Notes for Students

Malnutrition can result in obesity, which has been rising at an alarming rate. In the US for instance, while all states in 1990 had less than 14% obesity, figures started increasing. And by 2000, half of the country has more than 20% of its population obese. The trend continues. And in 2010, all states have at least more than 20% of their population obese. Many states across the country reached an alarming situation. More than 35% of American adults are obese. The trends Worldwide are no different. Obesity has nearly doubled across the globe. Obesity is one of today's blatantly visible public health problems, and increases people's risk to heart disease, stroke, and diabetes. In fact 65% of the world's population lives in countries where obesity kills more people than underweight. So good nutrition is essential for an overall healthy lifestyle and promoting it now is more important than ever. We have access to hundreds of nutrition and weight loss applications, and around 15% of adults with cell phones use health applications on their devices. These apps are mostly powered by the United States Department of Agricultural, or USDA, food database. The United States Department of Agricultural distributes nutritional information of over 7,000 food items including amount of calories, carbs, protein, fat, and sodium, among other nutrients. It is exactly this data that we will be analyzing in this exercise.

Step 1:- We'll start by reading in our dataset USDA.csv, which contains all foods in the USDA database in 100-gram amounts. You can find the file on https://storage.googleapis.com/dimensionless/Analytics/USDA.csv

Step 2:- Save the output to a data frame, and call it "USDA".

Step 3:- Read the structure of the USDA data frame and find out the number of observations and variables.

Step 4:- Obtain high-level statistical information about your dataset.
(Mean, Median, SD, Q1, Q3 etc)

Step5:- We realized that the maximum level of Sodium was 38,758 milligrams, which is very high. Find out which food this corresponds to?

Step 6:- Find out which foods contain more than 10,000 milligrams of sodium. To do so, you can create a new data frame, "HighSodium" and it will be a subset of our original dataframe "USDA".

Step 7:- How many foods exist in this new data frame? What are they, is "caviar" one of them ?

Step 8:- I thought "caviar" is well known to be among the top 10 foods with highest levels of sodium. But it doesn't appear in the above list. So find how much sodium it has in 100 grams.

Step 9:- Now, the value 1,500 milligrams seems to be very small compared to 10,000 milligrams or 38,000 milligrams, which are the values that we worked with so far with respect to sodium levels. But this doesn't seem to be a fair comparison. Maybe the best way to figure out how big this value is, is by comparing it to the mean and the standard deviation of the sodium levels across the data set. So find out the mean and Standard Deviation.

Step 10:- Well we got NA because we forgot to remove the non-available entries before computing our statistical measure. So remove the NA values and find SD again

.

Step 11:- Find z-score for sodium content in "Caviar" ?

Plotting the data

Visualization is a crucial step for initial data exploration. It helps us discern relationships, patterns, and outliers.

Step 1:- Create a scatter plot with Protein on the x-axis and Fat on the y-axis. And label the x-axis as "Protein", y-axis as "Fat" , and title the plot as " Protein Vs. Fat"

Step 2:- Create a histogram of Vitamin C content. Label the x-axis as "Vitamin C" and give it a title, "Histogram of Vitamin C Levels".

Step 3:- Create a box plot for Sugar with title "Box Plot of Sugar Levels"

Adding new variables

Step 1:- Suppose that we want to add a variable to our USDA data frame that takes a value 1 if the food has higher sodium than average, and 0 if the food has lower sodium than average. Create such vector of name “HighSodium” and attach it with USDA data frame.

Step 2:- Now repeat the same, and add the variables HighProtein, HighCarbs, HighFat, similarly to our data frame.

Step 3:- How many foods have higher sodium level than average?

Step 4:- How many foods have both high sodium and high fat level than average?

Step 5 :- Compute the average amount of iron sorted by high and low protein?
(Hint:-The tapply function takes three arguments, and groups the first argument according to the second argument, and then applies argument three.)

Step 6:-Find the maximum level of vitamin C in foods with high and low carbs?

Step 7:- Is it true that foods that are high in carbs have generally high vitamin C content?