

STATISTICAL ANALYSIS OF DIET, EXERCISE AND FITNESS VIA PEOPLE'S OWN VISUALISATION OF PARAMETERS

A project report submitted for the partial fulfilment of Semester-I,
Programing for Data Science – Big Data Analytics



Department of Computer Science
Ramakrishna Mission Vivekananda Educational
and Research Institute

Submitted By-

Saikat Patra

Ujjwal Chowdhury

Krishnakanta Maity



Student's Declaration

We, hereby declare that, the project work entitled ***“Statistical Analysis of Diet, Exercise and Fitness via people's own visualisation of parameters”*** is a record of an original work done by us under the guidance of Dr. Sudeep Mallick and this project work is submitted in the partial fulfilment of the requirements for the paper Programming for Data Science of Big Data Analytics.

We have collected the data from primary sources as per the requirement of my dissertation, which are appropriately referred in the report. All the computations involved in this dissertation are result of our own calculations on the data that we collected.

The formulae that are used in the dissertation are acknowledged providing appropriate reference of the source from which they are obtained. No part of the dissertation is been submitted to any other institution for the purpose of any degree/diploma.

Date: December 1, 2021

**Saikat Patra
Ujjwal Chowdhury
Krishnakanta Maity**

Acknowledgement

It is great pleasure for us to undertake this project. I am grateful to my project guide Dr. Sudeep Mallick.

This project would not have completed without his enormous help and worthy experience. Whenever we were in need, he was there behind us.

Although, this project report has been prepared with utmost care and deep routed interest I accept responsibility for any imperfection.

Table of Content

	Page No
Abstract	1
Introduction	2
Objectives	3
Data Set	4
Problem Statement	4
Questionnaire Design	5
Data Collection Methodology	6
Data Processing	7
Exploratory Data Analysis	8-14
Limitations	15
Future Scope	16
Reference	17
Appendix	18

Abstract

Choices about nutrition and exercise are very much linked with individual's health. Decisions about food intake also have great impact on an individual. Diet is the sum of food consumed by a person or other organism.

The word diet often implies the use of specific intake of nutrition for health or weight-management reasons. Although humans are omnivores, each person holds some food preferences due to personal tastes. Individual dietary choices may be more or less healthy. A healthy diet can improve and maintain optimal health. Exercise involves engaging in physical activity and increasing the heart rate beyond resting levels. It is an important part of preserving physical and mental health. Whether people engage in light exercise, such as going for a walk, or high intensity activities, for example, weight training. Regular exercise provides a huge range of benefits for the body and mind.

Food preferences are the evaluative attitudes that people express toward foods. Food preferences include the qualitative evaluation of foods, and also how much people like and dislike them.

Height and weight measurements are used to calculate measure of healthy versus unhealthy weights. Food preference and age can also contribute for health differences.

Introduction

In our day to day life, especially amidst the Covid situation, our diet and our exercise effect our fitness to a very large level which is very important to live a safe & healthy life. Now our main objective behind this study is to analysis how much our diet and exercise effects our fitness and write a report on it.

Objectives

The primary goal of our project is to understand the insight of food consumption, exercise and fitness. The main purpose of EDA is to look at data before making any assumptions. It can help identify obvious errors, as well as better understand patterns within the data, detect outliers or anomalous events, find interesting relations among the variables.

Our approach to measure food habits, fast food composition, physical activity, has greatly motivated our project for understanding of the key determinants and the health problems we currently facing.

Data Set

- **Data Source:** All data are collected through google [forms](#).
- **About Data:** This dataset contains 21 variables with timestamps. Description of variables are given below.

Data Description

Our data consists of 259 observations with the information of the following variables:

- ts: time of the responses
- age: age of the respondent
- sex: gender of the respondent
- work: work preference of the respondent
- phy_ff: rating of liking fast food (in 10 scale)
- phy_health: rating of healthiness (in 10 scale)
- phy_bw: rating of one's preference to maintain body weight
- phy_ex: rating of importance of exercise
- meal: number of meals in a day
- weight: weight of the respondent
- exercise: type of exercises
- fruit: number of meals contain fruits
- veg: number of meals contain vegetables
- cook: number of cooked plates
- spend: expenditure on fitness
- income: monthly family income
- gymtime: time spent in gym
- disease: whether suffers from any regular disease
- review: review of the survey
- rate: rate the project topics and question

Data Collection Methodology

We collected the primary data for our project using the self-assignment survey method using google forms. Our targeted audience are mostly UG and PG students.

Our questioner is created considering the dietary recommendations of individuals and their behaviour about exercise and fitness. We connected our targeted audience using Social Media and Instant messaging apps. The time frame for which we collect our data is 3 weeks.

Data Processing

Using tidyverse package we clean some outlier data and not complete cases from the data.

Exploratory Data Analysis



Figure 1: Age distribution of respondents

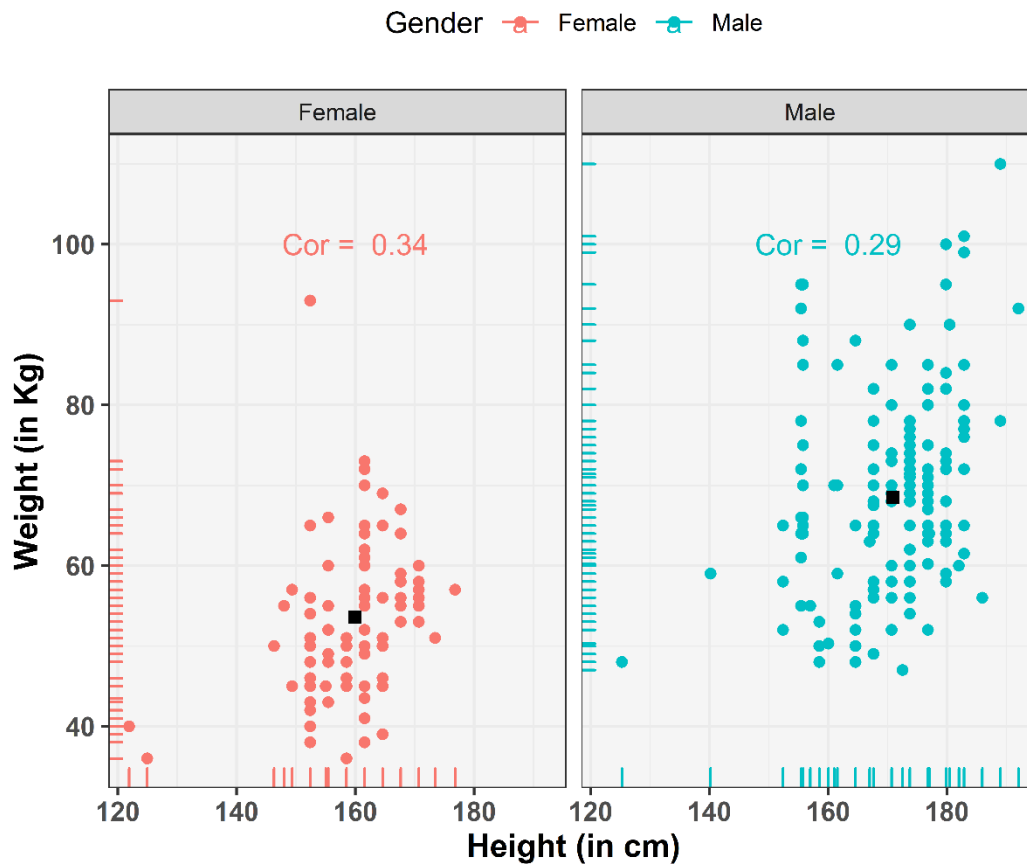
The above diagram shows nothing but the ages of our respondents. Now we are categorising the respondents according to their genders.

As we can see from the diagram all of our respondents are either male or female. Ages of the female respondents are marked red and the male ones are marked green.

Some other observations are:

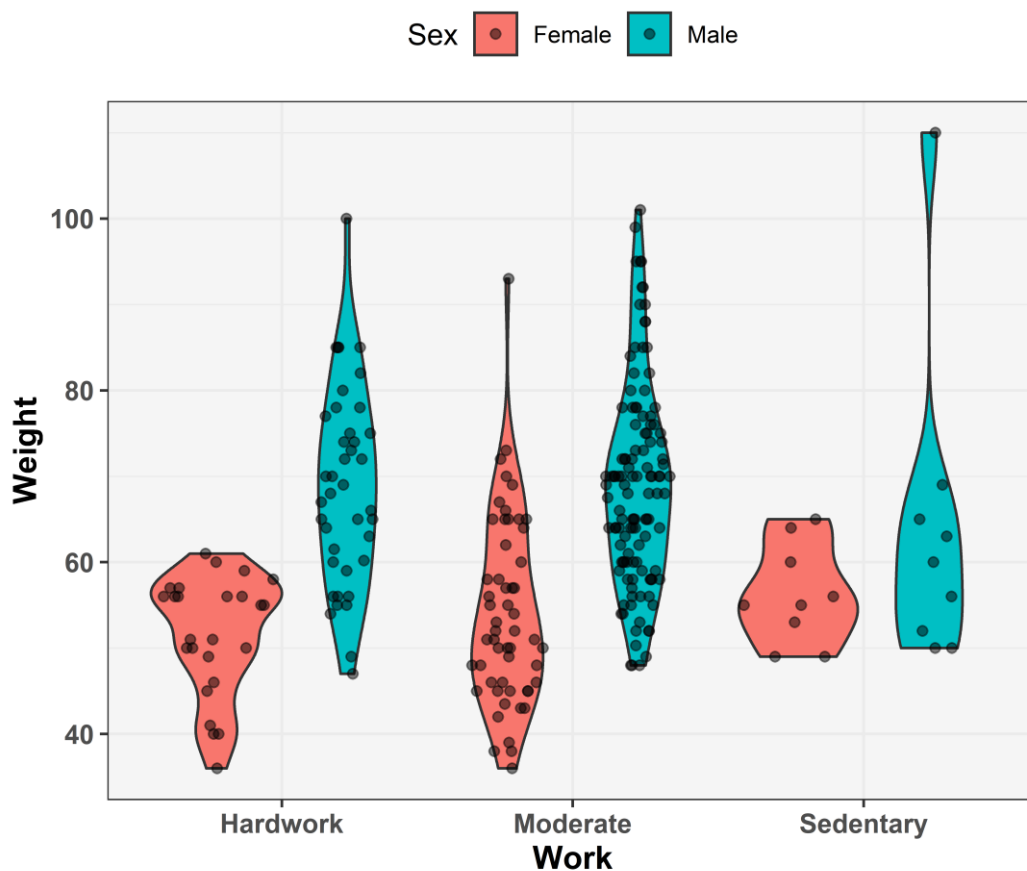
- There are total 3 outliers in the dataset, 2 of them are male and the other one is a female.
- Most of the data points lie in between 15 to 25 that means most of the respondents are aged in between 15 to 25.
- 35% of the total responses are from female and 65% are from male that means male respondents are almost double of the female respondents in this dataset.

Scatter Plot of Height and Weight



- ✚ It is clear that weight of female respondents are mostly in between 40kg to 70kg and height in between 145 cm to 175cm ; for males it is mostly in between 46 kg to 100kg and 155cm to 185cm respectively.
- ✚ Surprisingly, weights of males are scattered are more uniformly than females.
- ✚ Overall male respondents has more height and weight than females.
- ✚ From the correlations values, female's height and weights are highly correlated than males'.
- ✚ From the above graph it is evident that the correlation between height and weight is significantly higher for females as compared to their male counterparts. It is 0.34 for females as compared to 0.29 for males. It is supported by the popular belief that females tend to take care of their body more sincerely as opposed to their male counterparts.

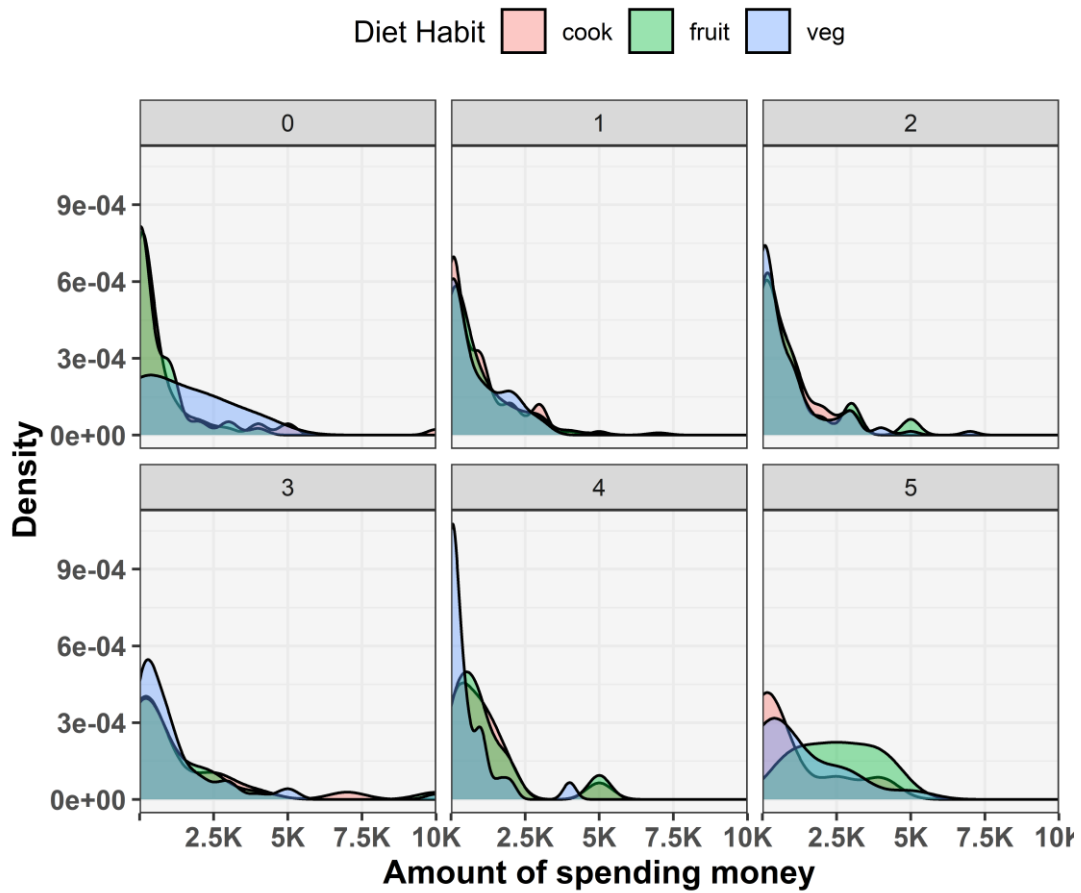
Gender-wise Work Preferences vs Weight



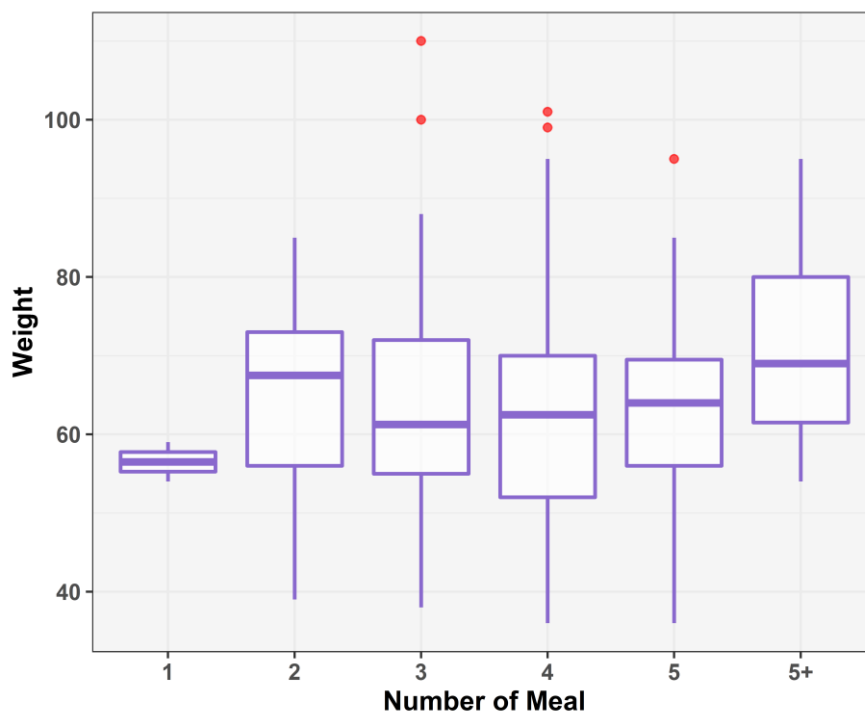
The above diagram is a violin plot of "Gender-wise work preference vs weight of the respondents". Violin plots are quite similar to the Box-plot with the addition of rotated density-plot on each side. Here the red area denotes the female responses and the blue area denotes the male responses. The other observations are:

- Most of the hardworking females are weighted between 40 to 60 kg whereas most of the sedentary females are weighted between 50 to 65 kg. But as we can see moderate working females' weights are distributed fairly in between 35 to 80 kg with a bit more density around 45 to 55 kg.
- Similarly if we try to interpret the male respondents' weights according to their work preference, most of the hardworking males' weights are almost fairly distributed between 50 to 100 kg with a bit more density around 60 to 80 kg. Moderate working males are weighted in between 50 to 100 kg in which case their weights are a bit more dense around 60 to 75 kg. Now we have an outlier in our dataset of male responses who is a sedentary person. Other than him, the rest of the sedentary males are weighted in between 55 to 80 kg.
- Now the graph tells us that most of the respondents in our dataset are moderate-workers.

Density Plot of Diet Habits

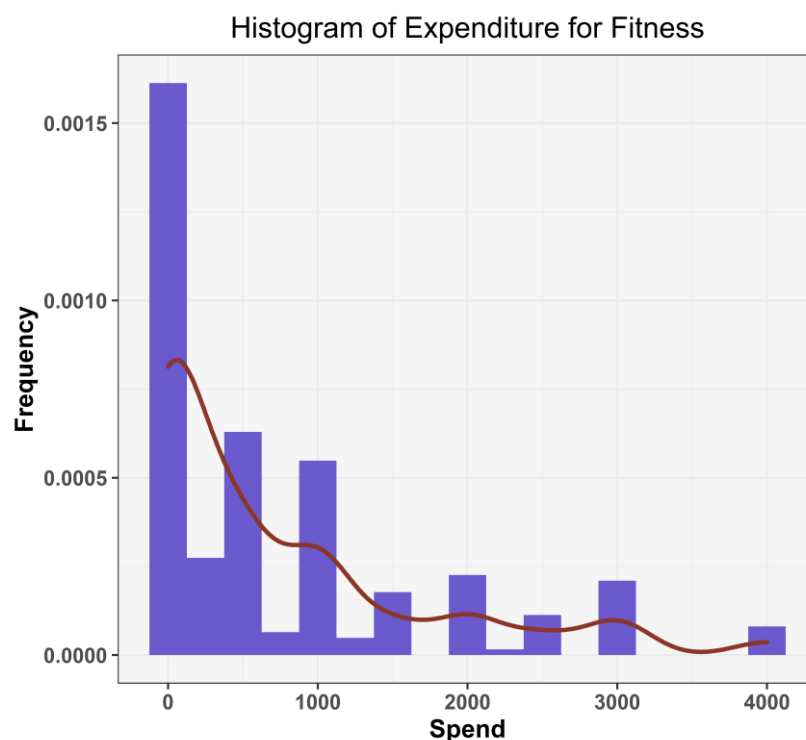


Boxplot of Weights (By number of meal)



The above diagram is the Box-plot of the numbers of meals taken by the respondents in a day vs their weights. The red dots are some outliers in the dataset. Now some other observations regarding the boxplots are:

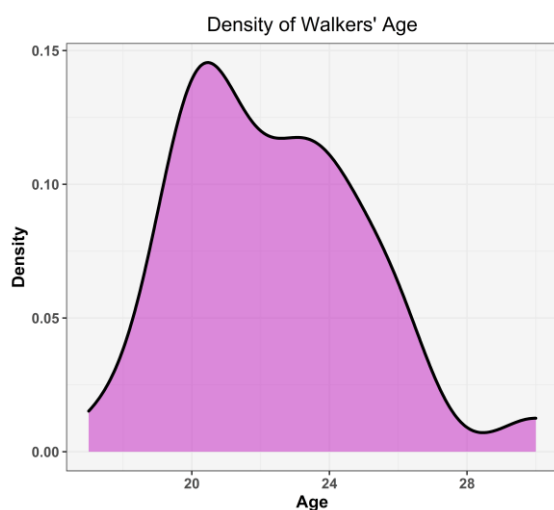
- ✚ The respondents who take 1 meal per day are weighted in between 57 to 59 kg with median 58 kg.
- ✚ Those who take 2 meals per day are weighted in between 58 to 73 kg with median around 68 kg, which means the weights are denser in between 68 to 73 kg.
- ✚ Those who take 3 meals per day are weighted almost similar to the persons who take 2 meals but their weights are denser in between 57 to 61 kg.
- ✚ Those who take 4 meals per day are weighted in between 52 to 70 kg with median at 63 kg and also the weights are distributed in that range quite fairly.
- ✚ Those respondents' weights who take 5 meals per day are in between 56 to 70 kg and the median weight is 64 kg. Those who take more than 5 meals per day are weighted in between 62 to 80 kg and the median weight is 69 kg.
- ✚ As we can see most of the respondents who consumes more than 5 meals per day are heavier than others which is quite natural. The other 4 categories (2, 3, 4, 5 no. of meals takers) are weighted quite similar. Two of the five outliers who take 3 meals per day are weighted 100 kg and 110 kg, two of them who take 4 meals are weighted around 100 kg and the other one takes 5 meals per day is weighted 95 kg.



The above diagram shows that most of the respondents either do not pay any money or pay less than 1000Rs. on fitness. The density curve decreases as the value of money spent on fitness increases that means no. of person spending more and more on their fitness decreases gradually.

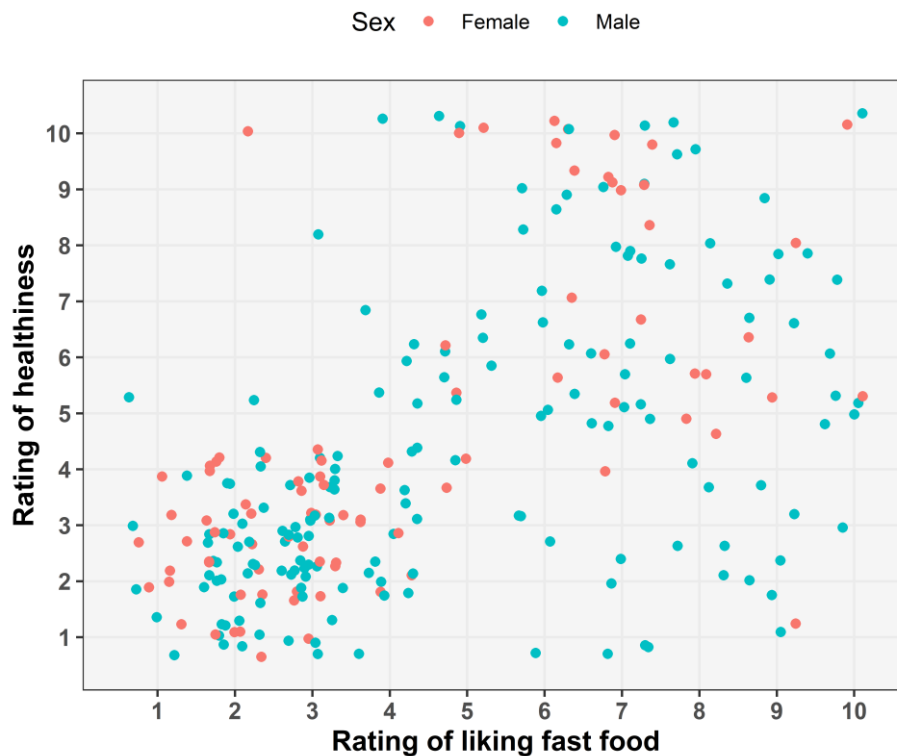


From the above graph, we can infer that 47% of the people have preferred walking as the primary mode of exercise while the second most preferred form of exercise is not known as 20% of the people have said others.

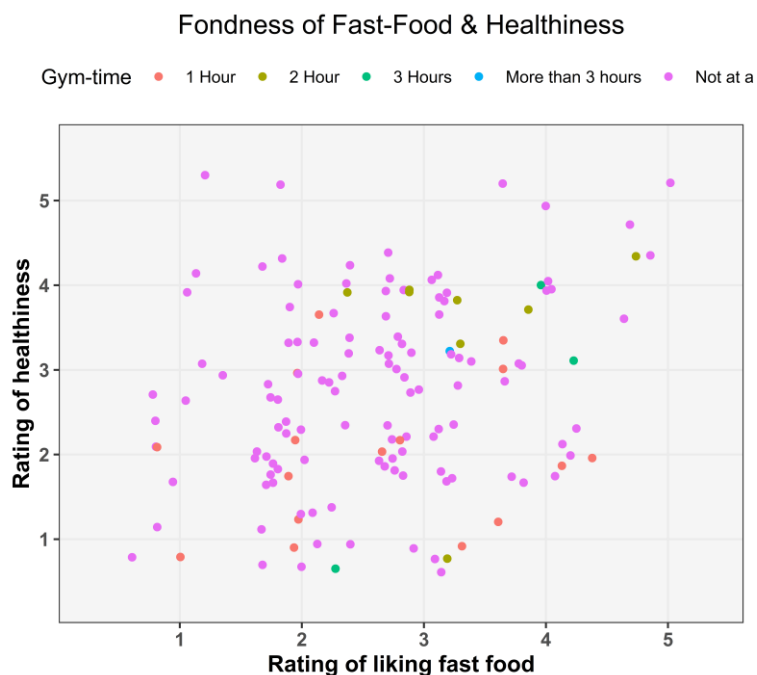


The age distribution of the respondents who have preferred walking as their primary mode of exercise have been given below. As opposed to the popular belief that young adults are more inclined towards intense exercises such as running or weight lifting, the respondents of walking have the age mostly in the range of 18-22!

Fondness of Fast-Food & Healthiness

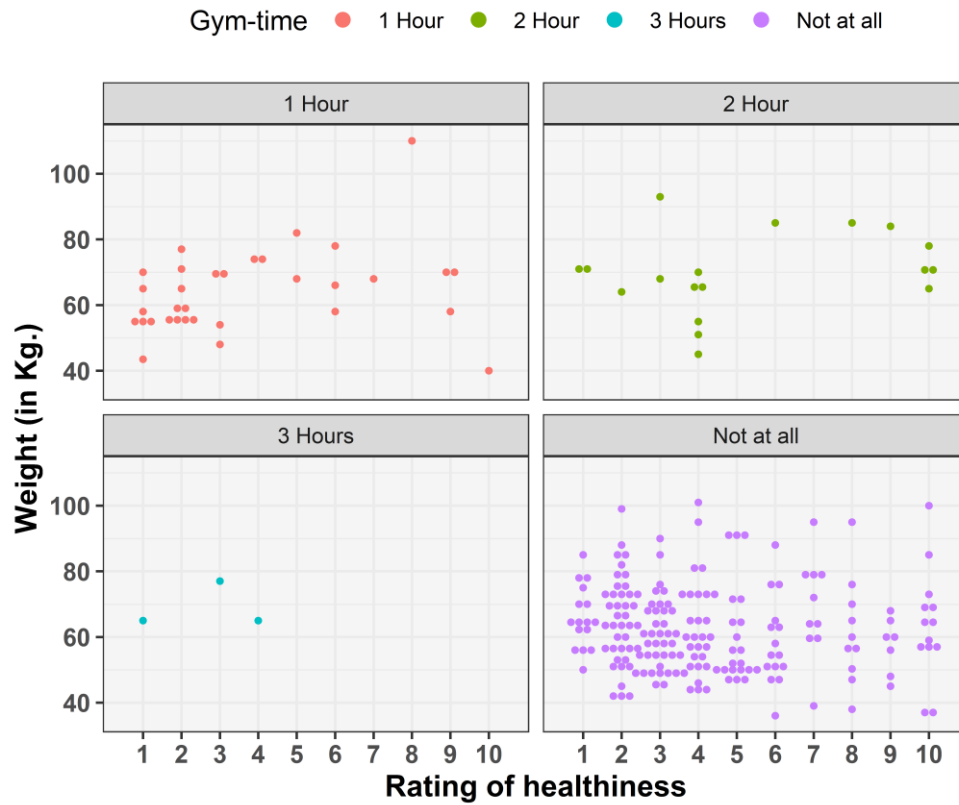


The above graph has suggested us with the finding that there is group cluster of males and females who have not rated their physical health very high despite not having high likability for fast food. The reason for this can be well understood from the graph below where we have linked this cluster's health parameter with the help of time spent in gym by these people.

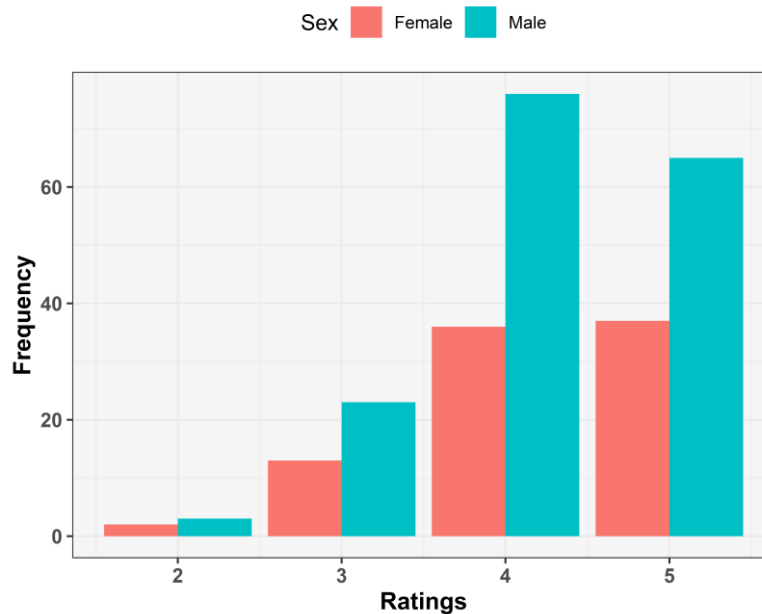


When we look closely at the dataset of the above cluster with respect to gym time we have found the rationale behind the unusual low rating of health despite not giving high rating for fast food preference and that is, most of the people making up the that cluster have reported that time spent in gym -"not at all". This might be a probable reason for this cluster behaviour.

Weight & Healthiness According to Gym-time



Bar Chart of Ratings of this study
(Grouped by gender)



As per the above graph, we can see that we have received more responses from the males for the positive ratings as compared to the female counterparts. Also, out of the total responses x % have given us a rating above 4.

Limitations

Our data set involves a smaller sample, hence the results cannot be accurately interpreted for a generalized population. Most of the inferences are focus on quantitative data that leads to some outliers. Our targeted audiences are mostly around the age 23, so there are some judgmental bias.

Not giving respondents enough options to respond properly as per their choice of others. We should have given a textual input to users so that they could have mentioned if their preferred options was not one of the choices provided by us.

Future Scope

With an ample amount of time and data-set we can create a model which give a clear picture of people's mind set about diet and fitness. With our analysis, people's fondness for fast food consumption also can be verified.

The changing health scenario globally has increased the challenges for public nutrition and people's behaviour about fitness. With the help of our project we can tackle some of those problems.


References

- [1] <https://r4ds.had.co.nz/>
- [2] Grolemund, G., & Wickham, H. (2017). *R for Data Science*. O'Reilly Media.
- [3] McKinney, W. (2017). *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython*.

Appendix

Google form was created for data collection. Snapshot of questionnaire is attached below:

Questions Responses 213 Settings



Diet **Exercise** **Fitness**

DEF Survey

We are students of Ramkrishna Mission Vivekananda Educational and Research Institute (Belur) pursuing M.Sc. in Big Data Analysis. This survey is based on people's habits on Diet and Exercise. Fitness developed by diet and exercise. We also collect some information on fitness.

Disclaimer: These data will not be used for commercial purpose. Also this survey is not collecting your any personal information.

What is your age? *

What is your work preferences? *

- ☐ Sedentary
- ☐ Moderate
- ☐ Hardwork

Physical condition *

This is like ratings. Give your preference out of 10.

	1	2	3	4	5	6	7	8	9	10
How p...	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
How m...	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
How m...	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
How i...	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

R-programming is used for exploratory data analysis. All codes are given below: (Codes can be found [here](#))