UKRAINIAN CATHOLIC UNIVERSITY

BACHELOR THESIS

Association of data-driven indentified dietary patterns with chronic disease prevention

Author:
Yana HOLOBORODKO

Supervisor:
Andrii CHEREPANIAK

A thesis submitted in fulfillment of the requirements for the degree of Bachelor of Science

in the

Department of Computer Sciences and Information Technologies
Faculty of Applied Sciences

Lviv 2024

Declaration of Authorship

I, Yana HOLOBORODKO, declare that this thesis titled, "Association of data-driven indentified dietary patterns with chronic disease prevention" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: _____

Date: _____

"What worries you masters you."

— John Locke, An Essay Concerning Human Understanding - Volume I

UKRAINIAN CATHOLIC UNIVERSITY

Faculty of Applied Sciences

Bachelor of Science

Association of data-driven indentified dietary patterns with chronic disease prevention

by Yana HOLOBORODKO

Abstract

Dietary patterns and different food groups are important factors in preventing and moderating the progression of a range of chronic diseases. This paper proposes to use data-driven methods in combination with prior knowledge of nutritional epidemiology 1) to find the most influential factors among food groups that determine population variance in diverse eating habits, 2) to identify dietary patterns, and 3) to find associations of these results with a number of chronic diseases among NHANES 2017 - March 2020 data taking into account, among other things, the respondents' gender. PCA was used as a dimensionality reduction technique, which showed that the amount of fruits (loading = 0.20 in joint for male and female data), eggs, beans and other non-animal protein (0.16), vegetables (0.26) and grains (0.24) consumed by respondents most explain the variance of dietary patterns in this population. While among men, the presence of alcohol in the diet plays a significant role in differentiating as well. Using K-means for clustering, meaningful for interpretation dietary patterns common to both sexes were named "Whole Foods" (mostly grains, fruits, non-meat protein and dairy products), "Unbalanced" (sandwiches, potatoes and sugars) and "Meat & Alcohol". The associations of dietary patterns and food groups were obtained by using RRR, which proved to be more effective than the usual linear regressions common for this task. The most sensitive to dietary habits in this study were waist-hips ration and high-density lipoprotein, which are among the indicators of obesity and dyslipidemia, respectively. Among the food groups, "Alcohol", "Meat", "Grains", "Sandwiches", and "Snacks&Sweets" had the strongest total associations with chronic disease indicators. Link to the implementation of all the methods can be found here.

Acknowledgements

Five pillars of iron held my sky throughout my studies, which ensured the creation of this thesis. This work came to life primarily thanks to my family, who gave me opportunities, motivation, freedom, and support; later, thanks to my girlfriends philology students, who made me feel warm and safe; my wonderful, every single one of my roommates, who shaped and molded me over the past four years; gym, which firmly kept my mental health; my nation and its soldiers, who instilled in me a sense of strength and dignity.

Contents

List of Figures

List of Tables

List of Abbreviations

BMI	**B**ody **M**ass **I**ndex
CHD	**C**oronary **H**eart **D**isease
COPD	**C**hronic **O**btrusive **P**ulmonary **D**isease
CRP	**C**-**r**eactive **P**rotein
CVD	**C**ardio**v**ascular **D**iseases
DBP	**D**iastolic **B**lood **P**ressure
DP	**D**ietary **P**attern
FMM	**F**inite **M**ixture **M**odel
HDL-C	**H**igh-**D**ensity **L**ipoprotein **C**holesterol
LDL-C	**L**ow-**D**ensity **L**ipoprotein **C**holesterol
LBXGH	**L**abaratory **B**lood E**x**amination **G**lyco**h**emoglobin
LBXTC	**L**abaratory **B**lood E**x**amination **T**otal **C**holesterol
LBDHDD	**L**abaratory **B**lood **D**irect **H**igh-**Density** Lipoprotein Cholesterol
MEC	**M**obile **E**xamination **C**enter
PCA	**P**rincipal **C**omponent **A**nalysis
RRR	**R**educed **R**ank **R**egression
SBP	**S**ystolic **B**lood **P**ressure
SSB	**S**ugar **S**weetened **B**everages
TC	**T**otal **C**holesterol
TG	**Trig**lyceride
TT	**T**reelet **T**ransform
T2D	**T**ype **2** **D**iabetes
UA	**U**rine **A**nalysis
WHR	**W**aist-**H**ip **R**atio

To my sister

Chapter 1

Introduction

1.1 Background & Motivation

A healthy diet is an individualised way of consuming food that optimises the amount of vitamins, minerals and other nutrients groups absorbed and gives a person the necessary energy to perform their daily tasks. Eating is a basic human need, which can have many variations. And an infinite number of such food combinations will be considered healthy, while many eating patterns and habits will be labeled as harmful.

As early as 1747, when humanity discovered that a sufficient amount of citrus fruit could help fight off scurvy, nutrition and its balance reached a new level of importance and began to be perceived as something more than just energy for life or enjoyment of flavours. That's why it's not the presence or absence of certain foods in the diet that is important, but rather the integrity of the diet and its ability to meet nutritional needs and be within the required calories. [18]

According to statistics, millions of people in the United States suffer from diseases caused or aggravated by physical inactivity and poor diet: obesity - just over 41 million[35], hypertension among US adults $\geq$20 years old - 46.7%[32], 11.6% of the population has diabetes, and almost 30% have prediabetes[3], 59.0% of non-Hispanic (NH) Black females and 58.9% of NH Black males had some form of CVD[32].

Consequently, avoiding certain eating habits and, conversely, adhering to others can often help prevent such disastrous statistics. And that's where dietary patterns (DPs) help to differentiate people's eating habits holistically and take into account not specific foods, but the balance of the diet as a whole, and analyse trends among the population. "Dietary patterns are defined as the quantities, proportions, variety, or combination of different foods, beverages, and nutrients in diets, and the frequency with which they are habitually consumed... Inappropriate dietary patterns are associated with risk of negative consequences in terms of diet-related chronic diseases, like cardiovascular disease, obesity, type 2 diabetes, and cancer."([48], p.1)

However, over the years, food trends and availability change and influence variations in people's DPs. Therefore, what are these most influential factors now and how can the benefits of the available data on the human diet be maximised is always an open question that requires more and more research. In addition, the use of different methods to recognize these patterns in the data yields different results. It is the qualitative search for patterns and the associations between them and food patterns that provide insight into how nutrition affects human health.

But before we look for the perfect key to humanity's nutrition, we should understand that a significant proportion of humanity is not concerned about their diet and does not realize its importance. According to Statista, only 44% of Gen Z members are actively trying to eat healthy, which, unfortunately, does not guarantee that

they have enough knowledge and are not victims of many misconceptions that appear on the Internet every day. While the highest rate among generations is among Baby Boomers, still only 58% of whom believe that they make a lot of effort to eat healthy.[7] Therefore, I hope that in the future, this work will contribute to raising awareness of existing, relevant and popular eating patterns among people in order to reduce mortality from preventable chronic diseases through good nutrition, improve overall health and well-being, and increase people's ability to more objectively assess their diets and recognize threats in their eating habits.

1.2 Goals & Contribution

This study aims to investigate DPs in the United States based on dietary data from the National Health and Nutrition Examination Survey and to explore the relationship between dietary habits and health status, especially the association with chronic diseases. In every similar study with the analogous goal, there is a moment of subjectivity in choosing a food group to focus on, a list of diseases selected for study, data from different populations and years, etc. Hence, this paper will try to focus on finding associations with obesity, dyslipidemia, hypertension and diabetes. As a result, the use of data-driven methods to recognize these patterns in this work will 1) provide a comprehensive overview of food groups variations in people's eating behavior and general trends, 2) cluster the population into mutually exclusive groups due to similarities in their diets, and 3) identify foods and food groups that potentially affect the exacerbation of chronic diseases. As for the gap that this study is trying to close, while performing the listed tasks, attention will be focused on the gender of the respondents, which is often omitted in similar studies. Also, the use of RRR has significant potential for finding associations of patterns with diseases, but its application for this task is very little described in publicly available sources.

1.3 Structure Of The Thesis

In order to achieve the desired result, the following objectives should be fulfilled within the framework of this work: i) review related research articles that would solve a similar problem to better understand the methods used, their advantages and limitations, the results obtained and how they are interpreted, ii) describe the data from the National Health and Nutrition Examination Survey to determine what features of the data contribute to the achievement of the goal and what may get in the way, iii) describe the theoretical part in detail explaining how the selected tools and approaches, such as PCA and RRR, work and the specifics of their implementation, iv) apply and adjust the models for recognizing dietary patterns and summarize the results and describe patterns' features, v) establish the relationship between chronic diseases and dietary patterns.

1.4 Challenges & Constraints

Certainly, this study has its limitations. First of all, its results Cannot be generalized due to the specific sample of respondents and demographics, respectively. When applying models such as PCA or RRR, it should be understood that the author of the study faces the difficulty of making a subjective choice about the number of components and the assignment of the correct cluster based on factor loadings and

food consumption frequencies. The interpretation of associations between DPs and health outcomes is also subjective. Regarding the data itself, it should be noted that all respondents' answers that are not the results of laboratory tests and measurements may contain bias. It is also difficult to take into account the importance of physical activity and lifestyle and habits in general in this study, although they certainly have a strong impact on health. In addition, the data for 2022-2023 had not yet been published at the time of completion of this diploma, so the latest available data is from before the Covid pandemic began. Therefore, the uncertainty from this factor should also be taken into account.

Chapter 2

Background Information & Related Works

2.1 Background Information

This section will review some of the necessary materials on which the selection of approaches for finding dietary patterns in the data and the selection of factors to be taken into account when analyzing relationships with chronic diseases will be based. The chapter will cover a review of 1) the statistical methods used to identify food patterns to understand the existing methods and justify the continued use of several of them, 2) scientific articles that support theories about the impact of the food patterns on the chronic diseases.

The outcome of this section should be 1) a clear justification of the chosen methods, the application and results of which in this context will be subsequently discussed in related works, and the principle of operation and implementation will be subsequently described in the methodology, 2) a list of chronic diseases that will be the focus of the study and highlighting the factors in nutrition that, according to existing scientific works, may have an impact on their progression or prevention.

2.1.1 Dietary Patterns Recognition Methods

Ever since ancient times, humanity has encountered the concept that certain foods can have an impact on an individual's health. First, Hippocrates of Kos wrote about the possibility of food being a medicine, then Theophrastus of Eresos recommended the consumption of garlic for a healthy heart, and then the Greek physician Pedanius Dioscorides wrote about the preventive properties of this product.[38] Indeed, for centuries, the impact of individual factors on human health has been studied in great detail. However, in recent decades, much more attention has been paid to the study of the impact of general features of human nutrition on the state of health. After all, often a single component does not have a sufficient or even noticeable impact on the development of chronic diseases. While a person's dietary habits may have a greater impact in the long term, the analysis of a single food does not provide any information about the cumulative effect, the impact of the combination and balance with other food components, and does not allow for an isolated assessment of the relationship between a component and a disease. Therefore, it is not surprising that the focus of research has changed in this vector and brought the recognition of DPs to a new level of importance.

Following this, there are currently many statistical approaches that have been tested to a greater or lesser extent to identify dietary patterns and assess their relationship with health status. In the article by Zhao, Li et al[46], 2021, the following 3 groups are proposed for consideration with an overview of statistical methods for

analyzing these patterns.

Investigator-driven methods

This method is based on assessing the quality of an individual's diet according to its compliance with general dietary recommendations for disease prevention. The method is based largely on prior knowledge and generally accepted notions of healthy eating, so it is subjective and the score may vary depending on the approach to its calculation. Among the most well-known existing indices are The Healthy Eating Index (HEI)[26], the Diet Quality Index (DQI)[33], the Healthy Diet Indicator (HDI)[20] and the Mediterranean Diet Score (MDS)[28]. The algorithm for calculating such a method is usually the same (Fig. 2.1), but the breakdown, scoring system and scale, index rank, and resulting relationships with diseases are different, as they are the result of subjective choices by the inventors of the indices.

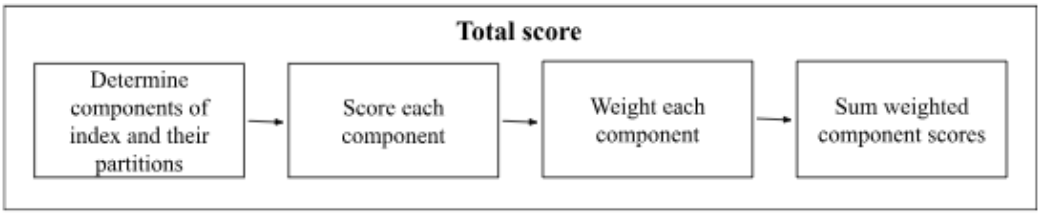

FIGURE 2.1: Algorithm of index creation.

A clear limitation of this approach is that this score does not characterize or provide information about the differences in components in human dietary patterns, so many indices that lie in a small range relative to each other may belong to completely different dietary patterns.[30]

Data-driven methods

Data-driven methods in nutrition are characterised by the same approach to obtaining information about human eating behaviour. First of all, they are used on the collected data on human food consumption, which are presented in the format of a frequency questionnaire, a food journal, or data resulting from a survey on a person's recollection of their diet over the past 24 hours. Secondly, these methods work mostly on the basis of the technique of reducing the number of dimensions and have many variations, some of which will be presented in the three subgroups below.

The most commonly used methods are Principal Component Analysis (PCA) and Exploratory Factor Analysis (EFA), which are very similar in nature, so they are rarely used together in the same study. There is also a degree of subjectivity in the choice of food groups, the threshold for the factor loadings, etc., and the combination of groups is linear, which makes it difficult to interpret the patterns obtained. However, the clear advantages of these approaches are their ability to show the prevalence of a given pattern in the population and to take into account and reflect different eating behaviours in terms of components, thus being easier to interpret.[46]

While the methods mentioned above show how each of the identified patterns applies to each individual, clustering methods assign individuals to the groups that are most characteristic of them. Zhao et al.[46] distinguish two categories of clustering methods in nutritional epidemiology: 1) traditional cluster analysis (TCA) and 2) finite mixture model (FMM). As for the second category, unlike TCA, FMM is

not popular in use because of its complexity, which grows out of proportion to the improvement in results, even compared to the conventional k-mean.

Among the data-driven methods for identifying dietary patterns, Treelet Transform stands out, which is essentially a combination of the strengths of PCA and hierarchical clustering. However, its advantage is that it solves the problem of the inability to analyze the quantitative indicators of components. This method also produces a set of factors, but due to the large number of zero weights, only a small part of the food groups is taken into account and makes it easier to interpret.

Hybrid methods

These methods are the result of a combination of priori knowledge and the use of existing data to find dietary patterns that could best explain the relationship between patterns and disease. This is where their name comes from, as they are by nature combinations of best practices from the previous two groups of methods.

This group includes many methods: reduced rank regression (RRR), data mining (DM), least absolute shrinkage and selection operator (LASSO), etc. All of them have their own implementation features and certain results of use (Fig. 2.2), but this work will focus on the RRR method, because it maximizes the possible benefit of the knowledge gained earlier about the influence of biological properties of the patterns in connection with the disease development. This prior knowledge is used to select intermediate response variables[46]. Since this paper aims to use the available NHANES 2017-2020 nutrition data, data-driven and hybrid methods are of particular interest. Given the difference between all approaches and their complexity, it is worth choosing two methods within the framework of one study that would allow us to get different results and interpretation and, in the long run, more insights. Therefore, in the following, mainly PCA will be considered, which reflects the distribution of the population among the identified patterns and takes into account all components, and in contrast, RRR, which takes into account the knowledge previously gained in the field of nutritional epidemiology and identifies dietary patterns with the goal of finding those associated with chronic diseases.

Methods group	Features		
	Description	Advantages	Limitations
Investigator-driven methods	Evaluates a person's diet according to its compliance with the recommendations and is represented by dietary quality scores	1) Takes into account nutritional recommendations for disease prevention based on scientific research; 2) scores are easy to interpret	1) Subjectivity of assessment; 2) lack of differentiation of patterns and their components
Data-driven methods	Based mostly on dimensionality reduction techniques and use food questionnaires and records		
PCA & EFA	PCA: replaces a set of possibly correlating food groups with uncorrelated PCs EFA: decomposes food groups into common factors and special factor	1) Reflect the characteristics of food patterns; 2) describe the distribution of the population in the consumption among patterns	1) Subjective in the choice of food groups, number of PCs, thresholds for factor loadings, etc.; 2) based on linear combinations that can complicate interpretation

Cluster analysis | TCA	Separate population into mutually exclusive clusters; "hard" clustering: one person - one cluster. Examples of TCA methods: k-means, Ward's minimum-variance method, flexible-beta clustering.	1)distinct subgroups; 2)intuitive results which can be as a dendrogram visually	1) Cluster is assigned with probability 1 or 0 and uncertainty is not considered; 2) subjectivity in food grouping, clustering algorithm, # of clusters, etc.; 3) fails to give numeric summary variables like factors or components
Cluster analysis | FMM	Based on a latent variable model. "soft" clustering: measures classification uncertainty by calculating a posterior probability of different clusters	1) More flexible than TCA; 2) allow the variances of food consumption frequencies to vary within and between clusters	1) Zero values can violate the distribution hypothesis and dealing with them increases model's complexity 2) algorithm tends to converge to local extremum and convergence speed is slow; 3) quality of the results does not compensate the complexity
TT	Produces a set of factors where only a few of the factor loadings of the food variable are non-zero, simplifying the explanation of factors. Combines PCA and hierarchical clustering benefits.	1)simplifies patterns differentiation with involving of small percentage of food groups; 2) Easy interpretable results and ability to visualise clustering tree	1)subjectivity in the cutting level of three choosing; 2) strong associations among food groups or highly correlated foods can cause inadequate diet complexity reflection
Hybrid methods	Combines prior knowledges on health outcomes and identifies dietary patterns based on data		
RRR	Chooses intermediate response variables based on prior knowledge and selects the best linear combination (similar to PCA) that explains the maximum variance among them	1) Considers prior knowledge about the dietary patterns influence on disease; 2) especially useful in deriving dietary patterns related to disease	1)uncertainty in the intermediate response variable selection; 2) the result is based on the response of the variables, so if a disease was not taken into account at the time of their determination, its relationship with dietary patterns will not be identified

DM	Uses many analysis tools for finding insights from large data-bases. One of the most common tools is decision tree induction for clustering.	1) Can determine extent to which factor in diet is influencing diseases; 2) does not need prior assumptions to find new hypotheses; 3) can consider non-dietary factors like lifestyle, etc.;	1) Requires expertise in selecting rules from a set of generated ones, which the model can find a lot of and can be very complex and difficult to interpret; 2) misclassification often occurs
LASSO	A regression-based method that adjusts the regression coefficients by penalising the absolute values. This, in turn, simplifies the model and helps to avoid overfitting.	1) Simultaneously performs variable selection and regularisation; 2) shrinkage method helps to select more relevant sets of group and this ease the interpreting	1) Almost never used for pattern recognition, so it is difficult to find evidence of its effectiveness and superiority over other methods

FIGURE 2.2: Features of the methods, their advantages and limitations.*Compiled by the author.*

2.1.2 Dietary Factors Influencing Chronic Diseases

In order to find relationships with chronic diseases, one needs to have prior knowledge of research on diseases and their associations with dietary habits. The list of chronic diseases that are usually investigated for the presence of this relationship usually includes T2D, CVD and cancer[42],[31]. The rather comprehensive work by Jayedi et al.[23] considers a few more diseases and is more detailed. A meta-analysis of many studies on the associations of empirically derived dietary patterns with chronic diseases revealed the following insights (Table 2.1):

1) Healthy diets may be associated with with lower risk of T2D, colorectal and breast cancer, fracture and the results of currently available studies provide moderate-quality evidence.

2) There are studies that suggest a link between CORD, depression, CVD mortality, colorectal adenoma, and CHD, but so far their results are low-quality evidence.

3) Respiratory disease, mental illness and site-specific cancers require more research.

In the following, only those variables from the dataset used in this study will be taken into account that are indicators for disease groups whose association with dietary factors has been suggested in previous studies. That is, there is no point in considering diseases whose associations have the status of "to be researched" at this time. In order to further justify the selection of the response variables for the RRR model, it is necessary to understand which factors to take into account for each disease.

Type 2 diabetes

According to research[47], important factors for T2D are age, hypertension, BMI, UA, TG, TC, LDL-C and HDL-C. Other studies have also chosen the following T2D-related biomarkers: adiponectin, leptin, TGs, and C-reactive protein (CRP)[22]; TG and hemoglobin[21].

Outcome	Healthy diet association evidence	Unhealthy diet association evidence
Type 2 diabetes	Moderate	Moderate
Fracture	Moderate	Moderate
Breast cancer	Moderate	Low
Colorectal cancer	Moderate	Low
Metabolic syndrome	Low	Moderate
CORD	Low	Low
Depression	Low	Low
CVD mortality	Low	Low
CHD	Low	Low
Colorectal adenoma	Low	Low
Respiratory disease	To be researched	To be researched
Mental illnesses	To be researched	To be researched
Lung, gastric, prostate, pancreatic cancer	To be researched	To be researched

TABLE 2.1: Robustness of results regarding the association of dietary quality and disease. Compiled by author based on[23]

CHD

Age, blood pressure, smoking, haemoglobin, high-density lipoprotein, Cholesterol and HbA1c are considered to be significant risk factors for coronary heart disease[4].

CVD

Smoking habit, increased body mass, physical (in)activity, arterial hypertension(blood pressure), dyslipidemia (dyslipidemias can change the values of TC, TG, LDL-C, or HDL-C[36]) and diabetes. [14]

Others

Fracture is not applicable as it is usually tested with bone scintigraphy, magnetic resonance imaging (MRI), or computed tomography (CT), the same as breast cancer is also not applicable due to specific tests. COPD and asthma are diagnosed via spirometry and this study does not have access to such data.

2.2 Related Works

Recognizing dietary patterns among available data from different populations is a frequently studied issue [15], [45], [12]. In order to conduct the present study, it is necessary to understand the previous results, the specifics of the analysis, the differences and the gaps in them.

That is why this section will review the previous application of the selected RRR and PCA in similar studies, compare their results and the factors that influenced this difference. It will also analyze which health outcomes earlier studies tried to correlate, how they approached this task, whether they succeeded, and to what extent.

2.2.1 Previous Studies Using RRR & PCA: Specifics & Results

Among the three methods that are distinguished among statistical methods and widely used to identify food patterns, namely investigator-driven, data-driven, and combinations of both methods, the data from the National Health and Nutrition Examination Survey will be best suited to data-driven and hybrid approaches. Its large volume can be best processed by machine, thus revealing existing food combinations in people's diets and the distribution of individuals among them.

Despite the existence of a large number of different methods, PCA is a classic approach to identifying dietary patterns among data and has remained one of the most common in the last decade. Burke et al. relied on this tool in their study [8], where they investigated dietary patterns among the Irish. As a result of applying this method, 5 distinctive dietary patterns were obtained, among which the "seafood-focused" pattern stood out among previous studies of Irish food preferences. According to the authors, this may be due to the observed increase in fish consumption among the nation as a whole. This confirms that people's eating habits change over time and should be studied periodically to understand the current state of affairs.

In contrast, Jacobs et al.[22] used the RRR only for deriving dietary patterns across the multiethnic of Hawaii and California. As their main goals were 1) identify associations of diet with type 2 diabetes and 2) do it separately for different ethnic, the result was one dietary pattern only, but it was the first work in the USA that tried to explain the difference of eating patterns influence among 5 ethnic groups. However, due to the same region of residence of this sample, the analysis showed very little difference between the diets of different ethnic groups.

Sauvageot et al.[37] combined two methods in their work and as a result obtained two very similar dietary patterns, which differed largely in the consumption of vegetables, meat and alcohol, as opposed to a pattern with a significant consumption of fruits, soups, vegetables, fish, etc. This study, unlike other similar previous ones, decided not to exclude people with diagnosed CVD from the population to prevent a biased sample. Although the sample excluded people who indicated that they were on a special diet, there is still a high probability that the impact of the diagnosis on dietary changes was not indicated by the respondents as a special diet.

These papers show a variety of results (Table 2.2) depending on the factors taken into account, the diversity of the population, its size and region of origin, and the different focus of the studies. No studies were found that applied PCA and RRR to the 2017-2020 NHANES to find associations with obesity, CVDs, hypertension, and diabetes.

Author, Year, Title	Used Methods	Main focus	Data	Results
Burke, et al., 2023	PCA, 15 food groups	1) Health outcomes: BMI, diabetes, hypertension, CHD 2) Socioeconomic profile, urban and rural diets differences 3) Comparison with self reported diets	957 adults; Ireland;	5 patterns: “Meat-Focused”, “Dairy/Ovo-Focused”, “Vegetable-focused”, “Seafood-Focused”, “Potato-Focused”
Sauvageot, et al., 2016	PCA, RRR, 45 food groups	CVRF*: 1) obesity, 2) hypertension, 3) diabetes, 4) dyslipidemia	2298 individuals; the Greater Region(Luxembourg, Wallonia (Belgium), Lorraine (France));	PCA and RRR resulted in 2 similar patterns each: “Prudent” “Animal protein and alcohol”
Jacobs, et al., 2017	RRR, 41 food groups	1)T2D* 2) biochemical markers 3) specifics across ethnic groups	10,008; MEC* (Hawaii and California)	1 pattern: low in processed and red meat, SSBs*, diet soft drinks, and white rice and high in whole grains, fruit, yellow-orange and green vegetables, and low-fat dairy.

TABLE 2.2: Related works overview.
Notes. CVRF - cardiovascular risk factors, T2D - type 2 diabetes, MEC - The Multiethnic Cohort included 215,831 African-American, Japanese-American, Latino, Native Hawaiian, and white adults living in Hawaii and California, SSB - sugar-sweetened beverage. Compiled by the author*

2.2.2 Associations of chronic diseases with identified patterns

Since one of the main goal of this study is to investigate the associations of chronic diseases with the identified dietary patterns, it is also important to consider the results of previous studies in this context.

According to Burke et al.[8], none of the patterns identified by PCA had a significant association with hypertension, diabetes, or coronary heart disease (CHD), except for the newly identified "seafood-focused" pattern, which was 5.4 times more likely to appear in the group with CHD that had such a diet. However, in the discussion, the author emphasizes that this result could be due to the fact that the population with diagnosed CHD could consume a special diet with an increased amount of lean fish, as recent studies show its positive effects on the cardiovascular system [44]. As for the other dietary patterns, the "vegetable-focused" diet was 1.9 times more likely to have a healthy body mass index (BMI), and the "meat-focused" diet was 1.46 times more likely to be associated with obesity.

As in the study by Sauvageot et al, in Jacobs et al. the use of RRR is supported by its ability to explain as much variation in response variables as possible. Therefore, the selected response variables are important in both papers (Table 2.3) and important to consider further in this work as well. However, the first paper focuses on diabetes, while the second paper takes a more comprehensive approach to the problem and, in addition, tries to find links between dietary patterns and obesity, hypertension, and dyslipidemia. Importantly, Sauvageot et al. emphasize the better ability of RRR to find patterns with stronger associations with cardiovascular disease.

Jacobs, et al., 2017	**Sauvageot, et al., 2016**
Obesity	
-	BMI(kg/m2), WHR
Hypertension	
-	Use of anti-hypertensive medication, n (%), SBP (mmHg), DBP (mmHg)
Diabetes	
TGs, leptin, CRP, adiponectin	Use of anti diabetic medication, n (%), FPG (mmol/L), Hba1c, n (%)
Dyslipidemia	
-	Use of serum lipid-lowering medication, n (%), TC (mg/dl), TG (mg/dl), LDL-C (mg/dl), HDL-C (mg/dl)

TABLE 2.3: Response variables used in related works. *Compiled by the author.*

Chapter 3

Data Overview

This chapter will cover the description of the data used for this study, discuss data preparation and important nuances when cleaning data and merging multiple sets. Since the data is taken from a large-scale, holistic study, it requires a large number of transformations before working, and an understanding of the appropriateness of using certain variables.

3.1 NHANES 2017-March 2020 Data Description

This thesis will use real-world data to reflect the latest possible relevant population trends and variations among the identified dietary patterns. The National Health and Nutrition Examination Survey (NHANES) conducted by the National Center for Health Statistics (NCHS) is an almost annual collection of data on the nutrition, health, and demographics of the US population.[1] At the time of writing the thesis, this is the latest available data from this particular source, as the outbreak of the COVID-19 pandemic has suspended the possibility of conducting this study for some time. That is why the data from 2017 to March 2020 was chosen, as the 2019-2020 sample itself is not sufficiently representative due to the interrupted data collection and 2021-2023 data has not yet been published.

NHANES is a publicly available data set that is widely used by researchers in the fields of public health, nutrition and dietetics, epidemiology, etc. Since the NHANES 2017-March 2020 is a large database with information on demographics, nutrition, laboratory tests, examinations, lifestyle, habits, mental health, etc., the following is a description of the tables that will be used in this paper.

3.1.1 Demographic Variables

Demographic data are important in this study primarily to observe differences in diet among the genders, to be able to select the right age group for the study (over 14 years), and to understand the general background of respondents having abroader picture. Demographic Variables and Sample Weights is a file containing information about 15560 people, namely:

- interview circumstances (interview period, language of the interview, use of a proxy or interpreter, etc;)
- pregnancy status;
- ratio of family income to poverty guidelines;
- and information about gender, age, race, place of birth, social status, etc;

After analyzing the feasibility of using some variables in this table, 13 columns (Fig. 3.1) were left out of the 29 available.

Variable	Description
SEQN	Respondent sequence number(15560 unique values)
INTRVW_STATUS	Interview and examination status of the participant(Interviewed only/Both interviewed and MEC examined)
GENDR	Gender of the participant (male/female)
AGE_YR	Age in years of the participant at the time of screening. Individuals 80 and over are topcoded at 80 years of age.
AGEMNS	Age in months of the participant at the time of screening. Reported for persons aged 24 months or younger at the time of exam (or screening if not examined).
RACE	1 - Mexican American, 2 - Other Hispanic, 3 - Non-Hispanic White, 4 - Non-Hisanic Black, 6 - Non-Hispanic Asian, 7 - Other Race - Including Multi-Racial
PERIOD	Six month time period when the examination was performed - two categories: November 1 through April 30, May 1 through October 31.
ORGCOUNTRY	Country of birth
TIMEUS	Duration of the participant's stay in the United States
EDUCTN	What is the highest grade or level of school completed or the highest degree received?
MARITALSTS	Marital status (Married/Living with Partner, Widowed/ Divorced/ Separated, Never married, Refused, Don't Know, Missing)
PREGNSTATUS	Pregnancy status for females between 20 and 44 years of age at the time of MEC exam.
INCOMEPOVERTY	A ratio of family income to poverty guidelines

FIGURE 3.1: Demographical Data Variables Description.*Compiled by the author*

3.1.2 Total Nutrients Intakes & Individual Foods

One of the main data required in this work is the nutritional characteristics of the food consumed by a person and records of his/her diet. The data files "Individual Foods" and "Total Nutrients" contain data on food and beverages consumed during the last 24 hours before the NHANES interview. In this analysis, the files are combined by the variable 'SEQN', which is the respondent's serial number. "Total Nutrients" contains information about the energy value of the food consumed, the amount of nutrients in it, and answers about the general characteristics of the respondent (such as the frequency of salt in cooking, the frequency of seafood consumption, etc.). In contrast, "Individual Foods" includes many more rows and more than one entry can be related to each respondent. From this file the information about each individual food consumed, the frequency and amount of it is important for the formation and characterization of DPs.

However, in Individual Foods, more than 7,400 food items are represented by the USDA food code[13], the share of each of which in the diet will be so small that it will not give an idea of more general nutritional features and, accordingly, will make it impossible to identify patterns. Therefore, in this case, the What We Eat in America (WWEIA)[2] categorization is used, where all foods are divided into 15 mutually exclusive groups and 48 subgroups based on similarity in use and nutrient content[43]. Based on the previous resources, and given the emphasis on finding associations between food groups that make up a dietary pattern and chronic diseases, all Food and Nutrient Database for Dietary Studies (FNDDS) codes have been regrouped into the slightly modified 46 food groups. More detailed description of food groups can be found in Appendix A.

These identified 46 food groups will be represented in the model as separate variables, the values of which will be the consumption percentage in the diet of each

respondent:

$$\text{Percentage}_{\text{group}} = \frac{\text{Kcal}_{\text{group}}}{\text{Total\_kcal}_{\text{day}}},$$

where $\text{Percentage}_{\text{group}}$ - the percentage of a specific food group in the diet, $\text{Kcal}_{\text{group}}$ - the calorie content of the specific food group ortion, $\text{Total\_Kcal}_{\text{day}}$ - the total number of calories consumed on that day. For each column, this will be a numerical value and thus will display the distribution of the diet for the day by group.

From "Total Nutrients" 55 columns about nutrients and food habits will be used, while from "Individual Foods" only USDA food codes and calorie content of portions will be used.

3.1.3 Laboratory & Examination Data

Since this work aims to find associations of certain patterns with chronic diseases such as diabetes and hypertension, chronic heart disease and dyslipidemia, obesity, referring to subsections 2.1.2 and 2.2.2, the RRR model requires certain variables that will provide some knowledge about the health status of the respondent in terms of these diseases. NHANES contains files with some data on both measurements and laboratory test results. Therefore, 11 indicators were taken from the data to be used as response variables for RRR (Fig.3.2).

Data File	Relevant Data
Body Measures[1]	BMXBMI - Body Mass Index (kg/m**2), WHRATIO - waist-hip ratio WHRATIO = BMXWAIST/BMXHIP
Blood Pressure[2]	Average of three consecutive systolic pressure measurements: $SBP_{avg} = \frac{BPXOSY1+BPXOSY2+BPXOSY3}{3}$; Average of three consecutive diastolic pressure measurements: $DBP_{avg} = \frac{BPXODI1+BPXODI2+BPXODI3}{3}$
Cholesterol - HDL[3]	LBDHDD - Direct HDL-Cholesterol (mg/dL)
Cholesterol - LDL & TG[4]	LBXTR - Triglyceride (mg/dL) LBDLDL - LDL-Cholesterol, Friedewald (mg/dL)
Cholesterol - Total[5]	LBXTC - Total Cholesterol (mg/dL)
Plasma Fasting Glucose[6]	LBXGLU (mg/dL)
Glycohemoglobine[7]	LBXGH (mg/dL)

FIGURE 3.2: Response variables for chronic diseases.* *1 - [6], 2 - [5], 3 - [9] 4 - [10], 5 - [11], 6 - [34], 7 - [17]*

3.2 Data Preparation

This section depicts the step-by-step processing of the data prior to principal component analysis and clustering (Fig. 3.3). Important steps included filtering the data based on the criteria of having information about a medical examination, age characteristics, unconfirmed pregnancy status in the demographic data, and no outliers in the Total Nutrients calorie count. The data were then combined with information on each meal and its characteristics with a previously prepared dataset containing information on 46 food groups and their consumption by each correspondent. The data on disease indicators from subsection 3.1.3 were added later before the RRR.

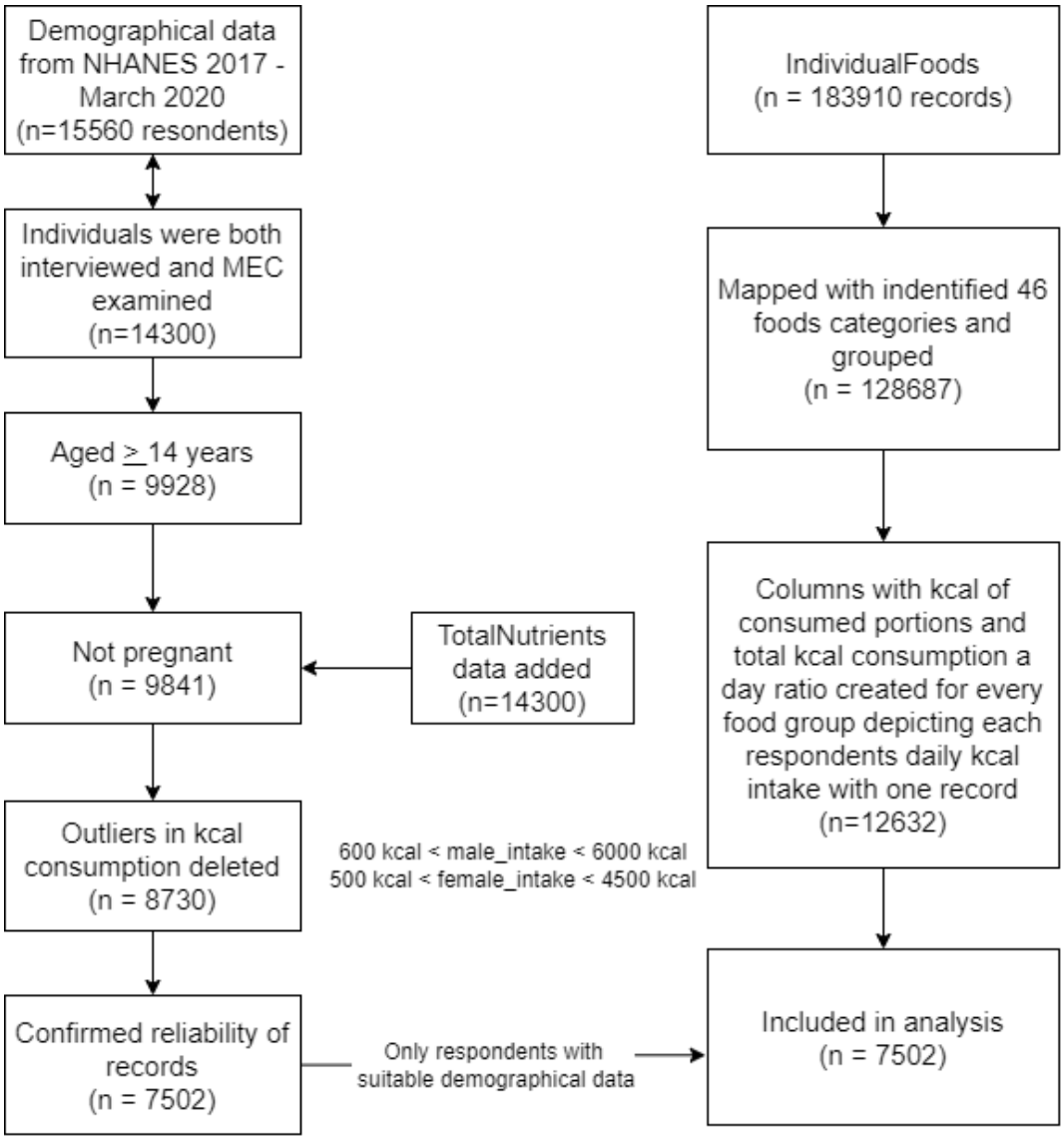

FIGURE 3.3: Data preparation flow

Chapter 4

Methodology & Theoretical Background

In the Background Information section 2.1, existing methods for detecting dietary patterns were already mentioned. The result of this section was the selection of the following two tools: PCA and RRR, which will be further used to find dietary patterns and determine their associations with hypertension, obesity, dyslipidemia and diabetes, respectively. The choice of these diseases is also the result of a review (2.1.2) of all potentially applicable diseases and was based on the availability of relevant information about respondents in the NHANES 2017-2020 data selected for use.

Briefly summarizing why these tools were chosen: 1) previous studies have conducted research on data from countries other than the United States or used other tools for these data; 2) the study expects to gain an understanding of the population distribution among all the resulting DPs obtained with the help of K-mean cluster analysis based on extensive PCA results; 3) using RRR, the study aims to select biomarkers that could potentially be influenced by DPs and analyze the associations with health outcomes. Thus, as a result, this thesis will describe those DPs and food groups that explain the decrease or increase in the risk of the above diseases.

Therefore, this section will briefly describe the selected tools, their working principle and limitations.

4.1 Methods Description

4.1.1 Principal Component Analysis

Scatter Coefficient & Psi-Index

Scatter coefficient is a metric that represents the size of the hypervolume in the feature space. If the variables have little correlation with each other, this volume increases. Therefore, a small value of this coefficient indicates that PCA is appropriate. Algebraically, the Scatter Coefficient is calculated by finding the determinant of the correlation matrix of the dataset. A higher determinant value indicates a larger hyper-volume and less correlation between variables. The maximum possible value is the number of dimensions in the space.[25]

The Psi-Index, on the other hand, should be large enough to confirm the appropriateness of the method. It is calculated as the sum of the squares of the difference between eigenvalues and 1: $\psi = \sum_i (\lambda_i - 1)^2$. And its large value indicates a significant correlation between the variables.[25]

Principal Component Analysis

PCA is a statistical method that takes a lot of data as input and produces non-correlated principal components, which are linear combinations of variables that explain as much of the variables' (in the context of dietary patterns, this means different food groups and nutrients) variance as possible (Fig 4.1).[46]

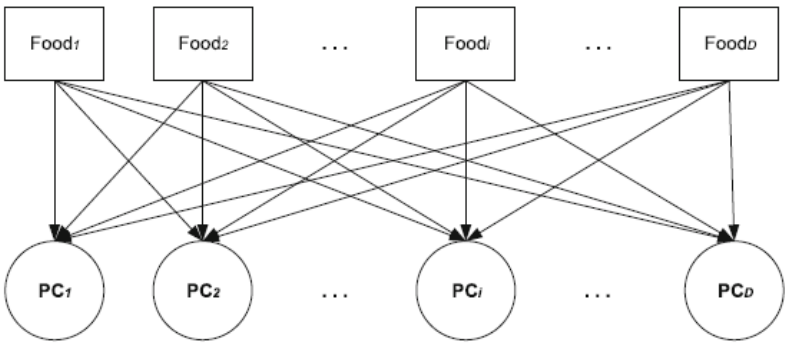

FIGURE 4.1: Principal Component Analysis [46]. *The principal component analysis with D food group variables and nutrients. Each PC is a linear combination of D food groups and nutrients.*

The objective of this algorithm is to find PCs that are orthogonal vectors, i.e., the variables are independent and cover different directions of variation. The vectors try to maximize the variance explained in the data in this way. They are calculated in turn and each subsequent vector tries to explain as much of the remaining variance as possible. Thus, in this work, PCs will describe various dietary characteristics that will differ in the frequency and amount of consumption of food groups based on their nutritional characteristics. If we represent PCs mathematically, each PC is described by the following formula: $PC_k = Xa_k$, where X is an np matrix in which each vector x_j corresponds to a specific p of all numerical variables and has n dimesions (according to the number of entities in the data), and a_k is the eigenvector whose elements are PC loadings. [24]

This paper will use the python library Scikit-Learn to apply this algorithm and remove noise and reduce the number of dimensions.

In this method, one of the main decisions is to choose the number of principal components (PCs) to be taken into account. Usually, researchers choose the first few

PCs that explain the variation the most. Each PC will have its own eigenvalue in the vector, which will indicate the percentage of explained variance in the data. If the eigenvalue of $PC > 1$, then this linear combination of food groups will explain more variation in the population than just one food group. But researchers still have the task of not only choosing the group that has the highest eigenvalue, but also the one that forms a meaningful dietary pattern.[16]

Limitations

This method is quite widespread in use and provides a broad picture of the distribution of food groups among food patterns, but it has significant limitations. For example, it is often difficult to interpret the resulting dietary patterns because all food groups and nutrients in each linear combination are taken into account. Secondly, the patterns obtained may not show any association with the disease of interest. Also, the researcher is required to make a subjective decision regarding the number of PCs, which significantly affects the result.

4.1.2 Reduced Rank Regression

RRR is a statistical method similar in nature to PCA, but aims to reduce the number of spaces in two sets of variables by determining linear functions for multiple predictor variables ($Food$) that maximize the variance in multiple response variables (M). In the context of dietary pattern recognition, if $Food_1, ..., Food_d$ and $M_1, ..., M_g$ are predictors and responses, then the first set of variables is a set of food groups consumed in grams or ratios, and the second set is health indicators in mg/dL for example. Each principal component is a linear combination of food groups that maximally explain the variance (V_{max}) in the intermediate response variables (Fig.4.2).[19]

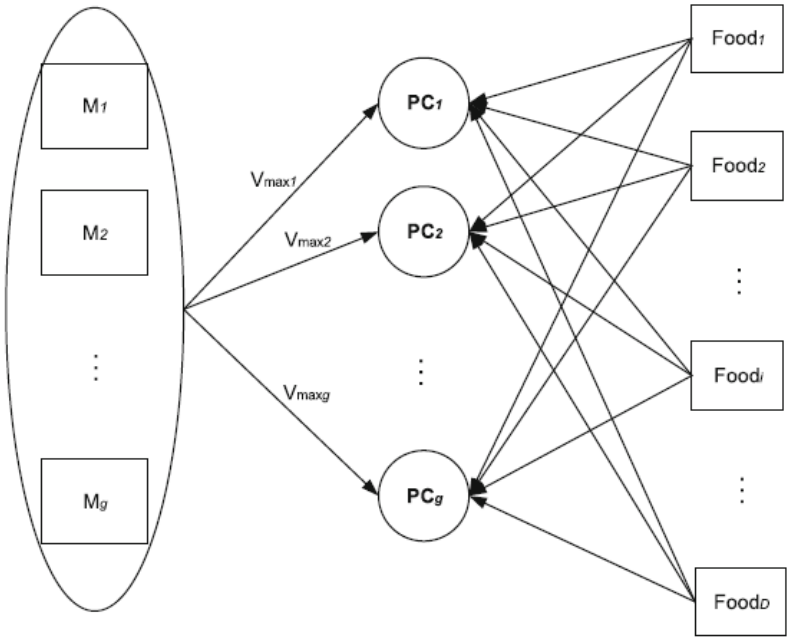

FIGURE 4.2: Reduced Rank Regression [46].*M - intermediate response variables, Food - food groups, PC - principal components corresponding to a dietary patterns which explain maximal variance, $D > g$*

The main point when using this tool is the selection of response variables, which should be equally important for the disease and in theory may be influenced by the

dietary patterns or certain food groups. Prior knowledge from research is required for the selection of response variables guided by a subjective decision.[46]

The main motivation for using this tool in this paper is the ability to determine associations with disease indicators of DPs and food groups. This can also be done using linear regression, but it has several drawbacks:

- The OLS estimate is equivalent to p independent univariate regressions. In other words, there no sharing information across outcome variables;

- There are $p \times q$ regression coefficients to estimate. For every outcome variable q new parameters have to be estimated.

The way to mitigate both effects is to impose a rank restriction on B:

- rank(B)=k is equivalent to having $p \times k$ linear combinations;
- rank(B)=k is also equivalent to writing $B^\top = UV$, where U is $p \times k$, U is $k \times q$, and both are of rank k. This means that at most (p+q)k regression coefficients must be estimated.

The implementation steps:

- Exclude food groups and nutrients that, in the subjective opinion of the researcher, will create noise and can not be significant;

- determine the response variables that will represent health status regarding certain disease;

- normalize and transform the data;

- use rrpack:rrr in R to run the reduced rank resgression and extract coefficients from model to summarize th associations.

Limitations

Despite the fact that this method is quite strong for recognizing dietary patterns that have associations with chronic disease, it does have some limitations. First of all, it relies entirely on prior knowledge, and if some health indicators are not included in the intermediate response variables or are not accurate enough, it is highly likely that this method will not be able to correctly identify the associations. Also, the reliance on prior knowledge will not allow finding more associations with diseases in the data and may lead to some omissions in the data.[46] Also, Hoffmann in his work[19] notes that since the coefficients of a factor score depend on the observed data, the method is not reproducible and cannot be generalized to a food frequency questionnaire data of different origins.

4.1.3 K-means

The k-means is one of the most common and easy-to-understand algorithms in unsupervised learning for clustering in data science, and it is also noted for its efficiency.

$X = \{x_1, ..., x_n\}$ - dataset in a d-dimensional Euclidian space in $\mathbb{R}^d$. $A = \{a_1, ..., a_c\}$ is a set of c cluster centers. Let $z = [z_{ik}]_{n \times c}$ indicates whether point x_i belongs to the k-th cluster, where $k = 1, ..., c$ and z_{ik} is binary. [39]
Then the k-means objective function is defined as:

$$J(z, A) = \sum_{i=1}^{n} \sum_{k=1}^{c} z_{ik} \left\| x_i - a_k \right\|^2$$

Iteratively updating the centers of the clusters and reallocating data points to them minimizing this objective function gives the groups of entities which are the closest one to each over. The algorithm stops when neither the clusters nor the entities change anymore and do not change their centroid assignment, respectively. [39]

To do this, the update equations are calculated at each step, where often the Euclidean ($\|x_i - a_k\|^2$) or Manhattan($|x_i - a_k|$) distances are used:

$$a_k = \frac{\sum_{i=1}^{n} z_{ik} x_{ij}}{\sum_{i=1}^{n} z_{ik}} \text{ and}$$

$$z_{ik} = \begin{cases} 1 & \text{if } \|x_i - a_k\|^2 = \min_{1 \leq k \leq c} \|x_i - a_k\|^2 \\ 0 & \text{otherwise} \end{cases}$$

In this case, the choice of k is important. However, it is expected that for a given dataset, the metrics may be ambiguous due to the large difference in food patterns and the inability to clearly divide into clusters. In this case, the Elbow Method can be used to test several values that are on the bend of the curve and select the clusters that are best suited for interpretation. The Elbow method is a heuristic function that displays the explained variance for the selected range of values on the graph and allows to narrow down the search for a better k.

Chapter 5

Proposed Solution

As for the proposed solution, the reasoning behind the chosen methods has already been discussed in previous chapters. To briefly summarize, the choice of PCA, K-means and RRR is supported by:

1. The feasibility of using data-driven methods to identify dietary patterns;
2. The ability of PCA to reflect the pattern distribution in the entire population, as well as the ability to select the most impactful factors;
3. Intuitiveness and effectiveness of using K-means to find similar features among the population's nutrition;
4. Potential in ability of RRR to reflect DPs' and food groups' associations with chronic diseases with respect to response variables;
5. Moderate complexity of implementing these algorithms and the possibility of applying them to existing NHANES data;
6. Lack of studies applying these tools to the latest US nutrition data for 2017-2020 and lack of focus on gender characteristics;

Regarding the methods chosen and the steps required to achieve the goal:

- NHANES data 2017-March 2020 about demography, 24-h diet recall and meal preferences, and examination and laboratory data will be used for identifying dietary patterns and finding the associations with factors pointing to chronic disease;
- The food codes for each individual meal in the respondent's answer to the dietary question will be classified, based on WWEIA assumptions and recommendations, into 46 mutually exclusive food groups to facilitate interpretation of factor components when using dimensionality reduction tools. Further, the proportion of these dietary groups in the total diet will be calculated for PCA and RRR analyzes;
- PCA analysis will be used to explain the variance among dietary habits. To do this:
  - The scatter coefficient and Psi-index will be used to assess the correlation of data in the dataset;
  - Scree plots will be used to choose the principal components;
  - Factor loadings will be analysed to understand the food groups influence.

- K-means will be used to obtain cluster which will represent DPs among joint dataset, as well as separate datasets for men and women;
- RRR analysis will be applied with BMI, WHRatio, LDL, HLD, blood pressure and total cholesterol as response variables;

Chapter 6

Experiments & Results

6.1 Reflecting Main Factors In Differentiation of DPs

6.1.1 Preliminary Preparation For Analysis

To begin with, a dimension reduction technique was applied to the prepared data with the distribution of daily calories and nutrients intake among 46 different food groups for each respondent (n=7502). The purpose of this step was to facilitate better visualization and understanding of the key factors that have the greatest impact on the distribution of the population by eating habits. The data was taken into account in two formats: with 46 food groups and with 22 (a generalized combination of some of the previous groups A.1). If meaningful results had been obtained from the first dataset, the effects of smaller and more specific groups would have been deduced. However, in the course of the experiment, it was decided to rearrange the foods into more general groups, since then their influence factor in the group is larger and the PCA results are more suitable for interpretation.

When preparing the data for PCA, food groups that appeared in less than 5% of respondents were excluded from both datasets (B.1, B.2). Thus, the group "Baby Foods & Formula" was completely eliminated, and in the dataset with 46 groups, groups 12 and 45 (Appendix A) were also eliminated, which in the second dataset were included in the group representing protein foods other than meat and fish products.

Next, the importance of gender in determining food patterns had to be understood. For men (n=3637) and women (n=3865), the distribution of all nutritional indicators and food groups consumed was displayed using violin plots [41], [40]. Given the shape of the distribution on the graphs, as well as the placement of the median and mean relative to each other, the overall distribution among men and women is almost identical and differs only proportionally. The Wilcoxon test was used to confirm or reject this assumption. The result of the test discarded the hypothesis that the data came from the same distribution, as almost 53% and 39% of the continuous variables in the two datasets had p-values above 0.05. Therefore, the datasets were further separated by gender.

Later, the feasibility of using the PCA method in this task was confirmed. As the heatmap visualization (Fig 6.1) showed that 45 out of 102 and 42 out of 80 variables in the dataset with 46 and 22 food groups, respectively, have a significant correlation >= 0.5 (excluding, of course, pairs of identical variables with a correlation of 1). Subsequently, two indicators were calculated that pointed to the fact that reducing the number of dimensions using PCA was a reasonable option. The Scatter coefficient was close to zero in both cases, indicating a high correlation between the variables, while the Psi-Index was high enough for all partitions of the dataset (joint, male and female) Table 6.1.

Dataset	46Data	46Data(M)	46Data(F)	22Data	22Data(M)	22Data(F)
Max PsiIndex	10100	9900	9900	6162	6006	6006
Psi Index	318.60	302.70	298.05	309.26	292.48	287.76
Scatter coefficient	4.6E-47	2.5E-47	1.0E-46	-1.1E-58	-7.2E-59	5.3E-57

TABLE 6.1: Scatter coefficients And Psi-Indexes.

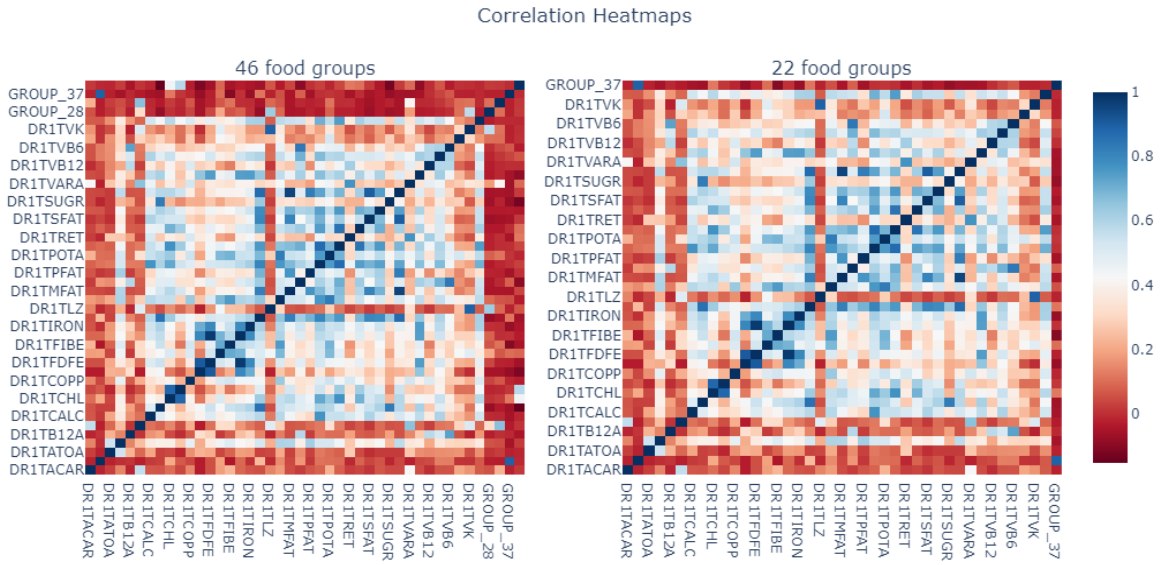

FIGURE 6.1: Food groups correlation matrix

6.1.2 Principal Components Analysis

The next important step is to select the number of principal components (PC). The traditional practice is to choose the number of PCs that would explain 95% of the variance in the dataset. However, in our case, the dietary patterns and nutrition of people in general are very diverse and multicomponent, so given the number of variables, even if we reduce the dimensions due to the correlation of many groups, it is difficult to get a small number of PCs to explain such a large percentage of variation.

Therefore, in order to discard features that create superfluous noise, only those PCs were selected that explained at least 3% of the variance (Table 6.2). In this study, PCA took into account both food groups and their macronutrients and micronutrients, as well as the frequency of meals, salt and water intake, etc. Therefore, the loadings with the highest modulus values were predictably mostly nutrients. In view of this, for each PC, only those factor loadings were displayed that were relevant to one of the food groups and also had a modulus >= 0.15 in order to identify the groups of greatest impact and widespread. Since the results for the dataset with 22 groups

were more meaningful, in the following analysis the dataset with 46 groups was not considered, as this number of groups was too large for meaningful interpretation.

	22Data	22DataMale	22DataFemale
PC1	No food groups	No food groups	No food groups
PC2	Fruits **(0.20)** Plant protein **(0.16)** Vegetables **(0.26)**	No food groups	Fruits **(0.22)** Vegetables **(0.33)**
PC3	Grains **(0.24)**	Fruits **(0.17)** Grains **(0.19)**	Dairy **(0.16)** Meat & Poultry **(-0.16)** Grains **(0.32)** Fats & Oils **(-0.19)**
PC4	Alcohol **(0.19)** Snacks & Sweets **(-0.25**) Meat & Poultry **(0.21)**	Alcohol **(0.30)** Snacks & Sweets **(-0.20)**	Fruits **(-0.15)** Dairy **(0.22)** Meat & Poultry **(0.20)** Snacks & Sweets **(-0.22)** Fats & Oils **(0.21)**
PC5	-	Alcohol **(0.42)** Dairy **(-0.18)**	-
CEV*	36%	46%	28%

TABLE 6.2: Principal Components explaining >= 3% of variance and their factor loadings.* *Cumulative Explained Variance*

As a result, PCA made it possible to reduce dimensionality and capture the main combinations of food groups that are consumed together and most differentiate diets among the respondents of this survey. Given that food groups have less influence on variations in people's diets than more general nutrients, the following trends can still be observed:

- Significant and representative for the diet is the consumption of fruits (positive loading in all three dataset splits), which in the women-only dataset and in the combined dataset move together with the "All vegetables" group from the mean;
- also the consumption of grains and cereals can be identified as a separate strongly influencing factor for dietary differentiation;
- the third group of factors is the combination of high meat consumption and, on the contrary, low consumption of salty snacks and sweets (in the dataset with men, alcohol was also significant).

Based on these results alone, it is not feasible to identify clear specific patterns of eating, but the results can be used to understand the influencing factors and visualize the groups obtained in the next stage of the study using K-means clustering.

6.1.3 K-means

Kmean was used to identify similar food patterns among the dataset. The data on respondents were taken into account, which reflected the percentage of food groups consumed by respondents relative to all food consumed by them per day. In this

way, an attempt to identify patterns was made based on the frequency and amount of consumption of foods and their common combinations.

One of the interim tasks was the choice of the number of clusters. Since the 22 food groups and the great variety among the quantities in which they could be consumed gives an endless number of combinations, it is quite complicated to choose the number of clusters unambiguously. That is why the Elbow Method Fig.6.2 used did not give a clear bend and it was on the verge of 3, 4, 5 clusters, which were used as k. The results for 5 clusters turned out to be the best for interpretation.

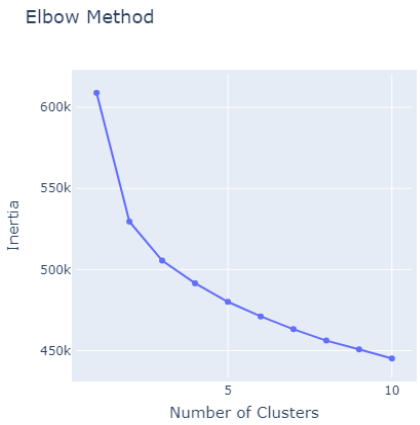

FIGURE 6.2: Elbow method.

A preliminary PC analysis was used to display the distribution of the resulting clusters across the dataset. Since it is clear from the table above Table 6.2 that PC1 does not allow to best trace the influence of food groups on the difference between DPs, the clustering was represented in two-dimensional space, where the ordinate axis and abscissa axis are PC2 and PC3 and the vectors indicate the direction in which the population is moving and the strength of the 22 food groups. For example, Fig. 6.3 shows that Cluster 0 can be described by increased consumption of non-animal protein, fruits, vegetables, and cereals. The other clusters are somewhat more difficult to distinguish in the image, but their location is more influenced by increased consumption of pizza and sugars (clusters 3 and 4) and the amount of meat, potatoes, sandwiches, and snacks in the diet.

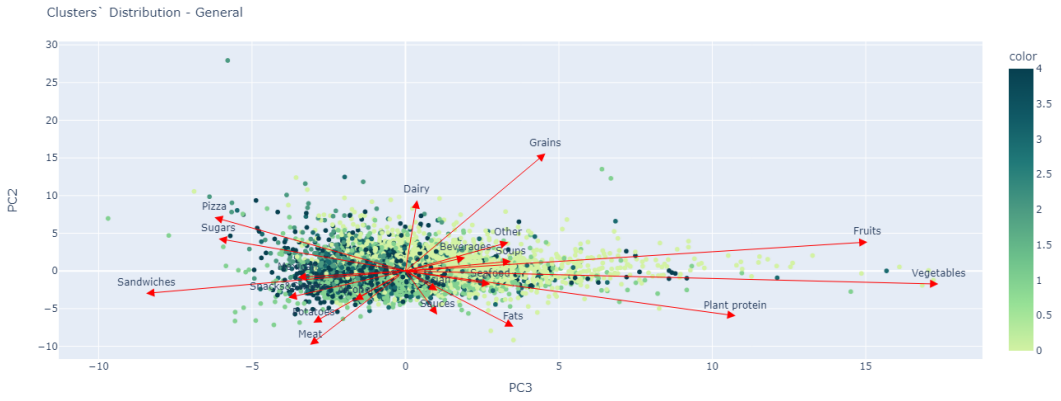

FIGURE 6.3: Clusters distribution visualisation using PC2 and PC3.

The obtained clusters were presented using the features' importance (C). As a result of comparing the received DPs between the three datasets, it was observed that the large clusters have mostly similar features. Therefore, for both genders, the following food patterns were identified:

1. "Whole Foods" C.1, which is characterized by a significant consumption of cereals, fruits, non-meat protein (mostly eggs, beans and vegetable meat substitutes) and dairy products, and vice versa, the absence of big amount of sandwiches, meat and potatoes in the diet;

2. "Unbalanced" C.2, which is characterized by high consumption of sandwiches (also includes burgers and hot dogs), potatoes and sugars (in the form of sugar as a condiment and as sugary drinks);

3. "Meat & Alcohol" C.3, also a large cluster characterized by high consumption of meat and the presence of alcohol in the diet.

Two other clusters covered a much smaller population and were very specific and related to the demographics of the sample. These were DPs characterized by a high consumption of pizza relative to all other categories, as well as a cluster characterized by mostly Mexican and Asian dishes. "Whole Foods", "Unbalanced" and "Meat & Alcohol" were identified not only in the overall dataset, but also for men and women in particular. A pattern called "Pescatarian" C.4 stood out as significantly different for women, which contained a lot of fruits, vegetables and seafood. While for men, a similar cluster was named "Asian" C.5, as it contained mostly Asian dishes and seafood.

6.2 Associations Of DPs With Chronic Diseases

6.2.1 Ordinary Least Squares

First of all, the association was estimated using the usual regression.linear_model from statsmodels library for Python. Chronic diseases were divided into four categories: obesity, hypertension, diabetes, and dyslipidemia. Ordinary least squares regression (OLS) was performed between the standardized data separately for each indicator which could potentially report these diseases in resondents (Section **??**) and their belonging to a particular cluster earlier obtained. Each health indicator was the dependent variable in separate models and the DP was the independent one. Table D lists biggest joint for both genders DPs and summarizes the models. Models with a p-value > 0.5 were considered to be statistically significant for the dietary relationship with the health factor. If we take into account the magnitude of the coefficients and R^2, it is clear that dietary patterns cannot have a direct strong impact on the disease. However, it is possible to conjecture about which dietary pattern is favorable for increasing undesirable indicators and which is vice versa. Thus, the "Whole food" dietary pattern shows an inverse association with obesity and elevated diastolic blood pressure (DBP). However, it cannot be said to have a clear impact on diabetes and cardiovascular disease. Low consumption of foods high in cholesterol, such as red meat, full-fat dairy, pastries, and sweets, showed "Unbalanced" dietary pattern as favorable for lowering the desired indicators that determine cardiovascular disease. "Meat & Alcohol" showed a positive association with hypertension, but the results were not sufficiently indicative of other disease groups.

6.2.2 Reduced Rank Regression

To obtain better results specifically in finding associations between DPS and 22 food groups and health indicators, the RRR method was applied. Using rrpack::rrr, coefficients were extracted first to trace the associations between health indicators such as BMI, waist-hip ration (WHR), SBP, DBP, glycohemoglobin (LBXGH), total cholesterol (LBXTC), direct HDL-cholesterol (LBDHDD) and dietary patterns, and then for each food group in particular. Other indicators were excluded from the model because two-thirds of the respondents were missing and this would have significantly reduced the dataset that could be used to track relationships. Coefficients were extracted from the models to describe the association between indicators in relation to food patterns and food groups. The results were presented in the form of heatmaps (6.4) for convenience.

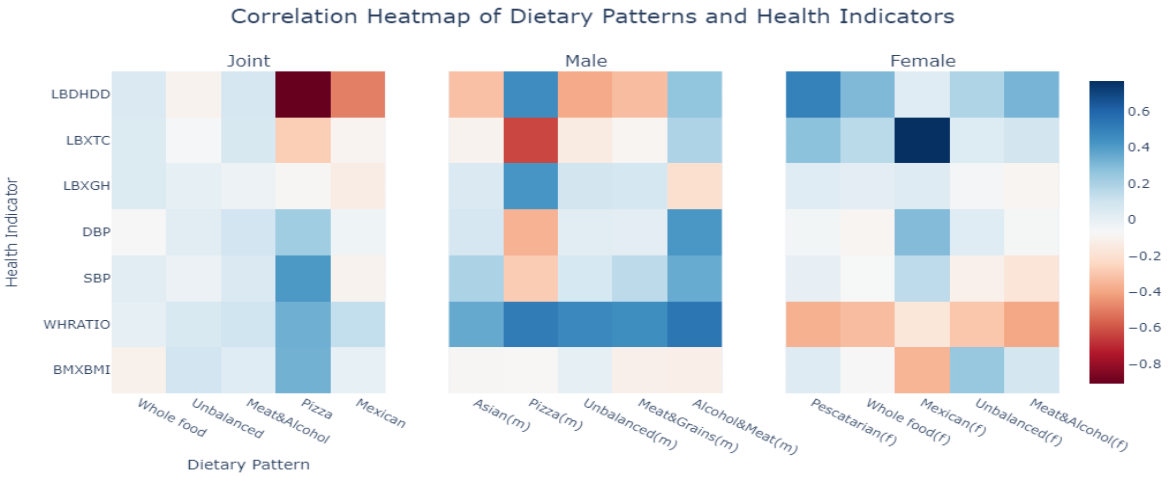

FIGURE 6.4: RRR Coefficient Matrix for Dietary Patterns.

All patterns from the joint, male and female datasets were included in separate heatmaps. It is more difficult to observe clear associations in the joint dataset, given that only a small cluster "Pizza" (n=576) has a strong negative association with high-density lipoprotein cholesterol and a positive association with indicators of hypertension and overweight. While everything is straightforward with the latter two chronic diseases, too low HDL count is bad because it can increase the risk of cardiovascular disease [29], which is not surprising, as to increase this indicator, the diet should include a high consumption of fruits, vegetables, legumes, fish, nuts, and olive oil [27]. The associations for the patterns separately identified for the two genders reflect the interrelationships more clearly. Thus, all the patterns characteristic of the male population have positive coefficients of influence on WHRATIO, which may indicate an overweight. The cluster "Alcohol & Meat(m)" together with "Pizza(m)" are also the most influential patterns if we summarize the absolute values of all their coefficients. However, "Pizza(m)" is identified among a very small number of respondents (n=296). "Alcohol & Meat(m) has a positive coefficient for almost all indicators. For women, the Mexican(f), Pescatarian(f), and Meat & Alcohol(f) patterns have the greatest impact on the indicators. In general, WHRATIO, unlike this indicator for men, remains almost insensitive to DPs for women. We assume that it is the proportion of food consumed, which is why, despite the similar distributional shape for the consumption of all food groups in men and women,

it was important to separate the datasets and trace the effects of nutrition in the sexes separately. Importantly, in the RRR analysis, WHRATIO and LBDHDD were the most influenced indicators by obtained DPs among all populations. However, women's patterns generally include more vegetables and fruits, so this may be the reason their patterns increase the so-called "good" HDL cholesterol, while men's patterns are more dominated by fast foods such as pizza, sandwiches, and higher meat consumption, which lowers this indicator.

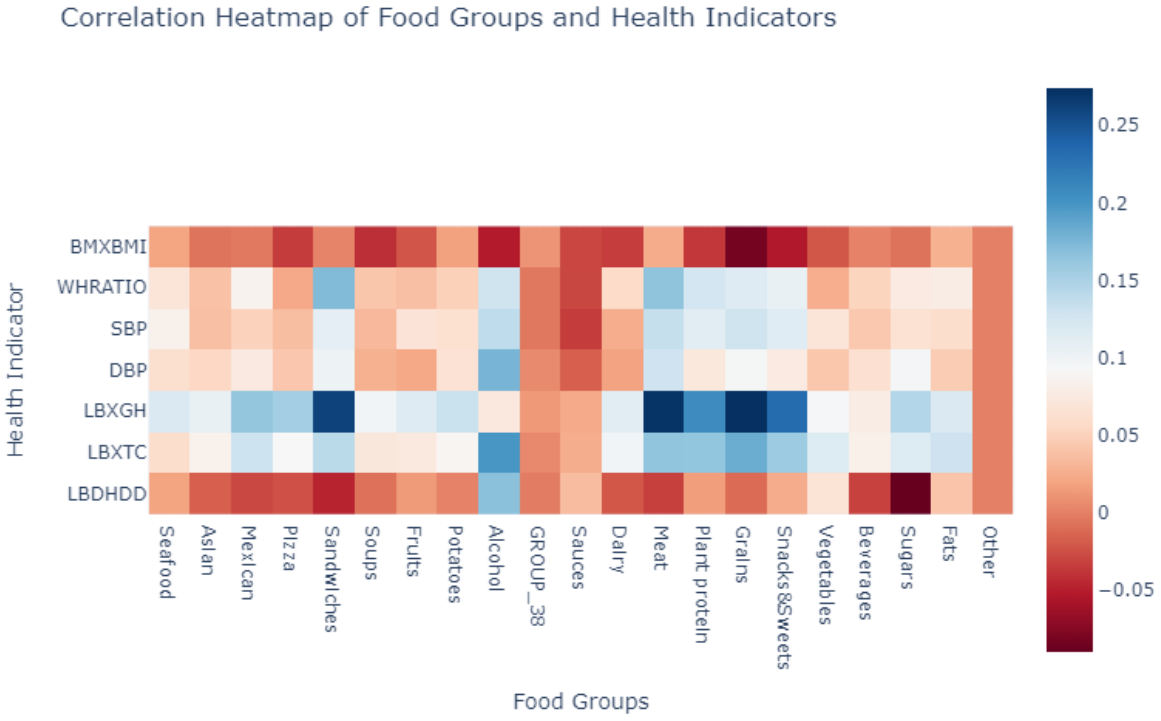

FIGURE 6.5: RRR Coefficient Matrix for Food Groups.

Another approach to assessing food patterns is to look at the associations between food groups and indicators in the initial dataset 6.5. In this case, patterns that contain less exposure to groups with unfavorable coefficients for indicators can be considered healthier. For example, according to the heatmap, increased consumption of meat, cereals, snacks and sweets, and sandwiches has a fairly large positive coefficient on glycohemoglobin, which can indicate diabetes. Among the food groups, "Alcohol" ($s = 5.63$), "Meat" ($s = 5.55$), "Grains" ($s = 5.36$), "Sandwiches" ($s = 5.03$), and "Snacks&Sweets" ($s = 4.5$) had the strongest total associations with chronic disease indicators, where s represents the sum of absolute values of all the coefficients for a certain food groups in the matrix.

Chapter 7

Conclusions

In this study, dietary patterns among the US population in the NHANES 2017-March 2020 data were investigated to understand the most influential factors in differentiating the diet of this sample, to identify popular dietary patterns, and to find their association with chronic diseases. PCA made it possible to reduce dimensionality and capture the main combinations of food groups that are consumed together and most differentiate diets among the respondents of this survey:

1. Fruits: the consumption of fruits emerges as a significant and representative dietary factor, presenting positive loading across all partitions of the dataset (joint, male, and female). Moreover, in the women-only dataset and the combined dataset, fruit consumption aligns closely with the "Vegetables" group.
2. Grains and cereals: Furthermore, the consumption of grains and cereals emerges as a distinct and influential factor contributing to dietary differentiation.
3. Meat: A notable pattern arises from the combination of high meat consumption combined with relatively low intake of salty snacks and sweets, underscoring the influence of these dietary choices.
4. Alcohol: In the dataset focusing on men, alcohol intake emerges as a significant factor contributing to variations in dietary patterns.

The Kmean clustering showed common patterns among both genders called:

- "Whole Foods" - characterized by a significant consumption of grains, fruits, non-meat protein (mostly eggs, beans) and dairy products, and vice versa, the absence of sandwiches, meat and potatoes in the diet;
- "Unbalanced" - characterized by high consumption of sandwiches (also includes burgers and hot dogs), potatoes and sugars (in the form of sugar as a condiment and as sugary drinks)
- "Meat & Alcohol - a large cluster with high consumption of meat and the presence of alcohol in the diet.

Establishing associations of the obtained dietary patterns with health indicators showed that the largest sum of absolute coefficients values (s), reflecting the total impact of nutrition, can be traced to the waist-hips ratio ($s_{joint} = 0.64, s_m = 2.35, s_f = 1.57$) and high-density lipoprotein ($s_{joint} = 1.63, s_m = 1.75, s_f = 1.35$), which are among the indicators of obesity and dyslipidemia, respectively. The "Meat & Alcohol" pattern, which is quite similar for both genders, turned out to be the most influential pattern among the three largest and has positive coefficients that raise most indicators. Among the food groups, "Alcohol" ($s = 5.63$), "Meat" ($s = 5.55$),

"Grains" ($s = 5.36$), "Sandwiches" ($s = 5.03$), and "Snacks&Sweets" ($s = 4.5$) had the strongest total associations with chronic disease indicators.

RRR analysis proved to be a much more effective and convenient tool for tracking associations of both DPs with chronic diseases and individual food groups than conventional linear regression. Thus, more specific individual clusters differing by gender proved to be more indicative of the relationship with chronic diseases. The associations of dietary patterns and food groups were compared using sums of absolute coefficients values (s) obtained by using RRR. The most sensitive to dietary habits in this study were waist-hips ration ($s_{joint} = 0.64, s_m = 2.35, s_f = 1.57$) and high-density lipoprotein ($s_{joint} = 1.63, s_m = 1.75, s_f = 1.35$), which are among the indicators of obesity and dyslipidemia, respectively. The "Meat & Alcohol" pattern, which is quite similar for both genders, turned out to be the most influential pattern among the three largest and has positive coefficients that raise most indicators. Women's patterns generally include more vegetables and fruits, so this may probably contribute to the fact that their patterns increase the so-called "good" cholesterol HDL, while men's patterns are more dominated by fast food such as pizza, sandwiches, and more meat consumption, which may potentinally reduce this indicator in 3 out of 5 patterns and, on the contrary, increases their WHR. The analysis of individual food groups showed a clear negative impact on all indicators of the "Alcohol"($s = 5.63$) group. Among the other food groups "Meat" ($s = 5.55$), "Grains" ($s = 5.36$), "Sandwiches" ($s = 5.03$), and "Snacks&Sweets" ($s = 4.5$) had the strongest total associations with chronic disease indicators.

The main advantage of this work is that it takes into account the difference in the influence of dietary habits of the two sexes and the different impact of nutrition on their health indicators, which has been mostly omitted in previous studies. Although this work has achieved the goal of identifying dietary patterns and underlying factors, there are still certain limitations. These results cannot be generalized, as the sample represents a specific demographic group in the United States and, for example, Mexican food may not be a separate important component for creating dietary patterns in other populations. Also, the data for 2021-2023 has not yet been released, so this survey was conducted in the pre-COVID19 period. Some decisions regarding the application of a particular approach and the choice of parameters are subjective, as they are the result of a synthesis of prior knowledge in the field and the researcher's decisions, which may be biased. Also, the RRR method is promising, but the researcher's little experience with it and the very limited number of sources that would thoroughly explain the use of this method in nutritionology limits the full use of the potential of this tool. In addition, the researcher is not an expert in the field of pathology.

In further research on this topic, it is important to take into account updated data and different demographic groups. Another important factor is to test various methods and approaches to find patterns and not be limited to common tools. Collaboration, especially in making subjective decisions about food groups, classification, identifying correlations with gender and demographics, etc., and the assistance of qualified experts in nutrition, etiology and pathogenesis, could be a great advantage for future research.

Appendix A

Food Groups

#	Food Group	Food Description	Food Categories' numbers
MILK & DAIRY			
1	Nonfat, Lowfat & Flavored Milk	Reduced fat, lowfat, nonfat milk Reduced fat, lowfat, nonfat flavored milk	1004, 1006, 1008, 1204, 1206, 1208
2	Whole & Flavored Milk, Milk Substitutes, Milk shakes	Whole milk, Whole flavored milk Milk shakes and dairy drinks Milk substitutes	1002, 1202, 1402, 1404
3	Cheese	Cheese Cottage/ricotta cheese	1602, 1604
4	Yogurt	Regular yogurt Greek yogurt	1820, 1822
PROTEIN FOODS			
5	Meats	Beef, excludes ground Ground beef Pork Lamb, goat, game Liver and organ meats	2002, 2004, 2006, 2008, 2010
6	Poultry	Chicken, whole pieces Chicken patties, nuggets and tenders Turkey, duck, other poultry	2202, 2204, 2206
7	Seafood	Fish Shellfish	2402, 2404
8	Eggs	Eggs and omeletes	2502
9	Cured Meats/Poultry	Cold cuts and cured meats Bacon Frankfurters Sausages	2602, 2604, 2606, 2608
10	Plant-based Protein Foods	Beans, peas, legumes Nuts and seeds Processed soy products	2802, 2804, 2806
MIXED DISHES			
11	Meat, Poultry, Seafood	Meat mixed dishes Poultry mixed dishes Seafood mixed dishes	3002, 3004, 3006
12	Bean/Vegetable-based	Bean, pea, legume dishes Vegetable dishes	3102, 3104
13	Grain-based	Rice mixed dishes Pasta mixed dishes, excludes macaroni and cheese Macaroni and cheese Turnovers and other grain-based items	3202, 3204, 3206, 3208
14	Asian	Fried rice and lo/chow mein Stir-fry and soy-based sauce mixtures Egg rolls, dumplings, sushi	3402, 3404, 3406
15	Mexican	Burritos and tacos Nachos Other Mexican mixed dishes	3502, 3504, 3506

16	Pizza	Pizza	3602
17	Sandwiches	Burgers Frankfurter sandwiches Chicken fillet sandwiches Egg/breakfast sandwiches Cheese sandwiches Peanut butter and jelly sandwiches Seafood sandwiches Deli and cured meat sandwiches Meat and BBQ sandwiches Vegetable sandwiches/burgers	3702, 3703, 3704, 3706, 3720, 3722, 3730, 3740, 3742, 3744
18	Soups	Soups	3802
GRAINS			
19	Cooked Grains, Cooked Cereals	Rice Pasta, noodles, cooked grains Oatmeal Grits and other cooked cereals	4002, 4004, 4802, 4804
20	Breads, Rolls, Tortillas	Yeast breads Rolls and buns Bagels and English muffins Tortillas	4202, 4204, 4206, 4208
21	Quick Breads, Bread Products, Ready-to-eat Cereals, Cereal And Nutrition Bars	Biscuits, muffins, quick breads Pancakes, waffles, French toast Ready-to-eat cereal, higher sugar (>21.2g/100g) Ready-to-eat cereal, lower sugar (=<21.2g/100g) Cereal bars Nutrition bars	4402, 4404, 4602, 4604, 5402, 5404
SNACKS & SWEETS			
22	Savory Snacks	Potato chips Tortilla, corn, other chips Popcorn Pretzels/snack mix	5002, 5004, 5006, 5008
23	Crackers	Crackers, excludes saltines Saltine crackers	5202, 5204
24	Sweet Bakery Products	Cakes and pies Cookies and brownies Doughnuts, sweet rolls, pastries	5502, 5404, 5506
25	Candys	Candy containing chocolateCandy not containing chocolate	5702, 5704
26	Other Desserts	Ice cream and frozen dairy desserts Pudding Gelatins, ices, sorbets	5802, 5804, 5806
FRUITS			
27	Fruits	Apples Bananas	6002, 6004, 6006, 6008, 600, 6011, 6012, 6014, 6016, 6018, 6010, 6020, 6022, 6024

		Grapes Peaches and nectarines Strawberries Blueberries and other berries Citrus fruits Melons Dried fruits Other fruits and fruit salads Pears Pineapple Mango and papaya	
VEGETABLES			
28	Green Leafy Vegetables	Spinach Lettuce and lettuce salads Cabbage	6409, 6410, 6413
29	Other Green Vegetables	Broccoli Other dark green vegetables String beans	6407, 6411, 6412
30	Red & Orange Vegetables	Tomatoes Carrots Other red and orange vegetables	6402, 6404, 6406
31	Other Vegetables	Onions Corn Other starchy vegetables Other vegetables and combinations Fried vegetables Coleslaw, non-lettuce salads Vegetables on a sandwich	6414, 6416, 6418, 6420, 6430, 6432, 6489
32	White Potatoes	White potatoes, baked or boiled French fries and other fried white potatoes Mashed potatoes and white potato mixtures	6802, 6804, 6806
BEVERAGES			
33	100% Juice	Citrus juice Apple juice Other fruit juice Vegetable juice	7002, 7004, 7006, 7008
34	Diet Beverages	Diet soft drinks Diet sport and energy drinks Other diet drinks	7102, 7104, 7106
35	Sweetened Beverages	Soft drinks Fruit drinks Sport and energy drinks Nutritional beverages Smoothies and grain drinks	7202, 7204, 7206, 7208, 7220
36	Coffee & Tea	Coffee Tea	7302, 7304
ALCOHOLIC BEVERAGES			
37	Alcoholic Beverages	Beer Wine Liquor and cocktails	7502, 7504, 7506

		Liquor and cocktails	
WATER			
38	Plain, Enhanced & Flavored Water	Tap water Bottled water Flavored or carbonated water Enhanced water	7702, 7704, 7802, 7804
FATS & OILS			
39	Animal Fats	Butter and animal fats Margarine Mayonnaise Cream cheese, sour cream, whipped cream Cream and cream substitutes	8002, 8004, 8006, 8008, 8010
40	Salad dressings and vegetable oils	Salad dressings and vegetable oils	8012
CONDIMENTS & SAUCES			
41	Condiments And Sauces	Tomato-based condiments Soy-based condiments Mustard and other condiments Olives, pickles, pickled vegetables Pasta sauces, tomato-based Dips, gravies, other sauces	8402, 8404, 8406, 8408, 8410, 8412
SUGARS			
42	Sugars	Sugars and honey Jams, syrups, toppings	8802, 8806
43	Sugar substitutes	Sugar substitutes	8804
BABY FOODS & FORMULAS			
44	Baby Foods, Beverages, Infant Formulas And Human Milk	Baby food: cereals Baby food: fruit Baby food: vegetable Baby food: mixtures Baby food: meat and dinners Baby food: yogurt Baby food: snacks and sweets Baby juice Baby water Formula, ready-to-feed Formula, prepared from powder Human milk	9002, 9004, 9006, 9007, 9008, 9010, 9012, 9202, 9204, 9402, 9404, 9602
OTHER			
45	Protein and nutritional powders	Protein and nutritional powders	9802
46	Other	Not included in a food category	9999

#	Food Groups	Description
1	Group 7	Seafood
2	Group 14	Asian
3	Group 15	Mexican
4	Group 16	Pizza
5	Group 17	Sandwiches
6	Group 18	Soups
7	Group 27	Fruits
8	Group 32	Potatoes
9	Group 37	Alcohol
10	Group 38	Water
11	Group 41	Sauces
12	Group 1, Group 2, Group 3, Group 4	Dairy
13	Group 5, Group 6, Group 9, Group 11	Meat & Poultry
14	Group 8, Group 10, Group 12	Eggs, Plant Protein & Beans
15	Group 13, Group 19, Group 20, Group 21	Grains & Cereals
16	Group 22, Group 23, Group 24, Group 25, Group 26	Snacks & Sweets
17	Group 28, Group 29, Group 30, Group 31	Vegetables
18	Group 33, Group 34, Group 36	Beverages
19	Group 35, Group 42, Group 43	Sugars
20	Group 39, Group 40	Fats & Oils
21	Group 44	Baby Food & Formulas
22	Group 45, Group 46	Other

TABLE A.1: 22 Food Groups Description

Appendix B

Food Groups Distribution

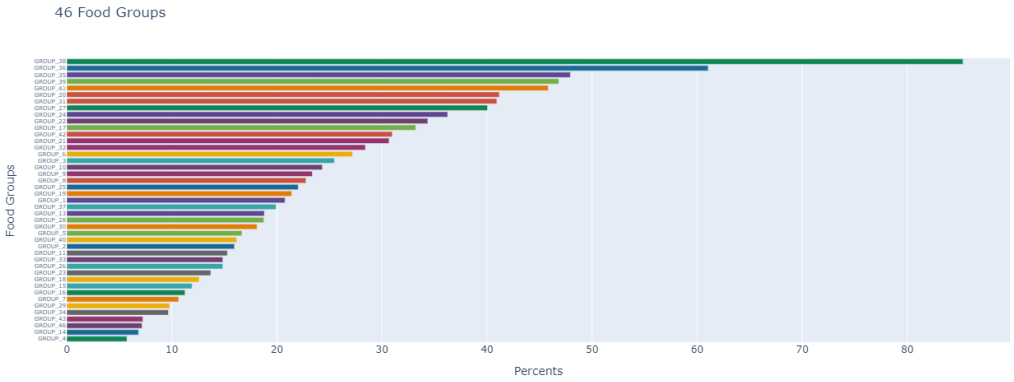

FIGURE B.1: Frequency of dietary group appearance among respondents (46 Food Groups)

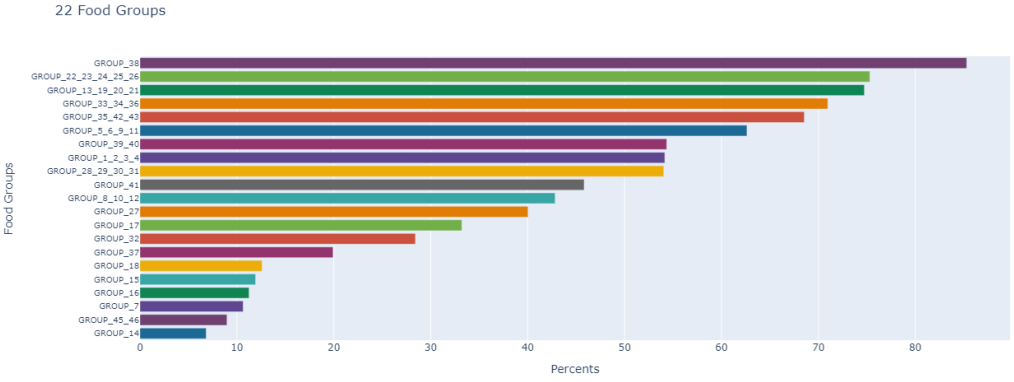

FIGURE B.2: Frequency of dietary group appearance among respondents (22 Food Groups)

Appendix C

Dietary Patterns

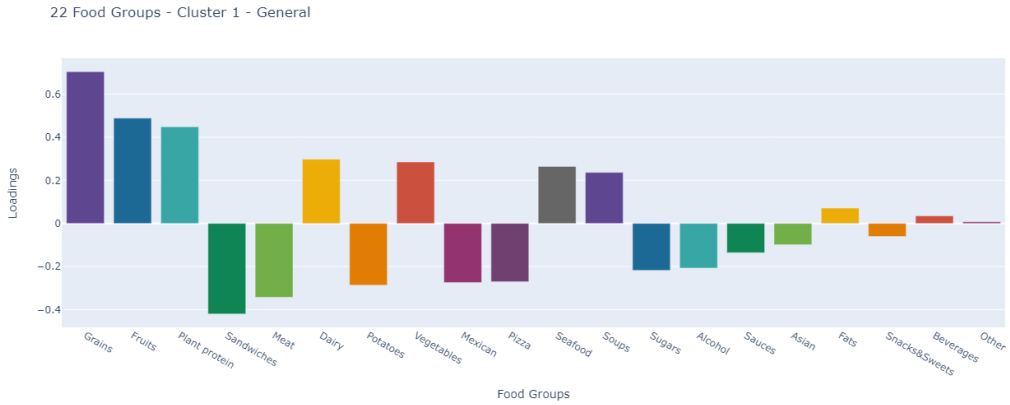

FIGURE C.1: "Whole Foods" dietary pattern's cluster loadings.

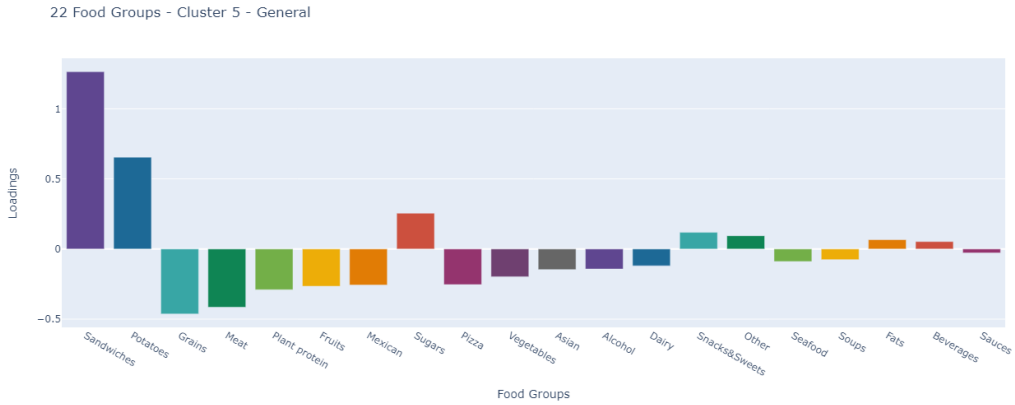

FIGURE C.2: "Unbalanced" dietary pattern's cluster loadings.

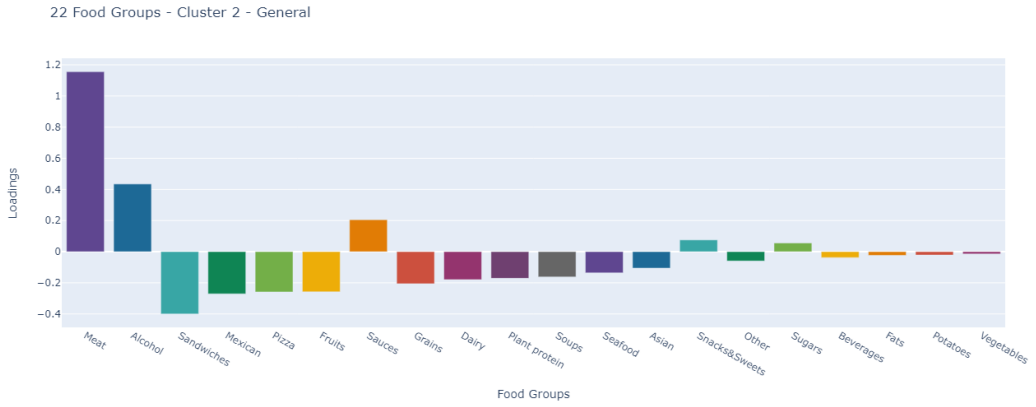

FIGURE C.3: Meat & Alcohol - cluster loadings.

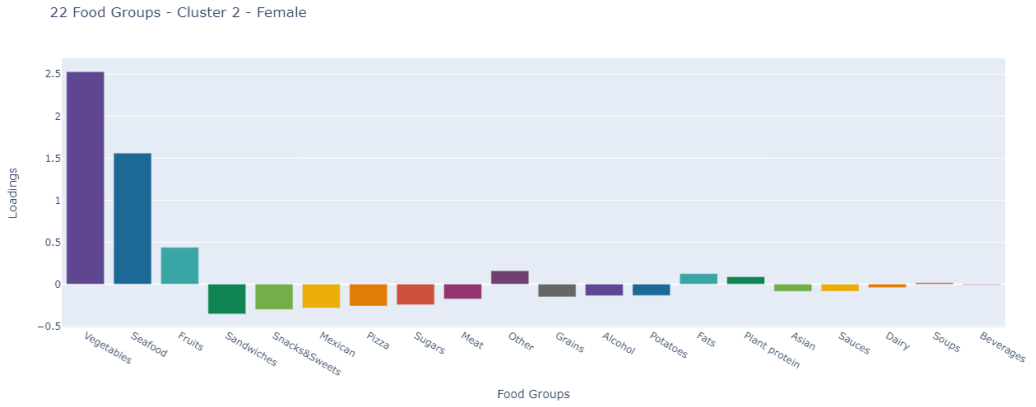

FIGURE C.4: "Pescatarian" dietary pattern's cluster loadings.

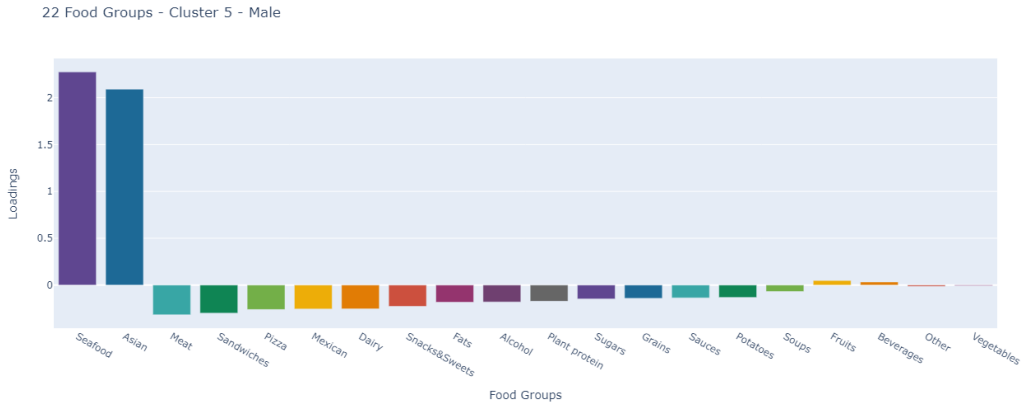

FIGURE C.5: "Asian" dietary pattern's cluster loadings.

Appendix D

OLS results

	Obesity		Hypertension		Diabetes		Dyslipidemia			
	BMI	WHRATIO	SBP	DBP	LBXGLU	LBXGH	LBXTC	LBXTR	LBDLDL	LBDHDD
Whole food (p, R^2,c)	.000, .007, -.082	.015, .001, -.032	p > .5	.000, .005, -.074	p > .5	.001, .002, .045	p > .5	.006, .003, .052	p > .5	p > .5
Unbalanced (p, R^2, c)	.000, .004, -.064	p > .5	p > .5	p > .5	p > .5	p > .5	.000, .003, -.052	.045, .001, -.038	p > .5	.000, .004, -.066
Meat & Alcohol (p, R^2,c)	p > .5	p > .5	.014, .001, .033	.000, .004, .066	p > .5	.020, .001, -.030	p > .5	p > .5	p > .5	.000, .003, .051

Bibliography

[1] Te-Ching; Davy Orlando; Ogden Cynthia L.; Fink Steven; Clark Jason; Riddles Minsun K.; Mohadjer Leyla K. Akinbami Lara J.; Chen. *National Health and Nutrition Examination Survey, 2017–March 2020 Prepandemic File: Sample Design, Estimation, and Analytic Guidelines*. 109. National Center for Health Statistics (U.S.), 2022. URL: https://stacks.cdc.gov/view/cdc/115434.

[2] American Diabetes Association. *DMR - Food Categories What are WWEIA Food Categories?* 2022. URL: https://www.ars.usda.gov/northeast-area/beltsville-md-bhnrc/beltsville-human-nutrition-research-center/food-surveys-research-group/docs/dmr-food-categories/ (visited on 03/29/2024).

[3] American Diabetes Association. *Statistics About Diabetes*. 2021. URL: https://diabetes.org/about-diabetes/statistics/about-diabetes (visited on 03/10/2024).

[4] Sawsan Babiker et al. "Logit model in prospective coronary heart disease (CHD) risk factors prediction in Saudi population". In: *Saudi Journal of Biological Sciences* 28.12 (2021), pp. 7027–7036. ISSN: 1319-562X. DOI: https://doi.org/10.1016/j.sjbs.2021.07.089. URL: https://www.sciencedirect.com/science/article/pii/S1319562X21006720.

[5] *Blood Pressure - Oscillometric Measurement (P_BPXO)*. URL: https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/P_BPXO.htm (visited on 03/29/2024).

[6] *Body Measures (P_BMX)*. URL: https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/P_BMX.htm#BMXBMI (visited on 03/29/2024).

[7] Katharina Buchholz. *Are Americans Trying to Eat Healthy?* 2023. URL: https://www.statista.com/chart/16796/us-interest-in-healthy-food/#:~:text=According%20to%20Statista%20Consumer%20Insights,they%20were%20pursuing%20the%20aim (visited on 02/17/2024).

[8] Daniel T. Burke et al. "Identifying Novel Data-Driven Dietary Patterns via Dimensionality Reduction and Associations with Socioeconomic Profile and Health Outcomes in Ireland". In: *Nutrients* 15.14 (2023). ISSN: 2072-6643. DOI: 10.3390/nu15143256. URL: https://www.mdpi.com/2072-6643/15/14/3256.

[9] *Cholesterol - High - Density Lipoprotein (HDL) (P_HDL)*. URL: https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/P_HDL.htm#LBDHDD (visited on 03/29/2024).

[10] *Cholesterol - Low-Density Lipoproteins (LDL) Triglycerides (P_TRIGLY)*. URL: https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/P_TRIGLY.htm#LBDLDLN (visited on 03/29/2024).

[11] *Cholesterol - Total (P_TCHOL)*. URL: https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/P_TCHOL.htm#LBXTC (visited on 03/29/2024).

[12] Julia R. DiBello et al. "Comparison of 3 Methods for Identifying Dietary Patterns Associated With Risk of Disease". In: *American Journal of Epidemiology* 168.12 (Oct. 2008), pp. 1433–1443. ISSN: 0002-9262. DOI: 10.1093/aje/kwn274. URL: https://doi.org/10.1093/aje/kwn274.

[13] *Dietary Interview Technical Support File Food Codes*. Accessed: 2024-02-22. URL: https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/P_DRXFCD.htm.

[14] Nola I. A. Francula-Zaninovic S. "Management of Measurable Variable Cardiovascular Disease' Risk Factors". In: *Current cardiology reviews* 14 (3), pp. 153–163. DOI: 10.2174/1573403X14666180222102312. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6131408/.

[15] Signe B Frederiksen et al. "Dietary patterns generated by the Treelet Transform and risk of stroke: a Danish cohort study". In: *Public Health Nutrition* 24.1 (2021), 84–94. DOI: 10.1017/S1368980019004324.

[16] Philip M. Gleason et al. "Publishing Nutrition Research: A Review of Multivariate Techniques—Part 3: Data Reduction Methods". In: *Journal of the Academy of Nutrition and Dietetics* 115.7 (2015), pp. 1072–1082. ISSN: 2212-2672. DOI: https://doi.org/10.1016/j.jand.2015.03.011. URL: https://www.sciencedirect.com/science/article/pii/S2212267215002725.

[17] *Glycohemoglobin (P_GHB)*. URL: https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/P_GHB.htm (visited on 03/29/2024).

[18] Hazel A.B. Hiza, Kristin L. Koegel, and TusaRebecca E. Pannucci. "Diet Quality: The Key to Healthy Eating". In: *Journal of the Academy of Nutrition and Dietetics* 118.9 (2018), pp. 1583–1585. ISSN: 2212-2672. DOI: https://doi.org/10.1016/j.jand.2018.07.002. URL: https://www.sciencedirect.com/science/article/pii/S2212267218313315.

[19] Kurt Hoffmann et al. "Application of a New Statistical Method to Derive Dietary Patterns in Nutritional Epidemiology". In: *American Journal of Epidemiology* 159.10 (2004), pp. 935–944. ISSN: 0002-9262. URL: https://doi.org/10.1093/aje/kwh134.

[20] Patricia Huijbregts et al. "Dietary pattern and 20 year mortality in elderly men in Finland, Italy, and the Netherlands: longitudinal cohort study". In: *BMJ* 315.7099 (1997), pp. 13–17. DOI: 10.1136/bmj.315.7099.13. eprint: https://www.bmj.com/content. URL: https://www.bmj.com/content/315/7099/13.

[21] Waleed Noori Hussein, Zainab Muzahim Mohammed, and Amani Naama Mohammed. "Identifying risk factors associated with type 2 diabetes based on data analysis". In: *Measurement: Sensors* 24 (2022), p. 100543. ISSN: 2665-9174. DOI: https://doi.org/10.1016/j.measen.2022.100543. URL: https://www.sciencedirect.com/science/article/pii/S2665917422001775.

[22] Simone Jacobs et al. "Dietary Patterns Derived by Reduced Rank Regression Are Inversely Associated with Type 2 Diabetes Risk across 5 Ethnic Groups in the Multiethnic Cohort12". In: *Current Developments in Nutrition* 1.5 (2017), e000620. ISSN: 2475-2991. DOI: https://doi.org/10.3945/cdn.117.000620. URL: https://www.sciencedirect.com/science/article/pii/S2475299122144598.

[23] Ahmad Jayedi et al. "Healthy and unhealthy dietary patterns and the risk of chronic disease: an umbrella review of meta-analyses of prospective cohort studies". In: *British Journal of Nutrition* 124.11 (2020), 1133–1144. DOI: 10.1017/S0007114520002330.

[24] Cadima J. Jolliffe I. T. "Principal component analysis: a review and recent developments. Philosophical transactions." In: *Series A, Mathematical, physical, and engineering sciences* 374.2065 (2016). DOI: https://doi.org/10.1098/rsta.2015.0202.

[25] Tiago Toledo Jr. *PCA 102: Should you use PCA? How many components to use? How to interpret them?* 2022. URL: https://towardsdatascience.com/pca-102-should-you-use-pca-how-many-components-to-use-how-to-interpret-them-da0c8e3b11f0 (visited on 04/27/2024).

[26] EILEEN T KENNEDY et al. "The Healthy Eating Index: Design and Applications". In: *Journal of the American Dietetic Association* 95.10 (1995), pp. 1103–1108. ISSN: 0002-8223. DOI: https://doi.org/10.1016/S0002-8223(95)00300-2. URL: https://www.sciencedirect.com/science/article/pii/S0002822395003002.

[27] Karla Paulina Luna-Castillo et al. "Functional Food and Bioactive Compounds on the Modulation of the Functionality of HDL-C: A Narrative Review". In: *Nutrients* 13.4 (2021). ISSN: 2072-6643. DOI: 10.3390/nu13041165. URL: https://www.mdpi.com/2072-6643/13/4/1165.

[28] Serrano-Martínez M. Wright M. Gomez-Gracia E. Martínez-González M. A. Fernández-Jarne E. "Development of a short dietary intake questionnaire for the quantitative estimation of adherence to a cardioprotective Mediterranean diet". In: *Eur J Clin Nutr* 58 (2004), 1550–1552. DOI: https://doi.org/10.1038/sj.ejcn.1602004.

[29] Thulesius H. O. Hillman M.-Svensson R. Landin-Olsson M.- Thunander M. Melin E. O. "Lower HDL-cholesterol, a known marker of cardiovascular risk, was associated with depression in type 1 diabetes: a cross sectional study." In: *Lipids in health and disease* 18.1 (2019), p. 65. DOI: https://doi.org/10.1186/s12944-019-1009-4.

[30] Suzen M. Moeller et al. "Dietary Patterns: Challenges and Opportunities in Dietary Patterns Research: An Experimental Biology Workshop, April 1, 2006". In: *Journal of the American Dietetic Association* 107.7 (2007), pp. 1233–1239. ISSN: 0002-8223. DOI: https://doi.org/10.1016/j.jada.2007.03.014. URL: https://www.sciencedirect.com/science/article/pii/S0002822307004415.

[31] Marian L Neuhouser. "The importance of healthy dietary patterns in chronic disease prevention". In: *Nutrition Research* 70 (2019), pp. 3–6. ISSN: 0271-5317. DOI: https://doi.org/10.1016/j.nutres.2018.06.002. URL: https://www.sciencedirect.com/science/article/pii/S0271531718302227.

[32] Cheng-X. Zhou W. He J.- Xiao S. Ouyang F. "Increased Mortality Trends in Patients With Chronic Non-communicable Diseases and Comorbid Hypertension in the United States, 2000-2019". In: *Frontiers in public health* 10, 753861 (2022). DOI: https://doi.org/10.3389/fpubh.2022.753861. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9309719/.

[33] Ruth E Patterson, Pamela S Haines, and Barry M Popkin. "Diet quality index: Capturing a multidimensional behavior". In: *Journal of the American Dietetic Association* 94.1 (1994), pp. 57–64. ISSN: 0002-8223. DOI: https://doi.org/10.1016/0002-8223(94)92042-7. URL: https://www.sciencedirect.com/science/article/pii/0002822394920427.

[34] *Plasma Fasting Glucose (P_GLU)*. URL: https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/P_GLU.htm (visited on 03/29/2024).

[35] Farberman R. "Trust for America's Health". In: (2023). URL: https://www.tfah.org/report-details/state-of-obesity-2023/#:~:text=Nationally%2C%2041.9%20percent%20of%20adults,percent%20and%2045.6%20percent%20respectively.

[36] Vanessa Costa Joana Correia-Anabela Bandeira Esmeralda Martins Helena Mansilha Mónica Tavares Margarida P. Coelho Sara Mosca Graça Araújo. "Dyslipidemia Diagnosis and Treatment: Risk Stratification in Children and Adolescents". In: *Journal of Nutrition and Metabolism* 2022.4782344 (2022). DOI: 10.1155/2022/4782344. URL: https://www.hindawi.com/journals/jnme/2022/4782344/.

[37] Leite S. Alkerwi A. Sisanni L.-Zannad F. Saverio S. Donneau A. F. Albert A. Guillaume M. Sauvageot N. "Association of Empirically Derived Dietary Patterns with Cardiovascular Risk Factors: A Comparison of PCA and RRR Methods." In: *PloS one* 11.8 (2016), e0161298. DOI: 10.1371/journal.pone.0161298.

[38] Angelos K. Sikalidis. "From Food for Survival to Food for Personalized Optimal Health: A Historical Perspective of How Food and Nutrition Gave Rise to Nutrigenomics". In: *Journal of the American College of Nutrition* 38.1 (2019). PMID: 30280996, pp. 84–95. DOI: 10.1080/07315724.2018.1481797. eprint: https://doi.org/10.1080/07315724.2018.1481797. URL: https://doi.org/10.1080/07315724.2018.1481797.

[39] Kristina P. Sinaga and Miin-Shen Yang. "Unsupervised K-Means Clustering Algorithm". In: *IEEE Access* 8 (2020), pp. 80716–80727. DOI: 10.1109/ACCESS.2020.2988796.

[40] *Violin Plots for 22 Food Groups*. URL: https://github.com/hoolooboooroodkoo/NHANES-Dietary-Patterns/blob/main/plots/violin_plots_22.png (visited on 04/27/2024).

[41] *Violin Plots for 46 Food Groups*. URL: https://github.com/hoolooboooroodkoo/NHANES-Dietary-Patterns/blob/main/plots/violin_plots_46.png (visited on 04/27/2024).

[42] Song M. Eliassen A.H. et al. Wang P. "Optimal dietary patterns for prevention of chronic disease". In: *Nat Med* 29 (2023), 719–728. DOI: 10.1038/s41591-023-02235-5. URL: https://www.nature.com/articles/s41591-023-02235-5.

[43] *What We Eat in America Food Categories 2017 - March 2020 Prepandemic*. URL: https://www.ars.usda.gov/ARSUserFiles/80400530/pdf/1720/Food_Categories_2017-March%202020%20Prepandemic.pdf (visited on 03/29/2024).

[44] Bo Zhang et al. "Fish Consumption and Coronary Heart Disease: A Meta-Analysis". In: *Nutrients* 12.8 (2020). ISSN: 2072-6643. DOI: 10.3390/nu12082278. URL: https://www.mdpi.com/2072-6643/12/8/2278.

[45] Junkang Zhao et al. "Exploring the association of dietary patterns with the risk of hypertension using principal balances analysis and principal component analysis". In: *Public Health Nutrition* 26.1 (2023), 160–170. DOI: 10.1017/S136898002200091X.

[46] Gao Q. Zhao H.-Chen Sh. Huang L. Wang W. Wang T. Zhao J. Li Zh. "A review of statistical methods for dietary pattern analysis". In: *Nutrition Journal* 20.37 (2021). ISSN: 1475-2891. DOI: https://doi.org/10.1186/s12937-021-00692-7.

[47] Ye Y. Zou N. Yu-J. Zou D. "Analysis of risk factors and their interactions in type 2 diabetes mellitus: A cross-sectional survey in Guilin, China". In: *Journal of diabetes investigation* 8.2 (2017), pp. 188–194. DOI: doi.org/10.1111/jdi.12549. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5334303/.

[48] Amra Ćatović. "Dietary Patterns". In: *Recent Updates in Eating Disorders*. Ed. by Ignacio Jáuregui-Lobera and José Vicente Martínez-Quiñones. Rijeka: IntechOpen, 2022. Chap. 5. DOI: 10.5772/intechopen.108367. URL: https://doi.org/10.5772/intechopen.108367.