Week 10: 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

Dataset description:

- 11 categories of food consumption (labeled healthy or unhealthy) of 153 countries.

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

- The obesity rate (%) and undernourished rate (%) of 153 countries.
- Percentages of COVID-19 confirmed/deaths of 153 countries.

My dataset was originally extracted from Kaggle. I chose 1 out of 4 data files for analysis, since I'm only interested in the amounts of food intake. I updated COVID-19 data to 10/31/2022 to keep up with its latest impact. More data preprocessing include:

1) Handle Missing Data: 9 countries miss the COVID-19 case, 3 countries miss data of the obesity rate, 7 countries miss data of the undernourished rate. I ultimately deleted them. Further, the undernourished rate of 44 countries were valued "<2.5"— I replaced them with "2" for later analysis.

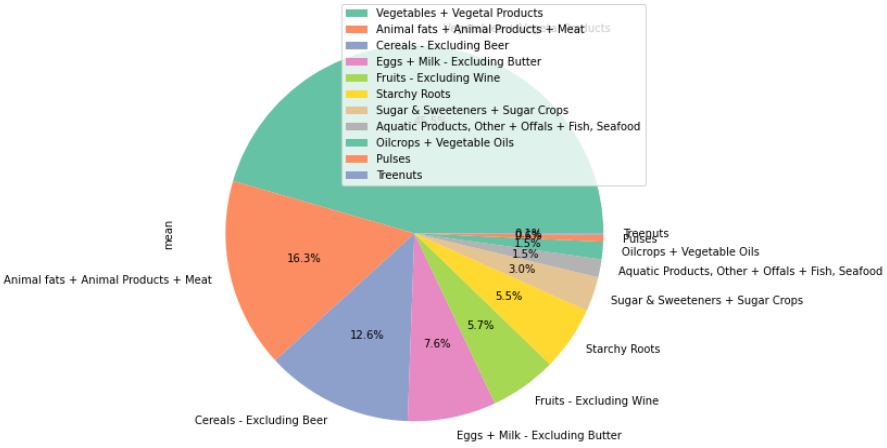
2) Categorize Food Data: I found it necessary to undergo feature selection and recategorization as well. Because some of the 23 different food categories overlapped and food can be categorized based on the nutritional element they have. By reviewing research on health studies, I decided to categorize food data according to suggestions from the U.S. Department of Health & Human Services (NIH, 2021).

EDA performed on the data:

1. General distribution

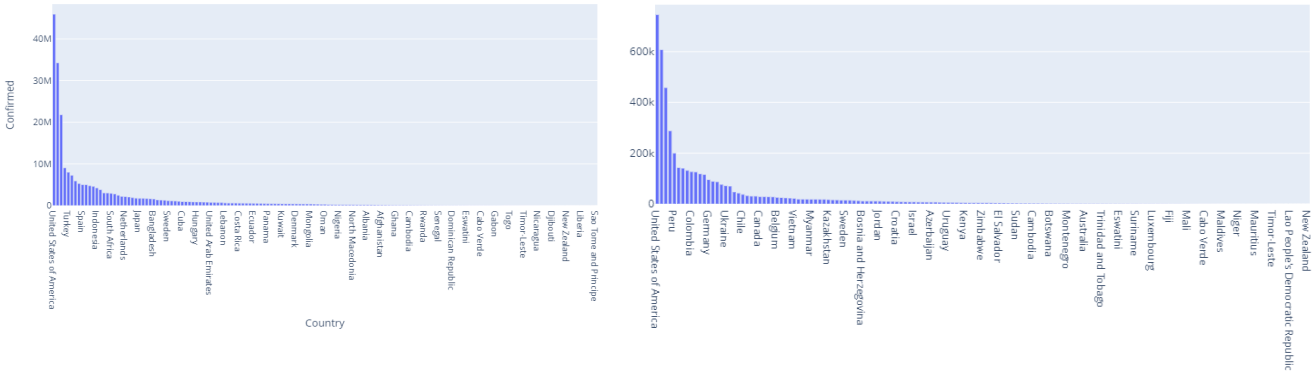
1.1 Food Consumption Distribution

As the figure shown below, *Vegetables + Vegetal Products* (45.5%), which are categorized as healthy food are the most consumed by people worldwide, followed by *Animal fats + Animal Products + Meat* (16.3%) and *Cereals - Excluding Beer* (12.6%).



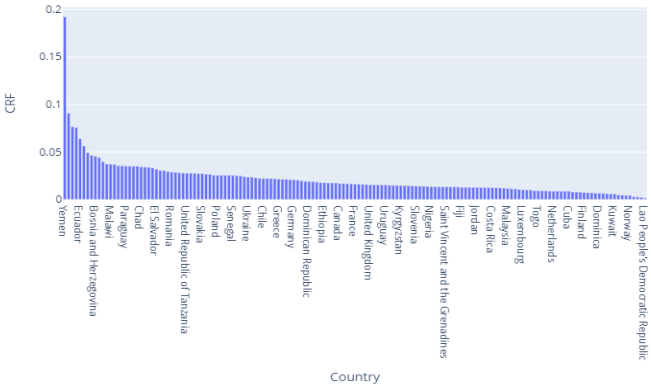
1.2 COVID-19 Case Rate

The United States of America has the most confirmed and deaths cases. The distribution of COVID-19 confirmed and death cases are shown below and separately



To better describe COVID-19 cases rate, I combine the diagnosed cases and the death using the concept of **Case Fatality Rate (CFR)**. I calculate the CRF of all countries, which are presented below. In the following analysis, I also use CRF besides COVID-19 case rate.

CRF = Number of Deaths / Number of Confirmed Cases (Ritchie et al., 2020)



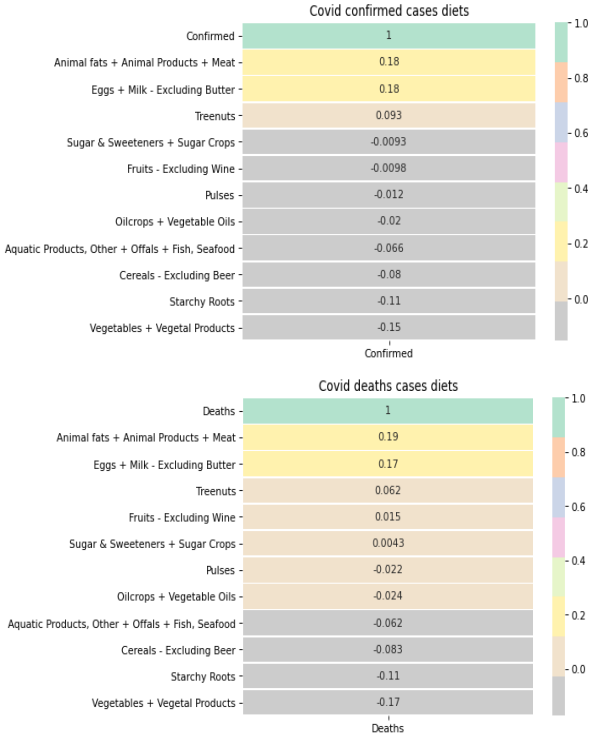
I consider the CRF of ***Yemen*** (19.22%) as an outlier and remove it in the following association analysis.

2. Association Detection

To answer my research question, I firstly analyze the association between food consumption, death cases and confirmed cases.

2.1 Food Consumption & COVID-19 cases

Generally, the relationship between food group consumption and countries' confirmed cases, and food group consumption and death cases, are very similar. The top correlation in both situations is *Animal fats + Animal Products + Meat*.

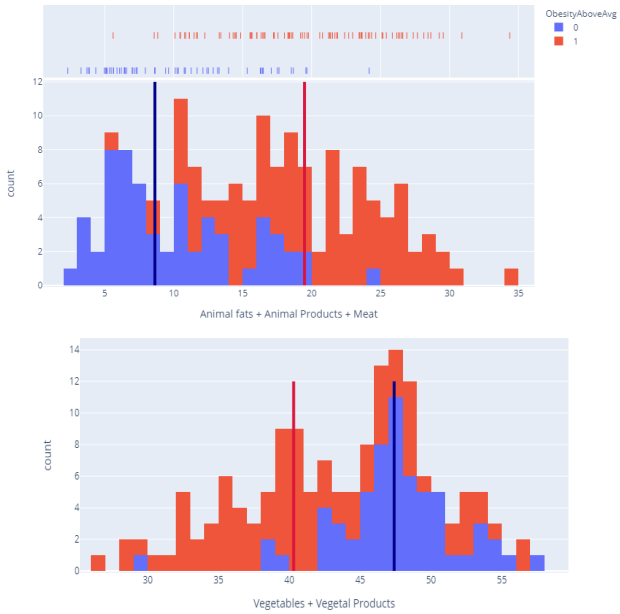


The correlation coefficient shows the relationship between diet and covid cases (both confirmed and death) is not strong, I need to do more exploration to find other potential patterns.

2.2 Food Consumption & Malnutrition

I start by exploring the most decisive food types: *Animal fats + Animal Products + Meat*. Research shows this category may cause obesity. Using the variables of obesity rates in our dataset, I find the world average obesity rate is 18%. I take it as a boundary and divide the world into *HOC (High Obesity Countries)* and *LOC (Low Obesity Countries)*. After calculations:

- HOC have a higher consumption of *Animal fats + Animal Products + Meat* (belongs to unhealthy diet) and lower consumption of *Vegetables + Vegetal Products* (belongs to healthy diet).

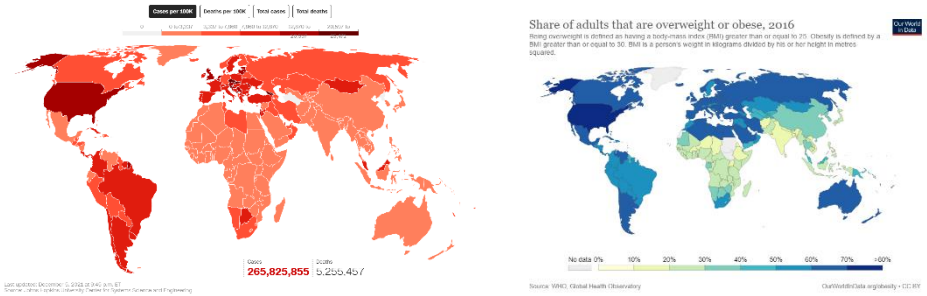


* Code HOC as 1, shown in red color (LOC as 0, in blue color).

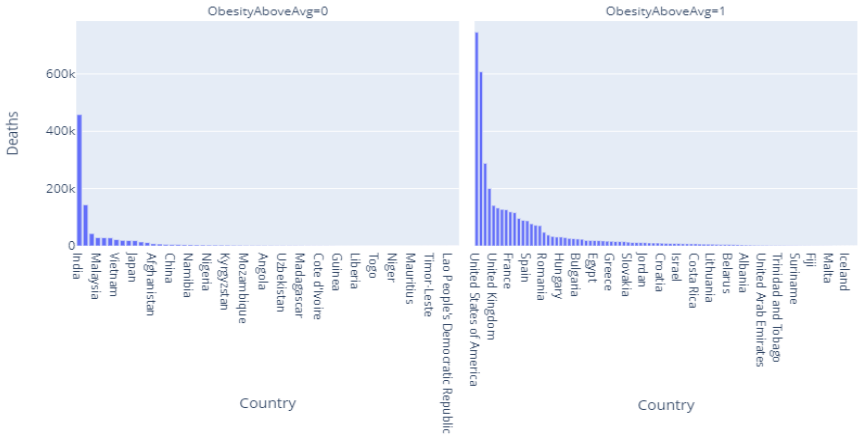
2.3 Obesity & COVID-19

By visualizing two variables in 3 different forms (mapping, bar chart, scatter plot chart), I can see similar patterns in distribution of COVID-19 cases and obesity.

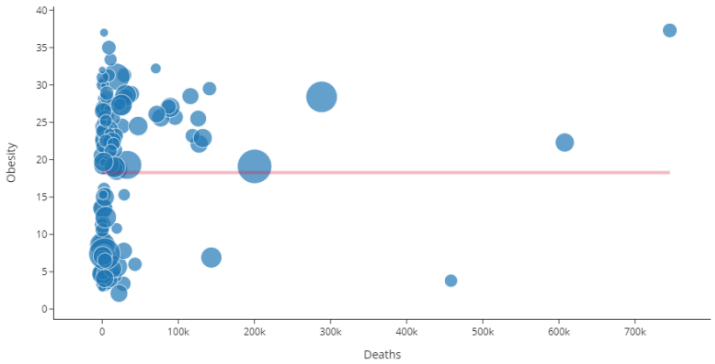
- Firstly, the left map shows the distribution of the COVID-19 cases, while the right shows the obesity. It is evident that dark areas are located in a similar place.



- Secondly, *HOC* have more COVID-19 deaths cases.



- Thirdly, *HOC* have higher CRF.



* x="Deaths", y = "Obesity Rate", size of the dot = "CRF"

I also analyzed the undernourished rate, but it doesn't have a strong correlation.

Final Recommendation:

Based on all the analysis results above, I can simply generalize the main observations:

1. Yemen's CRF (19.22%) is an obvious outlier.
2. Association between diet and COVID-19 is not strong (highest average correlation coefficient is around 0.18).
3. High Obesity Countries (HOC) have a higher consumption of Animal fats + Animal Products + Meat (belongs to unhealthy diet) and lower consumption of Vegetables + Vegetal Products (belongs to healthy diet).
4. High Obesity Countries (HOC) have more COVID-19 deaths cases and higher Case Fatality Rate (CFR).