Spring 2020

An Analytical Examination on the Effects of Vegetarian and Omnivorous Diets on C-Reactive Protein

Aletha Kleis
aletha.kleis@cwu.edu

An Analytical Examination on the Effects of Vegetarian and Omnivorous Diets on C-Reactive Protein

Aletha Kleis

Senior Capstone
Submitted in Partial Fulfillment of the Requirements for Graduation from
The William O. Douglas Honors College
Central Washington University

May, 2020

Accepted by:

Professor Department of Mathematics	May 19, 2020
Committee Chair (Name, Title, Department)	Date
Professor, Food Science & Nutrition, Department of Health Sciences	May 15, 2020
Committee Member (Name, Title, Department)	Date
	June 3, 2020
Director, William O. Douglas Honors College	Date

Introduction

Several books and documentaries alike profess the perils of eating meat on people's health and the environment. Doctors, scientists, nutritionists, and those who are generally well-informed, have presented the idea that eating too much meat can have a negative impact on health, which may overshadow the positive aspects. In an American society like ours, reducing meat consumption can prove to be very difficult without having a good cause to do so. It is important to provide people with the necessary facts to help them make the best-informed decisions can aid in decreasing the number of people who eat meat, which in turn benefits the publics' health.

The vegetarian or omnivorous diets' benefits or detriments can be measured through many lenses with a wide scope of variables to quantify them. This paper takes an analytical approach by synthesizing data of a national health survey to determine these health benefits or detriments of these diets and the quality of them. Before describing the exact hypothesis, it is necessary to familiarize all readers of the current research and discourse on these subjects.

Literature Review

There is a multitude of researchers who have found, through various methods, that eating less meat is better for your health when measured through a myriad of variables. In one study it is found that processed meat consumption is positively associated with stroke in men, and consumption of red and processed meats increases the risk of cerebral infarction, or stroke, in women as well (Larsson, Virtamo, & Wolk, 2011). Not only can reducing meat consumption decrease the chances of strokes, but scientists also consider the vegetarian diet as a viable option for weight loss and weight management (Farmer, 2009). Furthermore, multiple studies have found the negative cardiovascular effects of eating meat. In a research paper by Bovalino,

Charleson, and Szoeke the authors unveil from their analysis that there is a strong association between red and processed meat consumption and cardiovascular disease risk in women (2016).

In fact, a well-researched relationship is that of meat consumption and cardiovascular health. One study found that red and processed meat intake is strongly associated with increased mortality due to cardiovascular diseases (Sinha, Cross, Graubard, Leitzmann, & Schatzkin, 2009). Another study finds processed meats are associated with a higher incidence of coronary heart disease (Micha, Wallace, & Mozaffarian, 2010). Across the literature, many researchers and health experts generally accept the relationship between a reduction in meat consumption and a decreased risk of cardiovascular and coronary heart diseases.

The benefit of lowered cardiovascular disease risk can be seen by using a statistical approach, that combines statistical testing with the result of mortality or with a measurable health variable. One such variable for current heart health and future risk for heart disease is the level of C-Reactive Protein (CRP) in a simple blood sample. CRP is a protein made in the liver and is known as an acute phase reactant. This means that CRP can be released into the blood in the span of a few hours after a trauma, like a heart attack, or in the early stages of infection (C-Reactive Protein, 2020). By measuring the amount of CRP in the blood, inflammation due to acute or chronic conditions can be detected.

This measure is relied upon by doctors when testing for inflammatory diseases like rheumatoid arthritis and lupus, and as an indication of risk for heart disease (Mayo Clinic, 2017). Wong, Pio, and Valencia analyzed the levels of CRP and its relation to risk factors of coronary heart disease in respondents of the National Health and Nutrition Examination Survey (NHANES) (2007). They found that higher CRP levels are strongly associated with multiple major coronary heart disease risk factors, even after adjusting for age.

Cardiovascular diseases (CVD) and coronary heart problems (CHD) are extremely common in the older population, but it is not just a disease among the elderly. The American Heart Association estimated in 2013, that about 70% of men and women aged 60-79 and between 83% and 87% of men and women who are older than 80 years have cardiovascular diseases (Go, et.al, 2013). Since CRP values can be used to detect the presence of CVD, throughout this research age is treated as an important covariate included in analysis.

CRP doesn't distinguish between chronic inflammation associated with CVD or CHD and an acute inflammatory process which may have nothing to do with coronary and cardiovascular health. The non-specificity and sensitivity of CRP as an acute phase reactant makes it a difficult variable to rely upon, as there are many factors that can influence it. Despite this, CRP was chosen for this research for two main reasons: it is a useful variable in indicating a person's cardiovascular health, and to confirm the studies that claim vegetarians with a high-quality diet will have a lower CRP value.

The CRP values used in this paper were gathered by analyzing data from the National Health and Nutrition Examination Survey (NHANES), the nationwide health survey of adults and children in the United States, that many nutritionists and health researchers rely upon. NHANES includes interviews, health-related questionnaires, laboratory measurements, and thorough physical examinations. NHANES is the main program of the National Center for Health Statistics (NCHS), which is part of the Centers for Disease Control and Prevention (CDC) (Centers for Disease Control and Prevention, 2017). The survey collects the "health and nutritional status" of about 5,000 people per survey year across the United States (Centers for Disease Control and Prevention, 2017).

This thorough survey includes a myriad of the subjects' health variables: age, sex, dietary intake, responses to questions to sort vegetarians and omnivores, and variables distinguishing women who were pregnant or breastfeeding and many other pieces of information about each participant. Using a specially formulated population sampling, the NHANES data provides in-depth answers to "demographic, socioeconomic, dietary, and health-related questions … medical, dental, and physiological measurements, as well as laboratory tests administered by highly trained medical personnel" (Centers for Disease Control and Prevention, 2017). The amount of data available and the reliability of it from a government agency is why NHANES data was chosen for this research.

Computed after the survey, from a 24-hour dietary recall some participants take part in, is a value called the Healthy Eating Index (HEI). The 24-hour dietary recall is conducted by a skilled dietary interviewer using the Automated Multiple Pass Method (AMPM) developed by the USDA. The AMPM is a research-based, multiple-pass approach that employs 5 steps designed to enhance complete and accurate food recall while minimizing respondent burden.

From the 24-hour food logs, called the "What We Eat in America" dietary intake survey NHANES researchers are able to calculate a score that represents the quality of each person's diet. The HEI "is a measure of diet quality used to assess how well a set of foods aligns with key recommendations of the Dietary Guidelines for Americans" which is used to "help individuals (ages 2 years and older) … consume a healthful and nutritionally adequate diet" (U.S. Department of Agriculture, 2019). Each participant of the What We Eat in America portion of the survey receives a score from zero to one hundred, where a score of 100 means that person's food intake matched perfectly with key recommendations from the Dietary Guidelines for Americans.

The quality of each diet, or HEI, is measured by calculating a score for 13 components, or food groups, of a respondent's food intake. The insert in the appendix, "Average Healthy Eating Index-2015 Scores for Americans by Age Group, WWEIA/NHANES 2015-2016" shows this exact breakdown for the HEI-2015 calculation (U.S. Department of Agriculture, 2019). The way HEI is calculated is consistently updated every five years to reflect the most current conformance of the Dietary Guidelines for Americans recommendations. It is best practice to use the most current version of the HEI calculation, no matter the years of data a researcher is interested in. Thus, the HEI-2015 scoring calculation was used for this analysis.

NHANES surveys the entire US population, by covering 15 locations at each survey cycle from different regions of America. In an effort to retrieve the most reliable and accurate statistics, "NHANES over-samples persons 60 and older, African Americans, and Hispanics" (Centers for Disease Control and Prevention, 2017). This way of sampling ensures adequate numbers of subjects in these specific groups and allows for valid subgroup analysis. The National Center for Health Statistics (NCHS) "is working with public health agencies to increase the knowledge of the health status of older Americans" which NHANES has taken a primary role in" (Centers for Disease Control and Prevention, 2017). Given the reliability and quantity of data collected from NHANES for the particular variables of interest, CRP, vegetarianism, and quality of diet, this data source was the obvious choice research.

Hypothesis

Due to a lack of findings on the C-Reactive Protein levels of the vegetarian population compared to the omnivorous population, this paper seeks to fill that research gap, with an analysis of data collected from the National Health and Examination Survey for the survey cycle years 2007-2008 and 2009-2010. In addition, this study will examine the relationship between

the vegetarian and omnivorous population's CRP value, while considering age and the Healthy Eating Index.

Based on previous studies, the researcher hypothesized, that the vegetarian population would have a statistically significantly lower CRP value, and a statistically significantly higher HEI, while the opposite would be true for the omnivorous group. It was suspected that those with a high HEI score, following either diet, would also have statistically significant lower CRP values than those who followed either diet but scored a poor HEI value. Both were hypothesized assuming that corrections would be made for age, and omitting women who reported being pregnant or breastfeeding at the time of the survey, as is common in nutritional based studies.

Description of Data

As previously mentioned, NHANES surveyors seek out specific populations to survey for each cycle year to best represent the U.S. population of all ages. Due in part to this oversampling of some groups, survey non-response, and post-stratification, NHANES has also created survey weights to make the data "representative of the U.S. civilian non-institutionalized population" (Centers for Disease Control and Prevention, 2020). This is done by assigning a weight to each participant, which correlates to "the number of people in the population represented by" them (Centers for Disease Control and Prevention, 2020).

The two main survey weights are the interview weights and Mobile Examination Center (MEC) weights, which were both calculated from the participants of each survey. NHANES suggests "a good rule of thumb is to use "the least common denominator" where the variable of interest that was collected on the smallest number of respondents is the "least common denominator." The sample weight that applies to that variable is the appropriate one to use for

that particular analysis (Centers for Disease Control and Prevention, 2020). Thus, only the MEC weight was necessary and was recommended by the documentation for this analysis.

This MEC weight variable, along with the age variable was retrieved from the "Demographic Variables & Sample Weights" (DEMO) data table from NHANES. Other variables used in this study are from the following tables: "Diet Behavior & Nutrition" (DBQ), "C-Reactive Protein" (CRP), and "Reproductive Health" (RHQ). The table below, Table 1, shows the variables needed from each of these tables for both cycle years, which were necessary for the analysis.

Variables Pulled from Corresponding Tables

Tables	***Variables Pulled***
DBQ	SEQN, DBQ915
CRP	SEQN, LBXCRP
RHQ	SEQN, RHD143, RHQ200
Demo	SEQN, RIDEXPRG, RIDAGEYR, RIAGENDR, RIDRETH1, WTMEC2YR

Table 1: Variables Pulled from Corresponding Tables

From the DBQ table, question DBQ915 – "Self-perceived vegetarian" was pulled. Question DBQ915 is as follows: "{Do you/Does SP} consider {yourself/himself/herself} to be a vegetarian?" (Centers for Disease Control and Prevention, 2012). Survey participants were able to answer: yes, no and don't know. Additionally, some refused to answer, or their response was missing. Table 10 in the appendix shows an exact breakdown of responses collected for this question for both cycle years.

To separate the vegetarians from the omnivores, only those who responded yes to question DBQ were put into the vegetarian population, while the opposite was true for the

omnivorous population. Participants who answered “don’t know”, refused to answer or in the case their response was missing, were removed from the analysis altogether.

The following table, Table 2, defines the corresponding NHANES documentation file for each variable:

Variable Names and Definitions

Variable	*Definition from NHANES*
SEQN	Respondent sequence number
DBQ915	Self-perceived vegetarian
LBXCRP	C-reactive protein(mg/dL)
RIDEXPRG	Pregnancy Status at Exam
RIDAGEYR	Age at Screening Adjudicated
RIAGENDR	Gender of the sample person
RHD143	Are you pregnant now?
RHQ200	Now breastfeeding a child?
WTMEC2YR	Full Sample 2 Year MEC Exam Weight

Table 2: Variable Names and Definitions

There were adjustments made to clean the data and to reduce any skewness or significantly unexpected results. To start, to avoid any skewed results, all women who reported as being pregnant or lactating were excluded from this study, as this state may affect their food intake and CRP levels. Along with the DBQ table, the Reproductive Health (RHQ) table included counts of women who reported being pregnant as well as those who were breastfeeding at the time of the survey. This removed an additional 269 respondents from the total population for analysis from both cycle years, henceforth referred to as the total population.

Next, the data for analysis only includes people who were 20 or older, and who had a measured CRP value and HEI score. Removing people who were younger than 20 years old, took 6,157 observations out of the total population. The HEI score is calculated from the participant's 24-hour food log, which not all participants complete. Thus, 2,037 participants were removed because they didn't have an HEI score. Not all participants in the survey had a recorded CRP value either, and as it is an important predicting variable, the analysis performed on the population does not include those who had no recorded CRP. This removed an additional 418 observations from the total population. These were the only people not included in the analysis from the total and vegetarian populations.

With the quantity of data finalized, each variable of importance for this research can be analyzed with more detail. The figures below show the frequency and range of CRP, age, and HEI for each population. In Figure 1, it is easy to spot that most subjects who were vegetarian or omnivorous, have a CRP value between 0 and 1, creating a right skew. There were a few outliers in the omnivorous population, with a CRP value of 20, which were not shown in the histogram, but were kept for analysis. There was also a vegetarian outlier with a CRP value of 13.1, who was included in the histogram and for analysis.

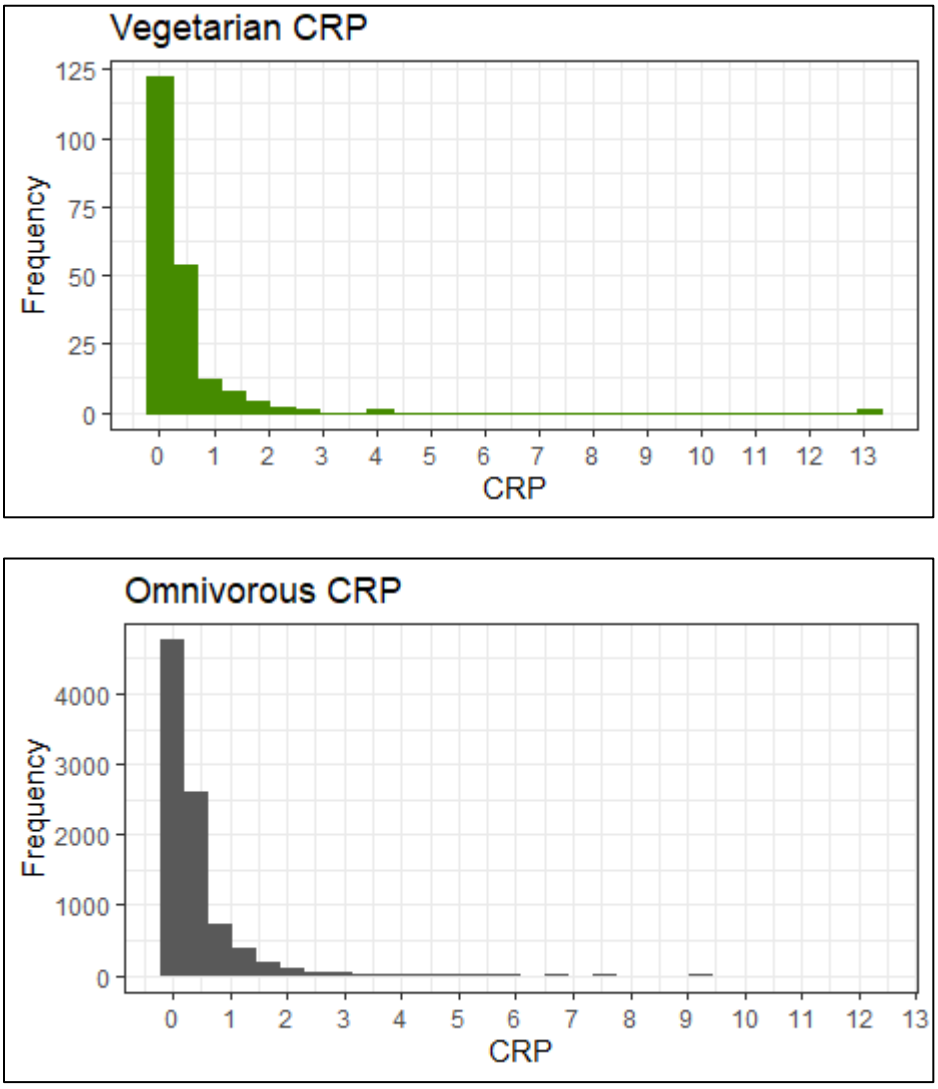

Figure 1: Vegetarian and Omnivorous CRP

While the histograms may not show it, the values of CRP found in each person can vary drastically. For example, in the total population of interest for this study, CRP ranged from 0.1 to 20. Those who have a very high CRP value may have recently had a heart attack or have pre-existing inflammatory diseases or conditions. The Mayo Clinic states that a normal CRP value is less than 10 milligrams per liter (2017). In this study, the average CRP value was 0.4283, with most participants having a CRP value between 0 and 1.

The age of vegetarians and omnivores was relatively evenly distributed, with what seems to be a larger population of older participants in both populations, as seen in Figure 2. However, the big column on the far right is the number of people 80 and above, whereas all of the other columns are just for people in two-year increments. Thus, the far-right column is not due to oversampling by NHANES but represents the demographics in the US.

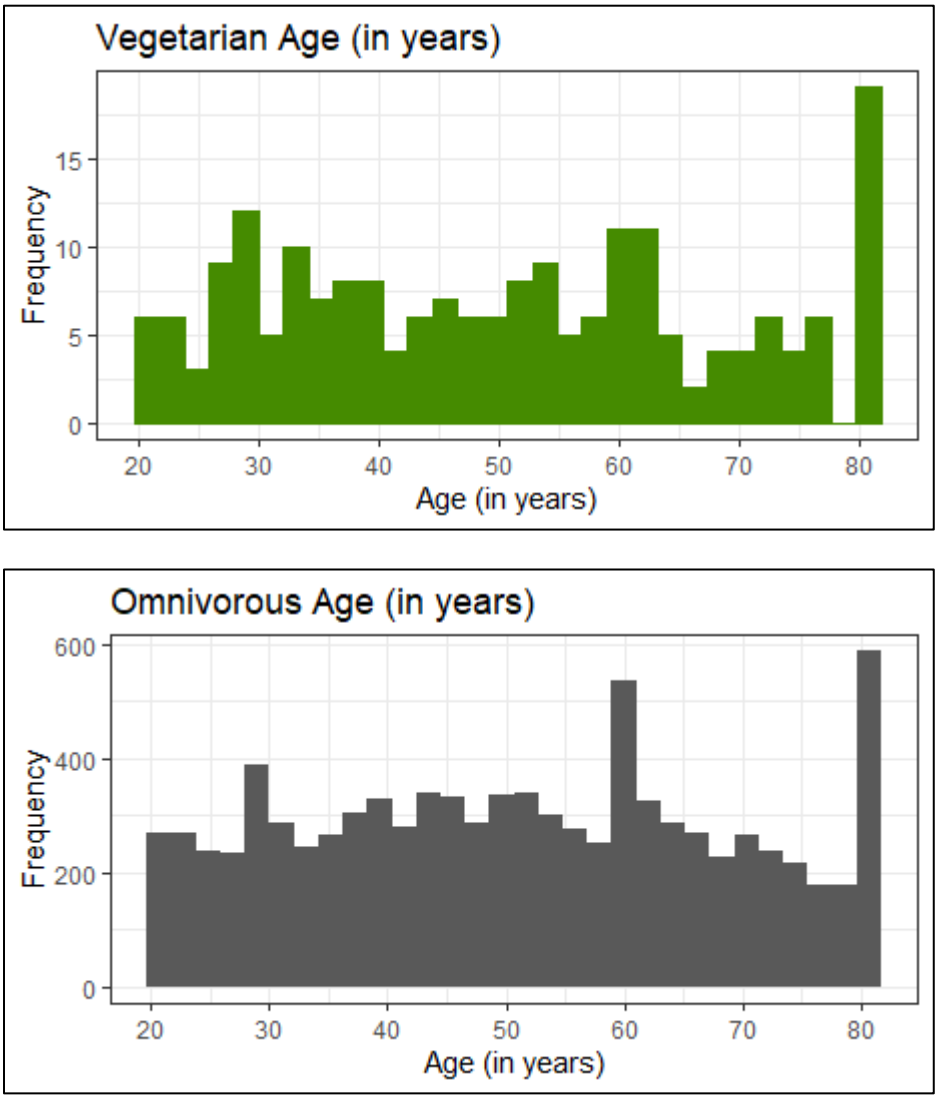

Figure 2: Vegetarian and Omnivorous Age (in years)

Finally, in Figure 3, the vegetarians appear to have higher HEI scores overall, than the omnivorous populations. On the other hand, the omnivorous population had a relatively even distribution of HEI scores, with quite a few more scoring very highly.

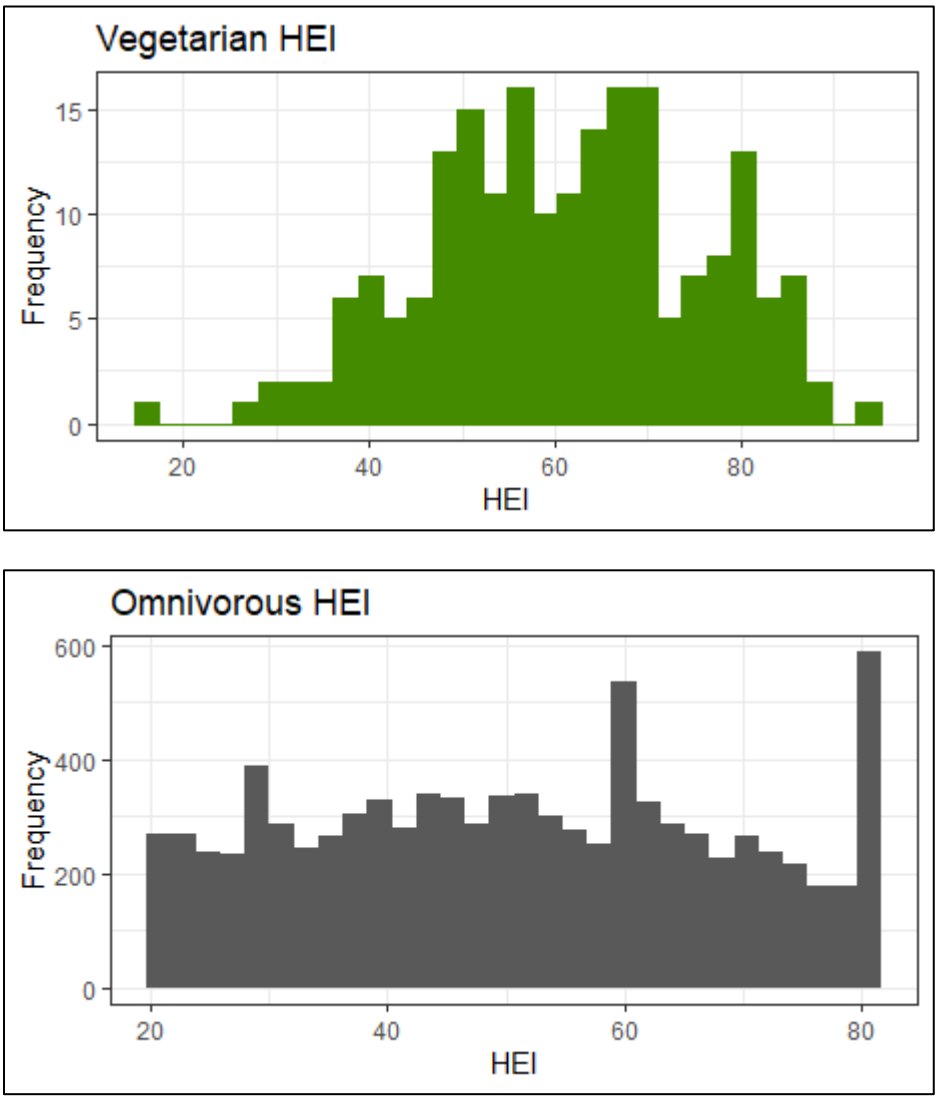

Figure 3: Vegetarian and Omnivorous HEI

Synthesizing the data in this way was imperative in the approach to gather results for the research questions, further detailed in a later section. To interpret this data and answer the

research questions, a procedure was followed, an explanation of which is given in the next section.

Methods

The data was gathered from NHANES using the statistical coding language R. R was chosen because it is well known for statistical computing and data analysis and is free software. However, in the nutrition field, people almost exclusively use SAS to analyze health data and, so, the downloadable NHANES data is formatted for SAS. To get around this, the R package *RNHANES* allows the user another way to retrieve and analyze the data (Susmann, 2016). Packages in R can be thought of as functions that are meant to aid the coder in their data analysis. The command `nhanes_load_data` was particularly useful from the *RNHANES* package. Using this command, all necessary variables from NHANES tables, described in Table 1 and Table 2 for the analysis were retrieved.

Once the data was retrieved there were further steps to clean the data and make the data usable for synthesis. After using the `nhanes_load_data` command and storing the data into data frames, each data frame had to be combined into one to contain the two cycle years. Selecting only the variables needed for this analysis was done with the *dplyr* which makes cleaning data much easier with straightforward commands (Wickham, François, Henry & Müller, 2020). Then additional columns or variables had to be added into the singular data frame as well.

Each collection of data is from a survey conducted in a span of two years. For example, the data used in this analysis is from the survey cycle years 2007-2008 and the cycle years 2009-2010. These cycle years were chosen because at the time this project began these were the most

recent years in which NHANES measured both CRP and asked the participants whether they considered themselves vegetarian.

Thus, after loading the CRP values for the cycle year 2007-2008, the CRP values for the cycle years 2009-2010 had to be joined below the rows of data for 2007-2008. Because this is the same variable, the R command `cbind` was used to combine the two sets of data together by row. Once each variable had both cycle years of data bonded together, each variable was then added into one large data frame with the R command `merge`. To continue the example, after all participants' CRP values were combined into one data frame, the other variables of interest, like age, were merged with that data frame by matching participants' unique ID number.

After all the data was in one data frame, the column for each participants' weight had to be modified. Both cycle years used in this analysis, 2007-2008 and 2009-2010, included a Mobile Examination Center (MEC) weight column that assigned each participant a survey weight as calculated by NHANES researchers. The MEC weights from both survey cycles represent the sampling technique and population from their corresponding years. Each participant's weight had to be adjusted to represent the "population at the midpoint of the combined survey period" (Centers for Disease Control and Prevention, 2020). As outlined and instructed by NHANES reports each participant's MEC weight was multiplied by 0.5.

Further adjustments were made to the data for analysis. As detailed in the Description of Data section, it was at this stage that participants were removed due to their age and pregnancy or breastfeeding status. Participants with missing HEI scores or CRP values were also removed. Then, the entire population was split into two: the vegetarians and the omnivores. Removing participants and splitting the data was done with the `filter` command from the *dplyr* package, examples of which are shown below (Wickham, François, Henry & Müller, 2020). The first line

removes all participants under the age of 20, and the next creates a subset of the data, containing only those who answered “yes”, identified with a 1, to the “Self-perceived vegetarian” question:

```
myData <- adjData %>% filter(age >= 20)
vegetarian.df <- myData %>% filter(vegetarian == 1)
```

After cleaning and configuring the data to become what was needed for analysis, a final column for each participant’s HEI score was added. To calculate this, the Food Patterns Equivalent Database (FPED), the 24-hour dietary log, and the demographics file for each survey cycle year had to be loaded. Then he column for each participant’s HEI score was added with the *hei* package created by Nagraj and Folsom and it’s simple command, `hei()` (2020).

Once the desired data frames were created, each was exported with the *rio* package for later use. The *rio* package has an `export()` command, allowing the user to easily export their data frame into a Comma Separated Values (.csv) file, an R Data Serialized (.rds) file, or a Stata Data (.dta) file (Chan, Chan, Leeper & Becker, 2018). After the data was fully cleaned and ready for analysis, it was exported into a .csv file for ease of reference during analysis.

To answer the research questions, there were two statistical tests used: weighted t-tests and weighted multiple regression analysis. The t-test was chosen because it identifies whether the means of two groups are significantly different. Multiple linear regression, or multiple regression, is used to predict a response variable, in this case, CRP, by using other explanatory variables.

To perform the weighted t-test, an additional package was needed. The *weights* package included a command for this test, among other weighted statistical tests (Pasek, 2020). A weighted t-test is easily done with this package and can be performed with the command: `wtd.t.test()`. The documentation for the *weights* package also suggests users with survey

data include the parameter `bootse = TRUE` in the command. What this does, is address the issues of the survey weights being used to "indicate probabilities of selection rather than the precision of estimates" (Pasek, Tahk, Culter & Schwemmle, 2020). An example using the `wtd.t.test()` command is below:

```
      wtd.t.test(vegetarians$HEI, omnivores$HEI,
weight = vegetarians$adjustedMEC,
weighty = omnivores$adjustedMEC, bootse = TRUE)
```

The exact t-test resulting from this code is Welch's Two-Sample t-test. This t-test does not assume the variances of the two groups are equal. The null hypothesis for the t-test is that the two populations' means are equal. With a p value of 0.05 or less, the null hypothesis is rejected for the alternative and it is concluded that there is a statistically significant difference in means.

Weighted multiple regression requires only a simple modification to R's `lm()` function, which is to include `weights =` in the arguments. Multiple regression analysis was used to identify the strength of diet, diet quality, and age on CRP. It was also used to find exactly how much these variables affected CRP values. The results from the multiple regression analysis and t-tests ultimately gave the answers to the research questions. The processes of coming to the conclusions are detailed in the following section.

Analysis

The table below shows the total number of men and women in both populations, after cleaning the data.

Number of Vegetarian and Omnivorous Men and Women

	Omnivorous Population	**Vegetarian Population**
Count	8,873	203
Men	4,361	78
Women	4,512	125

Table 3: Number of Vegetarian and Omnivorous Men and Women

To begin, the difference in means of the CRP values and HEI scores of the vegetarian and omnivorous populations were compared using a weighted t-test. Based on results from the weighted t-tests, with weighted multiple linear regression, there was then an attempt to predict CRP values using age and HEI. The findings of both tests are detailed in the paragraphs and tables to follow.

As reflected in the p values from the weighted t-tests in the table below, Table 4, there was not a significant difference between the average C-Reactive Protein values for the vegetarian and omnivorous populations. That is, an average CRP value of 0.39, between about 9,000 omnivores, is not significantly different from the average CRP value of 0.30 for about 200 vegetarians, with a p value just above the significance level, 0.07.

With a p value of less than 0.01, we reject the null hypothesis and conclude there is a significant difference in the age of the two populations. While it may sound counterintuitive, the fact that there is a significant difference between the age of the two populations is not a positive

indicator. There being a significant difference in the average age of the two populations means that the difference in the CRP values could be partly caused by the difference in age. To draw a statistically sound conclusion, the average age of the two populations should be relatively similar because it will mean the results of the analysis lie more heavily on other variables, such as HEI.

The outcome of the weighted t-tests shows there is a significant difference between the two means of the HEI scores; the omnivorous population has a mean HEI score of 53.19 and the vegetarian population has a mean HEI score of 61.34, with a p value of less than 0.01.

Descriptive Statistics for the Omnivorous and Vegetarian Population

	Omnivorous Population	**Vegetarian Population**	**p Value**
Count	8,873	203	---
Average Age (years)	48.04	44.70	< 0.01
Average CRP	0.39	0.30	0.07
Average HEI	53.19	61.34	< 0.01

Table 4: Descriptive Statistics for the Omnivorous and Vegetarian Population

The results from the weighted t-tests would seem to indicate that despite the outcome of the significantly different age of the populations, and the differences in the size of the populations, the average vegetarian's diet quality is better than the average omnivores. However, despite these things, the average CRP of the omnivorous and vegetarian populations is not significantly different, with a p value of 0.07. All these results culminate to say that being vegetarian and having a high-quality diet does not, statistically speaking, make a difference in a person's average CRP value.

Faced with these results about the two populations, the next step was to create a multiple regression model to attempt to predict a person's C-Reactive Protein value using their age and HEI score. The equations for the CRP values of the omnivorous population (CRP_o) and vegetarian population (CRPv), with their respective intercepts from the multiple regression analysis, are as follows:

$$CRP_O = 0.004(age) - 0.005(HEI) + 0.469$$

$$CRP_V = 0.006(age) - 0.002(HEI) + 0.155$$

In the weighted multiple regression analysis it seems that while the age and HEI of the two populations are significantly different, their effects on CRP are relatively small. Specifically, from the equations, we see that when predicting CRP for the omnivorous and vegetarian populations, for every year a person ages, their CRP value will increase by an average of only 0.004 and 0.006 milligrams per liter. What is more surprising is the similarly small effects of HEI on a person's CRP value. The literature on CRP values would suggest that the better quality of a person's diet, the lower their CRP value should be. In this case, based on the equations above, we see that a one-point increase in their HEI score will bring a person's CRP down by only an average of 0.005 and 0.002 milligrams per liter, for omnivores and vegetarians.

However, it is important to remember that any person's CRP value will most often be between 0 and 1 and that CRP is measured by milligrams per liter So, while these numbers are small, it is possible they may make a big difference, especially considering how small the units of CRP are measured by. For a deeper look at this and the previously-stated-conclusions, we can look at the p values of each variable and the adjusted r-squared values of the regression models above, given in Table 5.

Multiple Regression R-Squared and p values for Models

	Omnivorous Population	**Vegetarian Population**
Adjusted R-Squared	0.001	0.012
p value Age (years)	<0.01	0.04
p value HEI	<0.01	0.60

Table 5: Multiple Regression R-Squared and p values for Models

The adjusted r-squared values indicate that age and HEI predict about 1% of the variation in CRP for the vegetarian population. HEI is statistically insignificant in predicting CRP for the vegetarian model with a p value of 0.60, while age is significant in predicting CRP with a p value of 0.04.

Meanwhile, the adjusted r-squared value of essentially 0% in the omnivorous model means that age and HEI, while both are significant predictors of CRP, will not predict almost any of the variation of it. These low values of R^2, however, should not be taken to mean that there is no relationship between HEI and CRP_o, but that the strength of the linear relationship between them is weak. We will see below that there seems to be a moderately strong non-linear relationship between the variables.

While the increases and decreases to a person's CRP value from their age and HEI are very small, they're still important factors in predicting an omnivore's CRP value, but much less important in predicting a vegetarians CRP value. Along with this, other conclusions can be drawn from this analysis, but another key outcome is the wavering importance of the quality of a person's diet, and its effect on CRP. Based on the weighted t-tests, the average HEI was significantly higher in the vegetarian population than in the omnivorous population. This should mean that for the vegetarian population, HEI would be a statistically significant predictor of CRP

or, a stronger predictor of CRP than the omnivorous population, but surprisingly the opposite is the case. In fact, HEI is not a significant predictor in a person's CRP value in the vegetarian model, like it is in the omnivorous model. Furthermore, a one-point increase in HEI score will only lessen CRP_V by 0.002 milligrams per liter, as opposed to 0.005 in CRP_O.

The results from the weighted t-tests and weighted multiple regression, combine to give a glimmer of information about what is going on in the populations and their CRP values. A way to parse out more information from the data would be to compare the CRP values of those with the highest and lowest quality of diet, as measured by HEI, for both populations. This was done by separating the omnivorous and vegetarian populations into 10% increments based on their HEI scores, as seen in the graphs below (Figure 4). Each 10% of the populations are indicated on the x-axis, while the average CRP value for that decile, is on the y-axis. It is somewhat easy to see that in the omnivorous population, as diet quality and their respective HEI score increase, their average CRP value decreases. This is much less clear in the vegetarian population.

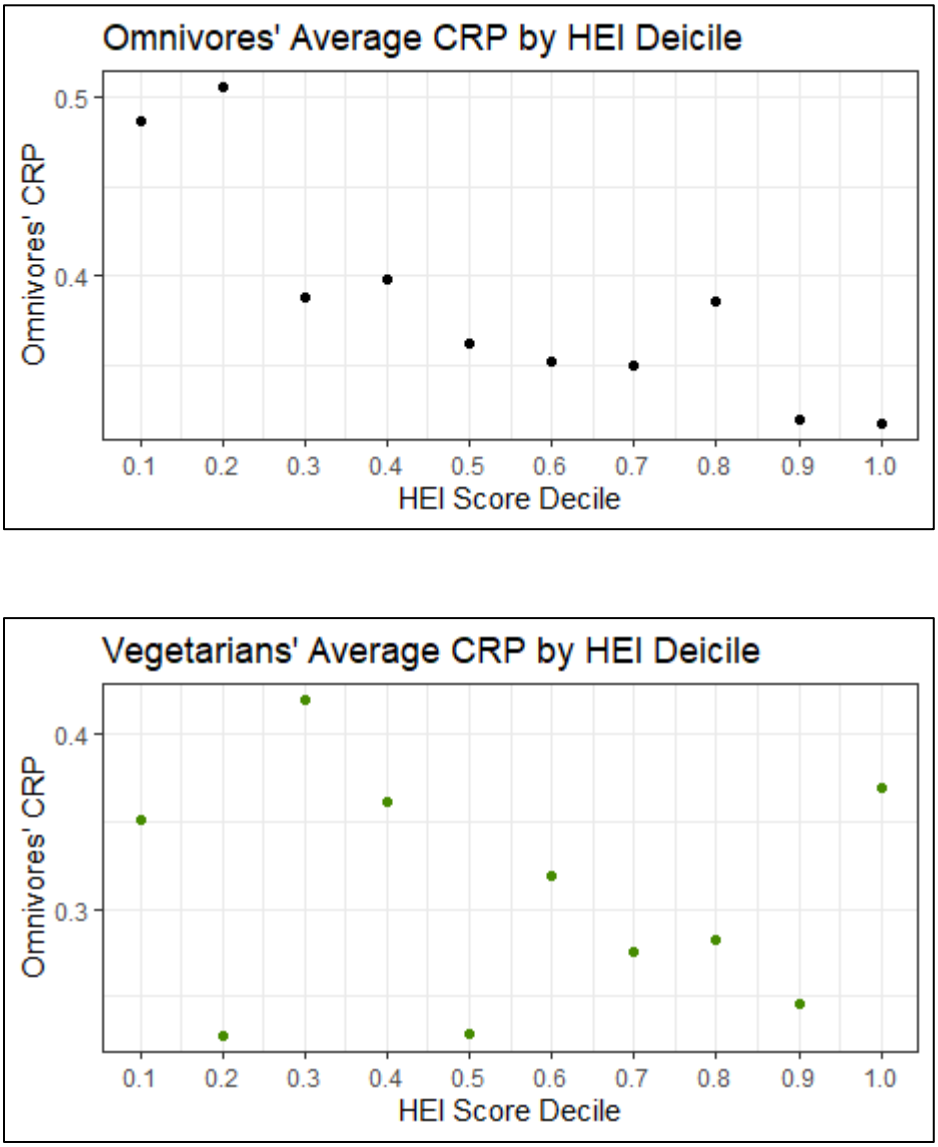

Figure 4: Vegetarian and Omnivorous CRP by HEI Decile

Though from the graphs in Figure 4 we can see there is not a linear relationship, as was done initially, weighted multiple regression analysis and weighted t-tests were performed on the two populations. This time the two populations were split by those with an HEI score in the bottom 30% and those with an HEI score in the top 30%. Below is the table with the results of the weighted t-tests for the omnivorous population, Table 6. The average age, CRP and HEI

score of the omnivores with an HEI score in the bottom and top 30% are significantly different, based on each having a p value of less than 0.01.

Descriptive Statistics for the <u>Omnivores</u> with an HEI Score in the Top and Bottom 30%

	Bottom 30%	**Top 30%**	**p Value**
Count	2,661	2,661	---
Average Age (years)	43.43	52.70	< 0.01
Average CRP	0.46	0.34	< 0.01
Average HEI	37.17	70.10	< 0.01

Table 6: Descriptive Statistics for the Omnivores with an HEI Score in the Top and Bottom 30%

Before splitting the populations by their HEI scores, the total omnivorous population and the total vegetarian population had a significant difference in their ages. In this case, with a p value of less than 0.01, as we can see in Table 6, there is a significant difference between the omnivores with an HEI score in the top 30% and the omnivores with an HEI score in the bottom 30%. Previously the difference in the average age of the two populations was not a positive indication and it still is not.

This distinction is important to make because from Table 6 we see that at a significance level of less than 0.01, there is a difference in the CRP averages between the omnivores with an HEI score in the top 30%, and the omnivores with an HEI score in the bottom 30%. In fact, it is a difference of almost 0.1 milligrams per liter, where the omnivores with an HEI score in the bottom 30% had a CRP value of 0.46, and the omnivores in the top 30% have a CRP value of 0.34. Meaning, despite that their age is on average older than the bottom 30%, the omnivorous with an HEI score in the top 30% still had a significantly lower CRP value. To further

understand this significance, weighted linear multiple regression analysis was performed on both omnivorous populations and resulted in the following equations:

$$\text{Top 30\% HEI Scores}, CRP_O = 0.004(age) - 0.005(HEI) + 0.477$$

$$\text{Bottom 30\% HEI Scores}, CRP_O = 0.004(age) - 0.008(HEI) + 0.578$$

From the equations to predict the CRP value of an omnivore with an HEI score in the top 30%, and the omnivores with an HEI score in the bottom 30%, we see again that age and HEI play a minuscule role. In these models, age increases a person's CRP value by about 0.004 per year they age as it did in the model for the CRP for the total omnivorous population. HEI scores play an equally small role in predicting CRP values, even when looking at those with the very best diet quality.

Recalling the original coefficient of -0.005 for the total omnivorous populations CRP value, when looking at the two populations of omnivores, the coefficient is either equal to -0.005 or greater, -0.008. HEI will decrease an omnivore's average CRP value by a little bit more, 0.003 milligrams per liter more if they have a lower HEI score. Thus, while the relationship is non-linear, it can still be pointed out that for the omnivores with an HEI score in the bottom 30%, increasing the quality of their diet by even one point decreases their CRP value by more than a one point increase of HEI in the omnivores with an HEI score in the top 30%.

When turning to the vegetarians with an HEI score in the top 30%, and the vegetarians with an HEI score in the bottom 30%, the results are less telling. Starting with the results from the weighted t-tests for the variables of interest in the vegetarian populations, there is a significant difference in the average of the two population's HEI score, as seen in the table below, Table 7. This is the only difference between the vegetarians with an HEI score in the top 30%, and the vegetarians with an HEI score in the bottom 30%.

Descriptive Statistics for <u>Vegetarians</u> with an HEI Score in the Top and Bottom 30%

	Bottom 30%	**Top 30%**	**p Value**
Count	61	61	---
Average Age (years)	45.44	47.10	0.61
Average CRP	0.33	0.30	0.68
Average HEI	44.35	76.99	< 0.01

Table 7: Descriptive Statistics for Vegetarians with an HEI Score in the Top and Bottom 30%

There is not a significant difference in the age of the vegetarians with an HEI score in the top 30%, and the vegetarians with an HEI score in the bottom 30%. More importantly, there is not a significant difference in their CRP values either. These results express that even for the vegetarians with poor diet quality, those with an HEI score in the bottom 30%, have almost the same average CRP value as those in the top 30%, as expressed in the p value of 0.68.

For further understanding, weighted multiple regression analysis is performed, and the equations for the two models for the vegetarian populations CRP values are given as follows:

$$\text{Top 30\% HEI Scores}, CRP_V = 0.004\,(age) - 0.002(HEI) + 0.251$$

$$\text{Bottom 30\% HEI Scores}, CRP_V = 0.001(age) - 0.009(HEI) + 0.327$$

From the model for the whole vegetarian population, the CRP value was increased by 0.006 for each year a person aged and decreased by 0.002 for each point increase in a person's HEI score. The effects of age in the above models are even less than the effects of age in the original model, especially for those with an HEI score in the bottom 30%. While the effect of HEI on CRP in the bottom 30% of vegetarians is quite a bit bigger and will decrease a person's CRP by 0.009 milligrams per liter. Again it is reiterated that a one-point increase in HEI will decrease the overall CRP

more if the participant has an HEI score in the bottom 30% than it will if the participant has an HEI score in the top 30%.

Thus, with all the combined analysis of the vegetarians with an HEI score in the top 30%, and the vegetarians with an HEI score in the bottom 30%, some theories can be posed. Based on the p-values of the significance of age and HEI in predicting the CRP value for the vegetarian populations, combined with the p-values of those variables from the weighted t-tests, one can say that age may not make a significant difference in CRP, but HEI will.

Results

With these findings, the researchers' hypothesis can be reassessed and reevaluated. Recall that the researcher expected the vegetarian population to have a statistically significantly lower CRP, and a statistically significantly higher HEI score. The researcher assumed the opposite would be true for the omnivorous population, and that for either population, those who had a high HEI score would have a statistically significant lower CRP value.

Regarding the first portion of the hypothesis, it turned out the vegetarians did not have a statistically significantly lower CRP value and a significantly higher HEI score. This can be immediately concluded from the weighted t-test results (Table 4). Though the data is non-linear, from the weighted multiple regression analysis, it is then found that HEI is not a statistically significant variable in predicting CRP in the vegetarian population (Table 5). Thus, it is found that the vegetarian population studied did not have a statistically significantly lower CRP value than the omnivores, but the vegetarians' HEI score is higher than the omnivores.

Regarding the second portion of the hypothesis, the study on the omnivorous population was more informative on the effects of HEI on CRP. From the weighted t-test, there was a significant difference in the CRP value of the omnivores who scored in the bottom and the top

30% of HEI (Table 6). For either population, the effects of HEI will decrease a persons' CRP value by more in the participants with an HEI score in the bottom 30%, than in the top 30%. In fact, HEI is a significant predictor of CRP in both omnivorous populations, with a p value of less than 0.01 (Table 7). Thus, statistically speaking having a higher HEI score will lower a person's CRP value.

Conclusion

The last result lends itself to the conversation of statistical significance versus clinical significance. Based on the findings, the statistical significance of HEI on CRP seems small, but in relation to the units of CRP is quite important. Recall that even despite their older age, the omnivores with an HEI score in the top 30% of HEI, still had a statistically significantly lower CRP value.

What could improve these findings and make the case of statistical versus clinical significance equal, is if the research was performed on a less volatile variable. CRP is highly sensitive and, contrary to current literature, evidently depends on many more things than the quality of diet and being vegetarian or omnivorous.

While the results aren't exactly what was hypothesized, two important conclusions can be made based on the NHANES data for 2007-2008 and 2009-2010. Firstly, there was a statistically significant difference between the CRP of the omnivores and the vegetarians. While the vegetarians had a higher HEI score than the omnivores, it was deemed to influence CRP in small amounts, even despite its small units of measure. Secondly, no matter the diet style, having a statistically significant higher HEI score will lower a person's CRP value, but will still only affect overall CRP by small amounts. Though, more research is needed to truly understand the clinical and statistical significance of this.

Average Healthy Eating Index-2015 Scores for Americans by Age Group, WWEIA/NHANES 2015-2016

Component	Maximum Points	Age Groups: All Americans (2+ years)	Children (2-17 years)	Adults (18-64 years)	Older Adults (65+ years)
Total HEI Score	**100**	**58.7**	**53.9**	**58.3**	**64.0**
Adequacy:					
Total Fruits	5	2.9	3.3	2.6	3.7
Whole Fruits	5	4.2	4.4	3.8	5.0
Total Vegetables	5	3.3	2.3	3.5	4.0
Greens and Beans	5	3.1	1.6	3.4	3.7
Whole Grains	10	3.0	3.3	2.7	4.0
Dairy	10	6.0	8.1	5.4	5.6
Total Protein Foods	5	5.0	4.7	5.0	5.0
Seafood and Plant Proteins	5	5.0	3.2	5.0	5.0
Fatty Acids	10	4.1	2.9	4.5	4.2
Moderation:					
Refined Grains	10	6.4	4.7	6.7	7.4
Sodium	10	3.7	4.4	3.4	4.0
Added Sugars	10	6.8	6.4	6.8	7.5
Saturated Fats	10	5.1	4.5	5.4	4.7

Due to rounding, HEI component scores in each age group may not add up precisely to the total HEI score of 100.

Notes: The Healthy Eating Index-2015 (HEI-2015) is a measure of diet quality used to assess how well a set of foods aligns with the *2015-2020 Dietary Guidelines for Americans*. The HEI-2015 includes 13 components that can be summed to a maximum total score of 100 points. The components capture the balance among food groups, subgroups, and dietary elements including those to encourage, called adequacy components, and those for which there are limits, called moderation components. For the adequacy components, higher scores reflect higher intakes that meet or exceed the standards. For the moderation components, higher scores reflect lower intakes because lower intakes are more desirable. A higher total score indicates a diet that aligns better with the Dietary Guidelines.

Sources:
Data—National Center for Health Statistics, *What We Eat in America/National Health and Nutrition Examination Survey, 2015-2016.*
Healthy Eating Index-2015 Scores—U.S. Department of Agriculture, Center for Nutrition Policy and Promotion, access https://www.fns.usda.gov/resource/healthy-eating-index-hei.

References:
Center for Disease Control and Prevention, *National Health and Nutrition Examination Survey Data, 2015-2016*. Hyattsville, MD: U.S. Department of Health and Human Services.

U.S. Department of Health and Human Services and U.S. Department of Agriculture. *2015-2020 Dietary Guidelines for Americans*. 8th Edition. December 2015. Available at http://www.health.gov/dietaryguidelines/2015/guidelines/.

References

Bovalino, Charleson, & Szoeke. (2016). The impact of red and processed meat consumption on cardiovascular disease risk in women. *Nutrition*, 32(3), 349-354.

Centers for Disease Control and Prevention (CDC). (2017, September 15). *National Center for Health Statistics (NCHS).* National Health and Nutrition Examination Survey About. https://www.cdc.gov/nchs/nhanes/about_nhanes.htm

Centers for Disease Control and Prevention (CDC). (2020, February 21). *National Center for Health Statistics (NCHS).* National Health and Nutrition Examination Survey Tutorials. https://wwwn.cdc.gov/nchs/nhanes/tutorials/Module3.aspx

Centers for Disease Control and Prevention (CDC). (2012, September). *National Center for Health Statistics (NCHS).* National Health and Nutrition Examination Survey Data. https://wwwn.cdc.gov/Nchs/Nhanes/2009-2010/DBQ_F.htm

Chan, C., Chan, G., Leeper, T. & Becker J. (2018). *rio: A Swiss-army knife for data file I/O*. R package version 0.5.16.

C-Reactive Protein (CRP). (2020, March 6). https://labtestsonline.org/tests/c-reactive-protein-crp

Farmer, B. (2009). *Comparison of nutrient intakes for vegetarians, non-vegetarians, and dieters: Results from the National Health and Nutrition Examination Survey 1999-2004*. Eastern Michigan University. https://commons.emich.edu/theses/150

Go, A. S., Mozaffarian, D., Roger, V. L., Benjamin, E. J., Berry, J. D., Borden, W. B., … Turner, M. B. (2013). Heart Disease and Stroke Statistics—2013 Update. *Circulation*, 127(1). doi: 10.1161/cir.0b013e31828124ad

Larsson, S., Virtamo, J., & Wolk, A. (2011). Red meat consumption and risk of stroke in Swedish men. *The American Journal of Clinical Nutrition*, 94(2), 417-21.

Larsson, S., Virtamo, J., & Wolk, A. (2011). Red meat consumption and risk of stroke in Swedish women. *Stroke*, 42(2), 324-9.

Mayo Clinic (2017, November 21). *C-reactive protein test*. https://www.mayoclinic.org/tests-procedures/c-reactive-protein-test/about/pac-20385228

Micha, R., Wallace, S., & Mozaffarian, D. (2010). Red and Processed Meat Consumption and Risk of Incident Coronary Heart Disease, Stroke, and Diabetes Mellitus A Systematic Review and Meta-Analysis. *Circulation*, 121(21), 2271-U52.

Nagraj, V. , Folsom, T. (2020). *hei: Calculate Healthy Eating Index (HEI) Scores.* R package version 0.2.0. https://timfolsom.github.io/hei/

Pasek, J., Tahk, A. Culter, G. Schwemmle, M. (2020). *weights: Weighting and Weighted Statistics*. R package version 1.0.1. https://CRAN.R-project.org/package=weights

Sinha, R., Cross, A., Graubard, B., Leitzmann, M., & Schatzkin, A. (2009). Meat intake and mortality: A prospective study of over half a million people. *Archives of Internal Medicine*, 169(6), 562-71.

Susmann, H. (2016). *RNHANES: Facilitates Analysis of CDC NHANES Data*. R package version 1.1.0. https://CRAN.R-project.org/package=RNHANES

U.S. Department of Agriculture (USDA). (2019, January). Center for Nutrition and Promotion (CNPP). *Healthy Eating Index.* https://www.fns.usda.gov/hei-scores-americans

Wickham, H., François, R., Henry, L., & Müller, K. (2020). *dplyr: A Grammar of Data Manipulation.* R package version 0.8.5. https://CRAN.R-project.org/package=dplyr

```
######
## PULLING NHANES DATA AND CREATING DATA FRAMES FOR ANALYSIS
install.packages('RNHANES')
library(RNHANES) #helps retrieve NHANES data
install.packages('dplyr')
library(dplyr) #helps clean data
install.packages('rio') #makes exporting easier
library(rio)


###### BRINGING IN DATA:
## EXPORT 2009-2010 DATA (post-fix is f)
fulldbqf <- nhanes_load_data("DBQ", "2009-2010") #vegetarian
(DBQ915)
fulldemof <- nhanes_load_data("DEMO", "2009-2010") #gender,
race/ethnicity, age, pregnancy (riagendr, ridreth1, ridageyr,
ridexprg)
fullcrpf <- nhanes_load_data("CRP", "2009-2010") #crp
fullrhqf <- nhanes_load_data("RHQ", "2009-2010") #pregnancy,
breastfeeding (rhd143, rhd200)
## EXPORT 2007-2008 DATA (post-fix is e)
fulldbqe <- nhanes_load_data("DBQ", "2007-2008") #vegetarian
(DBQ915)
fulldemoe <- nhanes_load_data("DEMO", "2007-2008") #gender,
race/ethnicity, age, pregnancy (riagendr, ridreth1, ridageyr,
ridexprg)
fullcrpe <- nhanes_load_data("CRP", "2007-2008") #crp
fullrhqe <- nhanes_load_data("RHQ", "2007-2008") #pregnancy,
breastfeeding (rhd143, rhd200)


###### DATA CLEANING:
install.packages("dplyr")
library(dplyr)
## REMOVE UNNECESARY COLUMNS
wantdbq.f <- fulldbqf %>% select(SEQN, DBQ915)
wantdbq.e <- fulldbqe %>% select(SEQN, DBQ915)
wantdemo.f <- fulldemof %>% select(SEQN, RIAGENDR, RIDAGEYR,
RIDEXPRG, WTMEC2YR)
wantdemo.e <- fulldemoe %>% select(SEQN, RIAGENDR, RIDAGEYR,
RIDEXPRG, WTMEC2YR)
wantrhq.f <- fullrhqf %>% select(SEQN, RHD143, RHQ200)
wantrhq.e <- fullrhqe %>% select(SEQN, RHD143, RHQ200)
wantcrp.f <- fullcrpf %>% select(SEQN, LBXCRP)
wantcrp.e <- fullcrpe %>% select(SEQN, LBXCRP)
## RENAMING COLUMNS
wantdbq.f <- rename(wantdbq.f, vegetarianQ = DBQ915)
```

```
wantdbq.e <- rename(wantdbq.e, vegetarianQ = DBQ915)
wantdemo.f <- rename(wantdemo.f, sex = RIAGENDR, age = RIDAGEYR,
pregnant = RIDEXPRG, MECweight = WTMEC2YR)
wantdemo.e <- rename(wantdemo.e, sex = RIAGENDR, age = RIDAGEYR,
pregnant = RIDEXPRG, MECweight = WTMEC2YR)
wantrhq.f <- rename(wantrhq.f, rhqPreg = RHD143, breastfeeding =
RHQ200)
wantrhq.e <- rename(wantrhq.e, rhqPreg = RHD143, breastfeeding =
RHQ200)
wantcrp.f <- rename(wantcrp.f, crp = LBXCRP)
wantcrp.e <- rename(wantcrp.e, crp = LBXCRP)


###### COMBINING
## BINDING BY ROW
dbq.original <- rbind(wantdbq.f, wantdbq.e)
demo.original <- rbind(wantdemo.f, wantdemo.e)
rhq.original <- rbind(wantrhq.f, wantrhq.e)
crp.original <- rbind(wantcrp.f, wantcrp.e)
sum(is.na(crp.original$crp))
attach(dbq.original)
attach(demo.original)
attach(rhq.original)
attach(crp.original)
## MERGING ALL TOGETHER
merge1 <- merge(dbq.original, demo.original, by = "SEQN")
full.original <- merge(merge1, crp.original, by = "SEQN")
#leaving out rhq to filter with later
attach(full.original) #starting with 17923 variables
#exported to Excel as "full.original"


######## REMOVING PREGNANT PEOPLE AND BREASTFEEDING PEOPLE IN
RHQ
attach(rhq.original)
#pregnant
length(which(rhq.original$rhqPreg == 2))
#1, pregant = 95
#2, not pregnant = 1454
#9, Don't know, = 37
sum(is.na(rhq.original$rhqPreg)) #5643
rhq.original$rhqPreg[is.na(rhq.original$rhqPreg)] <- 2
#assuming those with NAs are not pregnant, and those with a 9
are
95 + 37 #should be 132 people with a 1
#change all 9's to 1
rhq.original$rhqPreg[rhq.original$rhqPreg == 9] <- 1
```

```
length(which(rhq.original$rhqPreg == 1)) #132 1's
#breastfeeding
length(which(rhq.original$breastfeeding == 2))
#1, yes = 63
#2, no = 219
sum(is.na(rhq.original$breastfeeding)) #6947
rhq.original$breastfeeding[is.na(rhq.original$breastfeeding)] <-
2
#assuming those with NA are not breastfeeding, only keep 63 of
these people
#keeping only those who are either breastfeeding or pregnant
rhq.adjusted <- rhq.original %>% filter(rhqPreg == 1 |
breastfeeding == 1)
attach(rhq.adjusted) #data frame with participants who weren't
pregnant or breastfeeding
#exported to excel as "full.adjused1.csv"
## REMOVE FROM DATA
length(rhq.adjusted$SEQN %in% full.original$SEQN)
#193 matches; all from rhq.adjusted are in full.original
17923 -193
#after removing them there should be 17730 rows in new full
dataframe
library(dplyr)
full.adjusted1 <- anti_join(full.original, rhq.adjusted, by =
"SEQN") #full.adjusted now has 17730 rows
attach(full.adjusted1) #no breastfeeding or pregnant people from
rhq


###### REMOVING PREGNANT PEOPLE AND BREASTFEEDING PEOPLE IN DBQ
attach(full.adjusted1) #starting with 17730 rows
length(which(full.adjusted1$pregnant == 2))
#1, pregant = 26
#2, not pregnant = 2288
#3, possibly, = 50
sum(is.na(full.adjusted1$pregnant))
15366 + 2288 #should be 17654 people with a 2
full.adjusted1$pregnant[is.na(full.adjusted1$pregnant)] <- 2
length(which(full.adjusted1$pregnant == 2))
17730 - 17654
#should end with 17654 rows, removed 76 people (pregnant +
possibly pregnant)
full.adjusted2 <- full.adjusted1 %>% filter(pregnant == 2)
attach(full.adjusted2) #no breastfeeding or pregnant people from
rhq & no pregnant people from dbq
76+193
```

```
###### REMOVE YOUNG PEOPLE
attach(full.adjusted2) #starting with 17654 rows
length(which(full.adjusted2$age < 20)) #6157 people younger than
20
17654 - 6157 #should be 11497 people after removing younger than
20
full.adjusted3 <- full.adjusted2 %>% filter(age >= 20)
attach(full.adjusted3) #no breastfeeding or pregnant people from
rhq & no pregnant people from dbq & only people 20 or over


###### ADD HEI SCORE COLUMN
#https://www.rdocumentation.org/packages/hei/versions/0.1.0
install.packages("devtools")
library(devtools)
devtools::install_github("vpnagraj/hei")
library(hei)
fped_F <- get_fped("2009/2010", "both") #FPED
diet_F <- get_diet("2009/2010", "both") #dietary
demo_F <- get_demo("2009/2010") #demographic
hei_F <- hei(fped_F,diet_F,demo_F) #HEI scores for each
participant in 2009-2010
fped_E <- get_fped("2007/2008", "both") #FPED
diet_E <- get_diet("2007/2008", "both") #dietary
demo_E <- get_demo("2007/2008") #demographic
hei_E <- hei(fped_E,diet_E,demo_E) #HEI scores for each
participant in 2007-2008
#combine the two sets of HEI
hei.full <- rbind(hei_F, hei_E)
attach(hei.full)
names(hei.full)
#removing the age column so it doesn't get duplicated when
merging with full.adjusted3
hei.adjusted <- hei.full %>% select(-RIDAGEYR)
## ADD HEI COLUMN TO FULL DATASET
attach(full.adjusted3) #starting with 11497 rows, 7 variables
full.adjusted4 <- merge(full.adjusted3, hei.adjusted, by =
"SEQN")
11497 - 9460 #it removed 2037 people becasue they didn't do the
24-hour food log
attach(full.adjusted4) #no breastfeeding or pregnant people from
rhq & no pregnant people from dbq
#& only people 20 or over & added HEI column


###### REMOVE PEOPLE WHO DON'T HAVE A CRP VALUE
attach(full.adjusted4) #starting with 9496 rows
```

```
summary(full.adjusted4$crp) #418 people with NA for CRP
9496 - 418 #should end up with 9078 rows
full.adjusted5 <- full.adjusted4 %>% filter(crp >= 0)
attach(full.adjusted5) #no breastfeeding or pregnant people from
rhq & no pregnant people from dbq
#& only people 20 or over & added HEI column & removed
participants without CRP variable


###### ADJUST WEIGHTS
attach(full.adjusted5) #starting with 9078 rows, 8 variables
names(full.adjusted5)
full.adjusted6 <- full.adjusted5 %>% mutate(adjmec =
0.5*MECweight)
full.adjusted6 <- full.adjusted6 %>% select(-MECweight) #remove
original MEC Weights column
attach(full.adjusted6) #no breastfeeding or pregnant people from
rhq & no pregnant people from dbq
#& only people 20 or over & added HEI column & removed
participants without CRP variable & adjusted MEC weight column


###### SEPERATE DATAFRAMES
## OMNIVOROUS
attach(full.adjusted6) #all participants, 9078 rows
omnivores.df <- full.adjusted6 %>% filter(vegetarianQ == 2)
attach(omnivores.df) #8873 rows, omnivores only
## VEGETARIAN
attach(full.adjusted6) #all participants, 9078 rows
vegetarian.df <- full.adjusted6 %>% filter(vegetarianQ == 1)
attach(vegetarian.df) #203 rows, vegetarians only


###### EXPORT INTO EXCEL FOR FUTURE USE
install.packages("rio")
library(rio)
export(full.adjusted6, "Full Data.csv")
export(omnivores.df, "omnivores.csv")
export(vegetarian.df, "vegetarians.csv")


######
## CREATE TOP AND BOTTOM 30% HEI SCORE POPULATIONS
omnivores.df <- read.csv(file = file.choose())
vegetarian.df<-read.csv((file=file.choose()))
```

```
## START BY CREATING DECILES FOR OMNIVORES
library(dplyr)
hei1.omni <- omnivores.df %>% filter(HEI > quantile(HEI, 0) &
HEI < quantile(HEI, .1))
hei2.omni <- omnivores.df %>% filter(HEI > quantile(HEI, .1) &
HEI < quantile(HEI, .2))  %>% mutate(decile = 0.2)
hei3.omni <- omnivores.df %>% filter(HEI > quantile(HEI, .2) &
HEI < quantile(HEI, .3))  %>% mutate(decile = 0.3)
hei4.omni <- omnivores.df %>% filter(HEI > quantile(HEI, .3) &
HEI < quantile(HEI, .4))  %>% mutate(decile = 0.4)
hei5.omni <- omnivores.df %>% filter(HEI > quantile(HEI, .4) &
HEI < quantile(HEI, .5))  %>% mutate(decile = 0.5)
hei6.omni <- omnivores.df %>% filter(HEI > quantile(HEI, .5) &
HEI < quantile(HEI, .6))  %>% mutate(decile = 0.6)
hei7.omni <- omnivores.df %>% filter(HEI > quantile(HEI, .6) &
HEI < quantile(HEI, .7))  %>% mutate(decile = 0.7)
hei8.omni <- omnivores.df %>% filter(HEI > quantile(HEI, .7) &
HEI < quantile(HEI, .8))  %>% mutate(decile = 0.8)
hei9.omni <- omnivores.df %>% filter(HEI > quantile(HEI, .8) &
HEI < quantile(HEI, .9))  %>% mutate(decile = 0.9)
hei10.omni <- omnivores.df %>% filter(HEI > quantile(HEI, .9) &
HEI < quantile(HEI, 1))  %>% mutate(decile = 1)
## ADD COLUMNS FOR THE DECILE AND AVERAGE CRP
hei1.omni <- hei1.omni %>% mutate(decile = 0.1, meanCRP =
weighted.mean(hei1.omni$crp, hei1.omni$adjmec))
hei2.omni <- hei2.omni %>% mutate(decile = 0.2, meanCRP =
weighted.mean(hei2.omni$crp, hei2.omni$adjmec))
hei3.omni <- hei3.omni %>% mutate(decile = 0.3, meanCRP =
weighted.mean(hei3.omni$crp, hei3.omni$adjmec))
hei4.omni<- hei4.omni %>% mutate(decile = 0.4, meanCRP =
weighted.mean(hei4.omni$crp, hei4.omni$adjmec))
hei5.omni <- hei5.omni %>% mutate(decile = 0.5, meanCRP =
weighted.mean(hei5.omni$crp, hei5.omni$adjmec))
hei6.omni <- hei6.omni %>% mutate(decile = 0.6, meanCRP =
weighted.mean(hei6.omni$crp, hei6.omni$adjmec))
hei7.omni <- hei7.omni %>% mutate(decile = 0.7, meanCRP =
weighted.mean(hei7.omni$crp, hei7.omni$adjmec))
hei8.omni <- hei8.omni %>% mutate(decile = 0.8, meanCRP =
weighted.mean(hei8.omni$crp, hei8.omni$adjmec))
hei9.omni <- hei9.omni %>% mutate(decile = 0.9, meanCRP =
weighted.mean(hei9.omni$crp, hei9.omni$adjmec))
hei10.omni <- hei10.omni %>% mutate(decile = 1, meanCRP =
weighted.mean(hei10.omni$crp, hei10.omni$adjmec))
## COMBINE ROWS TOGETHER
heiOmni.df <- rbind(hei1.omni, hei2.omni,
hei3.omni,hei4.omni,hei5.omni,hei6.omni,hei7.omni,hei8.omni,hei9
.omni,hei10.omni)
```

```
## CREATE BOTTOM AND TOP 30% OF HEI SCORE POPULATIONS
bottom30.omni <- heiOmni.df %>% filter(decile <= 0.3)
top30.omni <- heiOmni.df %>% filter(decile >= 0.8)


#### SAME PROCESSES FOR VEGETARIANS
attach(vegetarians.df)
hei1.veg <- vegetarians.df %>% filter(HEI > quantile(HEI, 0) &
HEI < quantile(HEI, .1))
hei2.veg <- vegetarians.df %>% filter(HEI > quantile(HEI, .1) &
HEI < quantile(HEI, .2))  %>% mutate(decile = 0.2)
hei3.veg <- vegetarians.df %>% filter(HEI > quantile(HEI, .2) &
HEI < quantile(HEI, .3))  %>% mutate(decile = 0.3)
hei4.veg <- vegetarians.df %>% filter(HEI > quantile(HEI, .3) &
HEI < quantile(HEI, .4))  %>% mutate(decile = 0.4)
hei5.veg <- vegetarians.df %>% filter(HEI > quantile(HEI, .4) &
HEI < quantile(HEI, .5))  %>% mutate(decile = 0.5)
hei6.veg <- vegetarians.df %>% filter(HEI > quantile(HEI, .5) &
HEI < quantile(HEI, .6))  %>% mutate(decile = 0.6)
hei7.veg <- vegetarians.df %>% filter(HEI > quantile(HEI, .6) &
HEI < quantile(HEI, .7))  %>% mutate(decile = 0.7)
hei8.veg <- vegetarians.df %>% filter(HEI > quantile(HEI, .7) &
HEI < quantile(HEI, .8))  %>% mutate(decile = 0.8)
hei9.veg <- vegetarians.df %>% filter(HEI > quantile(HEI, .8) &
HEI < quantile(HEI, .9))  %>% mutate(decile = 0.9)
hei10.veg <- vegetarians.df %>% filter(HEI > quantile(HEI, .9) &
HEI < quantile(HEI, 1))  %>% mutate(decile = 1)
## ADD COLUMNS FOR THE DECILE AND AVERAGE CRP
hei1.veg <- hei1.veg %>% mutate(decile = 0.1, meanCRP =
weighted.mean(hei1.veg$crp, hei1.veg$adjmec))
hei2.veg <- hei2.veg %>% mutate(decile = 0.2, meanCRP =
weighted.mean(hei2.veg$crp, hei2.veg$adjmec))
hei3.veg <- hei3.veg %>% mutate(decile = 0.3, meanCRP =
weighted.mean(hei3.veg$crp, hei3.veg$adjmec))
hei4.veg<- hei4.veg %>% mutate(decile = 0.4, meanCRP =
weighted.mean(hei4.veg$crp, hei4.veg$adjmec))
hei5.veg <- hei5.veg %>% mutate(decile = 0.5, meanCRP =
weighted.mean(hei5.veg$crp, hei5.veg$adjmec))
hei6.veg <- hei6.veg %>% mutate(decile = 0.6, meanCRP =
weighted.mean(hei6.veg$crp, hei6.veg$adjmec))
hei7.veg <- hei7.veg %>% mutate(decile = 0.7, meanCRP =
weighted.mean(hei7.veg$crp, hei7.veg$adjmec))
hei8.veg <- hei8.veg %>% mutate(decile = 0.8, meanCRP =
weighted.mean(hei8.veg$crp, hei8.veg$adjmec))
hei9.veg <- hei9.veg %>% mutate(decile = 0.9, meanCRP =
weighted.mean(hei9.veg$crp, hei9.veg$adjmec))
```

```
hei10.veg <- hei10.veg %>% mutate(decile = 1, meanCRP =
weighted.mean(hei10.veg$crp, hei10.veg$adjmec))
## COMBINE ROWS TOGETHER
heiVeg.df <- rbind(hei1.veg, hei2.veg,
hei3.veg,hei4.veg,hei5.veg,hei6.veg,hei7.veg,hei8.veg,hei9.veg,h
ei10.veg)
## CREATE BOTTOM AND TOP 30% OF HEI SCORE POPULATIONS
bottom30.veg <- heiVeg.df %>% filter(decile <= 0.3)
top30.veg <- heiVeg.df %>% filter(decile >= 0.8)


###### EXPORT INTO EXCEL FOR FUTURE USE
library(rio)
export(bottom30.veg, "Bottom 30% Vegetarians.csv")
export(bottom30.omni, "Bottom 30% Omnivores.csv")
export(top30.veg, "Top 30% Vegetarians.csv")
export(top30.omni, "Top 30% Omnivores.csv")


######
## CREATING GRAPHS AND TABLES, AND PREFORMING TESTS
install.packages('ggplot2')
library(ggplot2)
install.packages('weights')
library(weights)
omnivores.df <- read.csv(file=file.choose()) #read in
omnivores.csv
attach(omnivores.df) #no vegetarians
vegetarians.df <- read.csv(file = file.choose()) #read in
vegetarians.csv
attach(vegetarians.df)


###### GRAPHS
## CRP HISTOGRAMS
library(ggplot2)
ggplot(vegetarians.df, aes(x = crp)) +
  geom_histogram(color = "chartreuse4", fill = "chartreuse4")  +
  labs(title="Vegetarian CRP", x ="CRP", y = "Frequency") +
  theme_bw() + scale_x_continuous(breaks = seq(0, 15, by = 1))
ggplot(omnivores.df[omnivores.df$crp < 13,], aes(x = crp)) +
  geom_histogram()  +
  labs(title="Omnivorous CRP", x ="CRP", y = "Frequency") +
  theme_bw() + scale_x_continuous(breaks = seq(0, 14, by = 1))
## AGE HISTOGRAMS
ggplot(vegetarians.df, aes(x = age)) +
  geom_histogram(color = "chartreuse4", fill = "chartreuse4")  +
```

```
  labs(title="Vegetarian Age (in years)", x ="Age (in years)", y
= "Frequency") +
  theme_bw() + scale_x_continuous(breaks = seq(20, 100, by =
10))
ggplot(omnivores.df, aes(x = age)) +
  geom_histogram()  +
  labs(title="Omnivorous Age (in years)", x ="Age (in years)", y
= "Frequency") +
  theme_bw() + scale_x_continuous(breaks = seq(20, 100, by =
10))
## HEI HISTOGRAMS
ggplot(vegetarians.df, aes(x = HEI)) +
  geom_histogram(color = "chartreuse4", fill = "chartreuse4")  +
  labs(title="Vegetarian HEI", x ="HEI", y = "Frequency") +
  theme_bw() + scale_x_continuous(breaks = seq(0, 100, by = 20))
ggplot(omnivores.df, aes(x = age)) +
  geom_histogram()  +
  labs(title="Omnivorous HEI", x ="HEI", y = "Frequency") +
  theme_bw() + scale_x_continuous(breaks = seq(0, 100, by = 20))
attach(heiOmni.df)
attach(heiVeg.df)
ggplot(heiOmni.df, aes(x = decile, y = meanCRP)) +
  geom_point(shape = 19) + theme_bw() +
  labs(title="Omnivores' Average CRP by HEI Deicile", x ="HEI
Score Decile", y = "Omnivores' CRP") +
  scale_x_continuous(breaks = seq(0, 1, by = 0.1)) +
scale_y_continuous(breaks = seq(0, 1, by = 0.1))
ggplot(heiVeg.df, aes(x = decile, y = meanCRP)) +
  geom_point(color = "chartreuse4", shape = 19) + theme_bw() +
  labs(title="Vegetarians' Average CRP by HEI Deicile", x ="HEI
Score Decile", y = "Omnivores' CRP") +
  scale_x_continuous(breaks = seq(0, 1, by = 0.1)) +
scale_y_continuous(breaks = seq(0, 1, by = 0.1))


###### INFORMATION FOR TABLES
#run each line for crp, HEI and age
weighted.mean(omnivores.df$age, omnivores.df$adjmec)
weighted.mean(vegetarians.df$age, vegetarians.df$adjmec)
library(weights)
wtd.t.test(omnivores.df$age, vegetarians.df$age, weight =
omnivores.df$adjmec, weighty = vegetarians.df$adjmec)
attach(bottom30.omni)
top30.omni <- read.csv(file=file.choose())
bottom30.omni <-read.csv((file=file.choose()))
top30.veg <- read.csv(file=file.choose())
attach(top30.omni)
```

```
attach(bottom30.veg)
attach(top30.veg)
#comparing
wtd.t.test(bottom30.omni$age, top30.omni$age, weight =
bottom30.omni$adjmec, weighty = top30.omni$adjmec)
wtd.t.test(bottom30.veg$age, top30.veg$age, weight =
bottom30.veg$adjmec, weighty = top30.veg$adjmec)
#same populations
weighted.mean(bottom30.omni$age, bottom30.omni$adjmec)
wtd.t.test(bottom30.omni$age, top30.omni$age, weight =
bottom30.omni$adjmec, weighty = top30.omni$adjmec)


######
## MULTIPLE LINEAR REGRESSION
#whole population
omni.lm <- lm(crp ~ age + HEI, weights = adjmec, data =
omnivores.df)
summary(omni.lm)
veg.lm <- lm(crp ~ age + HEI, weights = adjmec, data =
vegetarians.df)
summary(veg.lm)
#top and bottom 30% HEI scores
top30Omni.lm <- lm(crp ~ age + HEI, weights = adjmec, data =
top30.omni)
summary(top30Omni.lm)
bottom30Omni.lm <- lm(crp ~ age + HEI, weights = adjmec, data =
bottom30.omni)
summary(bottom30Omni.lm)
top30veg.lm <- lm(crp ~ age + HEI, weights = adjmec, data =
top30.veg)
summary(top30veg.lm)
bottom30veg.lm <- lm(crp ~ age + HEI, weights = adjmec, data =
bottom30.veg)
summary(bottom30veg.lm)


###### CITATIONS
citation(package = "RNHANES")
citation(package = "hei")
citation(package = "dplyr")
citation(package = "rio")
citation(package = "weights")
```