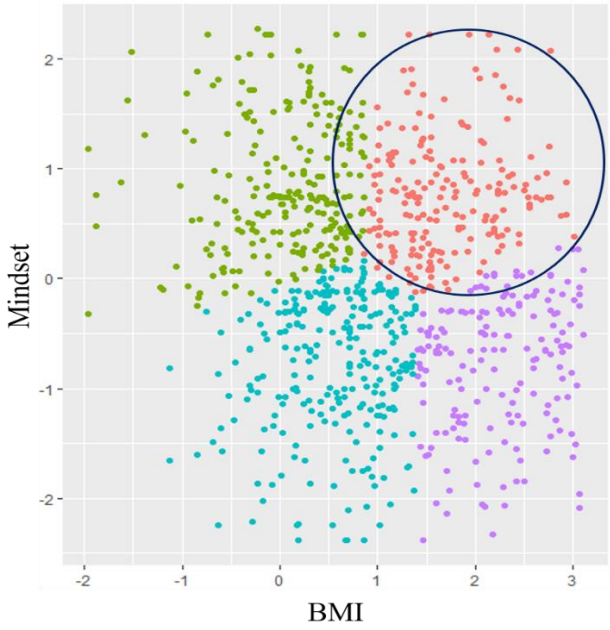
For health mindset dimension, we use the principal component generated from PCA process, and we scale it for k-means clustering.

Figure 6: BMI - health mindset clustering elbow plot



From the elbow plot and two dimensions clustering, we use a reasonable cluster number 4 in the k-means algorithm.

Figure 7: BMI - health mindset clustering plot



As shown in Figure 7, we cluster the participants into four groups, the cluster showed in green dots are the participants with a relatively good health mindset and relatively low BMI; the blue dots are the participants with a relatively poor health mindset but still have a low BMI; the purple dots are the participants with a poor health mindset and high BMI; and the red dots are the participants with a good heath mindset but high BMI, meaning suffering from the overweight and obesity problem. Among

these clusters, we are most interested in the cluster with red dots and will try to investigate why people have good health mindset but have overweight and obesity problem.

In the next step, we will run a regression analysis to find what factors impact BMI. We are setting BMI as the dependent variable, and setting exercise, leisure level, income, food choices, etc. as independent variables.

What are the factors contribute to high BMI (overweight)?

From the above clustering analysis, we divided the survey participants into four clusters. Among which, one group of participants with great health mindset but high average BMI (overweight) attracts our attention. Usually, we assume that people who have an active health mindset and better awareness about their health condition would have a better body shape and BMI. However, our findings have suggested that, for the cluster great health mindset and high BMI, this is not the case. So, there must be other important factors impacting this group of people and cause the overweight problem.

Considering the available variables in our dataset, we assume that the level of leisure-busy, gender, age, frequency of exercise, education level and income level will impact BMI. So, we use these factors as individual variable to do regression analysis to find out the important factors.

	bmi	leisure	mindset	age	exercise	education	income
bmi	1						
leisure	-0.0468	1					
mindset	-0.147	-0.0749	1				
age	0.165	0.198	-0.164	1			
exercise	-0.120	0.0839	0.291	-0.0333	1		
education	-0.0277	0.0155	0.0652	0.0555	0.156	1	
income	-0.0908	-0.0495	0.0988	-0.0130	0.173	0.367	1

Here we use the scaled score of leisure measures (scaleleisure) and health mindset measures (scalemindset) from the Principal Components Analysis (PCA) as the measure of leisure and mindset. The results of the model 1 shows that, besides mindset (scalemindset), level of leisure-busy (scaleleisure), gender (s2), age (s3), frequency of exercise (hq6) and income (d6) have significant impact on BMI index. The coefficient of leisure is -0.41, which means holding other factors fixed, relatively people who are leisure have a lower BMI than people who are busy. People who have

more time usually take better care of their health and wellness, for example, eating balanced meals and fresh foods, sleep well and have less stress. These are all important factors in keeping a good health status. The coefficient of mindset is -0.47, which means holding other factors fixed, people who have a better health mindset have a lower BMI than people who have not. Other coefficients in model 1 also indicate that, relatively, female have a lower BMI than male, older people have a higher BMI than young people, people who exercise more have a lower BMI, and people who are richer have a lower BMI, holding other factors fixed. The results of model 1 make good sense. Normally, female have less muscle ratio than male, so their BMI is lower; older people's metabolism level is lower and easy to gain weight than young people, so their BMI is higher; people who exercise more keep a better body shape and lead to lower BMI; and, people who are richer have more choices on healthy food and healthy life style, which lead to lower BMI.

Figure 8. Regression Analysis of BMI

Linear regression

Number of obs	= 853
F(7, 845)	= 9.82
Prob > F	= 0.0000
R-squared	= 0.0715
Root MSE	= 4.8585

bmi	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
Leisure	-.406943	.1675792	-2.43	0.015	-.7358634	-.0780227
Mindset	-.4691972	.1802195	-2.60	0.009	-.8229275	-.1154669
1.Gender	-.9569605	.3360442	-2.85	0.005	-1.61654	-.2973812
Age	.0630383	.0130452	4.83	0.000	.0374335	.0886431
Exercise	-.1533328	.0708916	-2.16	0.031	-.2924771	-.0141884
Education	.0262646	.1392456	0.19	0.850	-.2470432	.2995724
Income	-.2013383	.0910374	-2.21	0.027	-.3800243	-.0226524
_cons	25.44022	.8131816	31.28	0.000	23.84412	27.03631

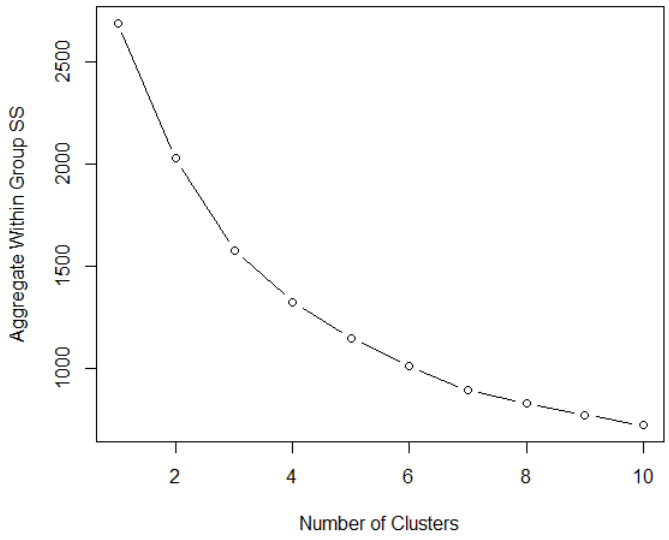
From this regression analysis, we know that besides health mindset, leisure also impacts the BMI index negatively. More precisely, people who are busier have higher BMI and suffer from overweight problem. This is the sub-health problem known as "obesity from overwork", which probably is the problem faced by many middle-age, stressed, long-hour workers, and probably is one of the main reasons why people with a great health mindset but are still overweight. Joelle Abramowitz (2016) studied the relationship between working hours and BMI index in the U.S and concluded that the longer the working hours, the higher the BMI. Since the "obesity from overwork"

problem is prevalent in US recently, we think it may bring many business opportunities, and businesses also can use their expertise to help consumers get rid of "obesity from overwork". So, we would like to introduce a third dimension into our clustering analysis and target our study on the specific group of people who have a greater health mindset, higher BMI (overweight), and lower leisure (busy).

BMI - health mindset - leisure clustering

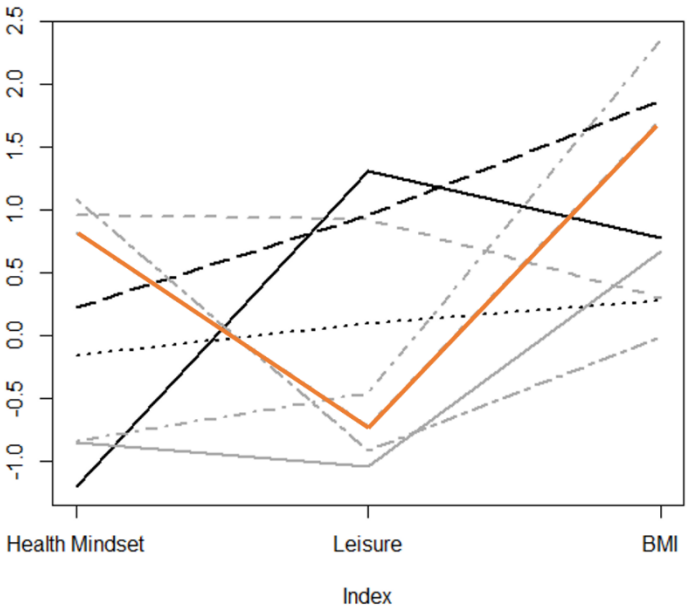
For leisure dimension, we use the principal component generated from PCA process, and we scale it for k-means clustering.

Figure 9: BMI – Mindset - Leisure clustering elbow plot



From the elbow plot and three dimensions clustering, we use a reasonable cluster number 8 in the k-means algorithm.

Figure 10: BMI – Mindset - Leisure clustering plot



As shown in Figure 10, we cluster the participants into eight groups, different cluster has different distribution on these three dimensions. Among these clusters, we are most interested in the cluster with red line highlight, which is the cluster with good health mindset, low leisure (busy), and high BMI (overweight and obesity). Next, we will focus on this target cluster and explore what factors affect their BMI.

How the foods choices / structures influence people's BMI?

In our survey dataset, among all available variables, foods choices are the most related factors for BMI. What and how much a person eats and drinks is an important factor of the body weight and composition. We are going to analyze whether some specific categories of foods contribute to higher or lower BMI, and whether there is heterogeneity for different groups of people.

Before we build model, we did many preliminary studies and found that there are four categories of foods impact people's BMI the most: frozen and canned food, carbohydrate, candy and sugar, and beer, which is consistent to our nutrition knowledges. We aggregated the frequency of people's answers in question 7 to get the variables indicate the frequency of have these categories of foods. For example, frozen and canned food includes all kinds of frozen food in question 7, and "CO: Canned soup", "DP: Canned fruit", "DQ: Canned vegetables", "DR: Canned pasta"; carbohydrate includes majority items in "Bakery" such as "DA: Bakery bread", "DB: Premade

desserts", "DC: Bagels" etc.

To study the relationship between specific foods category and BMI, we regress BMI on the aggregated frequency of the food category, adding the interaction items and keeping other factors controlled. The results of model 2 – model 5 shows the relationship between BMI and four specific foods categories: frozen and canned food, carbohydrate, candy and sugar, and beer. Because we would like to target our study on the specific cluster of people with a greater health mindset, higher BMI (overweight), and lower leisure (busy), we incorporated our target cluster into the model as a dummy variable that interacts with the other factors in our model. We care more about the interaction effect since it will give us the heterogeneity and characteristics of our targeted cluster, which will be the ground on which we can target them and make marketing strategy.

VARIABLES	Model 1 bmi	Model 2 bmi	Model 3 bmi	Model 4 bmi	Model 5 bmi	Model 6 bmi
Leisure	-0.407**	0.011	0.006	0.014	0.044	-0.013
	(0.168)	(0.177)	(0.177)	(0.176)	(0.175)	(0.178)
Mindset	-0.469***	-0.921***	-0.897***	-0.892***	-0.975***	-0.897***
	(0.180)	(0.183)	(0.181)	(0.183)	(0.177)	(0.185)
Female	-0.957***	-0.758**	-0.787**	-0.759**	-0.881***	-0.805**
	(0.336)	(0.315)	(0.314)	(0.314)	(0.319)	(0.325)
Age	0.063***	0.056***	0.056***	0.056***	0.054***	0.055***
	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)
Exercise	-0.153**	-0.152**	-0.157**	-0.153**	-0.159**	-0.153**
	(0.071)	(0.073)	(0.073)	(0.072)	(0.073)	(0.073)
Education	0.026	-0.121	-0.099	-0.113	-0.098	-0.102
	(0.139)	(0.144)	(0.144)	(0.144)	(0.144)	(0.144)
Income	-0.201**	-0.185*	-0.190**	-0.198**	-0.173*	-0.173*
	(0.091)	(0.086)	(0.085)	(0.084)	(0.088)	(0.087)
0Target*Age		0.000	0.000	0.000	0.000	0.000
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
1Target*Age		0.007	0.008	0.007	0.023	0.011
		(0.020)	(0.021)	(0.020)	(0.021)	(0.020)
0Target*Exercise		0.000	0.000	0.000	0.000	0.000
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
1Target*Exercise		0.167	0.138	0.128	0.120	0.135
		(0.144)	(0.145)	(0.144)	(0.144)	(0.141)
0Target*Education		0.000	0.000	0.000	0.000	0.000
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
1Target*Education		0.691***	0.668***	0.659***	0.620***	0.617***
		(0.232)	(0.232)	(0.235)	(0.230)	(0.227)
0Target*Income		0.000	0.000	0.000	0.000	0.000
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
1Target*Income		0.084	0.081	0.085	0.110	0.084
		(0.157)	(0.157)	(0.157)	(0.165)	(0.162)
Frozen and canned food		-0.0006**				0.000
		(0.000)				(0.000)
0Target*Frozen and canned food		0.000				0.000
		(0.000)				(0.000)

1Target*Frozen and canned food	0.0014***				0.001
	(0.000)				(0.001)
Candy and sugar		-0.001***			-0.001
		(0.000)			(0.001)
0Target*Candy and sugar		0.000			0.000
		(0.000)			(0.000)
1Target*Candy and sugar		0.003***			0.001
		(0.001)			(0.001)
Carbohydrate			-0.001***		-0.000
			(0.000)		(0.000)
0Target*Carbohydrate			0.000		0.000
			(0.000)		(0.000)
1Target*Carbohydrate			0.002***		-0.000
			(0.000)		(0.001)
Beer				-0.002**	-0.001
				(0.001)	(0.001)
0Target*Beer				0.000	0.000
				(0.000)	(0.000)
1Target*Beer				0.007***	0.004**
				(0.001)	(0.002)
Constant	25.906***	26.034***	26.086***	25.697***	26.077***
	(0.814)	(0.821)	(0.817)	(0.787)	(0.824)
Observations	853	853	853	853	853
R-squared	0.192	0.194	0.192	0.192	0.198

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

In model 2 – model 5, we have the following findings:

- In model 1, without target cluster interaction, education level has no significant impact on BMI, but in model 2 – model 5, when we focus on targeted cluster, the education level has significant positive impact on BMI. It means, for people who have a better health mindset, overweight and busy, the higher the education, the higher the BMI. The possible reason is, comparing to people with low education level, people with higher education are more likely to pursue a career in an office environment rather than manual labor, and relatively they work longer hours and potentially suffer from more mental pressure than labor workers. As a result, it is more common for people with higher education to have "obesity from overwork" problem, and that results in a higher BMI.
- In model 2, the coefficient of "Frozen and canned food" is significantly negative, which means, for other clusters, eating frozen and canned food frequently will not cause overweight problem. However, the coefficient of interaction item "1Target*Frozen and canned food" is significantly positive, and if we take into consideration the effect of "Frozen and canned food" on BMI for our target

cluster, the coefficient is positive. Which means that, though frequently eating frozen and canned food will not cause overweight problem for other clusters, it will cause overweight issue for our target cluster, who are busy, overweight, and have a better health mindset.

- In model 3-5, we have similar findings as model 2, the effect of "Candy and sugar", "Carbohydrate", and "Beer" on BMI of our target cluster is positive. Even frequently having "Candy and sugar", "Carbohydrate", and "Beer" will not cause overweight problem for other clusters, it will definitely cause overweight issue for our target cluster, people who has a greater health mindset, overweight and busy.
- In model 6, we put these four specific food categories together into our model, but only the effect of beer in our target cluster is significant. And the R-squared is 0.198, just slightly higher than the model 2-5 with only one specific food category in model. From the covariance matrix of the aggregated frequency of these four food categories, there are high correlation amongst these four food categories. So, we can infer that, first, though these four food categories are very different, the latent factors of why these foods categories cause overweight problem specific to our target cluster are same; second, we use aggregated frequency as the measure for food categories, the specific variance of specific category may be neutralized in the aggregation process. As a result, the model 6 has multilinearity problem.

	Frozencan	Candysugar	Carbohydrate	Beer
Frozencan	1			
Candysugar	0.708	1		
Carbohydrate	0.807	0.761	1	
Beer	0.586	0.558	0.543	1

We can solve this problem by finding the common latent factor behind these four food categories. In model 7, we use the main component from PCA of these four categories in regression. We can see that the common factor will increase the BMI index and cause overweight problem, and the impact is especially strong for our target cluster. This common factor is a component of those four categories of foods. It may be interpreted as unhealthy, high-calories foods.

Linear regression

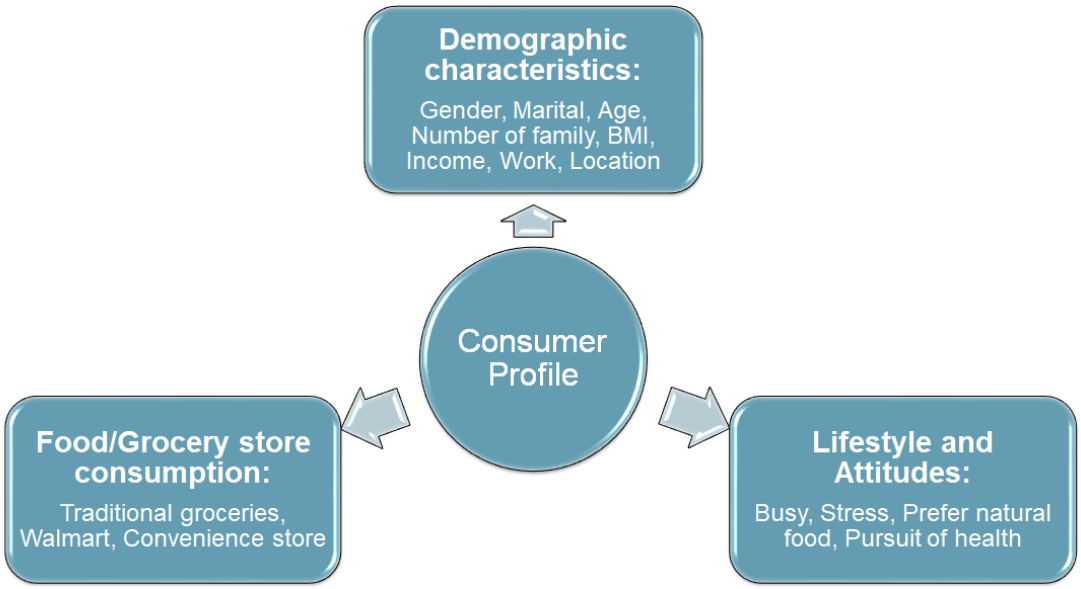
Number of obs	= 853
F(13, 839)	= 25.18
Prob > F	= 0.0000
R-squared	= 0.1917
Root MSE	= 4.5493

bmi	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
Leisure	-.0077241	.1775476	-0.04	0.965	-.3562137	.3407655
Mindset	-.8769611	.1842145	-4.76	0.000	-1.238536	-.5153858
1.Gender	-.8182458	.3138801	-2.61	0.009	-1.434328	-.2021634
Age	.0530318	.0132324	4.01	0.000	.0270593	.0790044
Exercise	-.1556107	.07254	-2.15	0.032	-.2979918	-.0132295
Education	-.1313543	.1446316	-0.91	0.364	-.4152365	.1525278
Income	-.1806791	.0959418	-1.88	0.060	-.3689932	.0076349
Commonfactor	-.3480398	.1204423	-2.89	0.004	-.5844434	-.1116362
Target#c.Age						
1	.0241389	.0210487	1.15	0.252	-.0171755	.0654533
Target#c.Exercise						
1	.1549119	.1464137	1.06	0.290	-.1324683	.4422921
Target#c.Education						
1	.7447365	.2328053	3.20	0.001	.2877873	1.201686
Target#c.Income						
1	.1128895	.1592026	0.71	0.478	-.1995926	.4253717
Target#c.Commonfactor						
1	.6963494	.1489896	4.67	0.000	.4039132	.9887855
_cons	25.62752	.7824142	32.75	0.000	24.0918	27.16323

From the above models, we can conclude that, people in our target cluster, who have good health mindset, overweight and busy, are especially responsive to the unhealthy, high-calories foods such as frozen and canned food, carbohydrate, candy and sugar, and beer. Though frequently having these foods will not cause overweight problem for other clusters, it will cause overweight issue for our target cluster. Though with a great health mindset, busy and overwork still led our cluster to obesity.

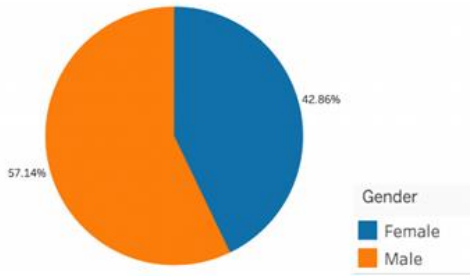
Analysis of Target Segment

According to our cluster results, we have decided to target customers who have good healthy mindset, in poor physical status and with limited leisure time. In this section, we will summarize the customer profile of our target segment based on their demographic characteristics, lifestyle and attitudes, food/grocery store consumption.

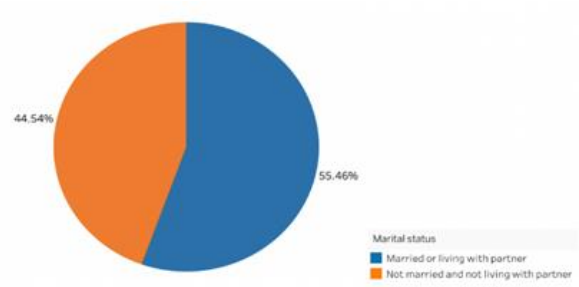


Demographic characteristics:

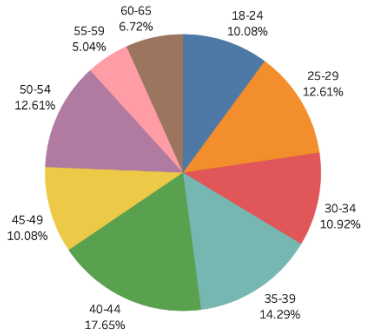
Gender: Among the 119 people in this segment, 68 (57.14%) of them are Male while 51 (42.86%) are female.



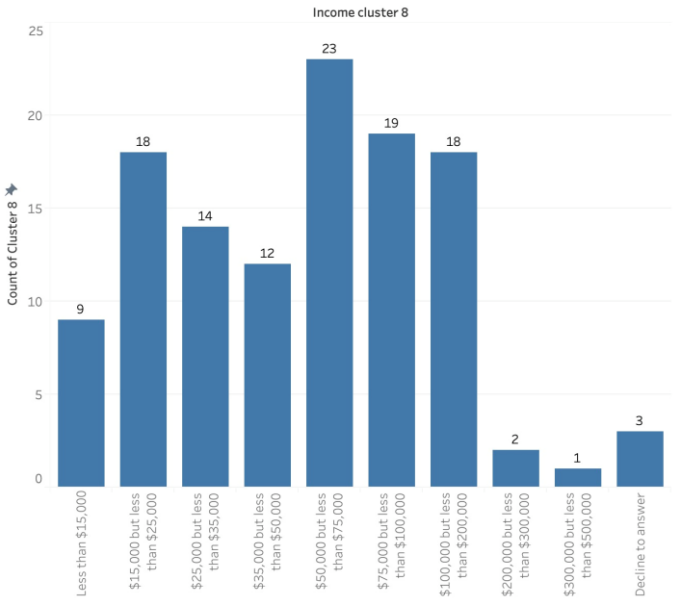
Marital Status: 66 (55.46%) people have married or living with partners while 53 (44.54%) people live alone.



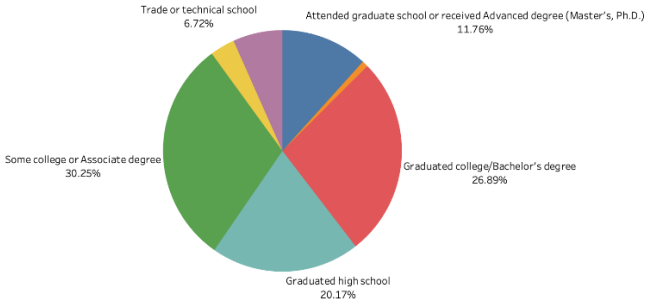
Age: The average age of cluster 8 is 39.91.



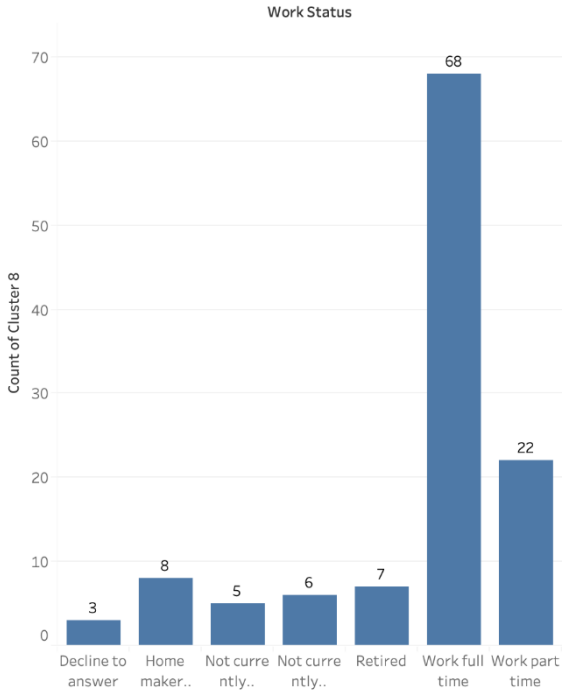
Income: 53 people in this group earned less than $50,000 per year, which accounts for 45% of the total group.



Education Background: 36 people achieved college or associate degrees, followed by graduated college/bachelor's degree (32), high school diploma (24). In addition, there are 14 people who received advanced degrees such as master's or Ph.D in this group.



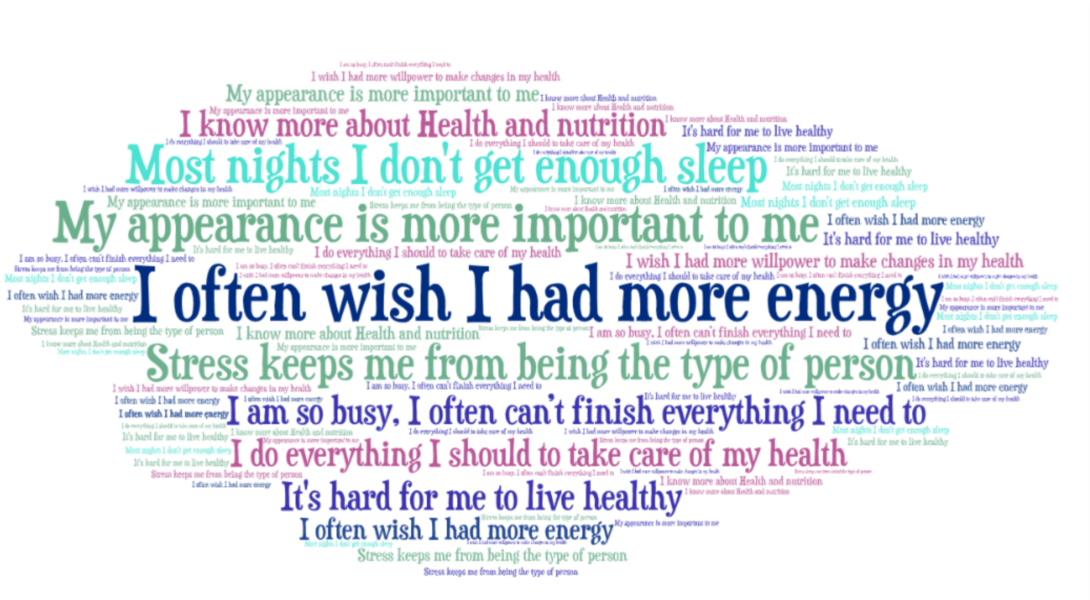
Work Status: 90 people have a job (68 people work full time, 22 work part time).



Location: 14 people come from New York, 13 come from California.

Lifestyle and Attitudes:

From the survey data, we found that generally people in cluster 8 care about their appearance and health, however, they have trouble living a healthy life. In addition, compared with other clusters, they seem to be more stressful and busier. On average, more than 30% people in this group chose “strongly agree” for questions like “ I often wish I had more energy”, “ I often wish I had more energy”, “ I am so busy, I often can’t finish everything I need to in a day”, “Stress keeps me from being the type of person I really want to be”, “Most nights I don’t have enough sleep” and “ I wish I had more willpower to make changes in my health”. We have created the following word cloud to highlight the attitudes that strongly resonated with our target cluster.

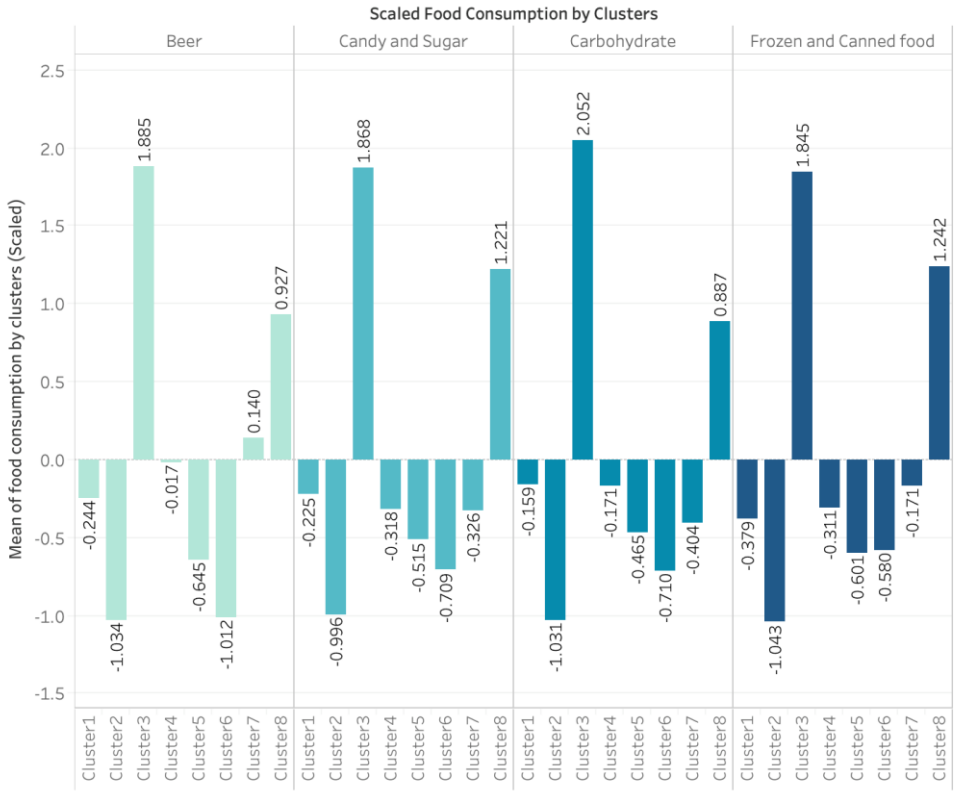


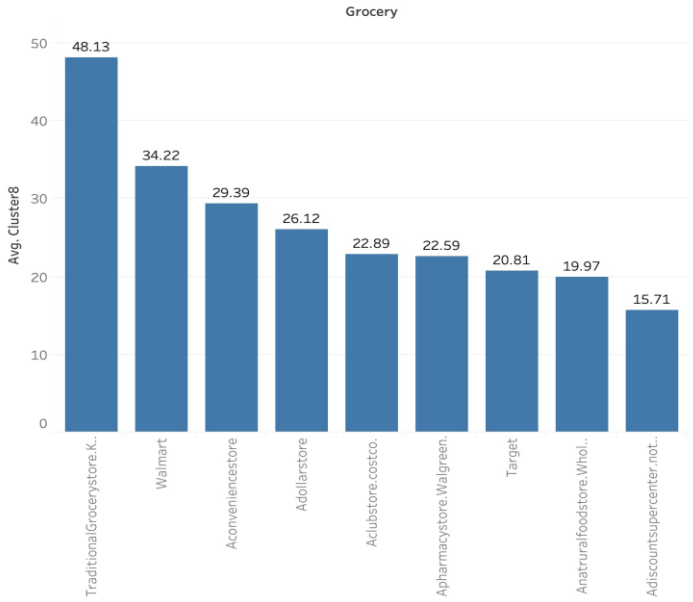
In terms of food choice and purchase, comparing to other clusters, our target cluster has shown a stronger agreement level towards questions like" I prefer cooking with fresh food rather than canned or frozen", " I am willing to change stores in order to save more money on my groceries", "I believe natural foods are both better for me and better for the environment", " I like to reward myself with something tasty", " I am willing to change stores in order to eat healthier", " I seek out variety in my everyday life", " I like to spend time in a grocery store exploring what is new", " When it comes to food, I'm primarily a price shopper", and " I enjoy trying different types of food".

In addition, based on the survey results, our target cluster has shown strong intention to follow healthy dietary guidelines like Low sodium, Low carb, Low calorie, Locally sourced, Diabetic etc.

In conclusion, people in cluster 8 live with a heavy workload every day. Though they care about healthy eating habits and lifestyle, they are not satisfied with their current health status, which have incentivized a strong will to make changes. Besides, people in cluster 8 like to explore grocery stores and are curious about new products. They care about tastiness as well as healthiness when it comes to foods.

Food/Grocery store consumption:





Based on the regression model we built, we found that Carbohydrate, Beer, Candy and sugar, Frozen and canned food are highly preferred by cluster 8. However, these food choices have significant negative effects on their BMI. Among them, Frozen and canned food, Candy and sugar are the two most frequently consumed food categories.

Compared with the lifestyles and attitudes they have; we have found that the behaviors of our target cluster do not align with their mindset. For example, although they have claimed to prefer fresh food over frozen or canned foods in one part of the survey, they have admitted to frequently consume these categories in other parts of the survey.

In terms of grocery store consumption, Traditional Grocery Store like Kroger or Safeway, Walmart, Convenience Store and Dollar Store are the top 4 grocery stores our target cluster frequently visit. And we also found that, on average, our target cluster go to grocery stores more frequently than other clusters. One possible reason could be that they go to grocery stores frequently but stay for a short time. In addition, exploring stores could be one way for them to relax if considered their lifestyle attitudes.

Strategy and Recommendation

According to our analysis on cluster 8 above, consumers in this group have a good mindset of health and have a strong willingness to lead a healthy lifestyle. However, the average BMI for this group is over 30, which indicates that the individuals in our target cluster are either exposed to the risk of obesity or experiencing obesity. We have found the tight schedule (i.e. being busy or having less leisure time), and preference for unhealthy and high-calories foods (e.g. frozen and canned food, carbohydrate, candy and sugar, and beer), have contributed to the inconsistency between the mindset and choices of our target consumer.

Further analysis on food choice and consumption frequency has given us insights into our target consumer's behavior. In addition to necessary consumption of food, consumers in this segment frequently buy Carbohydrate, Beer, Candy and sugar, Frozen and canned food, regardless the negative effects these food choices have on their BMI. Besides, they like to explore grocery stores and are curious about new food products.

Product features recommendation

Based on the analysis of the target consumer, we will focus our recommendation on providing a healthy alternative that fit into our target consumer's tight schedule. More precisely, we recommend SRG clients who are providing food products and services to design a new "healthy fast food" product focusing on the following features:

- Convenient. As our target consumers are busy and have limited time, this feature could be crucial when they make buying decisions. This product should require minimal preparation time.
- Healthy. This feature aims to resonate with our target consumer's great mindset of health and their willingness to be healthy. Our suggestion for food selection criteria is a combination of fresh material with high-protein and low-calorie/ low-sodium based on the nutritional requirements of the body.
- Flexible ordering system. This feature will help us better fit into the tight schedule of our target consumers. This feature should allow consumers to

place an order and also schedule order pick-up time ahead.

- A combination of traditional and innovative flavor. This feature aims to meet the target consumer's preference on tasty flavors.
- Fancy design with an attractive product name. As our target consumers like to explore grocery stores and are curious about new products, this feature should help us attracts their attention.
- Appropriate price. Considering the income group that our target consumers mostly fall in, we suggest that the products should be economic.

Marketing Strategies

- Providing delivery services through online ordering or app ordering. This service should also help us fit into our target consumer's tight schedule. This service could target first in metropolitan area with dense population.
- Collaborating with grocery stores such as Kroger or Walmart where our target consumers frequently visit. More precisely, we can collaboratively offer partner discounts, promotion event collaborations and etc.
- Email and social media marketing. Considering the age group of the target consumers, digital marketing should be an effective tool for branding and marketing.

Reference:

US Census Bureau. (2018, July 25). Income and Poverty in the United States: 2015. Retrieved from https://www.census.gov/library/publications/2016/demo/p60-256.html

Body Mass Index (BMI). (2020, April 10).

Retrieved from https://www.cdc.gov/healthyweight/assessing/bmi/index.html

Products - Data Briefs - Number 360 - February 2020. (2020, February 27).

Retrieved from https://www.cdc.gov/nchs/products/databriefs/db360.htm

Abramowitz, J. (2016). The connection between working hours and body mass index in the US: a time use analysis. Review of Economics of the Household, 14(1), 131-154.

Appendix

```
##Final Project Group 2

##Python code for clean dataset and cluster


import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

from sklearn import metrics


#select data

df = pd.read_csv('rawdata.csv')

leisure = df[['hQ1ra','hQ1rb','hQ1re','hQ1rp','hQ1rq','hQ3ra','hQ3rb']]

mindset = df[['Q30rae','Q30raf','Q30rah','Q30rak','Q30ral','Q30ran','Q30rao','Q30raq','Q30rat','Q30rbm']]

food = df.iloc[:,56:187]

demographic = df[['S3','D5','D6','D9','D10ra','D10rb','D11']]

exercise = df['Q6']

store = df.iloc[:,187:196]


#convert categorical to numerical

for i in range(mindset.shape[1]):

    mindset.iloc[:,i] = mindset.iloc[:,i].map({'Strongly agree':4, 'Agree somewhat':3, 'Disagree somewhat':2,'Strongly disagree':1})

mindset['Q30rbm'] = mindset['Q30rbm'].map({1:4, 2:3, 3:2,4:1})

for i in range(food.shape[1]):

    food.iloc[:,i] = food.iloc[:,i].map({'Every day':365,'Several times a week':156,'Several times a month':72,'Once a month':12,'Several times a year':5,'Once a year or less':1,'Never':0})

for i in range(store.shape[1]):

    store.iloc[:,i] = store.iloc[:,i].map({'Two times a week or
```

```
more':100,'About once a week':50,'About once every two weeks':24,'About once a month':12,'Less than monthly':6,'Not in the past several months':0})


exercise = df['Q6'].map({'Rarely or never':-2, 'Occasionally, but less than once a month':-1, 'A few times each month':1,'1 time per week':2,"2-3 times per week":3,"4 or more times per week":5})


demographic['D5'] = demographic['D5'].map({'Some high school or less':1, 'Graduated high school':2, 'Trade or technical school':3,'Some college or Associate degree':4,"Graduated college/Bachelor's degree":5,"Attended graduate school or received Advanced degree (Master's, Ph.D.)":6,'Decline to answer':0}).astype(int)

demographic['D6'] = demographic['D6'].map({'Less than $15,000':1, '$15,000 but less than $25,000':2, '$25,000 but less than $35,000':3,'$35,000 but less than $50,000':4,"$50,000 but less than $75,000":5,"$75,000 but less than $100,000":6,'$100,000 but less than $200,000':7,'$200,000 but less than $300,000':8,'$300,000 but less than $500,000':9,'$500,000 or over':10,'Decline to answer':0}).astype(int)

demographic['inch'] = demographic['D10ra'] *12 + demographic['D10rb']

demographic['BMI'] = demographic['D9'] /demographic['inch']/ demographic['inch'] *703

demographic['BMI'] = demographic['BMI'].round(2)


#cleandata without drop

cleandata = pd.concat([leisure,mindset,food,exercise,demographic,store],axis=1)


#drop outlier and unanswered value

df1 = cleandata[cleandata['BMI'] < 50]

df2 = df1[df1['BMI'] > 10]

df3 = df2[(~df2['D5'].isin([0]))]

df4 = df3[(~df3['D6'].isin([0]))]
```

```
df4.to_csv(r'/Users/apple/Desktop/df4.csv', index = False)


#correlation matrix

sns.heatmap(leisure.corr(),annot=True,fmt='.2f',cmap=
'coolwarm',annot_kws={'size':7, 'color':'black'})

sns.heatmap(mindset.corr(),annot=True,fmt='.2f',cmap=
'coolwarm',annot_kws={'size':7, 'color':'black'})


#kmeans

sns.set()

from sklearn.cluster import KMeans

kmeans.inertia_

wcss = []

for i in range(1,7):

    kmeans = KMeans(i)

    kmeans.fit(finalDf) #finalDf is from PCA result

    wcss_iter = kmeans.inertia_

    wcss.append(wcss_iter)

number_clusters = range(1,7)

plt.plot(number_clusters,wcss)

plt.title('The Elbow Method')

plt.xlabel('Number of clusters')

plt.ylabel

plt.show()


kmeans = KMeans(4)

kmeans.fit(finalDf)

identified_clusters = kmeans.fit_predict(finalDf)

data_with_clusters = finalDf.copy()

data_with_clusters['Cluster'] = identified_clusters


plt.scatter(data_with_clusters.iloc[:,0],
data_with_clusters.iloc[:,1],  c=data_with_clusters['Cluster'],
cmap='rainbow')
```

```
plt.xlim(-5,5)

plt.ylim(-6,6)

plt.xlabel('BMI')

plt.ylabel('pca mindset')

plt.show()


#export kmeans result

r1 = finalDf[(kmeans.labels_ ==0)]

r2 = finalDf[(kmeans.labels_ ==1)]

r3 = finalDf[(kmeans.labels_ ==2)]

r4 = finalDf[(kmeans.labels_ ==3)]

r1 = pd.DataFrame(data = r1)

r1['cluster'] = 1

r2 = pd.DataFrame(data = r2)

r2['cluster'] = 2

r3 = pd.DataFrame(data = r3)

r3['cluster'] = 3

r4 = pd.DataFrame(data = r4)

r4['cluster'] = 4


BMI_mindset = pd.concat( [r1,r2,r3,r4], axis=0 )

BMI_mindset['id'] = BMI_mindset.index

BMI_mindset.to_csv(r'/Users/apple/Desktop/BMI_mindset.csv',
index = False)


###Stata code for PCA and regression

1 /* Buan 6337 Project */
2 /* Group 2*/
3
4 ** PCA
5 use
"D:\Box\2020_Spring\Predictive_analytics_SAS\Final_project\project_rc
ode\retry\reg\dataset.
dta", clear
6
7 egen d1a = std(hq1ra)
```

```
egen d1b = std(hq1rb)
egen d1e = std(hq1re)
egen d1p = std(hq1rp)
egen d1q = std(hq1rq)
egen d3a = std(hq3ra)
egen d3b = std(hq3rb)

egen d30ae = std(q30rae)
egen d30af = std(q30raf)
egen d30ah = std(q30rah)
egen d30ak = std(q30rak)
egen d30al = std(q30ral)
egen d30an = std(q30ran)
egen d30ao = std(q30rao)
egen d30aq = std(q30raq)
egen d30at = std(q30rat)
egen d30bmrev = std(q30rbm)
gen d30bm = 0 - d30bmrev

pca d1a d1b d1e d1p d1q d3a d3b
predict leisure
egen scaleleisure = std(leisure)

pca d30ae d30af d30ah d30ak d30al d30an d30ao d30aq d30at d30bm
predict mindset
egen scalemindset = std(mindset)

** correlation
corr bmi scaleleisure scalemindset s3 hq6 d5 d6
ssc install logout
 logout, save(corr1) word replace: pwcorr bmi scaleleisure
scalemindset s3 hq6 d5 d6

** model 1
regress bmi scaleleisure scalemindset i.s2 s3 hq6 d5 d6, vce(robust)
est store Model1


** aggregate
** frozencannedfood candysugar carbohydrate beer alcohol
*q7_1raa-q7_1ral q7_6rco q7_9rdp q7_9rdq q7_9rdr
*q7_10rdz-q7_11reh
*q7_7rcv-q7_9rdh
*q7_13req-q7_13res
*q7_13req-q7_13rfb

** correlation
 logout, save(corr2) word replace: pwcorr bmi scaleleisure
scalemindset s3 hq6 d5 d6
frozencannedfood candysugar carbohydrate beer
 logout, save(corr3) word replace: pwcorr frozencannedfood
candysugar carbohydrate beer

** model 2,3,4,5,6
 regress bmi scaleleisure scalemindset i.s2 s3 hq6 d5 d6
frozencannedfood i.clu8#c.(
frozencannedfood s3 hq6 d5 d6), vce(robust)
est store Model2
regress bmi scaleleisure scalemindset i.s2 s3 hq6 d5 d6 candysugar
i.clu8#c.(candysugar s3
```

```
hq6 d5 d6), vce(robust)
61 est store Model3
62 regress bmi scaleleisure scalemindset i.s2 s3 hq6 d5 d6 carbohydrate
i.clu8#c.(
carbohydrate s3 hq6 d5 d6), vce(robust)
63 est store Model4
64 regress bmi scaleleisure scalemindset i.s2 s3 hq6 d5 d6 beer
i.clu8#c.(beer s3 hq6 d5 d6),
vce(robust)
65 est store Model5
66 regress bmi scaleleisure scalemindset i.s2 s3 hq6 d5 d6
frozencannedfood candysugar
carbohydrate beer i.clu8#c.(s3 hq6 d5 d6 frozencannedfood candysugar
carbohydrate beer),
vce(robust)
project_code.do - Printed on 5/7/2020 8:12:28 PM
Page 2
67 est store Model6
68 outreg2 [Model1 Model2 Model3 Model4 Model5 Model6] using reg, word
dec(3) replace
69
70 ** PCA for the commonfactor
71 corr frozencannedfood candysugar carbohydrate beer alcohol
72 pca frozencannedfood candysugar carbohydrate beer alcohol
73 predict Commonfactor
74
75 ** model 7
76 regress bmi scaleleisure scalemindset i.s2 s3 hq6 d5 d6 Commonfactor
i.clu8#c.(
Commonfactor s3 hq6 d5 d6), vce(robust)
77 est store Model7
78 outreg2 Model7 using table3, word dec(3) replace
```