**Is a Plant-Based Diet Overall Healthier Than an Omnivore Diet**

**Jill Mercer (jmm1037)**

Introduction

On October 16, 2019 “The Game Changer” [1] movie was released for streaming on Netflix. “The Game Changer” documented vegan diets among top athletes and compared them against athletes that maintained a more widely accepted, omnivore diet. Interviews were conducted with body builders, one such being Arnold Schwarzenegger, who would consume large amounts of animal products in order to attain the required protein to build muscle. Schwarzenegger has since switched to a plant-based diet and claims to still receive the required protein he had consumed during his lifting years.

After watching this documentary, I found it fascinating being a novice athlete myself, that there was such a stigma between plant-based and meat-based diets. From a young age, people are told to drink milk to get strong and eat eggs or soy-based protein drinks in order to rebuild muscle after a workout. These points aren’t without merit however, the health differences experienced between a plant and meat-based diet is interesting. Is a plant-based diet overall a healthier option than the typical omnivore diet?

Related Works

A research review was conducted to determine studies involving the health of vegetarians and vegans. This was performed in order to find any research previously undertaken that would help answer the question: Are vegans and or vegetarians healthier than people that eat meat? Sources for this review were retrieved through published articles from Google and Google Scholar. The focus for the research was to determine what data had been used to find results and to assess health risks in both plant-based and meat-based diets. For purposes of inclusion, the term plant-based will be used throughout the study as an umbrella term to mean vegetarian and vegan diets.

Vegetarians and vegans have a stigma of being overall healthier than the standard omnivore diet presented in the west. This idea is largely based on the ease in ability to get processed food from fast food chains or stores. Body mass index (BMI) is notably lower in individuals adopting a plant-based diet as can be seen in the study conducted by Nutrienet-Santé Study [2]. The Nutrienet- Santé Study is a large survey database that was launched online in France starting in 2009. There are limited public information and studies that have been run on vegetarian and vegan health that are not in survey format. The dataset for the Nutrienet study is not publicly available however, an overall summary of the data is.

As noted in previous research “Vegan Diet, Subnormal Vitamin B-12 Status and Cardiovascular Health” [4] there is a large association with Vitamin B-12 deficiency and a plant-based diet. Vitamin B-12 deficiency is associated with fatigue, anemia and memory loss. Vitamin B-12 is found largely in animal-based products such as eggs and milk, lower levels are found in people who adopt a vegan (no animal product) diet. There are many tests to prove how a supplement within these plant-based diets can correct this, although it is shown throughout gathered text that people who have plant-based diets are unaware of this deficiency [5]. Along with Vitamin B-12, plant-based diets are known to have a decrease in Vitamin D levels if inadequate fruits and vegetables are consumed.

A plant-based diet has been shown to decrease blood pressure in relation to meat eaters, notably mentioned in “Health Benefits and Risk Associated with Adopting a Vegetarian Diet” [3]. This decrease in blood pressure results in lower risks for, heart attack, strokes and heart disease among others. There are conflicting articles with this information over the years which will need to be focused on during this study.

Another analysis was done on the approaches on using survey data. There were three main elements that are described to ensure survey data is appropriately distributed: weights, stratification and clustering [8]. The most important method for data distribution is weight, as survey analysis tends to have some level of no response, which is not found in most data sets. Data that is acquired during a survey is a subset of a larger group in which the surveyors goal is to obtain an understanding of, when non response is received it can create a disbalance between variables that represent the larger population such as age and gender.

The Data

For the purposes of this study, an ideal dataset would include variables that show weight and or BMI of vegetarians and vegans. The set will also hold levels of B-12 and other vitamin deficiencies (i.e. Vitamin D and iron) as well as levels of vitamins that are higher in plant-based diets (Magnesium and Vitamin C) [2]. There will be a focus on blood pressure and overall health of individuals that participate in a plant-based diet vs those that have meat in their diet. A data set that included all these variables for omnivores, vegetarians and vegans would be preferred, as it would be easier to determine where there are strengths and weaknesses within each.

A dataset “Study of Current and Former Vegetarians and Vegans” from Faunalytics was chosen to contain the most information necessary to answer and cover the questions that many former research papers have attempted to answer. Datasets involving information about vegetarians and vegans are few, as there has not been a large medical test done which contains accurate results on the individuals and has then been made available to the public. Instead, most datasets are in a survey format and the results contained are completely self-answered and thus there may be some inaccuracies with the results.

This set contains 12,264 responses and after the data was cleaned, the dataset contained 11,429 data points. Each point contains 486 possible responses however this number will be limited when conducting the study as there are too many possibilities to report on. The original data file comes in a .sav format which can be opened using IBMs SPSS program or by importing into SAS. The contents have been exported to an excel file with a filter on the many possibilities of the variables.

For easier data manipulation, only select responses from the 11,429 respondents will be used. Each variable in the dataset is a question that was then answered categorically (1-x) where each number corresponds with a different answer, other questions were answered either yes, no or N/A. There were too many variables to list out (in total there are 486) all of which were a selection and contain largely nominal/categorical data with the remaining being ordinal data (any of the options that contain agree, disagree, neither). The dataset did not contain questions that were not defined by the survey and does not contain any medical readings such as blood pressure, vitamin b12 levels ect.

Ideally, a dataset that contained nominal medical findings in vegetarian/vegans would be preferred however, as it stands, there is not a dataset that contains this information on a large scale. The dataset in use for this study does ask the questions such as “did you experience protein/vitamin b12/ calcium deficiency while maintaining your diet”. There are many other questions that relate to whether the respondent experienced improved or worsening health from factors that induce medical issues that cause concern when switching to plant-based diets.

Data Preparation

The survey questions were split out into 5 diet plans; diet one (all diets), diet two (omnivores), diet three (unverified vegetarian recidivists & unverified vegan recidivists), diet four (unverified vegetarian recidivists) and diet 5 (unverified vegetarians). These survey results were then gathered in an all-encompassing former and current vegans’ spreadsheet, without labels as to which IDs are currently vegan or omnivore.

A filter was applied to eliminate any fields that did not contain an age and gender to associate with the respondent id. This limited the number respondents to 1,799 and although this is a large decrease in the original number of 11,429, it was deemed necessary in order to both process the data efficiently and create clusters based on age.

Survey data tends to weight nonresponses such as age and gender however, in the case of this populations (vegans and past vegans) the ratio of young, middle and older aged individuals are not known in a plant-based diet and thus makes it hard to weight. Instead, the approach of eliminating any non-response was used, as there was enough data to allow this.

Once unnecessary questions were removed from the Excel sheet the data was loaded into R. Additional cleaning was performed in order to get all questions in a similar format. any data that was not categorized on a numeric scale was manipulated to become numeric. For example, any question that was answered as “not at all” – “very much” was changed to represent 1-5 to allow for regression testing.

The main eight variables that were chosen to be the focus of this research were that of id, age, state, gender, number of times the respondent changed their diet between omnivore and plant-based and whether the individual had deficiencies in protein, iron and B-12. Summary statistics of mean, median, standard deviation and variance of major variables were performed in order to acquire a base understanding of the data. The results in figure 1 were obtained.

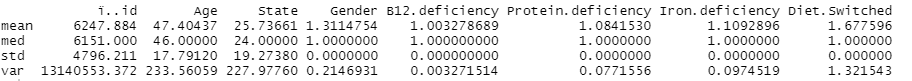


Figure 1

In the instance of gender, a 1 was coded to be a female and 2 was male, for all deficiencies a response of no was coded as 1 and a yes was 2.

Objectives

First, we will be looking at different ages within a plant-based diet. By doing this we will be able see the largest age group that has adopted this platform. The goal is to cluster different ages by the number of times that person has switched between plant-based and omnivore diets. This will allow us to get a good idea of the group that has the strongest and weakest diet retention. That group that is found to have the weakest retention will be deemed the omnivore group of study. Alternatively, the strongest retention will be our vegetarian/vegan group of study which will be referred to as plant-based. These groups will be the focal points in the analysis and will be used to determine what supplements they needed to take and if they experienced health issues while on their respective diet.

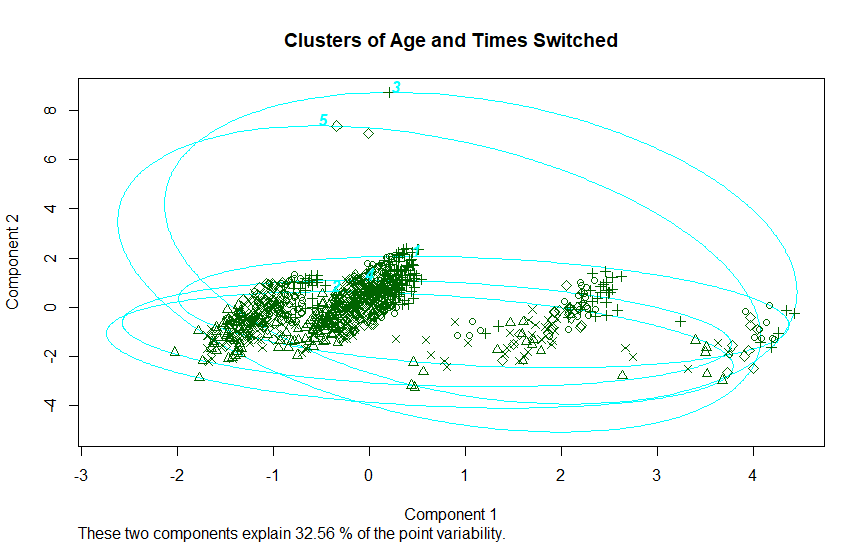
After the groups have been selected, a multiple regression will be performed on varying health risks that are associated with a plant-based diet. For example, a look into the correlation between the associated vegan/vegetarian/omnivores and low levels of vitamin B-12, protein deficiencies and iron deficiencies based on their age. Performing this regression will rule out factors that were likely caused by the age of the person. Lastly, summary statistics will be run on each group in order to see which group had the greatest number of deficiencies.

Methods

The first goal in determining if vegans are overall healthier, was to determine the age group with the best retention in maintaining a plant-based diets and those with the worst. Clustering was deemed the most effective method in assigning who would be in these group.

First, a K-means algorithm was performed on the dataset with 5 clusters. The number of clusters was chosen on the base that the data contained 5 possibilities for the number of times the diet was switched: 1 to 5 or more times. Data was consolidated in order to perform k-means clustering in R to include the variables of age, number of times switched and the id of the surveyor, as R is only able to perform k-means with numeric variable types. Data in the variable “number of times switched” was determined a non-applicable value if left blank as well as if “Don’t know” was selected. Any answer of “5 or more” was converted to 5 and deemed the maximum number of times switched.

The results produced five clusters with a mean age and mean diet switched as well as a size for each cluster. As shown in Figure 2, middle aged individuals tended to be grouped in clusters where diet had been switched more often (1.77 times). Younger individuals, aged 26, were grouped in a cluster with a mean switch of 1.48. These 5 clusters are shown in Figure 3.



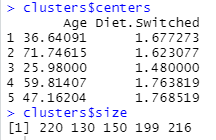


Figure 2

Figure 3

Figure 2 gives us the two clusters of focus, cluster 3 being our best diet retention and cluster 5 with the worst. Cluster 3 will now be referred to as the plant-based group and cluster 5 will be referred to as the omnivore group.

Secondly a hierarchical cluster was performed in order to compare the two clustering results. The number of clusters was kept the same at k=5. Figure 4 shows the results of this clustering method, each of the five clusters is grouped by a red rectangle. The agglomerative coefficient was found to be 0.9966, this number is said to show a good amount of clustering structure the closer it is to 1, indicating that the hierarchical clustering is appropriate.

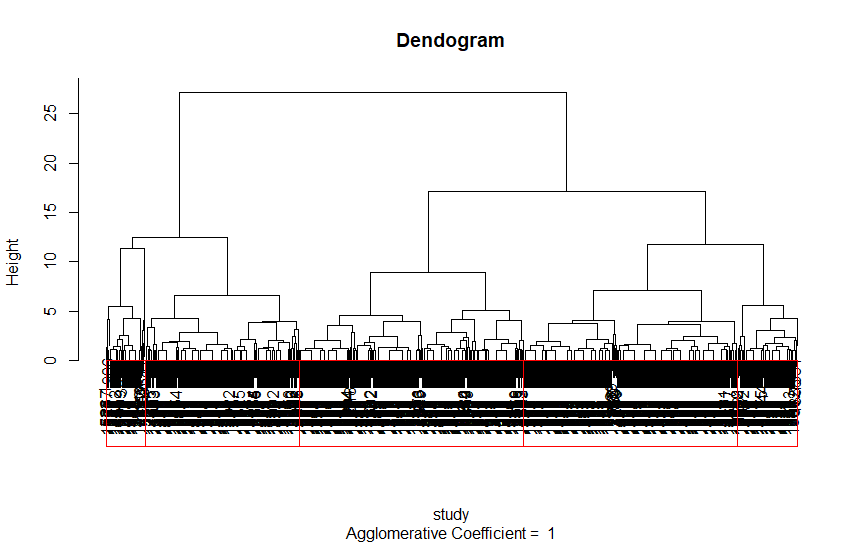


Figure 4

As the clustering results noted in Figure 4 resulted in strong clustering and since it was almost identical to that in the k-means clustering, the values in Figure 2 will be used.

After determining our clusters of focus, the plant-based and omnivore groups were split into two smaller datasets for analysis. A multiple regression was performed which looked at the dependent age variable’s correlation to the factors of diet switched, B-12 deficiency, iron deficiency and protein deficiency. Results for the regression of the plant-based group are shown in Figure 5 and the results for omnivores in Figure 6.

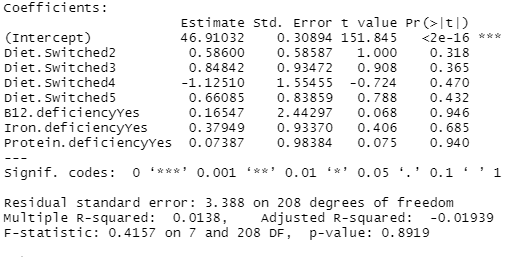
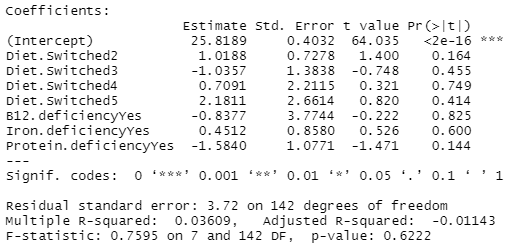


Figure 5 Figure 6

Since the diet switched variable is a categorical value, it appears 3 times in the regression models ranging from 2-5. Other variables in the regression are also categorical however, they only contained two factors, N/A or yes values, thus they appear in the regression with the factor yes beside them.

A regression was also performed without the independent variable of diet switched in order to observe if the R-square value increased. Results are shown in Figure 7 for plant-based and 8 for omnivore.

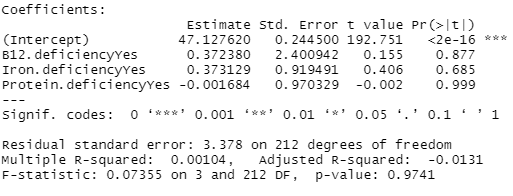
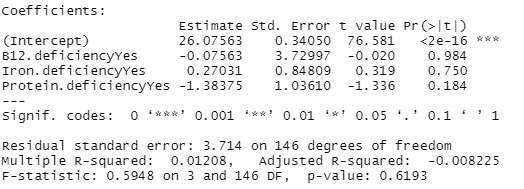


Figure 7 Figure 8

Finally, summary statistics were performed on the two groups to determine which had the highest average deficiencies in each of the three fields. Results of this analysis are found in Figure 9 for the plant-based diet and in Figure 10 for omnivores.

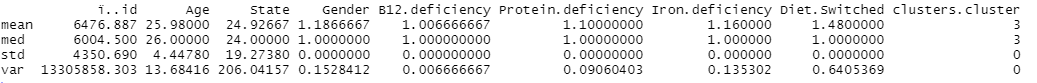


Figure 9

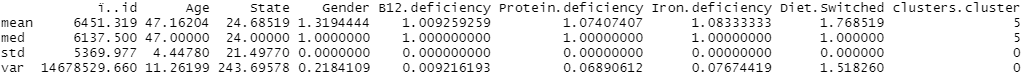


Figure 10

Limitations

One major hurdle was finding a dataset that was large enough and gathered the appropriate information to answer the question at hand. Medical data, on a large scale, is not available to the public that includes numbers associated to individual’s health. What was available was survey data, which entrusted individuals to report on whether they experienced vitamin deficiencies during the time of their plant-based diet. Several individuals, as reported previously, are unaware of having any deficiencies [5] thus the accuracy of the analysis is likely less than adequate.

Another limitation is that most data that is collected dealing with public health, tends to be in survey form. Survey data analysis requires alternative techniques as opposed to other data analysis. Results in survey data are not always independent random variables. In our data set, the option of vegan would rule out the option of omnivore and vice versa. In the article “A Survey on Survey Statistics: What Is Done and Can Be Done in Stata.” [8], two main challenges are referenced to deal with survey analysis; “obtaining correct point estimates and computing correct variances and standard errors”. As was discussed in the data preparation section, there were many N/A or “Don’t know” responses in the data set. In our case, nonresponses in the number of times switch, can create a heavier distribution in one diet over the other.

Our dataset also lacked a key element, whether the individual was practicing a plant-based or omnivore diet plan in the present. As this is a study of former and current vegetarians/vegans, an assumption was made based on the number of times an individual switched their diet. Based on the number of switches, the person was determined to be plant-based or omnivore. This technique limits calculations as to whether the individual was healthier, as it cannot be said with certainty that they are a certain diet. By basing the assumption on the number of switches it was assumed that they consumed more meat and dairy products the greater number of dietary switches.

The dataset also did not contain information that is critical in determining an individual’s health. Body mass index (BMI), overall feelings of health improvement or decline and many other elements are important in answering the question as to whether an individual is healthier over another. Unfortunately, this data is not readily available to the public on a large dataset basis. By limiting results to only deficiencies within each group, it prevented us from understanding what the plant-based group improved on that the omnivore group did not.

Lastly, due to the nature of the dataset, regression models were limited to those based on age and by id. Most of the dataset’s variables are categorical and cannot be used easily as a dependent variable in regression. By running a regression on age with the multiple deficiencies a person could experience through the diet plans, an attempt was made to eliminate factors that didn’t appear to be based on an individual’s diet, but by their age.

Results

First, let’s look at the k-means cluster that was created in Figure 2. The cluster that observed the best diet retention rate was found to be that of cluster 3, containing a mean diet switching rate of 1.48 and a mean age of 26. The cluster which observed the worst diet retention rate was found in cluster 5, consisting of a mean age of 48 and 1.77 number of diet switches. Neither of these values were that surprising, the younger the person the less likely they will have switched their diets, whereas the older the person the more chance for the switch to occur. Cluster 3’s results differ greatly from our larger data set mean. Age and diet rate in Figure 1 was found to be 48 and 1.677 respectively, cluster 5 is about the same values for both age and diet switched.

The two clusters were sorted out from the remaining 3, our data was left with 150 observations representing plant-based individuals and 212 observations representing omnivores. Each contained 9 total variables. We took this sample of individuals to represent a strong plant-based diet and an omnivore diet in which we would base their overall health.

After, a regression was performed on whether the individual experienced B-12 deficiency, protein deficiency or iron deficiency and was run against the individuals age. Regression could not be performed based on the number of times the person had switched their diet as the data variable diet switched is a factor value and cannot be used to regress on. Regression was performed again with the same factors baring diet switched. The results in Figure 5 versus figure 7 are interesting in that the r-square value decreases by 0.02 indicating that for plant-based diets the number of times the diet was switched played a role in the age. Another interesting note about the regressions is that from Figure 6 versus Figure 8 the r-square value improves without the variable of switches. In omnivore diets the age appears to be more a factor of dietary deficiencies.

All variables p-values were high, baring protein deficiency in Figure 7 at 0.184 and diet switched of 2 in Figure 5 at 0.164. This indicates that protein deficiency in plant-based individuals and having a diet switch of 2 is a moderate predictor of the individuals age. Notably, all 4 multiple regressions were not strong predictors of age.

Lastly, there were the results of the summary statistics of both groups. As cluster 3 is our plant-based diet, it was important to look at the mean of all deficiencies within the data. Cluster 3 maintained an overall higher mean in all vitamin deficiencies apart from vitamin B-12. Protein and iron were evaluated at 1.1 and 1.16 respectively, noted in Figure 9. Our omnivore group noted a mean of 1.074 and 1.083 for protein and iron respectively noted in Figure 10.

The closer our values were to 2 in the plant-based group, indicates that a greater volume of individuals answered yes to the resulting deficiency. What is most interesting is our plant-based group tended to experience less B-12 deficiencies than our omnivore group. This could be the result of limitations within our assumed omnivore group or simply due to the size difference within our clusters. Cluster 3 consists of 150 datasets whereas cluster 5 has 216. The value for B-12 is close within each group, 1.0067 compared to 1.009, only being off by .003.

Discussion

There are many studies available that show scientific proof of whether eating a plant-based diet is better for you than a typical omnivore. Through our study, it was largely undetermined whether our plant-based group was overall healthier. Our study has shown that with our test omnivore and plant-based datasets that overall deficiencies were higher in the plant-based data.

What this data failed to account for are other factors that contribute to an individual’s health. The dataset lacked BMI information, daily exercise, blood pressure and the persons interpretation of their overall health status. In a study by the Nutrition Society [9], obesity tends to be significantly lower in people of a plant-based diet. This decrease in obesity leads to lower rates of type 2 diabetes as well as decreases in cardiovascular disease, heart disease and strokes within plant-based diets.

With the data given we can conclude that the plant-based group contains higher deficiencies in areas such as protein and iron, whereas they seem to attain the appropriate levels of vitamin B-12. There are vitamins that are lacking in their diet, however the risks associated with a greater meat diet seems to carry larger risks than those with a plant- based diet. With this knowledge in mind, by adding supplements of the appropriate vitamin deficient area within the diet, a plant-based diet appears to be an overall healthier choice.

Appendix:

**References**

1. Psihoyos, Louie, director. *The Game Changer*. Sundance Films, 2018.
2. Allès, B.; Baudry, J.; Méjean, C.; Touvier, M.; Péneau, S.; Hercberg, S.; Kesse-Guyot, E. Comparison of Sociodemographic and Nutritional Characteristics between Self-Reported Vegetarians, Vegans, and Meat-Eaters from the NutriNet-Santé Study. Nutrients **2017**, 9, 1023.
3. Pilis W, Stec K, Zych M, Pilis A. Health benefits and risk associated with adopting a vegetarian diet. *Rocz Panstw Zakl Hig* (2014) 65(1):9–14.
4. Woo, K.S.; Kwok, T.C.; Celermajer, D.S. Vegan Diet, Subnormal Vitamin B-12 Status and Cardiovascular Health. Nutrients **2014**, 6, 3259-3273.
5. Key, T., Appleby, P., & Rosell, M. (2006). Health effects of vegetarian and vegan diets. *Proceedings of the Nutrition Society,* *65*(1), 35-41. doi:10.1079/PNS2005481.
6. Clarys, P.; Deliens, T.; Huybrechts, I.; Deriemaeker, P.; Vanaelst, B.; De Keyzer, W.; Hebbelinck, M.; Mullie, P. Comparison of Nutritional Quality of the Vegan, Vegetarian, Semi-Vegetarian, Pesco-Vegetarian and Omnivorous Diet. Nutrients 2014, 6, 1318-1332.
7. Funalytics, *Study of Current and Former Vegetarians and Vegans*,2014. Funalytics. Web. <https://faunalytics.org/dataset-study-of-current-and-former-vegetarians-and-vegans/>.
8. Kreuter, Frauke, and Richard Valliant. “A Survey on Survey Statistics: What Is Done and Can Be Done in Stata.” *The Stata Journal: Promoting Communications on Statistics and Stata*, vol. 7, no. 1, 2007, pp. 1–21., doi:10.1177/1536867x0700700101.
9. Appleby, P., & Key, T. (2016). The long-term health of vegetarians and vegans. *Proceedings of the Nutrition Society,* *75*(3), 287-293. doi:10.1017/S0029665115004334

**Code**

#Libraries Used

library(survey)

library(cluster)

library(fpc)

library(matrixStats)

#read in the file, making sure all blanks are read as NA

questions=read.csv("formatted.csv", header=T, na.strings=c("","NA"))

str(study)

dim(study)

names(questions)

#Shows the amount of missing data

aggr(questions)

#removes the junk considered to be NA

questions$Diet.Switched[questions$Diet.Switched=="Donâ???Tt know"] = NA

questions$Diet.Switched=factor(questions$Diet.Switched)

#columns to run the cluster on

cols=c(2,8)

#check to make sure there are or are not NA values

is.na(questions[,cols])

#Run KMeans (five clusters as there are at most 5 times switched)

set.seed(10)

clusters=kmeans(na.omit(questions[,cols]),5)

#looks at the cluster created

clusters$centers

clusters$size

table(na.omit(questions$Diet.Switched), clusters$cluster)

study=na.omit(questions[,cols])

attach(study)

#plots of the clusters

clusplot(na.omit(questions),clusters$cluster, color=FALSE, col.study=2, labels=5, main="Clusters of Age and Times Switched")

plot(na.omit(questions[c("Diet.Switched","Age")]), col=clusters$cluster)

names(questions)

plotcluster(study,clusters$cluster,pch=clusters$cluster)

#Summary/Descriptive Stats

test=data.matrix(na.omit(questions))

summary(test)

mean=colMeans(test)

med=colMedians(test)

std=colMads(test)

var=colVars(test)

summ=rbind(mean,med,std,var)

summ

#Dendogram Cluster

set.seed(23)

questionsdag=agnes(study,diss=FALSE,metric="euclidian")

plot(questionsdag, main="Dendogram")

groups=cutree(questionsdag,k=5)

rect.hclust(questionsdag, k=5, border="red")

questionsdag$ac

plot(study$Age,study$Diet.Switched)

test=diana(study)

rect.hclust(test, k=5, border="red")

grouped=cutree(as.hclust(test),k=5)

plot(grouped)

test$dc

plot(test)

groups

questionsdag$ac

head(groups)

#combine the cluster association as a column in questions survey data

length(study[,1])

clusters$cluster

completeData=data.frame(na.omit(questions), clusters$cluster)

#Elminates all groups that are not group 3 or group 5

cluster3=completeData[+ grep(3, completeData$clusters.cluster),]

cluster5=completeData[+ grep(5, completeData$clusters.cluster),]

#Running regressions

summary(lm(Age~Diet.Switched+B12.deficiency+Iron.deficiency+Protein.deficiency, data=cluster3))

summary(lm(Age~Diet.Switched+B12.deficiency+Iron.deficiency+Protein.deficiency, data=cluster5))

summary(lm(Age~B12.deficiency+Iron.deficiency+Protein.deficiency, data=cluster3))

summary(lm(Age~B12.deficiency+Iron.deficiency+Protein.deficiency, data=cluster5))

#Summary stats on 2 cluster datasets

clusters3=data.matrix(cluster3)

mean=colMeans(clusters3)

med=colMedians(clusters3)

std=colMads(clusters3)

var=colVars(clusters3)

summ=rbind(mean,med,std,var)

summ

clusters5=data.matrix(cluster5)

mean=colMeans(clusters5)

med=colMedians(clusters5)

std=colMads(clusters5)

var=colVars(clusters5)

summ=rbind(mean,med,std,var)

summ