TAMPEREEN TEKNILLINEN YLIOPISTO
TAMPERE UNIVERSITY OF TECHNOLOGY

SOUMYA DAS

LIFESTYLE ANALYSIS USING DIGITAL FOOTPRINTS

Master of Science thesis

Examiner:
Senior Research Fellow. Hannu Niemi-
nen,
Professor. Ari Visa
Examiner and topic approved by the
Faculty Council of the Faculty of
February
on 14th February 2018

ABSTRACT

SOUMYA DAS: Lifestyle Analysis using Digital Footprints
Tampere University of Technology
Master of Science thesis, 52 pages, 0 Appendix pages
September 2018
Master's Degree Programme in Information Technology
Major: Data Engineering
Examiner: Senior Research Fellow. Hannu Nieminen,Professor. Ari Visa
Keywords: Digital footprints, Digital Health Revolution(DHR), Food Frequency Questionnaire(FFQ), Food Categories, Physical Activity Questionnaire(PAQ)

Digital Revolution has changed how the world used to function. Using digital devices and sensors every activity we do in our daily life can be tracked and can be stored as data. This thesis work aims to search how digital footprint gathering infrastructure can be used to collect digital footprint data specially related to health care, and a case study was done using data from Digital Health Revolution(DHR) study to analyze lifestyle of the participants in the study.

Digital footprints of shopping and physical activity of persons participating in DHR study were collected. Also their daily diet and physical activity was recorded using Food Frequency Questionnaire(FFQ) and Physical Activity Questionnaire(PAQ). Comparative analysis was done between shopping and FFQ data. Physical activity data gathered using Withings activity app were compared with PAQ data. Boxplots and and median plots were used to show differences between the values. For many of the food categories shopping done by the participants corresponded with their intake, but for some food categories such as coffee, spreads, pastry and desserts there were high differences in shopping and consumption. No correlation was found between physical activity data gathered from sensors and PAQ(Physical Activity Questionnaire). From the results it can be concluded that the data collected has several limitations. The sample size was small, and if more data could have been gathered using various digital footprint gathering infrastructure, we could have more coherent results. Further using background study information, an ideal digital footprint gathering system, is proposed which will be able to collect all information required for more accurate life style analysis. Using the system results could most probably be improved.

PREFACE

This thesis work on "Lifestyle analysis using digital footprints" helps me complete my master's studies in Information Technology from Tampere University of Technology. This work was done at the end of my studies.

I would like to thank Dr.Hannu Nieminen, for his excellent guidance, and helping me select and understand the topic, and for sending quick feedbacks while I was writing the thesis. I would also like to thank Antti Kallonen for helping me with various coding problems and for helping me with the Digital Health Revolution data. I would also like to thank Dr.Riitta Sallinen for helping me to categorize the shopping data into food groups, with her expert knowledge in nutrition.

Finally I want to thank my parents and my friends, who have constantly encouraged me and were supportive during my studies at Tampere University of Technology.

Tampere, 19.9.2018

Soumya Das

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND SYMBOLS

SCI	Social and Community Intelligence
RFID	Radio Frequency Identification
GPS	Global Positioning System
MIT	Massachusetts Institute of Technology
HITECH	Health Care Information Technology for Economic and Clinical Health
KP	Kaiser Permanente
FFQ	Food Frequency Questionnaire
USDA	United States Department of Agriculture
NNR	Nordic Nutrition Recommendation
GNKQ	General Nutrition Knowledge Questionnaire
FCQ	Food Choice Questionnaire
FPQ	Food Preference Questionnaire
HEI	Healthy Eating Index
HFII	Healthy Food IntakeIndex
MMM	My Meal Mate
CDC	Center for Disease Control and Prevention
ASM	American College for Sports Medicin
HHS	Health and Human Services
PA	Physical Activity
SB	Sedentary Behavior
FIMM	Finnish Institute of Molecular Medicine
HiLIFE	Helsinki Institute of Life Science
DHR	Digital Health Revolution
PAQ	Physical Activity Questionnaire
app	Application

r	correlation coefficient
$\overline{x}, \overline{y}$	median
n	sample size
$\sum$	summation
x_i, y_i	observation
Q1	Lower Quartile
Q3	Upper Quartile

1. INTRODUCTION

Healthiness is next to godliness, hence leading a healthy lifestyle is very important for every person. Since public health crisis like obesity and diabetes are growing more among the urban population due to their often sedentary lifestyle, it is important for individuals to focus more on healthy eating and physical activity[12]. Health can be defined as overall mental and physical state of an individual when he/she is free from disease, whereas wellness is the optimal physical and mental health of a person. [13]. The wellness wheel(see Figure 1.1), describes the dimensions of wellness. The dimensions are interconnected and are very important for an individual in order to have a balanced lifestyle.

Figure 1.1 *Wellness Wheel*

Emotional wellness deals with ways to minimize emotional stresses. Spiritual wellness teaches to balance needs of an individual. Social wellness helps developing relationships, communication skills, and teaches to respect each other. This study deals mostly with physical wellness. Physical wellness is connected to an awareness about healthy eating, physical exercise, sleep, and recovery [13] [14].

According to World Health Organization(WHO) records in 2014, 39 percent of adults above 18 years are over weight. Technical solutions are emerging to inspire people to participate in more physical activity. Research based as well as commercial sectors are also developing wellness based applications since it is increasingly becoming popular.Mobile applications enabled to track user activity to maintain wellness diaries are becoming very popular [15]. Using these kinds of applications user activity can be collected as digital footprints. Digital footprint data can be used for personalized health care solutions. Also shopping data of individuals can be tracked from large retail data, generated by individuals while shopping. With the advancement of data storage technologies, different formats of data can be stored. For example DiabeticLink is a diabetes big data platform created by University of Arizona and University of Taiwan aimed for personalized health care and diabetic treatment. The platform helps individuals by allowing exchange of disease information, experiences and also integrates information related to drug side-effects, electronic health records, etc.

The purpose of this thesis was to conduct a comparative analysis of user shopping or user activity with respect to their reporting in Food Frequency Questionnaire(FFQ) [16] and Physical Activity Questionnaire(PAQ). In Chapter 2 background of this thesis, digital footprints and various system architectures to collect digital footprints are discussed. I have also discussed different food pyramids, food plate and nutrition which is aimed to understand the food categorization. Food products bought by the participants while shopping were categorized into different food categories. In Chapter 3, there is discussion about the data.The shopping data collected from S-group shops and the physical activity data collected from Withings activity app are digital footprints created by the participants who participated in DHR project. They also reported about their daily intake and daily physical activity in the questionnaire data collected from FFQ [16] and PAQ. In Chapter 4, Methods - the data preparation steps are discussed in details. How the shopping data and questionnaire data was processed and mapped to food categories. I have also described the processing steps of activity data. In Chapter 5, Results shopping and food consumption reporting were compared statistically, using box plot and median plot visualization. In the Chapter 6, Discussion: I have discussed limitations of the project and solutions to over come those limitation were also discussed and new ways to collect more data was suggested.

2. BACKGROUND

2.1 Digital Footprint

Digital footprints are collections of personal digital records created by the trail of activities we do in everyday life [17]. The technological advancement in the field of sensing, computing, storage and communication has enabled us to gather data about our daily activities. Nowadays most of the mobile devices are equipped with sensors.Due to large scale deployment of sensors and sensor networks in private buildings, outdoor environments and public facilities, technically collecting digital footprints is not a big challenge, but collecting them legally and ethically can be very challenging. Mobile devices carried by people have built in sensors capable of generating a huge amount of digital footprints . Automobiles are also equipped with GPS and tracking devices, which can be a potential source for data collection. Apart from this, person's social media accesses, web page accesses, emails and instant messages provide information about preferences and give insight about their environmental context. Ability to capture abundant community data has given birth to a new research field called **social and community intelligence SCI**.[1] [2] The aim of this research area is to reveal the pattern of an individual's or group's societal behavior. Heterogeneous data collection from multiple data sources has formed the digital footprints. Compiling these digital footprints creates a picture of an individual's daily life activity and enables researchers to create computational models of human behavior and to enable innovative services into areas such as human health, public safety, city resource management and transport management [1][18].
In order to understand social and community intelligence system functions, we need to know the architecture of the system.

2.1.1 General Architecture of SCI System

Social and Community Intelligence platforms increase the capacity to collect and analyze digital footprints. Since SCI system collects real-life and real-time data, data

sensing and inference is one of the most important features of the system. SCI system must have an infrastructure which has heterogeneous devices and software integrated and provides capabilities for rapid development,deployment and evaluation. The data collected by SCI system come from heterogeneous sources. The main data sources are mobile and wearable sensors data related to individual, movement, and infrastructure sensor data gives insight about the environment they live in.Data from social media and other social networks provide input on individuals preferences and relationships. Technologies used in SCI system include data-mining, machine learning and artificial intelligence to understand the patterns of the data collected by SCI system.

The general architecture of a social and community intelligence system(SCI) consists of five layers. The five layers are as follows:

- Pervasive sensing layer: This layer contains the main sources to collect information from the environment. The main tools to collect information are mobile devices and user-centric wearable sensors capable of sensing activities, interactions and recording location information. For recording indoor activities static infrastructure, for instance surveillance cameras, environmental sensors, indoor positioning sensors and radio frequency identification(RFID) can be utilized for monitoring and detecting human movements in the real world.

- Data anonymization layer: Since data privacy is a very important issue for every organization both private and research-based, the framework is designed so that the data released are anonymized and privacy protection algorithms are applied on the data. If users personal digital data gets shared or revealed it will lead to privacy risk of users or persons providing their information.

- Hybrid learning layer: In hybrid learning layer machine learning and data-mining techniques are applied to convert low level single-modality sensing to high-level features. The goal of the hybrid layer is to mine patterns of individual's behavior and to extract social community intelligence.

- Semantic inference layer: This layer uses logic based approach to aggregate features. In this layer statistical learning is complemented and uses explicit rules to effectively associate the hybrid learning with social and community intelligence based on expert domain knowledge.

- Application layer: The application layer in SCI system includes a variety of potential services, which can be enabled by the available SCI. An application installed on mobile devices can directly run on remote servers(web application) and it communicates with mobile devices via wireless gateways.

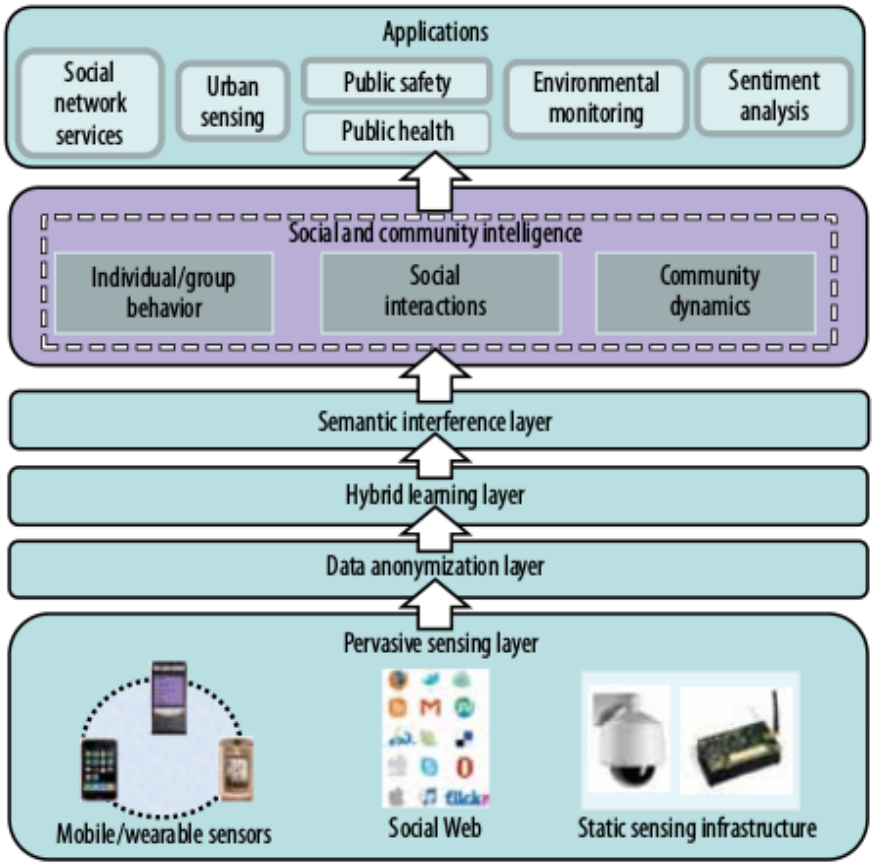


Figure 2.1 *Architecture of SCI system [1][2]*

Since mobile devices are carried by users most of the time,mobile/wearable sensors help to collect sensing data in real time and help in predicting their daily activity. Some of the most popular sensor devices are RFID(Radio frequency identification), GPS(global positioning system) and accelerometer. Sensor based activity recognition devices can be divided into two categories - devices which can be worn by humans on their body and devices which are placed on objects.Wearable sensors which are attached to the human body generate signals when humans perform activities like walking, running and exercising. Object based activity recognition is

based on real world observations- for instance complex physical motions like grooming, cooking, phoning, toileting and washing hands. These types of activities can be recorded by object based activity recognition system.

The main idea behind collecting sensor based activity data is to build mathematical models on activities and feeding the real time sensors the model for predicting human activities. RFID readers used in wristbands are used to track objects touched by the user in daily life. Recordings of the objects touched are used as data for machine learning methods, which can be used to learn a model for recognizing daily activities ranging from brushing teeth to complex ones like medications and cooking meal.GPS based sensors are used to monitor location based activity.Some of the applications using GPS based sensors are trip plan derivation based on GPS traces, significant location identification, route prediction etc.Sensor based applications have been used in elderly care, health care, habitat monitoring and tracking human interactions.

Data collected by the pervasive systems need to be secured, for which data privacy features or data anonymization are very necessary for the SCI systems.For instance, revealing a person's position records may lead to exposing someone's interest. Data anonymization and user control answer privacy protection questions such as: who is asking for data? how much the data reveal the user's identity and how long will the data be retained? The aim of data anonymization is to prevent from revealing user data. K-anonymity algorithm used by Metrosense[2] is an example of data anonymization method used.The method generalizes a users position, which contains k-users hence anonymizing their identity.

Other than the data anonymization, there are more to privacy issues, for instance enhancing user control and decision making. User control to share data empowers people to decide themselves what personal information would they like to share and with whom they would like to share. For instance if someone tracks his heart rate every day, it is unnecessary to share the information with everyone but with the doctor. Researchers are trying to find methods so that users could manage their data by access control and data management tools. Data trust issues are also handled in data anonymization layer - how much of the data source can be trusted. For example Twitter data is sometimes unreliable due to casual and unmediated text.

Social and community pattern mining helps to identify set of characteristics and behaviors in a community. By pooling the digital traces of individual users and

mining their pattern, various social and group behaviors can be extracted.The key of SCI pattern mining is to identify user similarity in services such as social network services, urban sensing and city resource management, human health, sentiment application and public safety.[1] [2]

2.1.2 Applications of SCI Systems

Earlier the social network analysis was mostly based on relational data obtained from surveys. But in last two decades the boom of internet applications like chat boxes, photos over social media, email and other instant messaging system has generated a huge amount of data, on which social network analysis and knowledge discovery is being done. Some examples like ArterMiner[19] harvest on personal profile information to summarize event info space, time and theme.Similarly popular social media twitter also does reporting by real time mining of natural disasters like earthquakes[20] and the mood of the citizens and social media users [21].

Wireless sensors are also used to leverage community sensing and to help with urban problems like monitoring traffic, planning and better public utilities. Data collected from multiple sources for example cellphones, buses, and taxis are used to improve urban lifestyle. Video surveillance has greatly improved public safety and helps protecting people from crimes and disasters.

SCI also helps in tracking disease outbreaks across populations. The epidemic of seasonal influenza is a major public health concern. With the help of SCI like system, it is possible to estimate how a disease is spreading in a region. SCI also promotes personal well-being. Community sensing enables people to log their physical activities and track the amount of their food intake, sense their mental status in real time, and to record their social interactions - all these records are important for people to improve their health management.[1]

Since the health and wellness sector is growing constantly, there is a need for public digital services in the field of health care. Health care services e.g hospitals, occupational health care are having online services, and the commercial wellness services and lifestyle related digital services are also expanding, making it easier to collect user data in the digital age. Data of individuals are no more collected through a centralized system but collected using the interaction of various devices e.g smart phones, internet and credit cards. These transactions create digital footprints of

our daily lives. Mobile phones have strong social function, and they are one of the prime sources to collect wellness data. Wellness applications such as wellness diary[22] are used to record wellness related information. Mobility enables users to record self-observations and have a graphical feedback from the data, making wellness monitoring pervasive. Häkkilä et al.[15] discusses about designing some health apps of future. Examples include, Shopping Receipt Guidance and Health Money [15]. Shopping Receipt Guidance has the aim to change shopping habits, suggesting healthier alternative products when unhealthy groceries are being shopped. The target is to affect the purchasing patterns next time when shopping is done. Similarly, Health Money is used to reduce the purchase of unhealthy foods. Foods are priced based on unhealthiness. If an individual is signed up for health money, he has to pay more for unhealthy food purchase.
Connected Dentist [15] is also a proposed health app which provides brushing data from smart toothbrush to dentist, and the dentist can assess brushing habits and guide the individual with better dental care.

Electronic health records are revolutionizing the health-care industry. The Health Care Information Technology for Economic and Clinical Health(HITECH) adopted plans to meaningfully use electronic health records. The aim is to computerize a large proportion of US health records. In the article [23] it is discussed how KP(Kaiser Permanente's) manages its electronic health records for 8.7 million members. In 2010 KP health connect was implemented, in which information related to physician's office, hospital, radiology laboratory and pharmacy are all seamlessly connected for 454 medical offices and 36 hospitals across nine states. The KP health connect picture replaces conventional x-ray films and is capable of archiving pictorial health records. This system reduces water use for machines that process x-ray machines for imaging procedures. Since the hospital x-ray machines run 24 hours a day and seven days a week, for the maintenance of x-ray machines, it requires almost 71 million gallons of water annually. Using KP health connect and picture archiving system saves a lot of maintenance cost and resources.
Since x-ray films are composed of 57 percent of plastic, digitizing and archiving x-ray images avoid 68 tons of plastic waste each year, thereby digitizing records also help in creating an environment-friendly system. Usage of electronic health records has reduced green house gases, decreased paper and plastic consumption.

2.2 Healthy Eating

The project aims to measure healthy life style and wellbeing of people, by collecting and analyzing the digital footprints of people from their daily shopping and physical activity. Shopping information and the Food Frequency Questionnaire(FFQs) [16] answered by the participants, gives an idea of the participant's eating habits. It is possible to measure the healthy eating from the food participants have regularly. In order to understand healthy eating, it is important to understand nutrition, different food pyramids that make a healthy diet.

2.2.1 Nutrition

Nutrition is the study of how food works in our bodies and what are the different sources of energy. In nutritional studies different nutrient functions are discussed, and nutrition is regarded as building blocks of life. Some of the most essential nutrients of life are carbohydrates, proteins, fats(lipids), fiber, vitamins, minerals and water. The nutrient absorption starts in our body as soon as digestion starts.Good nutrition means getting the right amount of nutrients from healthy foods in the right combinations. [24]

Nutrition is important since it is one of the keys to develop and maintain good health. Good health is defined as a state of complete physical, mental and social wellbeing. Since to maintain good health and wellbeing a person should eat a balanced diet. Food provides our bodies with energy. Taking a wide variety of foods provides the right amount of nutrients for good health. Enjoying a healthy diet and having a good lifestyle can be one of the great cultural pleasures of life. The foods and dietary patterns that promote good nutrition are discussed in the next section Food guides and Recommendation, where we discuss different food guides and food pyramids and different food consumption patterns recommended to have a healthy diet.[25]

2.2.2 Food guides and Recommendations

Food Guides : A food guide is a reference for understanding the amount of food we should consume from various food groups for a healthy diet. [26] Food guides are usually graphically represented along with messages of dietary guidelines. In the

food guide recommended food groups are represented by their suggested proportion for a good diet. Information and guidelines related lifestyle is typically included, for example how much physical activity is recommended, or warnings related to alcohol consumption. Mostly the food guides are in the shape of food pyramids or food plates, but some food guides are very specific to country's indigenous population's food habits, and the nation's nutrition communication and education strategy.[27]

Below some well-known food pyramids and food plates are discussed :

- **USDA food guide :** MyPyramid released by USDA served as food guide from the year 2005 to 2011.[3] MyPyramid food guide was designed to educate consumers about their lifestyle in consistence with January 2005 Dietary Guidelines. This guideline was produced jointly by the Department of Health and Human Services. The symbol of MyPyramid had a picture of a person climbing stairs which represented physical activity and the measuring quantities are in cups and ounces instead of servings. MyPyramid consisted of six food groups - Grains, Vegetables, Fruits, Oils, Milk, Meat and beans. It is recommended in MyPyramid that at least half of the grains consumed must be whole grains. More emphasis is given for the intake of dark green, orange, dry beans and peas. It is also recommended to take a variety of whole fruits over fruit juices. Oils from various sources like fish, nuts and vegetables are recommended for intake. Milk includes fluid milk and other milk based products. More emphasis for the intake of low fat, lean meat, fish, beans, peas, nuts and seeds is given in the category of meat and beans. MyPyramid also contains the recommendation of physical activity. There are recommendations for children, adolescents, and pregnant women. It is important to engage in regular physical activity and to reduce sedentary lifestyle and, to enhance psychological wellbeing and to strive towards healthy body weight. It is recommended to do 30 minutes of physical activity for adults and 60 minutes of physical activity for teenagers and children everyday. More the duration of physical activity more benefits. Pictures below depict MyPyramid food group recommendations and also six food groups of MyPyramid in detail.[3]

  Myplate is the current nutrition guide for USDA. It has four sections which consist of 30 percent of grains, 40 percent of vegetables, 10 percent of fruits and 20 percent of protein accompanied by dairy i.e a glass of milk or yogurt.[4]

Figure 2.2 *Six food groups of MyPyramid along with MyPyramid symbol [3]*

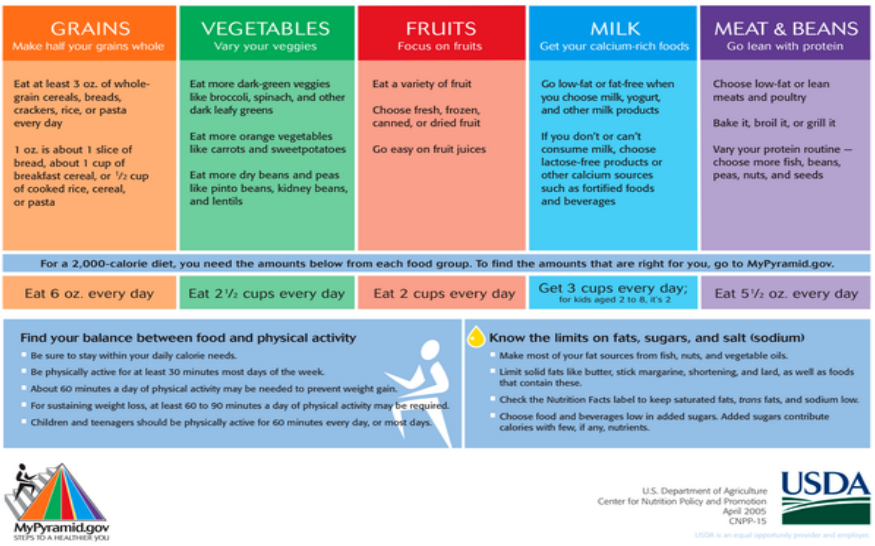


Figure 2.3 *MyPyramid miniposter with sample food group recommendations[3]*

Figure 2.4 *MyPlate Model [4]*

- **Nutrition Recommendations in Finland:** In Nordic countries, nutrition recommendations have been made since the early 1980s. In Finland, the National Nutritional Council [5] is responsible for Nutrition Recommendations. The State Nutrition Advisory Board published its first recommendation in 1987 and their revised versions in 1998, 2005 and 2013. The recommendations are mostly for the Finns and the board follows mostly Nordic Nutrition Recommendation (NNR) [28]. Components promoting healthy diet include for instance vegetables, berries, fruits, leguminous plants and whole grain cereals as well as fish, vegetable oils and vegetable oil based spreads, nuts, seeds and fat-free and low-fat milk products are recommended for intake. The food triangle or food plate suggests an intake of half kilograms of vegetables each day, and reduced consumption of red meats(e.g beef, pig meat and sheep meat) and especially that of processed meats and food containing a lot of saturated fats. In the food triangle components of a good diet are presented according to their relative weight in the whole diet. Finns usually have a recommended diet that is varied, versatile, reasonable and tasty. Variety and versatility in diet guarantee the supply of all kinds of nutrients, which is aimed to combat obesity.

  The other way of recommending a diet is the plate model, where half of the dish must comprise of fresh and cooked vegetables, a quarter for potatoes, pasta or rice and a quarter for meat or fish. The below picture of Finnish plate model shows a detailed listing of the items. Since half of the Finns eat at least one meal outside their home, it is important for the mass caterers to implement healthy eating. Nutrition recommendation affects the wellbeing of most of the citizens, healthy eating helps people to adopt a rational lifestyle. [5]

Food Groups	Foods/Food Products
Vegetables,Fruit and berries	Vegetables,potatoes,berries and fruits
Whole Grain Products	Cereal products,rye bread, whole grain porridges
Fish and Fish Products	Salmon, Baltic herring and pike
Meat,Meat Products and Egg	Red meats(beef,lamb and pork), Meat Products (sausages,bacon and different cold cuts), skin less poultry meats-chicken and turkey and eggs
Milk and Dairy Products	fat-free or low-fat liquid dairy products (yoghurt and other cultured milk products) ,flavoured yoghurt

Table 2.1 *Food groups and Foods in Finnish Plate [5]*

Figure 2.5 *Finnish Food Pyramid [5]*

Figure 2.6 *Finnish Plate Model [5]*

2.2.3 Ways to measure Healthy eating

Dietary patterns consist usually of certain food groups, food items, and beverages consumed with specified habitual frequencies. Dietary patterns are typically identified in large scale population studies, in which diets are often described at a very general level, as the dietary assessment method most often used is the food frequency questionnaire (FFQ). Food Frequency Questionaire is a method commonly used in epidemiological studies to assess an individual's dietary intake of foods and nutrients consumption [29]. FFQ consists of a long list of food items to capture a variety of food choices with defined portion and frequencies (per day, per month, and so on) that the person interviewed has consumed [30].
The questions asked in FFQ are self-administered, in the interview questions are asked about the usual frequency of consumption of items from list of food items. Although the absolute intake is an estimate, it is still possible using FFQ questions to quantify the amount of healthy food intake by an individual. In FFQ questions, some of the general questions asked aim to assess how many times a day a person is having a food or beverage. Often also there are selectable options for the range for the frequency of intake. Questions about portion size are also asked , and the participants are also asked to respond in terms of amounts or any specified units. Sometimes questionnaires include portion size images to help respondent to answer

more accurately. FFQ also includes frequency of consumption of some foods at a particular time of the year. FFQ lists are sometimes very specific according to the locality,i.e the specific food and beverage items are valid for a certain population, but not valid for other populations [29] [16] .
Using FFQ questionnaire an individual can be ranked based on their levels of nutrient intake. Since the aim of most nutritional studies is to assess the risk of common chronic diseases, FFQ gives an overview of the individual's nutritional intake. FFQ serves as a long-term nutritional assessment tool. Other than FFQ there are other eating behavior questionnaires for instance General Nutrition Knowledge Questionnaire(GNKQ), Food Choice Questionnaire(FCQ), Food Preference Questionaire(FPQ)[31] .

The diet pattern can also be quantified using scoring algorithms such as HEI (Healthy Eating Index) [30] [6], used mostly to measure US diet pattern.Healthy Eating Index (HEI) is used to measure the quality of diet and also to monitor whether the dietary intake are conforming to Dietary Guidelines for Americans or not.Its main purpose is to monitor diet quality of US population.It further checks the relationship between diet and health-related outcomes. The original version of HEI (i.e HEI 2005) consisted of 10 components with an optimal score of 10 without any energy adjustment whereas HEI 2010 which is a revised version of HEI consists of 12 components (see Figure 2.7). The score is calculated by comparing with a target amount of food or nutrient to be consumed each day relative to energy intake rather than absolute amounts. In the scoring pattern, 100 is the maximum number which corresponds to an optimal dietary pattern. Fruits (with emphasis on whole fruits), vegetables with specified targets for dark green and orange vegetables, whole grains, dairy, and then seafood and plant proteins are the recommended food groups. When an individual's intake values are closer to the recommended intake amount, his scores are higher. Whereas there are limits for diets like solid grains, empty calories (i.e solid fats, alcohol, and added sugars), having items like solid grains and empty calories reduces the score. Harvard Healthy Eating plate is also a similar measurement platform for healthy eating. Similar to HEI, there is also Healthy Food Intake Index(HFII)[32]. This diet quality index was developed to measure healthy eating based on Nordic Nutrition Recommendation food guidelines. HFII comprises of 11 components - vegetables,fruits and berries, high-fiber grains, fish, low fat milk, low fat cheese, cooking fat, fat spread, snacks, sugar sweetened beverages and fast food [32][6].

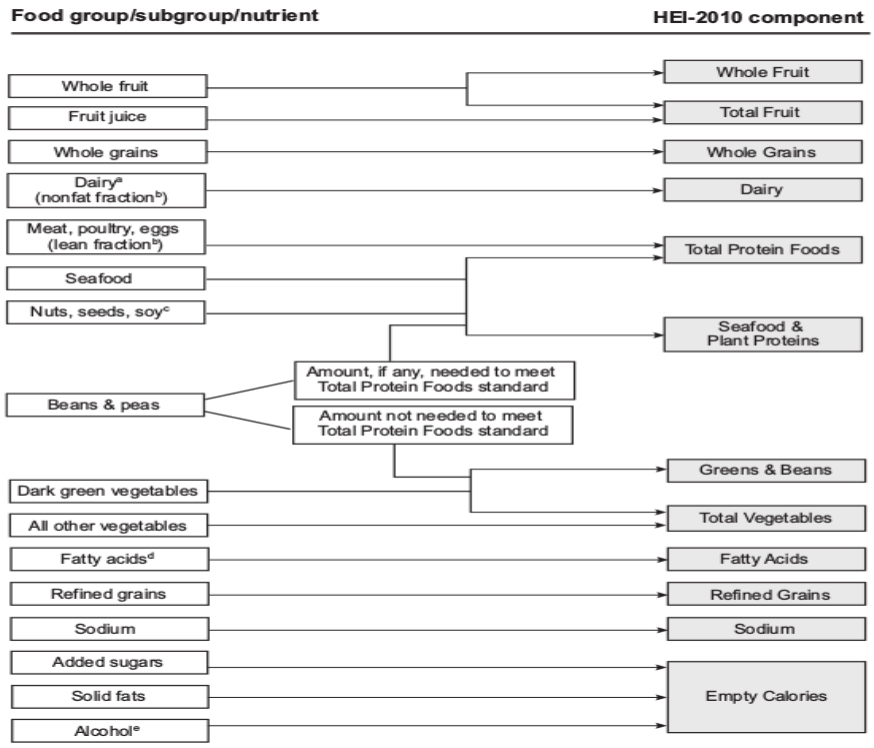


Figure 2.7 *Food groups and subgroups of the USDA Food Patterns, alcohol, and nutrients that contribute to the components of the Healthy Eating Index-2010 [6]*

Mobile Applications to Measure Healthy Eating

Healthy eating can also be monitored using apps. There are many apps related to nutrition, diet and weight control available in iPhone, Android, Nokia, and Blackberry. For example Pebble Diabetes Diary app [7] available in app stores(such google play and apple app store -itunes) can be used in watches and other wearable devices. Some of the features present in the app are shown in below figure(Figure 2.7).

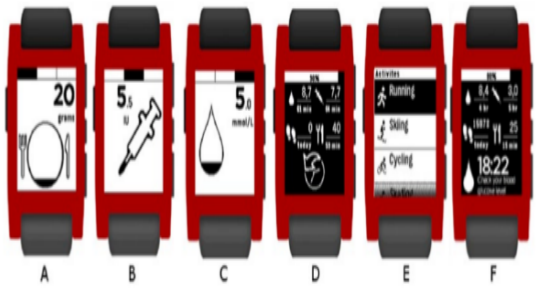
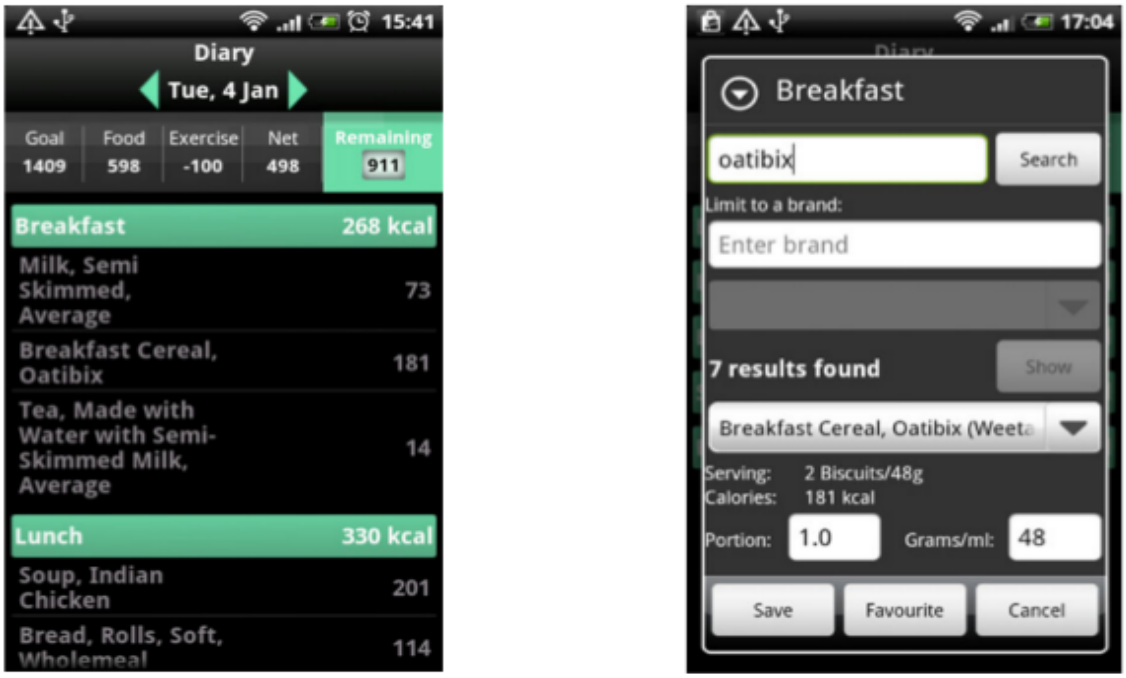


Figure 2.8 *(A-F) Entry of carbohydrate intake, insulin units, and blood glucose measurements are done using the middle button. Previous registrations are displayed using the right lower button. Specific physical activities are recorded using the upper-right button. Automatic blood glucose measurement reminder is set to 90 minutes after meals as a default.[7]*

In figure 2.8, display screen A : shows the measure of carbohydrate consumption, B : shows the amount of insulin taken, C : shows the measure of blood glucose, D : shows the previously recorded values of all the measures, E : displays different physical activity measures and F : shows the measure of blood glucose level. Some features that are expected from wearable food/diabetes diary apps are for example : the app must give blood glucose level reminders and also the time when meal should be taken, access to previously collected data, and ability to record new entries directly [7].

My Meal Mate(MMM) [8] is a mobile app which is used as a self monitoring tool for food and drink intake. MMM features consist of an electronic food diary in which users can select and log foods from 4000 items commercial food database. MMM has the usability feature to enhance the process by storing favorite meal combinations, taking photographs of the food, and memory prompt feature to store recently used items. Food items recorded by the application are further uploaded to a secure administrator website for macro nutrient intake analysis. The picture below shows the data entry page of My Meal Mate app.

Figure 2.9(a) shows the foods and the amount of calories consumed by an individual during breakfast and lunch and also his amount of calorie goals and how much more he needs to consume to meet his goal for the day. The app also keeps track of exercise and calculates the net amount of energy remaining based on the calorie burnt after exercise. Though MMM database contains default portion sizes for each item, users are still encouraged to enter correct portion sizes manually. Figure 2.9(b) shows the



(a) Data Entry Page of My Meal Mate

(b) Search Page for finding a food

Figure 2.9 *My Meal Mate [8]*

page of the app to manually enter the items with accurate proportion.

2.3 Physical Activity

Physical activity and exercise is an important contributor to a healthy lifestyle. Having a sedentary lifestyle may lead to obesity, weakness and lack of endurance, and it as a consequence may lead to poor health, which fosters disease development. Regular exercise has many benefits: for example it prevents developing coronary heart diseases, stroke, diabetes obesity and high blood pressure. It also helps in building bone strength and improves balance and flexibility, thereby preventing diseases related to bones, for instance osteoporosis, arthritis and reverse age-related problems by developing muscle mass and muscle strength.

Healthy living means that both physical and mental health are well balanced and functioning well together. In many instances physical and mental health are closely linked, hence good or bad one directly affects the other. Mental health is based on whether a person is having enough sleep or not, mind exercises like solving a puzzle, or doing something that interests(hobby) during leisure and having fun like

shopping, going for activities like fishing and networking with friends.

2.3.1 Activity Recommendations

Physical inactivity is one of the major reasons behind growing public health problems, and hence it is one of the contributing factors behind several chronic diseases. Recognizing these health hazards and sedentary lifestyle patterns some public healthcare societies for instance Center for Disease Control and Prevention(CDC), American College for Sports Medicine(ASM)[33], and UKK Institute in Finland[9] have come forward with physical activity recommendations for the public.[33]

UKK Institute founded in 1980 is responsible for promoting physical activity in Finland. The institute provides training and support to people for having a healthy lifestyle. UKK Institute's physical activity recommendation is provided using UKK Institute's Physical activity pie chart(see Figure:2.5) for adults aged between 18 to 64. The physical activity includes moderate-intensity aerobic physical activity of two and half hour of activity per week, vigorous intensity activity of one hour fifteen minutes per week, and muscle strengthening and balance training twice a week. The recommendation suggests dividing physical activity to at least three days per week.For beginners moderate physical activity- like walking, cycling, Nordic walk, heavy house or yard walk for 2.5 hours per week is enough. People aiming to develop further their physical fitness should do heavier exercises for 1 hour 15 minutes or more during the week, which includes more strenuous physical activity for instance uphill climbing, climbing stairs, running, cross- country skiing and water running.[9]

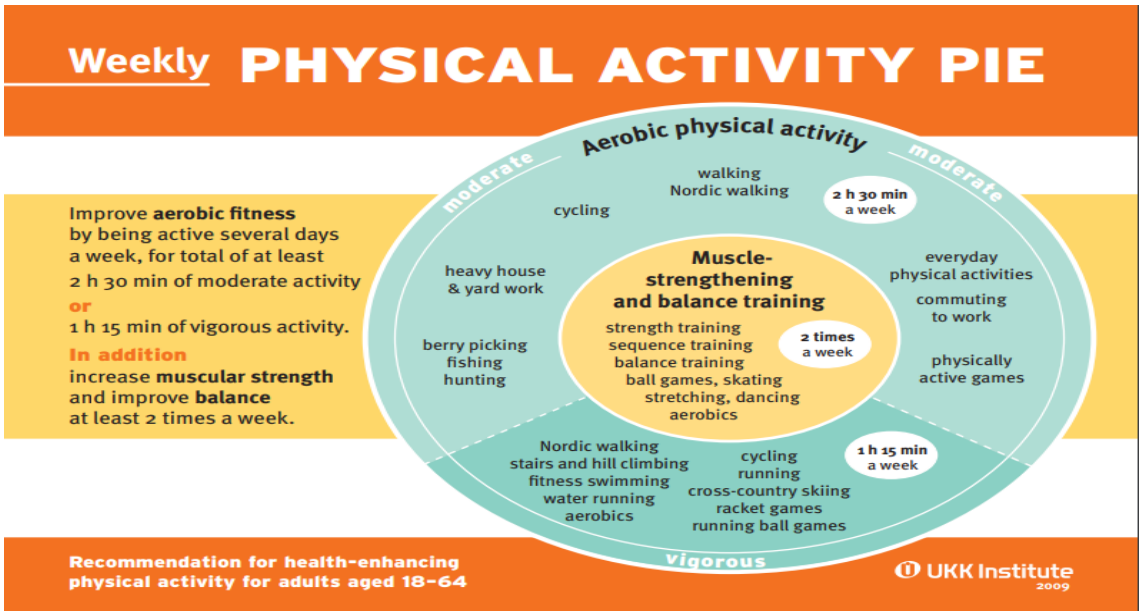


Figure 2.10 *UKK Physical Activity Pie [9]*

Center for Disease Control and prevention(CDC) , American College for Sports Medicine(ASM) recommends that every US adult should be engaged in 30 minutes of moderate-intensity physical exercise everyday. The 2008 Physical Activity Guidelines for Americans provides science-based guidance for Americans of age 6 and older to improve their health through appropriate physical activity. Physical activity guidelines are joint effort of Health and Human Services(HHS) [34] and US Department of Agriculture(USDA) [6] [34] to provide guidance for being physically active and to have a healthy diet. There are different guidelines prescribed for different age groups. Children and adolescents should engage in 60 minutes of physical activity daily which includes aerobic, muscle strengthening and bone strengthening exercises. Adults are advised not to be inactive and gradually increase their span of exercise from 150 minutes per week to 300 minutes every week, which includes vigorous-intensity aerobic physical activity. Minimum of 150 minutes of exercise per week could include e.g walking. Also exercise helping to maintain balance is beneficial especially for older adults[34].

2.3.2 Measuring Physical Activity

Developments in technologies like smartphones and smart watches have made it possible to collect physical activity measurements data. It was estimated that by

2015,500 million people out of 1.4 billion smart phone users will be using health related applications in their smart phones. Since it has been observed that customers spend around 80 minutes per day in smart phone applications, health apps in smart phones are believed to have great potential as an assistance to change health behavior. At end of 2010, there were 17,000 health applications available for download.One such application was 10,000 Steps [35] which engaged participants in physical activity programs and helped them to self monitor behavior.Physical activity included step counting pedometer to monitor physical activity levels.The iStepLog [35] application used in the study allowed the participants of 10,000 steps program to record their daily physical activity on their mobile device. In the experiment researchers were able to monitor how much time the participants spent using the application and how often they used the application. Figure 2.11 provides a list of well known wearable solutions for health and fitness. The applications are typically designed to monitor the total number of steps or minutes of activity. In the figure PA means - physical activity and SB - sedentary behavior. After the data is collected by the wearable devices there are procedures to measure these data i.e to quantify the amount of physical activity done, intensity of physical activity and sedentary behavior among the participants.[10]

Brand	Model	Where worn	Display/ compatibility	Measures	Possible measures[a]	Food/Weight tracking
Basis		Wrist	Display, personal computer, iOS[b], Android	PA[c], Steps, Heart rate, Skin temperature, Perspiration, Sleep		
BodyMedia	Fit	Upper arm	Personal computer, iOS, Android	PA, Steps, Sleep	Heart rate, Weight	Food, Weight, Balance
Fitbit	Force	Wrist	Display, personal computer, iOS, Android	PA, Steps, Sleep, Stairs, Distance, Calories	Weight	Food, Weight, Balance
Fitbug	Orb	Multiple	Personal computer, iOS, Android	Steps, Distance, Calories, Sleep	Weight	Food, Weight, Balance
Gruve		Waist	Personal computer	Calories, Activity zones		
Ibitz	Unity	Waist	iOS	Steps, Distance, Calories		Weight
Jawbone	Up24	Wrist	iOS, Android	PA, Steps, Sleep, SB[d]		Food, Weight, Balance
Lumo	Back	Waist	iOS, Android	Posture, Steps, Calories, Distance, SB, Sleep		
Misfit	Shine	Multiple	iOS, Android	Steps, Calories, Distance, PA, Sleep, "Points"		Food, Weight
Nike	Fuelband SE	Wrist	Display, personal computer, iOS, Android	PA, Steps, "Hours won", Calories, "Nikefuel"		
Polar	Loop	Wrist	Display, personal computer, iOS, Android	PA, Steps, Calories, SB, Sleep	Heart rate	
Striiv	Play	Waist	Display, personal computer, iOS, Android	PA, Steps, Stairs, Distance, Calories		Weight
Withings	Pulse	Multiple	Display, personal	PA, Steps, Sleep, Resting heart	Weight,	Weight

Figure 2.11 *List of Activity Monitor [10]*

Fitness Stats Tracking gadgets, such as Nike Feulband, Fitbit One, Bodymedia are equipped with embedded sensors such as accelerometer, gyroscope, and GPS. They track people's steps taken, stairs climbed, calorie burned, hours slept, distance travelled, and quality of sleep.[36] Human activity monitoring applications are also helping elderly care. For example smart phones which have the ability to detect fall, can also help older people from a dangerous situation. Life routine reminder reminds them when to take medicine etc. Activity recognition also has applications in youth care, including for example monitoring the sleeping status of infants and predicting their demands of foods and also used to detect autism spectrum disorder [36]

3. DATA AND METHODOLOGY

3.1 Digital Health Revolution

The Digital Health Revolution(DHR) [37] study is a part of the multidisciplinary DHR program. which has a unique approach of utilizing personal data in the future service structures. The objective of the program is to propose future health-care strategies, which will be evolving towards allowing individuals to make use of their personal data to control and improve overall health and wellness.

The DHR Study focuses on next-generation personalized health-care research by integrating longitudinal multi-omics profiling,digital monitoring, and other personal data. Finnish Institute of Molecular Medicine Finland(FIMM, HiLife, University of Helsinki, Finland) [38] conducted in the DHR study program a study between September,2015 and January,2017 where health related personal data was collected from 97 subjects of age group between 25-59. The subjects came from a clientele of private occupational health-care service. The participants in the study were not allowed to have any previously diagnosed serious diseases. However, individuals with elevated risk factors for chronic disease were included.

Over a 16 month period, the participants donated blood, urine, saliva and stool samples five times. Anthropometric and physiological measurement, health-related questionnaires, fitness tests, Quantified Self-tracking with sensors, and food purchasing data was systematically collected, and clinical chemistry, genomics, proteomics, metabolomics, and gut microbiome analysis were performed. Key actionable data was interpreted and returned to participants in real time via personal web account and web-based tools.

In this thesis, questionnaire data about healthy eating, shopping data and physical activity data from questionnaires and activity sensors was utilized.

In total there were 97 participants, 34 of them were single and 63 were family mem-

bers. 30 were male and 67 were female. The age of female participants ranged between 26 to 60 and male participants belonged to age group of 27 to 60. All the participant details were collected and stored in a json file. The attributes `subject_id` and `user_id` present in the json file were used as keys to match the google ids to other datasets for instance questionnaires and lab values. [38].

3.2 Data

3.2.1 Shopping Information

Shopping data used in this thesis contains information about the items study participants had bought from all S-group department stores over a 6-month period. Target of the analysis was to compare this information to the responses to a questionnaire about nutritional habits to see, how well the automatically collected digital footprint information could be used to describe the nutritional habits.

Shopping records were present in 83 files. The files consisted of information for example when the product was purchased i.e the date and time of the purchase,the place where the product was purchased and type of the product(tuoteryhmä) or the name of the product(tuotenimi) and also the cost of the product(ostot). All the shopping files were in comma separated (.csv) file format. The files were named based on the customer id encoded in the file name, which helped to track a particular customer's buying information.

Out of the 97 persons in the dataset, from 44 person's shopping information had been collected. Among them, 25 belonged to family, having family size of 2-7 people and the other 19 were single. Product purchase information was available for 35 participants. Participants having tuoteryhmä(type of product purchased) information were only selected for analysis. Among these 35 participants 16 were single and 19 were family member. 427 different types of products (tuoteryhmä) were purchased by the participants which consisted of different kinds of products for instance food products, clothing, cleaning products, cleaning chemicals and cleaning equipment, auto chemicals, deodorants, oral care products, flower care instruments etc. From all these categories food products were only selected and mapped to the food categories [38].

3.2.2 Category Selection

21 food categories [38] were defined for the study and S-group and FFQ questionaires were mapped into these 21 food categories(see Table 3.1). Food categories included vegetables,potatoes, grain products, meat and meat products, poultry, dairy products,pastry, desserts,etc. Initially, the food categorization was attempted following HEI- 2010 [6], but since the products were purchased from Finnish supermarket S-group shops, the categorization was aligned more according to Finnish food guide [5] and with the help of nutrition specialist Riitta Sallinen of Finnish Institute of Molecular Medicine, FIMM, HiLIFE, University of Helsinki.

Food Categories
Alcohol
Candy and chocolate
Coffee
Dairy products
Pastry and desserts
Eggs
Fish and seafood
Fruit and berry juices
Fruits and berries
Grain products
Meat and sausages
Nuts and almonds
Pasta and rice
Potatos
Poultry
Ready made foods
Soft and energy drinks, juices
Sports supplements
Spreads and oils
Sugar and added sugar
Vegetables

Table 3.1 *Food Categories*

Beers, wines and ciders were categorized as "Alcohols". Liquorice, sweets and other

chocolate products were categorized as "Candy and Chocolate". Yogurt, cheese, cream and ice cream were categorized as "Dairy products". "Fish and seafoods" consisted mainly of Fish cereals, smoked fish, canned fish and fresh fish. Juices and crushes as "Fruits berries and Juices". "Fruits and berries" contained mostly the whole fruits like apple, banana, pears, grapes, citrus fruits, other frozen fruits as well. Fresh bread, dry bread, flakes, wheat flour, cereal products, other flour ingredients(muut jauhot/ainekset), muesli, other bakery products such as leipomo / paistopiste, leipomotuotteet were classified under the category "Grain products". "Meat and sausages" were mostly meat products like kestomakkarat - kind of sausage, kokolihavalmisteet- full meat products, pork, smoked meat products, beef. "Nuts and almonds" included nuts and mantel. Category "Potatoes" included potato and frozen potatoes. In the category "Poultry", there were poultry meat only(siipikarjanliha). Prepared meals or meals which are ready to eat were mapped to "Ready made food". "Sports and energy drinks" were juices, mieto siideri, sima, soft drinks(virvoitusjuomat), and water(vedet)."Sports supplements" included - mostly sports nutrition such as energy bars. Category "Spreads and oils" consisted of vegetable oils and fats. Sugar was mapped to "Sugar and added sugar". There is a list of items under the category vegetables which are shopped including for instance - cultured mushrooms, frozen vegetables, vegetables, canned vegetables, root crops, legume vegetables, spices and salads.

3.2.3 Activity Data

The activity data were collected from Withings [39] activity sensors. Withings is a consumer electronics company known for designing connected devices. It is now owned by Nokia. Some of the well-known devices designed by Withings to monitor health and activity are blood pressure monitor, smart baby monitor, and Pulse O2 which contains pedometer and heart rate monitor. The device from which we collect our data is Activite - which is an activity tracking watch. The watch is compatible both with iphone and android. The watch has the ability to track sleep, swimming, walking and running. The dataset collected from Withings activity watch , consisted of, date, day of the week, calories burnt, distance traversed, type of activity done (intense or moderate), and step count. Steps are based on user's motion. Calories is a part of health mate app, another Withings product which provides detailed information regarding activity and metabolic calories, as well as total calorie expenditure.

3.2.4 Questionnaire Data

Questionnaire data is the feedback collected from the participants regarding their eating habits using FFQ [16] and also information about their daily exercise. FFQ data included answers to questions about the amount of food consumed by participants on a daily or weekly basis and monthly basis. Questions about different food groups were represented by different question codes. For example - `DHR07-93-` was the question code for FFQ. DHR07-93-2D was the question code for "Perunaa tai perunaruokia" - potatoes or potato related foods, DHR07-93-2O was the question code for "Broileria, kalkkunaa tai niistä valmistettuja ruokia" i.e chicken or food related to turkey. There were 51 types of answer alternatives for describing different types of food, which were mapped to 21 food categories [38] as listed in table 3.1 ,to compare between shopping data and reporting data.

FFQ question code DHR01-27 provided information related to family size. There were participants who were single and participants who were part of a family. Shopping done by a single person and their reporting varied much from people who were part of a family, since persons who were family members most probably often had done the shopping for the entire family as well. Family size of the participants varied from 2-7 people and average family size was 5.

Questionnaire related to activity was encoded as `DHR04-`. As an example, the questionnaires included questions related to whether the work done by the person was physically challenging or not ? How much gym exercise the person does and how much exercise has been done by the person in average per week during the last 6 months.

Amount of exercise done by the participants were calculated using Table 3.2. Level of doing exercise was categorized as strenous, mild and participants who were doing very light or no exercise were in the category of sedentary exercise or sedentary life style. The strength of exercise was numerically encoded between the range 0-6. Strenous exercise or very high of physical activity was encoded as `4-6`. Mild physical activity which included jogging, nordic walks or walking with high pace ranged between `1-4`. Sedentary lifestyle was `0-2`.

Strength of Exercise	**Encoding**
Sedentary Exercise	0-2
Mild Exercise	1-4
Strenuous Exercise	4-6

Table 3.2 *Exercise Range*

3.3 Methods

3.3.1 Processing of the shopping files

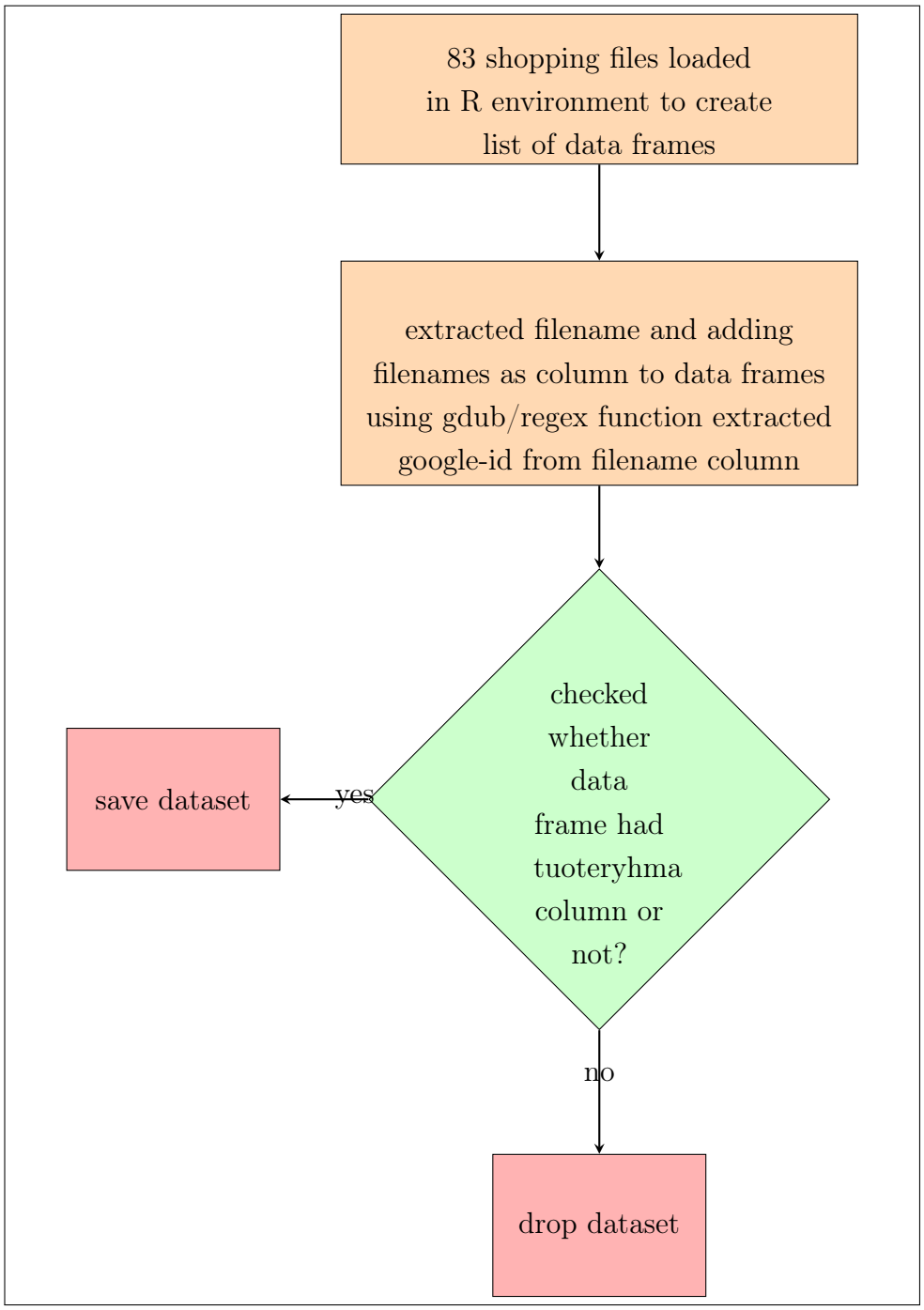


Figure 3.1 Data Extraction

The data processing and analytics were done using statistical tool R in unix environment. For processing the files and text processing various R functions like `regex, gsub`, and package like `dplyr` were used to merge datasets. Most of the visualization was done using `plotly` package.
Shopping files were loaded as a list of data frames in R environment forming a large list of 168 elements of size 5.5MB. A new column was added, to all the data frames present in the list. Names of the csv files were extracted and added as new column to the data frames. Since the file name column contained the entire name of the csv file which contained details of the session and google-id. Only the google-id was extracted from the column using `gsub()` function which belongs to regex or regular expression package. List of data frames was processed using `sapply` function. Three columns were chosen from the data frame **google-id**, **tuoteryhma**(product type) and **ostot** (amount spent on buying a product). A condition was set to select dataframes containing **tuoteryhma** and other dataframes were dropped. Data frames containing **tuoteryhma** was saved as Rdata file with the same file name. Figure 3.1 , shows a pictorial representation of the file processing steps.
The saved datasets consisted of both, food items and non-food items purchased by the participants. Food items were extracted from the list of items and were mapped to the food groups. Foods purchased from S-group shops were categorized into food groups present in Table 3.1. Figure 3.1 also provides some examples showing how the mapping of foods was done to their respective food groups.

```
_s_group_categories.csv
TUOTERYHMÄ,Category_RS
"OLUET <4,7%","Alcohol"
"VIINIT, SIIDERIT <4,7%","Alcohol"
"LAKRITSIT","Candy and chocolate"
"MAKEISPUSSIT","Candy and chocolate"
"MUUT MAKEISET","Candy and chocolate"
"PASTILLI- JA RAERASIAT","Candy and chocolate"
"SUKLAAVALMISTEET","Candy and chocolate"
"KAHVIT","Coffee"
"JOGURTTI","Dairy products"
"JUUSTO","Dairy products"
"MAITO JA KERMA","Dairy products"
"MUUT HAPANMAITOVALM.","Dairy products"
"JÄÄTELÖT","Dairy products"
```

Figure 3.1 *Mapping of products to Food groups*

There were 427 types of products purchased by the participants, 73 of them were food products that matched our category list. These 73 products were categorized into our 21 categories. `merge()` function was used to map shopping list containing

product list(tuoteryhma) and the respective food categories(see figure 3.1). Personalized computation was done for 35 participants. Calculations for singles were done separately from family members, to find the difference between shopping pattern of singles and family members.

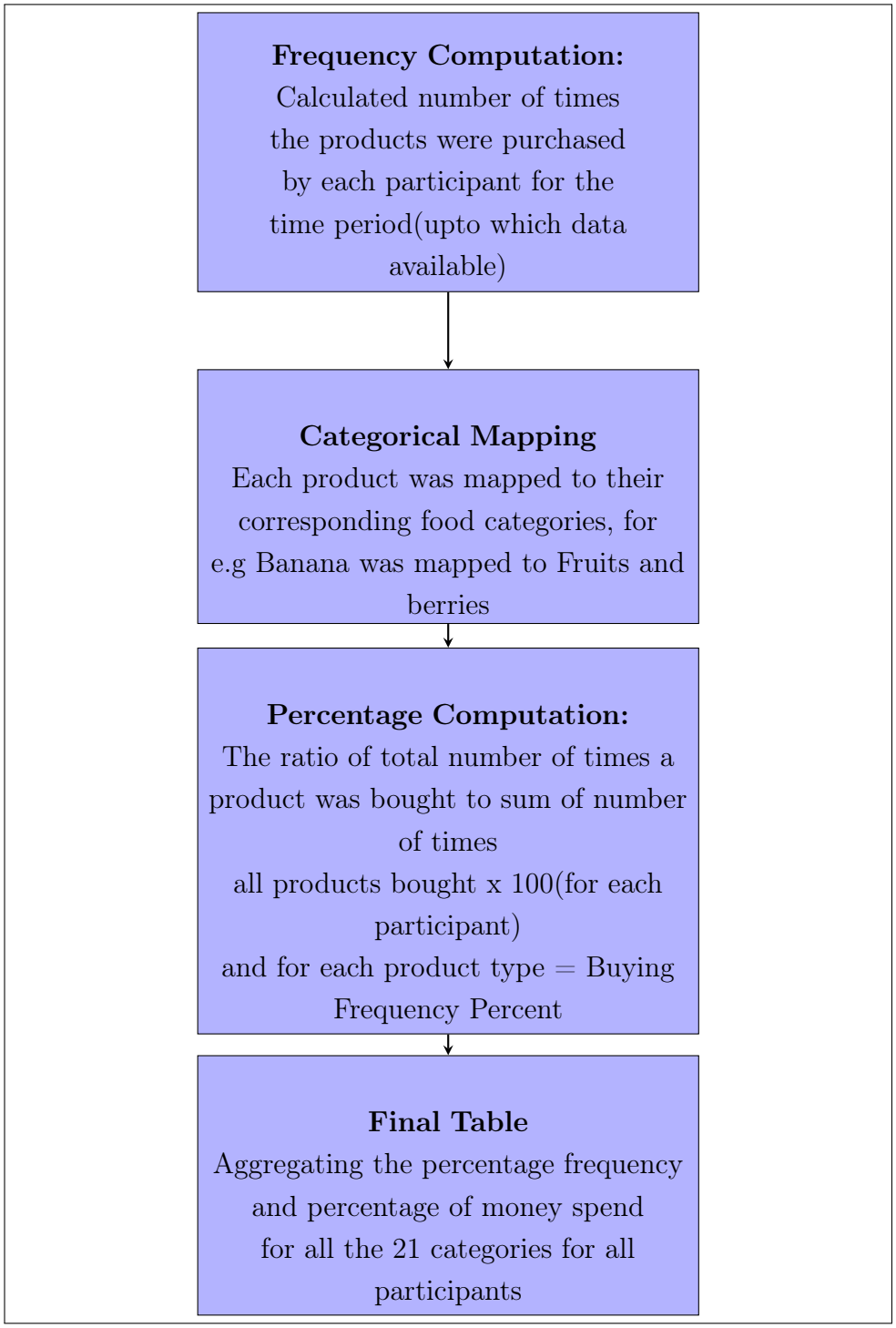


Figure 3.2 Mapping Food Categories and Percentage Calculation

Figure 3.2 is a pictorial representation of how the buying frequency percentage and money spend on each category was computed. Using the `table()` function, frequency of product purchase was calculated for the participants individually, and using `as.dataframe()` converted the frequency computation output to dataset. The dataset consisted of google-id, product type (tuoteryhma), number of times the product was bought by the participant(frequency) and also the amount of money invested in buying the product.

	TUOTERYHMÄ	OSTOT	freq
ATER.AIN,	LIEMIVALM	12.18	8
	BANAANIT	35.76	39
	DEODOR&KOST.PYYHK	4.10	2
	EINESVALMISTEET	2.36	2
	ELIMET	3.19	1
	ETIKAT	16.39	7

Figure 3.2 *Frequency table of a participant*

The frequency table consisted of all the products purchased by the participants. Food products were extracted from the list using food category mapping. Using the S-group food mapping table(see figure 3.1) , the food products were extracted from participants shopping list using `merge()` function i.e by merging S-group food mapping table with each participants shopping list. Figure 3.3 shows the mapping of the food items to food groups.

	TUOTERYHMÄ	RS	OSTOT	freq
	BANAANIT	Fruits and berries	10.43	6
	EINESVALMISTEET	Ready made food	19.28	5
HED&MARJA&MEHUPAKASTEET		Fruit juices	8.02	1
	JÄÄTELÖT	Dairy products	32.47	6
	JÄLKIRUOAT	Pastry and desserts	1.19	1
	JOGURTTI	Dairy products	1.38	1

Figure 3.3 *Mapping of Food Categories*

Frequency percentage of products being bought and money spend percentage on each product was calculated. The equation for calculating the frequency percentage of the products bought was :

$$\text{Frequency Percentage}(\%) = \frac{\text{Buying frequency of a product}}{\text{Sum of buying frequency of all products}} \times 100 \qquad (3.1)$$

Similarly,the percentage of money spend on buying the products was calculated using the equation:

$$\text{Money Spend Percentage}(\%) = \frac{\text{Money spend on buying a product}}{\text{Total amount spend in buying all products}} \times 100 \qquad (3.2)$$

After percentage calculation, the products were aggregated based on the food category they belonged to. The aggregation was done using `aggregate()` function. The resulting dataset consisted of category percentage showing how frequently a category of food was shopped and how much was invested on a category by each customer(see figure 3.4).

Category_RS	freq	OSTOT
Alcohol	1.0319917	2.4259623
Candy and chocolate	0.2063983	0.2789797
Coffee	0.7223942	1.0178555
Dairy products	12.1775026	13.7632635
Eggs	3.0959752	1.4833553
Fish and seafood	1.9607843	5.6364299

Figure 3.4 *Aggregate Category table*

Single participants were separated from participants who were family member. A separate dataset for single participants and another dataset for family members was created.

3.3.2 Processing Activity Data

Activity data collected from the Withings activity device consists of 21 variables and had activity observation data for 88 people. Figure 3.5 shows a section of sensor data collected from the activity app. The observation consisted of how much distance was covered and how many steps did the participant walk on a particular date. From these 21 columns , google-id, date and steps columns were extracted.The data type of date column was factor, it was converted to date(yyyy-mm-dd) format using `as.Date()` function.

X_clazz	date	distance	steps	timezone
WithingsActivity	2016-02-20	4546.757	5230	Europe/Helsinki
WithingsActivity	2016-02-21	6148.051	7114	Europe/Helsinki
WithingsActivity	2016-02-25	5823.042	6564	Europe/Helsinki
WithingsActivity	2016-03-01	12216.64	13611	Europe/Helsinki
WithingsActivity	2016-02-28	8031.599	8979	Europe/Helsinki
WithingsActivity	2016-03-05	2094.981	2483	Europe/Helsinki

Figure 3.5 *Activity Table*

Activity data were recorded for a span of 1 year from `2016-02-10` to `2017-02-11`. In activity analysis also data of 35 participants were taken into account, and singles were separated from the family members. Activity for first 30 days was considered. Since the participants started their activity on different dates, the data set for each participant was rearranged according to ascending order of date and first 30 observations were only considered. The steps column was converted from factor data type to numeric using `as.numeric()` function, and were aggregated for each google-id, in order to find out how many steps did each participant traverse during first 30 day time span.

3.3.3 Questionnaire Data Processing

In the questionnaire data, the focus is FFQ [16], family size and activity feedback. There were different question codes that represented different types of questionnaire feedback collected. FFQ questions mostly started with question code pattern `DHR07-93-`. `DHR01-27-` represented the questions regarding family size. There were single persons reporting the questionnaires as well persons with family size of 2-7 people. Questionnaire data collected before the wellness coaching was used. The questionnaire data didn't have google-id , thus to identify the participants feedback, the questionnaire dataset was mapped with `subject_id` dataset which contained information of all the participants using `user_id` as a common key to merge two datasets.

Like in the shopping food categorization the FFQs were also categorized based on what type of food the question was about. Similar to shopping categorization these types of food in FFQ questions were categorized into 21 food categories(see Table 3.1) based on what food product the question was about. Figure 3.6 is an example of how FFQs were categorized. There were questions about 46 food types, and these questions were mapped to 21 food groups.

question	label	RS_Categ
DHR07-93-2A	Tuoreita kasviksia, raasteita tai kasvissalaatteja	Vegetables
DHR07-93-2B	Keitettyjä lisäkekasviksia	Vegetables
DHR07-93-2C	Kasvisruokia	Vegetables
DHR07-93-2D	Perunaa tai perunaruokia	Potatos
DHR07-93-2E	Tuoreita hedelmiä	Fruits and berries
DHR07-93-2F	Hedelmäsäilykkeitä	Fruits and berries
DHR07-93-2G	Marjoja	Fruits and berries
DHR07-93-2H	Pähkinöitä tai manteleita	Nuts and seeds
DHR07-93-2I	Ruis- tai kokojyväleipää	Grains

Figure 3.6 *FFQ Food Categorization*

In FFQ question code `DHR07-93-2` queried about how many times a day/week/-month did the participant have the following foods. To answer the question, participants gave the measure in per day basis, some in weekly basis and some on monthly basis. To have uniformity in the measure of intake and ease of calculation, intakes were estimated in monthly scale. Table 3.3 shows the conversion of intakes to monthly scale.

Intake Value (Finnish)	Intake Value (English)	Average Value
1-3 krt/kk	1-3 times/month	2 times/month
Kerran/viikko	once /week	4 times/month
2–4 krt/viikko	2-4 times/week	12 times/month
5–6 krt/viikko	5-6 times/week	22 times/month
1 krt/vrk	once /day	30 times/month
2-3 krt/vrk	2-3 times/day	75 times/month
4-5 krt/vrk	4-5 times/day	135 times/month
>=6 krt/vrk	>=6 times/day	180 times/month

Table 3.3 *Consumption Values Reported in Questionaire Feedback*

Other than the FFQs, questionnaire dataset asked questions related to activity as well. Similar to FFQs there were no google-ids for activity questionnaire. Google-ids for activity questionnaires were mapped from subjectid dataset, as was done for FFQs. Question code `DHR04-41, 42, 43, 44, 44A, 44B, 44C and 45` , asked questions such as - How stressful is a person's job physically? How much amount of time is spent by a person in gym training ? How often did a person practice exercise for the last six months on average per week? There were multiple choice questionnaires to best describe exercise habits of a person, for example whether a

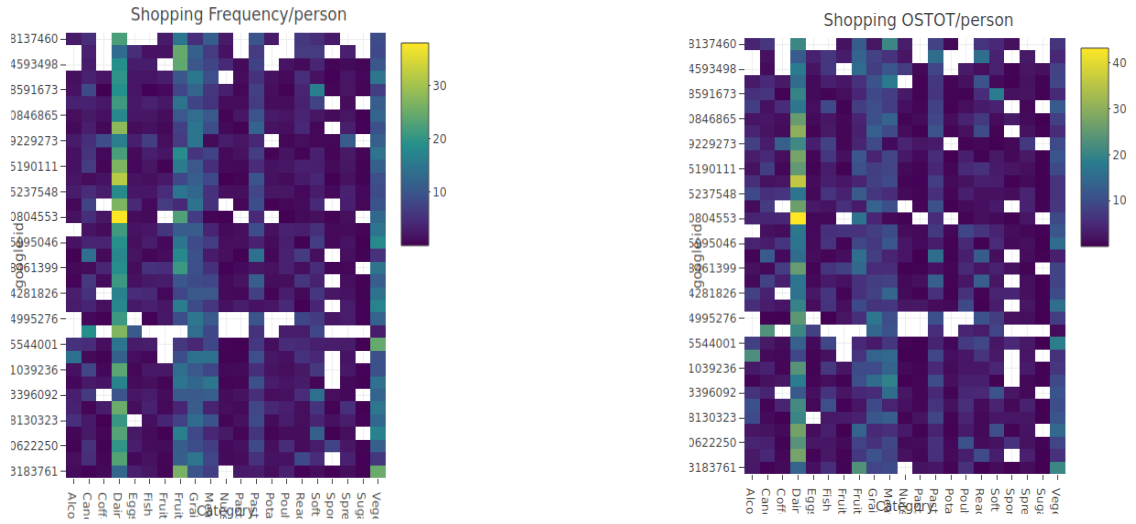
person is engaged in walking or not? how much does he walk? Whether a person is engaged in demanding physical activity or not? etc. The amount of activity done by a person is answered using coded values ranging from 0-6. Indexing of activity is done to quantify the amount of exercise done based on questionnaire results. 4-6 stands for strenuous exercise, 1-4 for mild exercise whereas low coded values 0-2 stand for minimum exercise or sedentary lifestyle.

4. RESULTS

In DHR Study [37] , 96 people participated and provided their information regarding shopping and answered the questionnaires. Physical activity records were available from 88 persons, 35 persons provided their information both for shopping and also participated in FFQ [16] feedback. Activity data for these 35 participants were also available. Thus, all needed shopping, activity and questionnaire data was available from 35 persons. Due to small sample size, descriptive statistics methods like box-plots and median plots were mostly used for analysis.

4.1 Shopping Information and FFQ Comparison

Personalized shopping trend was observed using heat maps which displayed, the most frequently bought food categories for every participant and on which food categories they invested more money.



(a) frequency of buying food categories

(b) money spend on food categories

Figure 4.1 *Heatmap representation of shopping*

Food categories were plotted against the google-ids. google-ids were plotted in the

y-axis and in the x-axis were the food categories. Color coding represents the number of times a product was bought by a person. Number of times a food category is bought is in percentage scale(0-100) and amount invested in buying food category is also represented in percentage scale(0-100). Higher the number of products bought or more the amount of money spend more color tends are towards green and yellow. The heat map shows buying trend for some participants.From the heatmap it is visible that participants have been buying dairy products quite often and money spent heat map also supports the fact that they have spend more money on buying dairy products. Other than dairy products, food categories like fruits and berries, grains, vegetables were frequently bought by the participants.The white regions in the heat map showed that some of the food categories were not purchased by the participants.

4.1.1 Boxplot Comparison

Box plot visualization explains the data more descriptively. Box plot provides a summary statistics of the dataset. The summary statistics include five numbers - sample minimum, sample maximum, median, first quartile and third quartile. Minimum is the lowest data point in a dataset, Maximum is the highest data point in the dataset. Median is the middle value if there are odd number of samples in an ordered dataset. Median for odd number of data samples can also be written as :

$$(\frac{n+1}{2})th \quad \text{observation} \tag{4.1}$$

For even number of data samples, median is average of :

$$(\frac{n}{2})th \quad \text{and} \quad (\frac{n+1}{2})th \quad \text{observations} \tag{4.2}$$

Lower quartile `(Q1)` or first quartile is 25 percentile of dataset i.e 25 percent of numbers in a dataset is below lower quartile, and 75 percent of numbers in the dataset is above first quartile. The third quartile or the upper quartile `(Q3)` or 75 percentile of the data i.e 75 percent of the data in the dataset is below this number [40][41] [42].

Boxplots are very useful when there is large number of observations involved and when data from two different datasets are compared. Boxplots can be used to compare distributions, since in boxplots center, spread of the data and overall range

of the data is visible in details [43].
Figure 4.2 shows a boxplot in details, the ends of the box are upper and lower quartiles, and the box spans the interquartile range. Median is marked by the vertical line present in the middle of the box. Whiskers are the two lines outside of the boxplot that extends to show highest and lowest observations in the dataset. Outliers are represented using a dot [43].

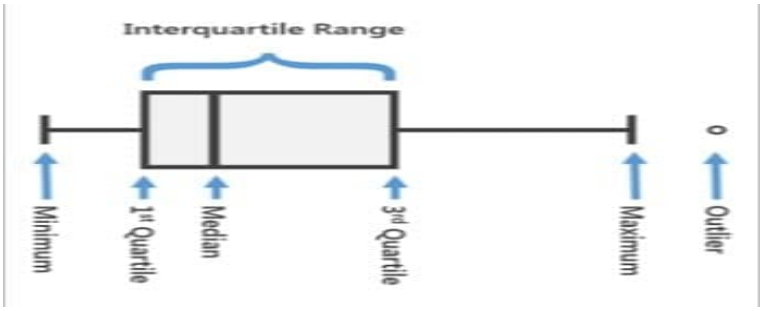


Figure 4.2 *Boxplot [11]*

Boxplot(figure 4.3) representation shows the average distribution of the shopping data of the participants who are single. Buying frequency of the food groups and money spend on buying different food groups were in percentage scale. The plot shows how frequently the participants bought food categories like dairy products, fruit and berries, vegetables, meat and sausages, etc. The frequency of buying also corresponds with the money spent on buying, for most of the categories. For categories like alcohol, dairy products, fish and seafood, meat-sausages and poultry more amount of money was spent relative to their buying.

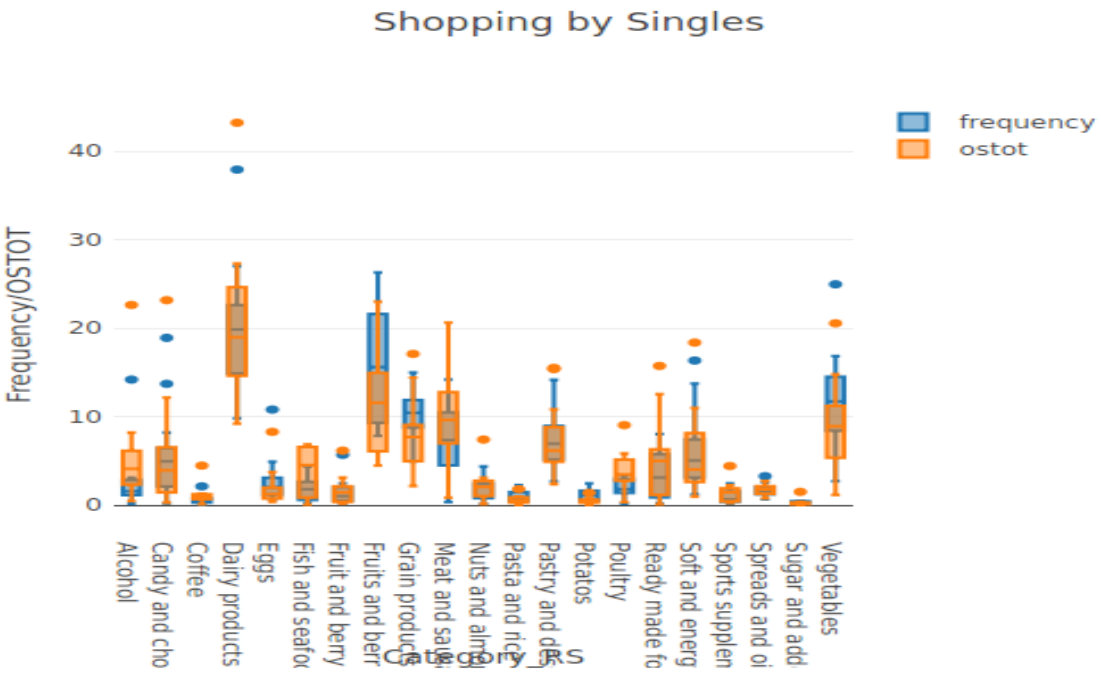


Figure 4.3 *Descriptive Shopping Details for Singles*

FFQs answered by the participants were also calculated and were transformed to 0-100 scale(see figure 4.4), so as to maintain uniformity of the measure and to have comparative analysis with the shopping information.

Category_RS	value
Alcohol	0.5917160
Candy and chocolate	0.5917160
Coffee	8.8757396
Dairy products	15.3846154
Eggs	3.5502959
Fish and seafood	1.1834320
Fruits and berries	10.6508876
Grain products	11.2426036

Figure 4.4 *Section of FFQ values for singles*

A comparative boxplot(see figure 4.5) was plotted to show the difference between the self-reported nutritional habits and categorized shopping data. For some of the categories, the buying frequency corresponded well with the self-reported food consumption. But for some categories for example coffee, spreads, alcohol, candy and chocolates, sugar and ready-made food, there was a large difference between self-reported information and information derived from the shopping data.

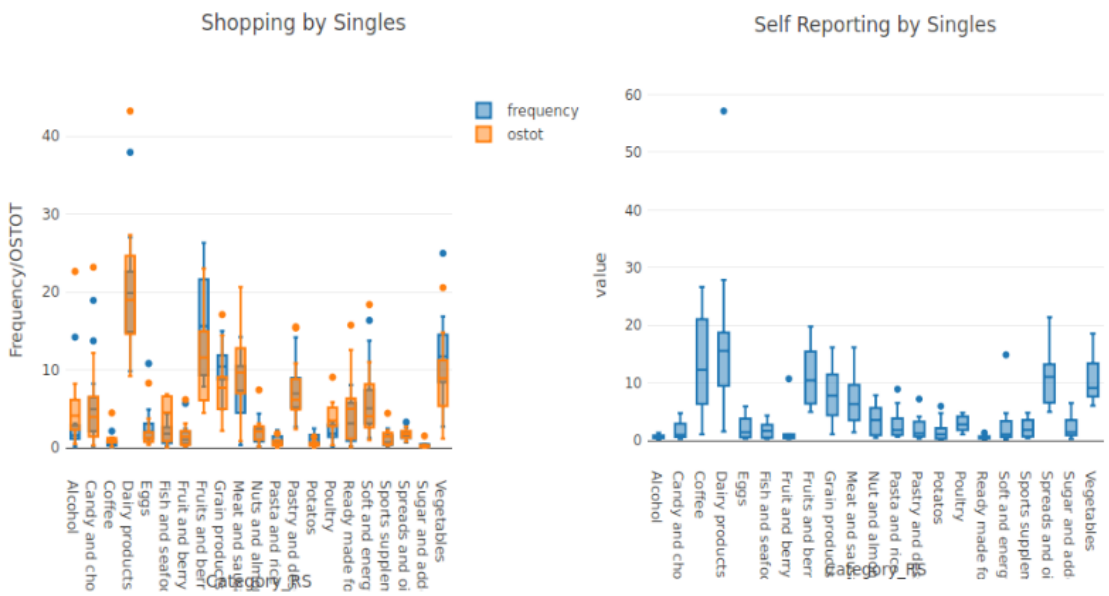


Figure 4.5 *Boxplot comparison for singles*

Consumption of dairy product corresponded, what participants reported. Reporting on coffee consumption differed from participants spending to buy coffee from S-group stores.

Amount of spreads bought, and money spent on buying spreads did not match with the consumption. Monthly consumption of spreads by the participants was quite high whereas they were not frequently bought by the participants. Some of the categories for example - Alcohol, candies and chocolates, pastry and desserts, which were frequently bought and a greater percentage of money was spent on buying them. But the participants reported to have not consumed them as proportion to their buying. Though very small amount of money was spent on buying sugar and frequency of buying sugar was very small. But the plot showing food consumption, showed participants consumed high amount of sugar.

Family members also had a similar reflection on buying and reporting, but the shopping amounts were more than the single shopping.

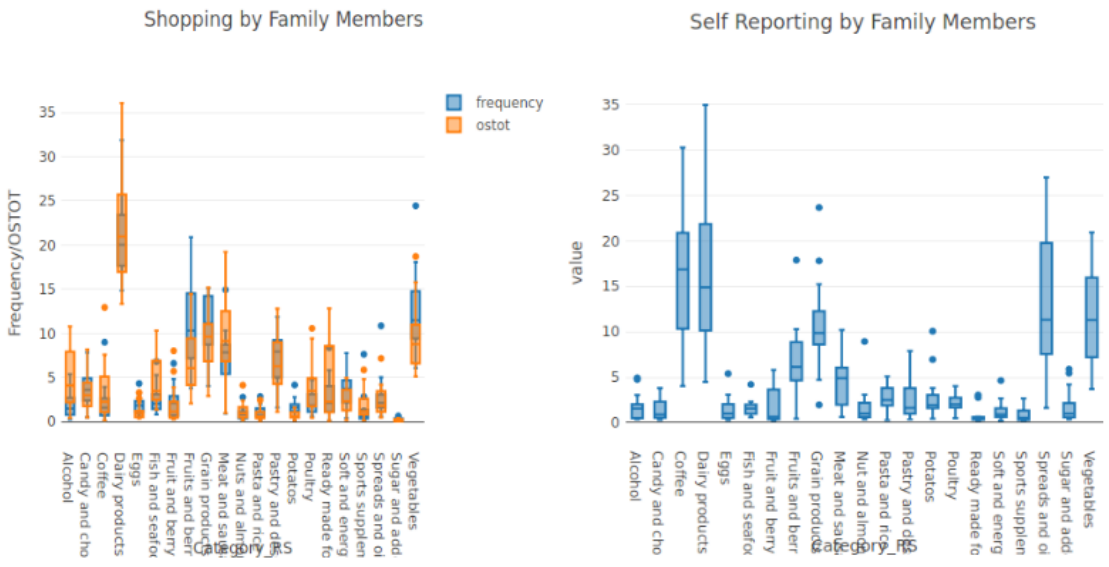


Figure 4.6 *Boxplot comparison for family members*

For example, family members spend more money on buying alcohol,dairy products than single members. Maybe because they had to buy not only for themselves but also had to buy for other family members. Fruit berry juices, dairy products were consumed more by the family members than the singles. Family members consumed more coffee than the singles. Most of the singles consumed 6-21 percent of coffee among all the food groups consumption, whereas most of the family members consumed around 10-20 percent of coffee among all food groups. Spreads were consumed more by the family members than the single participants. Family members bought spreads in more amount(around 1-5 percent of their total shopping), whereas spreads comprised of 1-2 percent of total shopping done single participants.

4.1.2 Median Plot

Median difference measures the absolute difference between the medians in two different groups of data. It is difficult to compare the differences between the data from two different groups with different measures. Median difference is a way to measure effect size [44].

Median plots were used to visualize the average difference for all the food categories. The average difference between the food categories gave a clear picture how the shopping information collected differed from the reporting done by the participants. In Figure 4.7 the median table contains the median values for frequency of food

categories bought and amount of money spend by the participants for buying all food categories. It also contains the median values for food intake reporting done by the participants in FFQ for all the food categories.

	Category_RS	freq	OSTOT	value	freq_val	ostot_val
1	Alcohol	1.5612161	4.1125560	0.6230530	0.93816315	3.48950304
2	Candy and chocolate	4.9514053	3.9421196	0.9731794	3.97822594	2.96894025
3	Coffee	0.5809497	0.9228695	12.2661272	11.68517750	11.34325767
4	Dairy products	19.8646072	18.9820046	15.5398413	4.32476585	3.44216323
5	Eggs	2.0293187	1.2420356	1.4268748	0.60244391	0.18483922
6	Fish and seafood	1.8000643	4.5266637	1.6905118	0.10955251	2.83615188
7	Fruit and berry juices	1.0245349	1.4298959	0.7104796	0.31405537	0.71941635
8	Fruits and berries	15.6118143	11.5821675	10.4536489	5.15816543	1.12851857
9	Grain products	10.4417822	7.7050970	7.7851676	2.65661461	0.08007062
10	Meat and sausages	7.3423075	9.6471759	6.3260341	1.01627341	3.32114185
11	Nuts and almonds	0.9819967	2.0694105	3.5629454	2.58094864	1.49353485
12	Pasta and rice	0.7233273	0.4843208	1.8314404	1.10811312	1.34711966
13	Pastry and desserts	6.9565101	6.1684148	1.2524850	5.70402509	4.91592980
14	Potatos	1.1503698	0.4371965	1.0783340	0.07203578	0.64113747
15	Poultry	1.8018018	3.4707372	2.7931873	0.99138555	0.67754988
16	Ready made foods	3.1268837	4.9980610	0.5420054	2.58487824	4.45605556
17	Soft and energy drinks, juices	5.0535240	4.0414855	1.0657194	3.98780463	2.97576610
18	Sports supplements	0.6479064	1.4597312	1.8543313	1.20642491	0.39460015
19	Spreads and oils	1.5028707	1.7683578	11.0476190	9.54474840	9.27926126
20	Sugar and added sugar	0.3048780	0.1754666	1.4285714	1.12369338	1.25310485
21	Vegetables	11.7302932	8.9118982	9.1201759	2.61011736	0.20827767

(a) Median table for singles

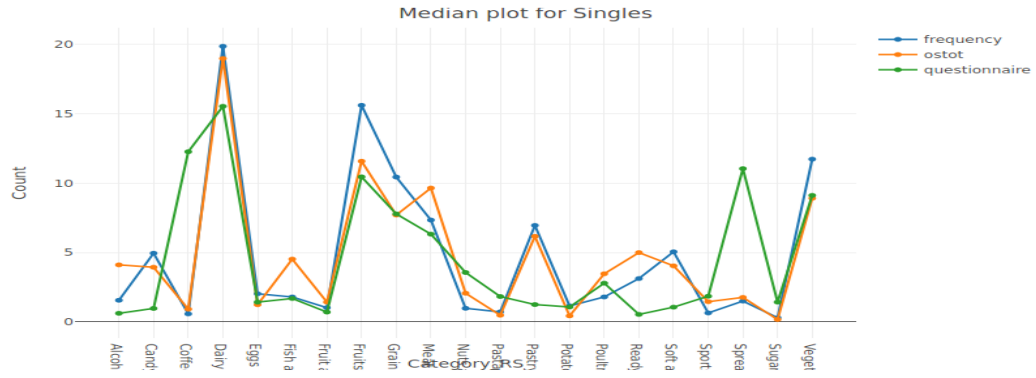
	Category_RS	freq	OSTOT	value	freq_val	ostot_val
1	Alcohol	1.5069968	4.1184409	1.5625000	0.05550323	2.55594093
2	Candy and chocolate	3.6124795	3.0173300	0.9154099	2.69706956	2.10192007
3	Coffee	1.5907677	2.2699284	16.8539326	15.26316489	14.58400415
4	Dairy products	20.0278164	20.9602390	14.8876404	5.14017596	6.07259856
5	Eggs	1.8541559	0.9043662	1.0178117	0.83634422	0.11344555
6	Fish and seafood	2.0661157	3.4782102	1.5625000	0.50361570	1.91571021
7	Fruit and berry juices	2.1342846	1.1596258	0.6608494	1.47343520	0.49877631
8	Fruits and berries	10.3305785	6.0687341	6.1433447	4.18723380	0.07461063
9	Grain products	11.1526946	9.6274537	9.8654709	1.28722376	0.23801712
10	Meat and sausages	7.8571429	9.1567640	4.9140049	2.94313794	4.24275905
11	Nuts and almonds	0.8116883	0.9823702	1.0178117	0.20612339	0.03544152
12	Pasta and rice	1.0714286	0.5706173	2.5080822	1.43665361	1.93746491
13	Pastry and desserts	7.9470199	6.2963808	1.6310992	6.31592067	4.66528163
14	Potatos	1.6160393	0.6387002	1.9448947	0.32885531	1.30619450
15	Poultry	1.8595041	3.4902500	2.0202020	0.16069789	1.47004795
16	Ready made foods	2.1528525	2.2820294	0.4494382	1.70341433	1.83259122
17	Soft and energy drinks, juices	3.7463977	2.3303573	0.9280742	2.81832345	1.40228301
18	Sports supplements	0.6568144	1.1837528	0.5356759	0.12113850	0.64807688
19	Spreads and oils	2.1528525	1.8148450	11.3281250	9.17527247	9.51327998
20	Sugar and added sugar	0.2674399	0.1265886	1.0101010	0.74266109	0.88351237
21	Vegetables	11.4906832	8.7887439	11.3022113	0.18847193	2.51346742

(b) Median table for family

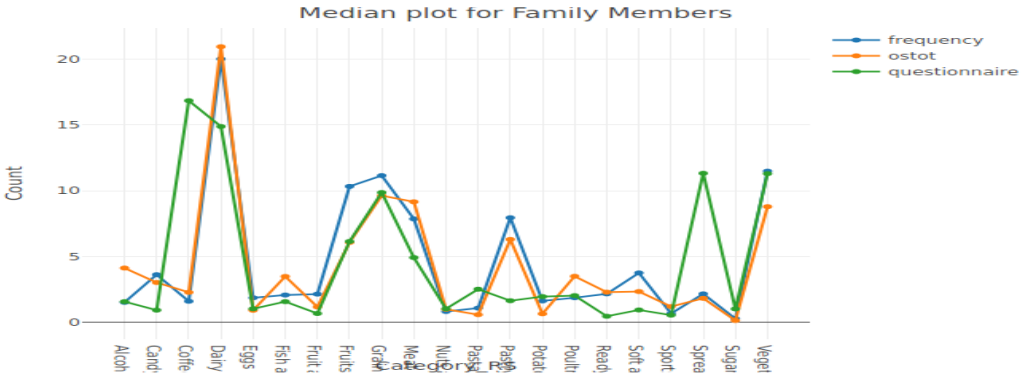
Figure 4.7 *Median Table*

Figure 4.7(a) and (b) shows the median values for shopping i.e column freq is the median percentage frequency of shopping and column ostot is the median percentage

amount of money spend on shopping all the food groups by the participants who are single and family members respectively. Column value contains the median values of the food intake reporting done by individuals for all the food groups. Column 4 and 5 are the median difference between frequency and feedback value, money spend(ostot) and feedback value respectively. The differences calculated were absolute differences [45]. Figure 4.8(a) and (b) visualizes the differences.



(a) Median plot for singles



(b) Median plot for family members

Figure 4.8 *Median plot*

Figures 4.8 shows on which food groups average participants have invested more. The plot compared the shopping and the food intake reporting done by the participants. Participants have bought products like dairy products, fruits and berries, vegetables, grain products, meat and sausages frequently and they also spend most of their money buying these food groups. For single participants there was high difference between shopping and consumption of food categories such as coffee, spreads and oils, pastry and desserts(see figure 4.9).

Category_RS	freq	OSTOT	value	freq_val	ostot_val
Coffee	0.5809497	0.9228695	12.2661272	11.68517750	11.34325767
Spreads and oils	1.5028707	1.7683578	11.0476190	9.54474840	9.27926126
Pastry and desserts	6.9565101	6.1684148	1.2524850	5.70402509	4.91592980

Figure 4.9 *Top 3 categories with high differences between buying and reporting (Single participants)*

Family members reported to have consumed more amount of coffee, dairy products, pastry and desserts than the singles. Figure 4.10 shows the differences in shopping and food consumption for family members. Figure 4.10(a) shows the food categories for which there is a high difference between the buying frequency and consumption and Figure 4.10(b) shows the top three food categories. For buying these food categories considerable amount of money was spent, but the consumption reporting of these food categories did not correspond to the amount spend for buying these food categories.

Category_RS	freq	OSTOT	value	freq_val	ostot_val
Coffee	1.5907677	2.2699284	16.8539326	15.26316489	14.58400415
Spreads and oils	2.1528525	1.8148450	11.3281250	9.17527247	9.51327998
Pastry and desserts	7.9470199	6.2963808	1.6310992	6.31592067	4.66528163

(a) Top 3 categories with high differences between buying(i.e frequency of buying) and reporting (Family Members)

Category_RS	freq	OSTOT	value	freq_val	ostot_val
Coffee	1.5907677	2.2699284	16.8539326	15.26316489	14.58400415
Spreads and oils	2.1528525	1.8148450	11.3281250	9.17527247	9.51327998
Dairy products	20.0278164	20.9602390	14.8876404	5.14017596	6.07259856

(b) Top 3 categories with high differences between money spend and reporting (Family Members)

Figure 4.10 *Categories with high median differences*

4.2 Measuring the step counts with respect to PAQ

Calculating correlation between the steps and the coded values was an effective measure to understand how much truly the Physical Activity Questionnaire(PAQ) was answered by the participants. Pearson's correlation was calculated between step count and exercise feedback, as answered by the participants in terms of coded values.

Pearson's correlation is the measure of association between two variables X and Y. The correlation value ranges from +1 to -1. Higher the correlation strength between two variables the correlation coefficient is closer to +1 this is also called positive correlation. Lesser the correlation strength, value of the correlation coefficient is closer to -1. It is also called negative correlation or anti-correlation. If the correlation value is 0 then there is no association between two variables [46].
So for a dataset containing n variables

$$\{x_1, ..x_n\}$$

and another dataset containing n variables

$$\{y_1, ..y_n\}$$

correlation coefficient between these two data set can be calculated using the formula:

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{(\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2})(\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2})} \tag{4.3}$$

where n is the sample size. x_i ,y_i are the represented as the single data.$\overline{x}$ is the mean i.e

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{4.4}$$

`corr()` function in R uses (4.3) to calculate the correlation between two variables. By default `corr()` function calculates Pearson's correlation. Calculating the correlation between step counts and coded value for all the 35 participants, Pearson's correlation coefficient was -0.002112618. That means there was no correlation between step counts and participant's physical activity reporting.

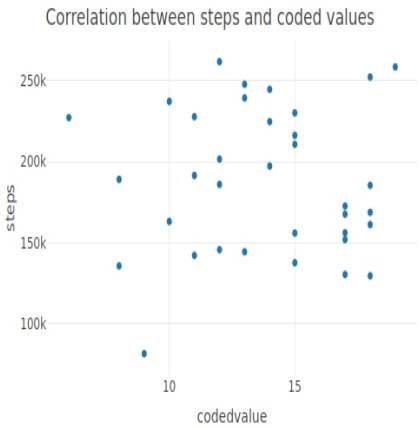


Figure 4.11 *Correlation plot between steps and coded values*

5. DISCUSSION

Since 35 participants were only available for the comparative analysis, the sample size was very small, and it was very hard to predict and conclude based on small sample size. In the results sections, there were some food categories which did not have any correspondence with the participant's intake and their buying.

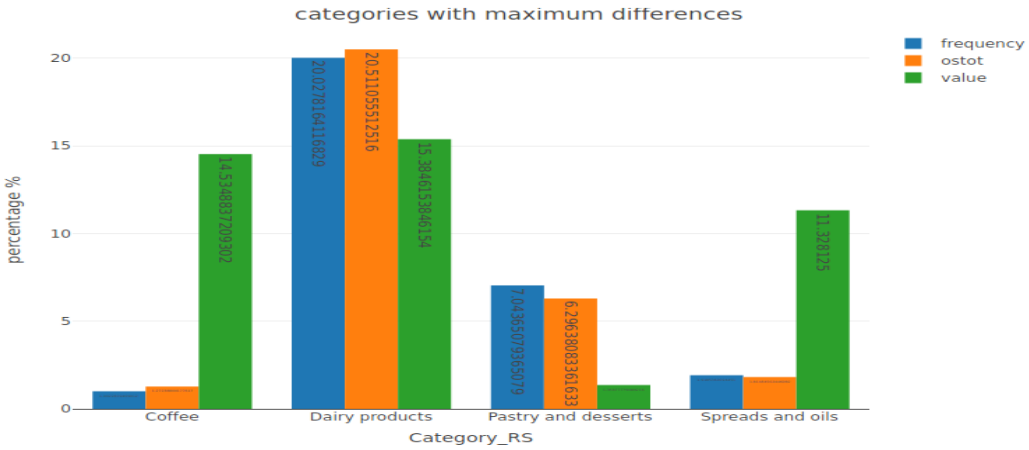
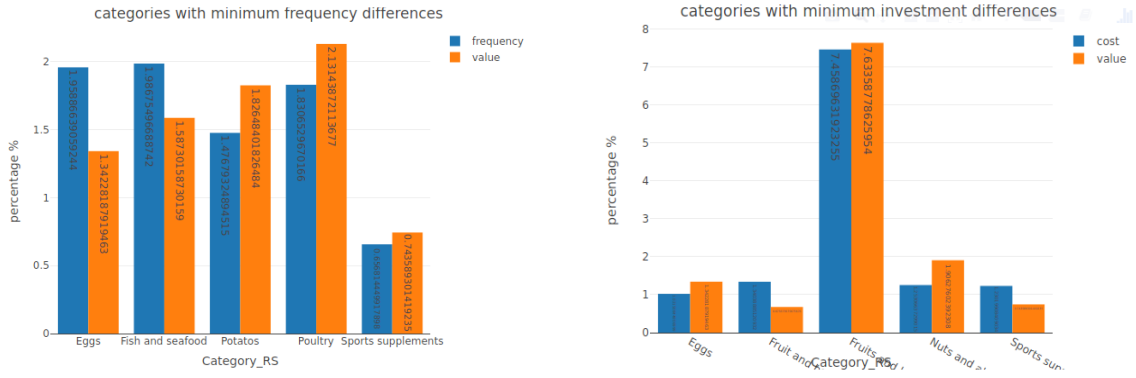


Figure 5.1 *Food Categories with high difference between shopping and consumption by participants*

Figure 5.1 shows the participants coffee consumption was around 13 percent higher relative to what they shopped from S-group shops. The participants spend 1.27 percent of their entire shopping on buying coffee, even the frequency of buying coffee was approximately 1 percent of the entire shopping. Whereas the participants reported to have consumed 14.53 percent of coffee from entire consumption. This may be because they have consumed coffee not only from their buying but also from other places may be coffee shops/cafe, some people also drink a lot of coffee at workplace. Similar trends were spotted for the food category spreads and oils. Participants spend 1.8 percent of their buying on spreads and 1.9 percent was the buying frequency of spreads from all the food categories bought. Whereas spreads and oils accounted about 11.32 percent of total consumption. This may be because spreads are quite often used in cooking and baking foods. When oils are bought they are in per litre amount but are used in very small quantity to fry or bake foods. Also

the [5] suggests most of the Finns have one of their meals outside. At the same time food categories like dairy products, pastry and desserts were bought quite frequently by the participants, even the investments were quite high on buying those food categories. But the participants reported they did not consume the categories much relative to their buying. 20 percent of the shopping was spent buy the participants on dairy products where as they had 15 percent of dairy product in their diet. 6 percent of expenditure of all the food categories was on Pastry and Desserts but the participants reported to have consumed 1.3 percent pastry and dessert from all the food groups. This may be either due to false reporting by the participants or measurement errors while filling the FFQ [16]. One of the major limitations of an FFQ is the measurement errors which are because of inaccuracies in food frequency reporting and also because of portion size estimation [47] [48]. Food groups which corresponded buying and consumption are eggs, fish and seafood, potatoes, poultry and sports supplements (see figure 5.2).



(a) categories with minimum frequency difference

(b) categories with minimum investment difference

Figure 5.2 *Categories with minimum differences*

There was no correlation between the step counts and the PAQ coded value feedback given by the participants. The pedometer is designed to count each step taken by a person. Present day pedometers use electronics and software to measure a person's step. The amount of distance travelled is calculated directly by GPS receiver [49]. In PAQ the participants answered feedback regarding general exercise which consisted of other strenuous activity other than walking, running, etc. This can be a reason why the step counts were not correlated with feedback given by the participants.

5.1 Limitations

From the results, limitations of the analysis were easily figured out. Some of the limitations are listed below which hindered the analysis:

- Available data sample was very small. Analysis work was done based on 35 participants data. It is very difficult to conclude an analysis based on such a small sample.
- Data from one shopping chain was only available, which doesn't give a proper picture of the shopping and what people are having throughout the study period. If the whole digital footprint of consumption for instance detailed product level shopping information from all stores and restaurants,were collected - food consumption could have been estimated more accurately. For example, it would have then been possible to answer questions like why did shopping for categories like coffee, spreads and pastry-desserts, etc did not correspond to FFQ answered by the participants.
- Some of the products were purchased in terms of weight and some in terms of piece. These kind of purchases create scaling problem for buying analysis. Thereby the purchases were calculated based on approximation.
- Since family size varied for most of the family members, it was the potential cause of noise in estimation.
- FFQ [16] data was also an approximation and was based on verbal answers and questionnaires filled by people. This way of obtaining FFQ didn't provide a transparent picture of consumption. A true picture of participant's daily food consumption was missing.
- Only steps count was collected using Withings [39] activity app. More activity related information such as motion detection, apps capable of measuring exercise should have been used to gather data for an ideal analysis. This can be considered as a positive exploration from the analysis as well that we should not depend on pedometer data and steps count for activity analysis.

5.2 Proposal: Ideal ways to record data

Using the knowledge of SCI[1][2] system , and limitations faced while collecting data in this project, and unavailability of proper data which affected the analysis, an architectural proposal for ideal data gathering system can help solve problem in future.

For gathering shopping related information, all the shopping done by the participants must be recorded. It should not only come from one source(for example S-group). [17] were able to gather data from 15 sources - *banking, education, energy, fitness, groceries, health care, housing, insurance, library, mobility, municipality, police, retail, telecommunication and web*. With the aid of SCI system (see Figure 2.1) it is possible to collect the required data. More use of static sensing infrastructure for example image analysis of purchased products can help with a better specification of the amount being bought. [47] [48] suggested about gathering non-quantitative FFQ data for example with the aid of photographs, we will be able better validate FFQs. Apps like My Meal Mate(MMM) [8] can be used to collect FFQs [16], which will be able to gather portion size questions along with non-quantitative FFQs. Only data from pedometers or step counts data is not enough for measuring the activity and comparing with Physical Activity Questionnaire. Many people spend time in vigorous activities like swimming, cycling, gardening and gymnastics. Pedometer records movement of the body in the vertical direction only, but movements done while biking, swimming cannot be recorded using pedometer [50]. Hence more motion capturing sensors and static sensing infrastructure[1] are required to record physical activity.

6. CONCLUSIONS

The study mostly focused on analyzing the digital footprints data of the participants and how much the information gathered from the questionnaires correlated with the digital footprints data. Due to small sample dataset prediction analysis was not done. Descriptive statistical analysis was done, for example boxplot representation visually represented the distribution of shopping data(how frequently did the participants buy the food categories and how much amount of money they have been spending on buying each food category) and the FFQ data(how frequently are the participants consuming the food groups). Comparative median table and plots showed the differences between shopping data and questionnaire data more clearly. For example coffee,candy and spreads buying information did not match with data gathered from FFQ. Also the activity data collected from Withings sensor app did not correspond with Physical acitivity reportings done by the participants. The step counts data was not correlated with the physical activity feedback by the people.To bridge the gap between limitations of the data were recognized. A proposal for data gathering infrastructure is suggested in Chapter 5.2 based on the knowledge we have from the architecture of digital footprints data gathering system.

BIBLIOGRAPHY

[1] D. Zhang, B. Guo, B. Li, and Z. Yu, "Extracting social and community intelligence from digital footprints: an emerging research area," in *International Conference on Ubiquitous Intelligence and Computing*. Springer, 2010, pp. 4–18.

[2] D. Zhang, B. Guo, and Z. Yu, "The emergence of social and community intelligence," *Computer*, vol. 44, no. 7, pp. 21–28, 2011.

[3] MyPyramid, "Mypyramid — Wikipedia, the free encyclopedia," 2010, [Online; accessed 21-May-2018]. [Online]. Available: https://en.wikipedia.org/wiki/MyPyramid

[4] MyPlate, "Myplate — Wikipedia, the free encyclopedia," 2010, [Online; accessed 6-June-2018]. [Online]. Available: https://en.wikipedia.org/wiki/MyPlate

[5] "Evira-finnish food safety authority." [Online]. Available: https://www.evira.fi/en/foodstuff/healthy-diet/nutrition-recommendations-for-all/

[6] P. M. Guenther, K. O. Casavale, J. Reedy, S. I. Kirkpatrick, H. A. Hiza, K. J. Kuczynski, L. L. Kahle, and S. M. Krebs-Smith, "Update of the healthy eating index: Hei-2010," *Journal of the Academy of Nutrition and Dietetics*, vol. 113, no. 4, pp. 569–580, 2013.

[7] E. Årsand, M. Muzny, M. Bradway, J. Muzik, and G. Hartvigsen, "Performance of the first combined smartwatch and smartphone diabetes diary application study," *Journal of diabetes science and technology*, vol. 9, no. 3, pp. 556–563, 2015.

[8] M. C. Carter, V. Burley, C. Nykjaer, and J. Cade, "'my meal mate'(mmm): validation of the diet measures captured on a smartphone application to facilitate weight loss," *British Journal of Nutrition*, vol. 109, no. 3, pp. 539–546, 2013.

[9] "Ukk institute guidelines for physical activity." [Online]. Available: http://www.ukkinstituutti.fi/en/

[10] E. J. Lyons, Z. H. Lewis, B. G. Mayrsohn, and J. L. Rowland, "Behavior change techniques implemented in electronic lifestyle activity monitors: a systematic content analysis," *Journal of medical Internet research*, vol. 16, no. 8, 2014.

[11] "Boxplot-diagram." [Online]. Available: https://pro.arcgis.com/en/pro-app/help/analysis/geoprocessing/charts/box-plot.htm

[12] F. B. Hu, "Globalization of diabetes: the role of diet, lifestyle, and genes," *Diabetes care*, vol. 34, no. 6, pp. 1249–1257, 2011.

[13] "Health wellness: Definition and dimensions." [Online]. Available: https://study.com/academy/lesson/health-wellness-definition-and-dimensions.html#transcriptHeader

[14] Jennifer Johnson, "Are you balanced? here are the 7 kinds of wellness you need." [Online]. Available: https://www.mindbodygreen.com/0-6795/Are-You-Balanced-Here-are-the-7-Kinds-of-Wellness-You-Need.html

[15] J. Häkkilä, A. Colley, V. Inget, M. Alhonsuo, and J. Rantakari, "Exploring digital service concepts for healthy lifestyles," in *International Conference of Design, User Experience, and Usability*. Springer, 2015, pp. 470–480.

[16] FFQ, "Ffq — Wikipedia, the free encyclopedia," 2010, [Online; accessed 2-March-2018]. [Online]. Available: https://en.wikipedia.org/wiki/Food_frequency_questionnaire

[17] O. Gencoglu, H. Similä, H. Honko, and M. Isomursu, "Collecting a citizen's digital footprint for health data mining," in *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*. IEEE, 2015, pp. 7626–7629.

[18] K. Shilton, "Four billion little brothers?: Privacy, mobile phones, and ubiquitous data collection," *Commun. ACM*, vol. 52, no. 11, pp. 48–53, Nov. 2009. [Online]. Available: http://doi.acm.org/10.1145/1592761.1592778

[19] J. Tang, R. Jin, and J. Zhang, "A topic modeling approach and its integration into the random walk framework for academic search," in *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*. IEEE, 2008, pp. 1055–1060.

[20] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: real-time event detection by social sensors," in *Proceedings of the 19th international conference on World wide web*. ACM, 2010, pp. 851–860.

[21] J. Bollen, H. Mao, and A. Pepe, "Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena." *Icwsm*, vol. 11, pp. 450–453, 2011.

[22] E. Mattila, J. Pärkkä, M. Hermersdorf, J. Kaasinen, J. Vainio, K. Samposalo, J. Merilahti, J. Kolari, M. Kulju, R. Lappalainen *et al.*, "Mobile diary for wellness management—results on usage and usability in two user studies," *IEEE Transactions on information technology in biomedicine*, vol. 12, no. 4, pp. 501–512, 2008.

[23] M. Turley, C. Porter, T. Garrido, K. Gerwig, S. Young, L. Radler, and R. Shaber, "Use of electronic health records can improve the health care industry's environmental footprint," *Health affairs*, vol. 30, no. 5, pp. 938–946, 2011.

[24] "What is nutrition?" [Online]. Available: http://whatisnutritiontips.com/

[25] "National health and medical research council." [Online]. Available: https://www.nhmrc.gov.au/health-topics/nutrition

[26] "Canada's food guide: What is food guide serving." [Online]. Available: https://www.canada.ca/en/health-canada/services/food-nutrition/canada-food-guide/food-guide-basics/what-food-guide-serving.html

[27] "Food-based dietary guidelines." [Online]. Available: http://www.fao.org/nutrition/education/food-dietary-guidelines/background/food-guide/en/

[28] W. Becker, N. Lyhne, A. N. Pedersen, A. Aro, M. Fogelholm, I. Phorsdottir, J. Alexander, S. A. Anderssen, H. M. Meltzer, and J. I. Pedersen, "Nordic nutrition recommendations 2004-integrating nutrition and physical activity," *Scandinavian Journal of Nutrition*, vol. 48, no. 4, pp. 178–187, 2004.

[29] D. Shahar, I. Shai, H. Vardi, A. Brener-Azrad, and D. Fraser, "Development of a semi-quantitative food frequency questionnaire (ffq) to assess dietary intake of multiethnic populations," *European journal of epidemiology*, vol. 18, no. 9, pp. 855–861, 2003.

[30] C. C. Tangney, B. A. Staffileno, and H. E. Rasmussen, "Healthy eating: How do we define it and measure it? what's the evidence?" *The Journal for Nurse Practitioners*, vol. 13, no. 1, pp. e7–e15, 2017.

[31] "Eating behaviour questionnaires." [Online]. Available: http://www.ucl.ac.uk/iehc/research/behavioural-science-health/resources/questionnaires/eating-behaviour-questionnaires/#fcq

[32] J. Meinilä, A. Valkama, S. B. Koivusalo, B. Stach-Lempinen, J. Lindström, H. Kautiainen, J. G. Eriksson, and M. Erkkola, "Healthy food intake index (hfii)–validity and reproducibility in a gestational-diabetes-risk population," *BMC public health*, vol. 16, no. 1, p. 680, 2016.

[33] S. N. Blair, M. J. LaMonte, and M. Z. Nichaman, "The evolution of physical activity recommendations: how much is enough?" *The American journal of clinical nutrition*, vol. 79, no. 5, pp. 913S–920S, 2004.

[34] "Physical activity guidelines for americans." [Online]. Available: https://health.gov/paguidelines/

[35] M. Kirwan, M. J. Duncan, C. Vandelanotte, and W. K. Mummery, "Using smartphone technology to monitor physical activity in the 10,000 steps program: a matched case–control trial," *Journal of medical Internet research*, vol. 14, no. 2, 2012.

[36] A. Anjum and M. U. Ilyas, "Activity recognition using smartphone sensors," in *Consumer Communications and Networking Conference (CCNC), 2013 IEEE*. IEEE, 2013, pp. 914–919.

[37] "Digital health revolution." [Online]. Available: http://www.digitalhealthrevolution.fi/

[38] H. N. Antti Kallonen, Soumya Das and R. Sallinen, "Comparison of food frequency questionnaire data and actual shopping records for the assessment of food intake," *WIS2018*, 2018.

[39] Withings, "Digital footprint — Wikipedia, the free encyclopedia," 2010, [Online; accessed 14-April-2018]. [Online]. Available: https://en.wikipedia.org/wiki/Withings

[40] "Median." [Online]. Available: http://www.mathcaptain.com/statistics/median.html

[41] "What is a box plot and when to use it." [Online]. Available: https://chartio.com/resources/tutorials/what-is-a-box-plot/

[42] Quartiles, "Quartiles — Wikipedia, the free encyclopedia," 2010, [Online; accessed 22-June-2018]. [Online]. Available: https://en.wikipedia.org/wiki/Quartile

[43] "Constructing box and whisker plots." [Online]. Available: https://www.statcan.gc.ca/edu/power-pouvoir/ch12/5214889-eng.htm

[44] "Effective sample size: Definition, examples." [Online]. Available: http://www.statisticshowto.com/effective-sample-size/

[45] "Mean difference / difference in means (md)." [Online]. Available: http://www.statisticshowto.com/mean-difference/

[46] Pearson correlation coefficient, "Pearson correlation coefficient — Wikipedia, the free encyclopedia," 2010, [Online; accessed 15 July 2018]. [Online]. Available: https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

[47] J. E. Wong, W. R. Parnell, K. E. Black, and P. M. Skidmore, "Reliability and relative validity of a food frequency questionnaire to assess food group intakes in new zealand adolescents," *Nutrition journal*, vol. 11, no. 1, p. 65, 2012.

[48] M. Watanabe, K. Yamaoka, M. Yokotsuka, M. Adachi, and T. Tango, "Validity and reproducibility of the ffq (ffqw82) for dietary assessment in female adolescents," *Public health nutrition*, vol. 14, no. 2, pp. 297–305, 2011.

[49] Pedometer, "Pedometer — Wikipedia, the free encyclopedia," 2010, [Online; accessed 15 July 2018]. [Online]. Available: https://en.wikipedia.org/wiki/Pedometer#Nokia_Step_Counter

[50] L. E. Voorrips, A. C. Ravelli, C. Petra, A. Dongelmans, P. Deurenberg, and W. A. van Staveren, "A physical activity questionnaire for the elderly." *Diet and physical activity as determinants of nutritional status in elderly women*, p. 43, 1991.