Project description

The project I work on is to the dietary behavior of human beings. A saying, “you really are what you eat”, emphasizes how important the food one takes in every day shapes a human. We need energy along with nutrition to support the function of your cell, your muscles, your bones and your brain. Hair loss, lack of energy, depression, and not being able to focus are common symptoms of unbalanced dietary. In addition, the deficiency of certain nutrients including some micronutrition could lead to fatal diseases. Comparing to the efforts on curing diseases, no matter who we are looking at, either we as individuals, or government and healthcare providers, if the patterns of food intake could be identified such that the nutrient deficiency could be predicted in certain groups, certain diseases could be prevented at the first place.

However, even a lot of us are aware of the daily nutrient intake level, being able to eat accordingly is almost unrealistic. It is a tedious job to keep track of the food consumption and compare with the suggested amount. Moreover, eating pattern could be more than just a personal choice; geographic location, religion, culture, weather, family income, and where the meals take place might limit the choice of food, which will also affect the amount of nutrients intake. For example, in the northeastern region of the nation, the so-developed corn industry makes the corn the most accessible choice; cold wave massively reduces the production of agricultural products, which could lead to a more meat-based dietary.

The datasets I use are National Health and Nutrition Examination Survey (NHNES) from Centers for Disease Control and Prevention (CDC). I specifically used the dataset of the one-day interview, which contains the information of what the interviewees ate in the past 24 hours. The datasets have more than 80 columns including what kind of food, which meal, where the interviewees ate, home or restaurants, the amount of the common nutrients’ intake, and so on. I have collectively examined the dietary over three continuous years and identify the change of food intake and the corresponding nutrients level over time.

The individuals considered in the survey are humans from 0 to 150. I first filtered the data by excluding the breast-fed only infants since infants cannot choose the food source, which in my case could be considered as an outliner. There are several correlations that I would like study within the datasets. First is the effect of culture. The datasets contain a column called DR1LANG, which is the information of language respondent used mostly. I calculated the correlation matrix and plotted the heatmap in order to find whether this term is correlated to any nutrition. The plot suggests that the language is not correlated to the nutrition takes in; on the other hand, the correlation amongst different nutrients are quite large, which comes from the fact that when we take in food, there is almost no such food that contains only one nutrient; in most cases, we are taking in multiple nutrients simultaneously. In addition, from the heatmap, I can see that protein and magnesium are correlated to each other. This provides useful information since we are more aware of what food contains protein but not magnesium; for the group of people who have muscle twitches and cramps, a common symptom of magnesium deficiency, a good diet suggestion would be to consume more meat or eggs.

To get the idea of whether the nutrients taken in daily is sufficient or not, I compare the average values with the reference. The reference data of the suggested daily nutrients intake is obtained from this website: <https://health.gov/dietaryguidelines/2015/guidelines/appendix-7/>. However, the data reported here are categorized by gender and age, which are not provided in the NHNES datasets. I took the average of female and male with age ranging from 14 to 51+. I focused on the values of DR1IPROT (protein), DR1ICARB (carbohydrates), DR1IFIBE (fiber), DR1ICALC (calcium), DR1IMAGN (magnesium), DR1IIRON (iron), DR1IVARA (Vitamin A), DR1IVD (Vitamin D), and DR1IVC (Vitamin C). I calculated the average of those columns for the three continuous years grouped by the meal (breakfast, lunch, and dinner) and divide by the reference values to get the fraction. Then I plotted the stacked bar chart. According to the chart, the conclusion I get is that for majority of the interviewee did not take in sufficient amount of nutrients. The percentage distribution among three meals are close to the suggested 30%, 40%, 30% ratio of breakfast, lunch, and dinner. Among all the nutrients I have investigated, Vitamin D deficiency is the most severe. However, Vitamin D could be produced from cholesterol when skin is exposed to the sun, yet the calculation of that amount is not direct. This means the actual Vitamin D intake should be larger than the value suggested but the number is hard to predict.

There are several insights I gained from the analyses. First is in general, the nutrients we obtained from three meals do not reach the requirement, which means taking dietary supplementary is suggested regardless. In addition, there is no obvious growth in the nutrition level for the past three years; the deficient intake of macronutrients including fiber and other minerals and vitamins might cause health issue. The future works could lie in the following directions. First, the datasets do not have the information of age and gender; yet the two parameters affect the amount of daily nutrients required. A dataset containing above information will benefit the prediction of the models and help explain the trends in the data. In addition, the variability in each nutrient especially macronutrients are huge; our common sense tells the protein required for a sports player per day will be much larger than that for a white-collar worker. If the data could be grouped in a way such that the analysis could be done separately on people with various protein intake level for example, then the characteristics of occupations could be seen, and more guideline could be provided based on the differences in each group. In addition, for Vitamin D calculation, if the amount of time exposed to the sun every day could be provided, along with the relationship between the cholesterol conversion to Vitamin D and the exposed time, then the evaluation of Vitamin D level could be revised.