Week 9: Deliverables

Group Name: Fight on Healthy diet
Name: Sijing Liu
Email: sijingli@usc.edu
Batch code: LISUM13: 30
Country: U.S.
College: University of Southern California
Specialization: Data Science

Problem description:

Does Healthy Diet Help Prevent COVID-19?

On March 11, 2020, the World Health Organization declared COVID-19 a global pandemic. Since then, the worldwide recorded death rate as a result of the illness has surpassed five million (Roberts, 2021). Research shows that the epidemic growth rate for disease spread depends on many factors, including biological, demographic, and social factors. However, dietary risks during the pandemic are void of investigation, given the acknowledged impact of food on health outcomes.

According to the World Health Organization (WHO), eating a healthy diet is very important during the COVID-19 pandemic (WHO, 2021). Now more than ever, we need to prioritize what we put into our bodies to reduce the susceptibility to and long-term implications from the illness. The relationship between dietary habits and diseases has been extensively investigated. However, most of the associations focus on chronic non-communicable diseases (Afshin et al., 2019). Therefore, through this project, I aim to fill this void to make clear the relationship between dietary habits with communicable disease, like COVID-19.

Overall, this project will look into a dataset that measures the nutrition of several food groups, a variety of eating styles, obesity and undernourishment rates, and data on COVID-19 cases from 170 countries. I hope to conduct exploratory data analysis (using descriptive statistics), machine learning (mainly association analysis and prediction), and data visualization to learn more about how diet ultimately influences the contraction and survivability rates of COVID-19. My main objective is to answer the following questions: **Are countries with healthier eating habits less impacted by COVID-19? Does a healthy diet ultimately help prevent COVID-19?**

GitHub Repo link:

https://github.com/Sijing98/Internship22Fall/tree/main/Project%20-%20Fight%20on%20Healthy%20diet

Data Understanding:

My dataset is the following existing data from Kaggle (Dataset owner: Maria Ren). It contains three main parts:

- **Percentage of fat intake**, **food (kg) intake**, **energy (kcal) intake**, **protein intake** from **23 categories of food** of **170 countries** worldwide
- The **obesity rate (%)** and **undernourished rate (%)** of **170 countries** worldwide
- **Percentages of COVID-19 confirmed/deaths/recovered/active cases** (of the total population) of **170 countries** worldwide (updated to 02/06/2021, I plan to further update them to 09/30/2022)

As stated in last week's Data Intake report, the data storage location is https://www.kaggle.com/datasets/mariaren/covid19-healthy-diet-dataset?resource=download.

Data cleansing and transformation done on the data:

1. Reduce Redundant Data

The original dataset contains four data files (percentage of fat intake, food (kg) intake, energy (kcal) intake, protein intake). I decide to choose **two out of four** data files for analysis: the percentage of **food (kg) intake** and the percentage of **energy (kcal) intake.** This is because I'm interested in the amounts of food categories being consumed, rather than the specific macronutrients (fat, protein, and carbohydrates) provided by such foods.

2. Update the Data and Handle Missing Data

The original data regarding COVID-19 cases is retrieved up to 02/06/2021. However, COVID-19 isn't going away and this data changes daily. To keep up with the latest impact, I plan to further update them to 10/31/2022. Unfortunately, I only found two out of four attributes updated to this date: **COVID-19 cases ('Confirmed' and 'Deaths')** of data. I will delete the other original two attributes ('Recovered' and 'Active' cases) simply because they are outdated.

Among the remaining two attributes ('Confirmed' and 'Deaths'), 9 countries miss the relevant updated data. And there is also other missing data — 3 countries miss data of the obesity rate, 7 countries miss data of the undernourished rate. I ultimately deleted the corresponding rows of 17 countries since I feel as though this data is miniscule in comparison to the entirety of the world and imputing data could potentially skew further analysis.

Further, the undernourished rate of 44 countries is valued "<2.5"— I replaced them with numerical "2" for later analysis.

3. Categorize Food Data

In addition to handling missing data and updating the dataset with new information, I found it also necessary to undergo feature selection and recategorization as well for three reasons.

Firstly, the raw data itself looked into 23 different food categories — some of these categories overlapped. I need to make the data cleaner and avoid repetition.

Secondly, it is not comprehensive to analyze a certain food's impact. It is also meaningless to find the relationship between single kinds of food and rate of COVID-19. A healthy diet is made of a variety of foods, which contains different nutritional elements. Almost all the components of the diet play important roles in health and disease. So I need to consider as many categories as possible.

Thirdly, food can be categorized based on the nutritional element they have. Much evidence proves that intake of specific food groups that contain certain types of nutrients positively impacts health and helps the prevention of diseases. By reviewing relevant researches on health studies, I decide to take suggestions from the U.S. Department of Health & Human Services (NIH, 2021). As a result, I make food groups as follows:

Healthy	- Aquatic Products, Seafood, Offals, Other/Fish - Cereals - Eggs/Milk - Fruits - Pulses - Starchy Roots - Tree Nuts - Vegetables/Vegetal Products
Unhealthy	- Animal Product/Animal Fats/Meats - Oil Crops/Vegetable Oils - Sugars & Sweeteners/Sugar Crops

Finally, I decided to omit the 'Alcohol', 'Spices', 'Stimulants', and 'Miscellaneous' features from the data because they did not fall into our designated categories of a healthy versus unhealthy diet. In addition, these features each contributed to very small portions (mostly less than 1%) of dietary intakes in all countries. As previously mentioned, I also omitted the 'Recovered' and 'Active' attributes as I was unable to find up to date information to populate these columns like I did with the 'Confirmed' and 'Deaths' attributes.

Remark:

Since my process of data cleansing and transformation mainly consist of re-searching, re-categorizing and merging data based on relevant references, I didn't code to deal with due to the not that big data scale as well. Missing data are directly omitted according to my detailed explanation above. The only replacement is for the undernourished rate of 44 countries valued “<2.5” to numerical “2”. I'll upload the processed data to GitHub for review. Thanks.