Eating Habits of People in America

Abhishek Gupta (guptaa10@rpi.edu)

Rensselaer Polytechnic Institute, Tetherless World Constellation, Troy, NY, United States,

Abstract

The traditional eating patterns that many in the United States are actually following do not agree with the Dietary Guidelines. A comparison is drawn in Figure1. About three-fourths of the population has a low vegetable, fruit, dairy and oils eating pattern. More than half of the population meets or exceeds the guidelines for total grain and protein foods, but within each of these food groups they do not follow the recommendations for the subgroups. Most Americans go beyond the recommendations for added sugars, saturated fats, and sodium.

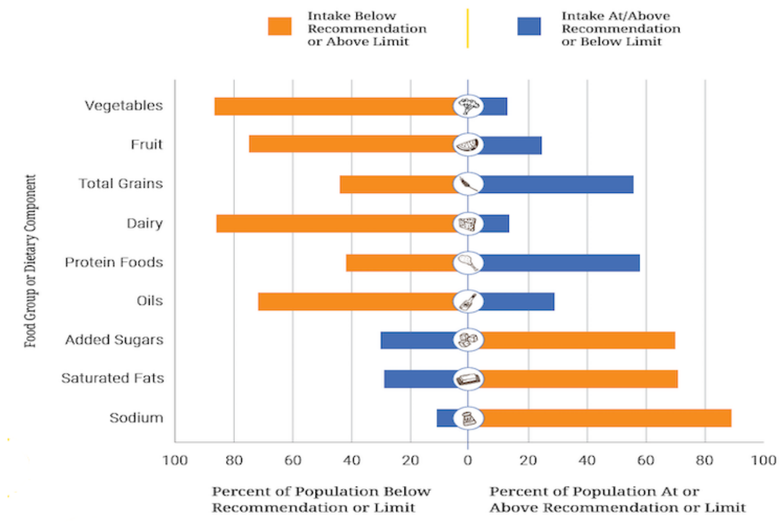

Figure1: Dietary Intakes Compared to Recommendations

Furthermore, many eating patterns are too high in calories. Compared to calorie needs, calorie intake over time is best assessed by measuring body weight status. The high percentage of the overweight or obese population means that many in the U.S. are over-consuming calories.

Motivation

Obesity in the United States (US) has become a serious health problem: nearly 35% of Americans have obesity. Obesity is not just a problem of "girth control"; it is now considered a chronic disease by the American Medical Association.

The aim of my project is to analyse the diets of people in America and compare them with those of the obese people. I will use amount of calories as a basis of comparison. I will also analyse the relation of calorie value with other nutritive values of diets of people in America.

Data Description

My analysis consist of two datasets, the first dataset contains the commonly eaten food items by people in America. The second dataset is a weekly analysis of eating habits of Obese people.

The first dataset consists of commonly eaten food items with attributes such as solid fats, added sugars, calories, saturated fats, oils, alcohol content, meats, etc.

	Milk	Meats	Soy	Drybeans_Peas	Oils	Solid_Fats	Added_Sugars	Alcohol	Calories	Saturated_Fats
0	0	0	0	0	0	105.6485	1.57001	0	133.65	7.36898
0	0.29393	0	0	0	0	130.99968	95.20488	0	267.33	9.0307
0	0.2516	0.0962	0	0	0	213.06672	96.1034	0	368.52	15.2884
0	0.38233	0	0	0	0	170.39808	123.83793	0	347.73	11.7467

The second dataset consist of categories of food eaten by Obese people on a weekly basis. Some of the food categories of this dataset include meat/fish, no color vegetables, fruits, grains, sweets etc.

food_type	times_per_week	number_of_males	number_of_females	Male_percentage	Female_percentage	Total_percentage	Total_number
Meat/fish	1	8	16	17.8	24.2	21.6	24
Meat/fish	4	25	36	55.6	54.5	55	61
Meat/fish	12	10	11	22.2	16.7	18.9	21
Meat/fish	21	2	3	4.4	4.5	4.5	5
beans/tofu	1	13	10	28.9	15.2	13.5	23

Sponsors:

Exploratory Analysis

1. Import data:

R language (R)

A. Food Items dataset

```
Food_table<-read.csv(file.choose(),header = TRUE)
Food_table<-data.frame(food_table)
```

B. Obesity dataset

```
ObeseData<-read.csv(file.choose(),header = TRUE)
ObeseData<-data.frame(ObeseData)
```

2. Inspect data:

Several methods are used to inspect the dataset.

A. Get an overview of the dataframe (df):

```
head(Food_table)
tail(Food_table)
> summary(Food_table)
```

```
  Food_Code                    food_type                          Display_Name
Min.   :    7258   Snacks             :606   Cheese pizza, thick crust       :   7
1st Qu.:27214100   Meat/fish          :496   French fries, deep-fried        :   7
Median :54403090   Sweets             :220   Raw tomatoes                    :   7
Mean   :52961704   milk/milk products:190   Chocolate-covered candy         :   6
3rd Qu.:72901282   Fruits             :148   Dietetic chocolate-covered candy:   6
Max.   :94210100   Grains             : 84   Ground beef (75% lean, regular) :   6
                   (Other)            :270   (Other)                         :1975
```

```
> dim(food_table)
[1] 2014   27
```

B. Explore the dataset:

i. Histogram plot

```
>ggplot(food_table+geom_bar(mapping=aes(x=Food_table$food_type), fill = "magenta")
```

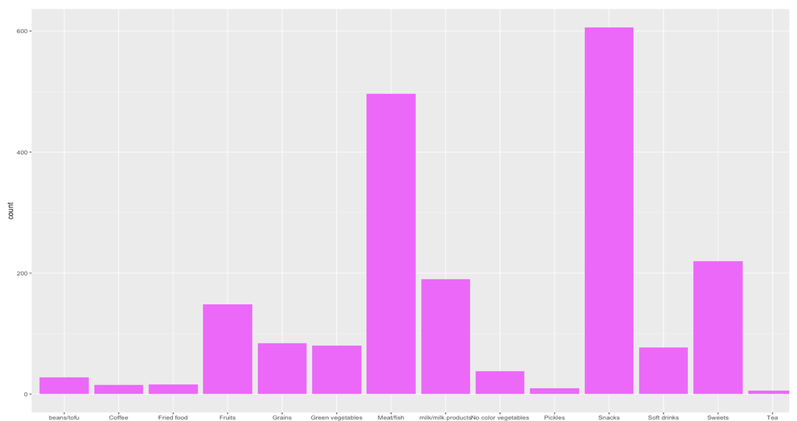

Figure2: Frequency of various categories of food items eaten

ii. Box plot

```
>boxplot(Food_table$Saturated_Fats,Food_table$Alcohol,Food_table$Solid_Fats,Food_table$Added_Sugars,Food_table$Calories, col = "Green")
```

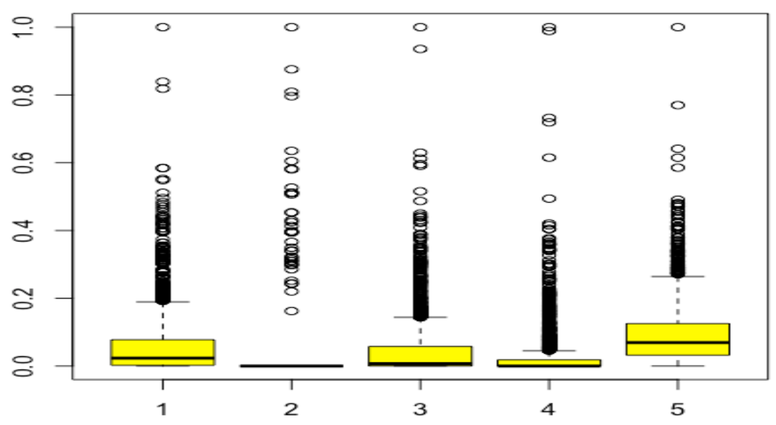

Figure3:Distribution of different attributes in the Food_table dataset

iii. Heatmap

```
> heatmap(data, col = terrain.colors(256))
```

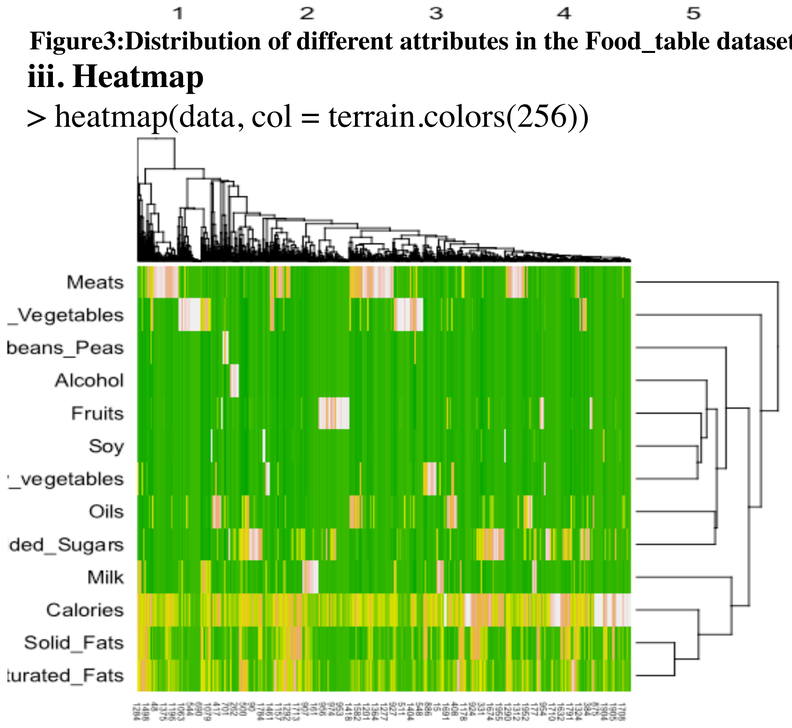

Figure4: Heatmap for different variables in the Food_table dateset

Data Preprocessing

3. Data Manipulation:

The dataset of food items were not categorised according to the food categories, so a column of food type was added to the dataset to provide a basis of comparison between the two datasets.

4. Data Cleaning:

The dataset of food items contained some missing values and NA's which were removed from the dataset.

```
> df.dropna()
> df.dropna(subset = ["colname"])
```

The data was normalised before performing computations so as to make different attributes lie in the same range.

```
> nor <-function(x) {(x-min(x))/(max(x)-min(x))}
```

Results

The Analytics were performed on the Obese people eating habits dataset and the food items dataset. The results gave the following values for weekly consumption of calories by Obese people from different food categories.

```
> avg_calories_meat_fish = sum(meat_fish$Calories)/nrow(meat_fish)
> average_times_meat_fish_intake = sum(average_foodtype_intake[0:4])/111
> weekly_calories_meat_fish =
avg_calories_meat_fish*average_times_meat_fish_intake
[1] 1198.741
```

Food Category	Average no. of times eaten per week	Total calorie intake per week
Meat/ fish	5.630631	1198.741
Beans/tofu	4.54955	685.8777
Milk	4.225225	529.3827
Green Vegetables	8.621622	516.1476
Colorless Vegetables	13.32432	856.7751
Fruits	9.459459	669.7129
Grains	13.61261	1490.579
Sweets	8.905797	1213.377
Coffee	5.521739	428.2624
Tea	3.985507	156.2385
Snacks	5.362319	913.5161
Soft Drinks	4.101449	649.021
Pickle	4.231884	90.22483
Fried food	5.956522	1252.437

Figure5: Calorie intake by Obese people

Model Implementation

```
set.seed(100)  # setting seed to reproduce results of random sampling
trainingRowIndex <- sample(1:nrow(food_table_norm),
0.8*nrow(food_table_norm))  # row indices for training data
trainingData <- food_table_norm[trainingRowIndex, ]  # model training data
testData  <- food_table_norm[-trainingRowIndex, ]
```

1. Linear Regression:

```
lmMod <- lm(food_table_norm$Calories ~ food_table_norm$Saturated_Fats
+ food_table_norm$Solid_Fats , data=trainingData)
distPred <- predict(lmMod, testData)
```

2. KKNN:

```
##extract training set
food_table_train <- food_table_norm[ran,]
##extract testing set
food_table_test <- food_table_norm[-ran,]
##extract 12th column of train dataset because it will be used as 'cl' argument
in knn function.
food_table_target_category <- food_table[ran,12]
##extract 12th column if test dataset to measure the accuracy
food_table_test_category <- food_table[-ran,12]
##load the package class
library(class)
##run knn function
pr <-
knn(food_table_train,food_table_test,cl=food_table_target_category,k=10)
##create confusion matrix
tab <- table(pr,food_table_test_category)
accuracy <- function(x){sum(diag(x)/(sum(rowSums(x)))) * 100}
accuracy(tab)
```

The calorie estimate is one of the most important indicators of a person's diet. I have made the calorie estimate of food items based on the other nutritive values of food items.

Glossary:

Python – A programming language, capable of processing data/statistical analysis
R – A program to process data and perform statistical analysis
Package (P) or Library (R): software package to be loaded to perform extra tasks
Pandas – An useful data manipulation package in python
Df, dataframe – Data manipulation structure in R & python pandas

Resources:

Obesity Data Survey: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4877769/
Factors leading to Obesity: https://www.hsph.harvard.edu/obesity-prevention-source/diet-lifestyle-to-prevent-obesity/
Stats on Obesity: https://health.gov/dietaryguidelines/2015/guidelines/appendix-2/
Read excel files in R: https://www.datacamp.com/community/tutorials/r-tutorial-read-excel-into-r
R deal with missing data: https://www.statmethods.net/input/missingdata.html
R visualization: https://www.analyticsvidhya.com/blog/2015/07/guide-data-visualization-r/