
# Exploratory Data Analysis

## Does Healthy Diet Help Prevent COVID-19?

16-Nov-2022


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

# Agenda

- Background
- Problem description
- Dataset description
- Data cleansing and transformation
- EDA
- EDA Summary
- Proposed modeling technique

# Background – Diet & COVID-19?

- 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, we aim to fill this void to make clear the relationship between dietary habits with communicable disease, like COVID-19.

- This project takes the problem of COVID-19 and aims to better understand how the illness is impacted by the intake of a healthy versus an unhealthy diet. We want to be able to better advise those around us on how what they are putting into their bodies might affect their livelihoods during these pandemic times.

# Problem description

- The worldwide recorded deaths from COVID-19 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.

- This project will look into a dataset that measures several food groups, malnutrition (obesity and undernourishment) rates, and data on COVID-19 cases. Our main objective is to answer the following questions:
  - **What's the possible nexus among food diet, malnutrition and COVID-19?**
  - **Does a healthy diet ultimately help prevent COVID-19?**

# Dataset description

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

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

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

# Data cleansing and transformation

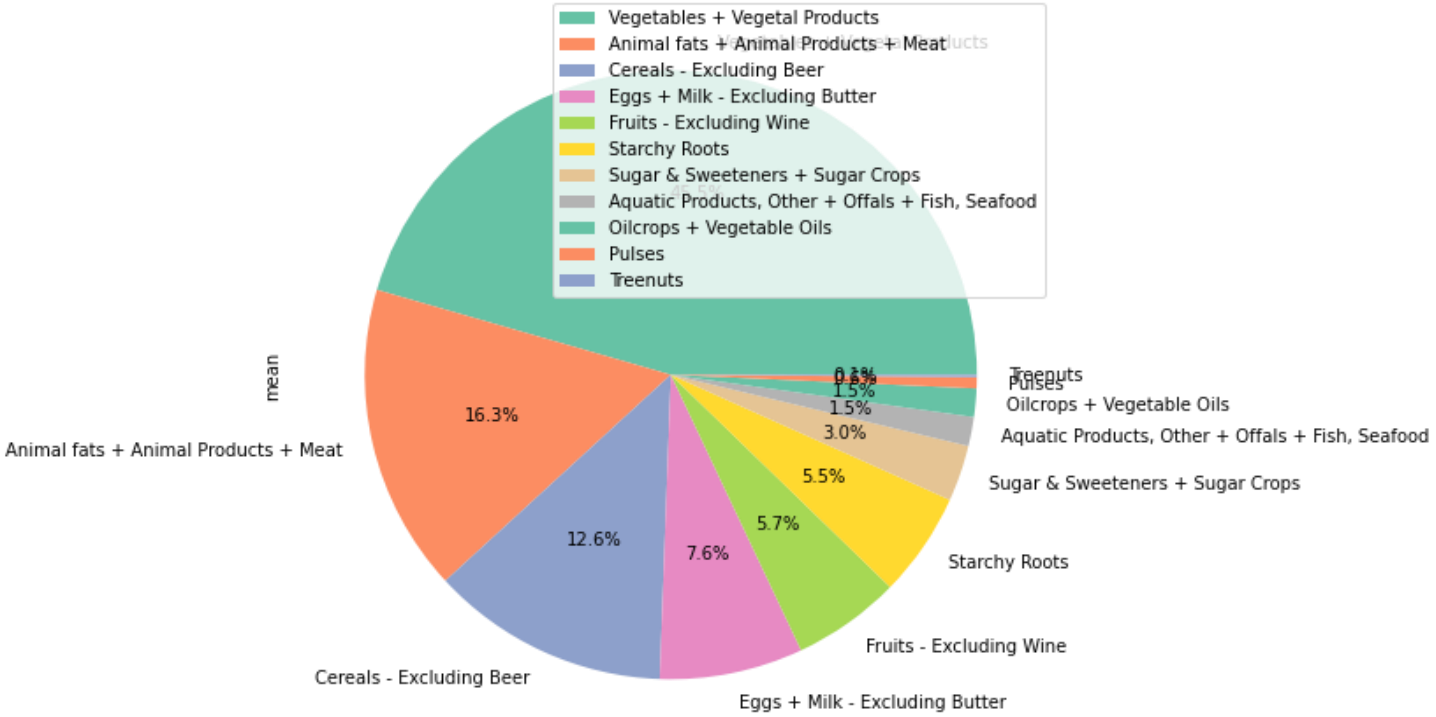
- My dataset was originally extracted from Kaggle. I chose 1 out of 4 data files for analysis, since I'm only interested in the amounts of food intake. I updated COVID-19 data to 10/31/2022 to keep up with its latest impact. More data preprocessing include:

  - **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

- **Food Consumption Distribution**

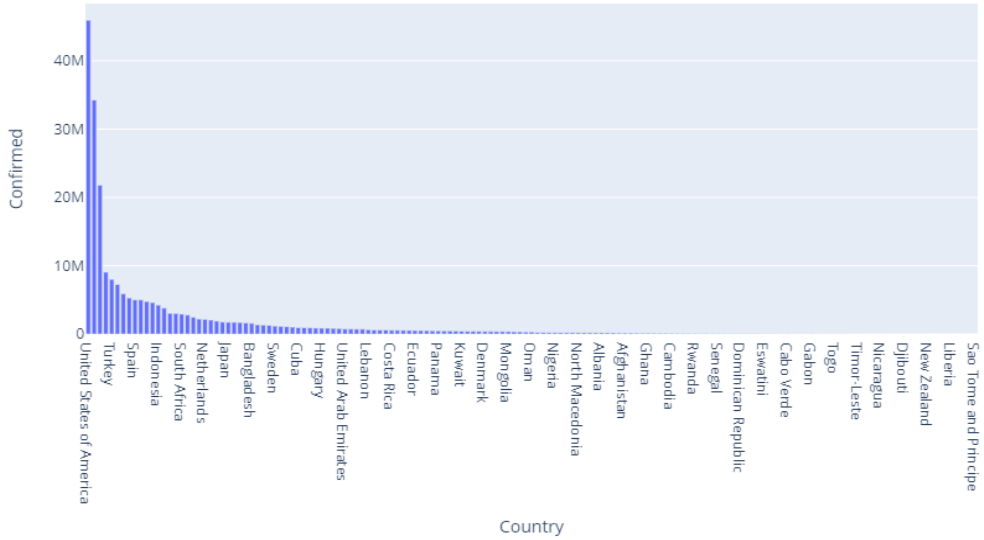
  *Vegetables + Vegetal Products* (45.5%), which are categorized as healthy food are the most consumed by people worldwide, followed by *Animal fats + Animal Products + Meat* (16.3%) and *Cereals - Excluding Beer* (12.6%).

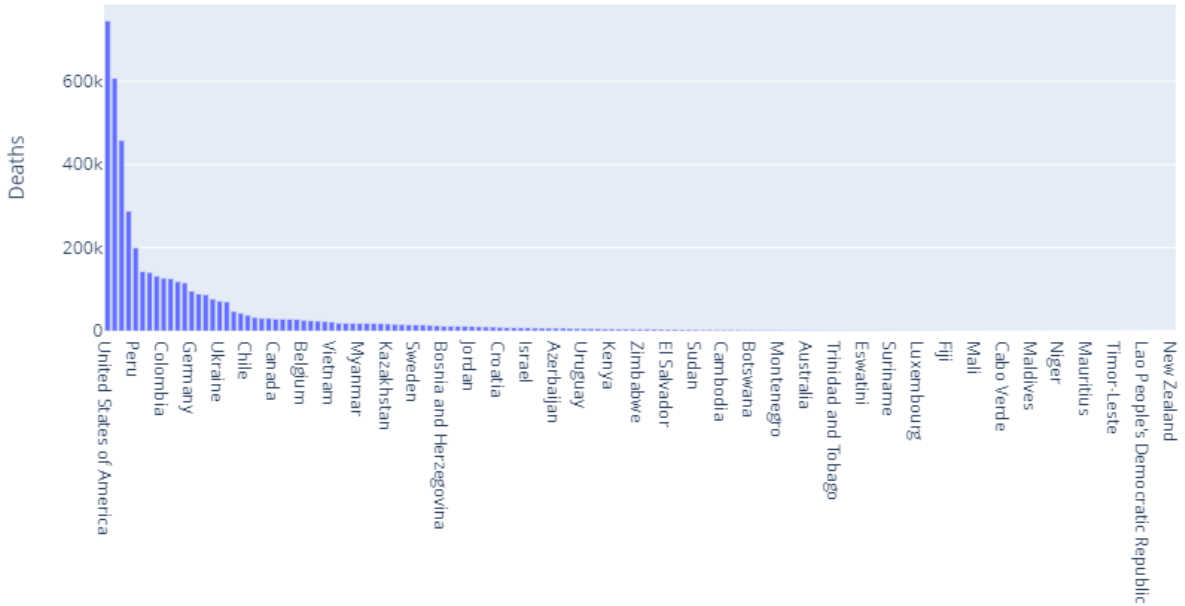


# EDA

- **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.



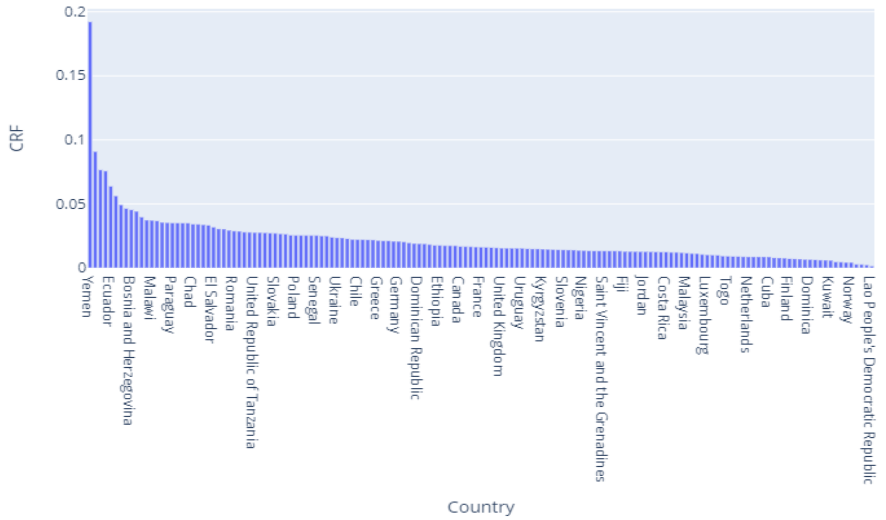
# EDA

- **COVID-19 Case Rate**

  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)

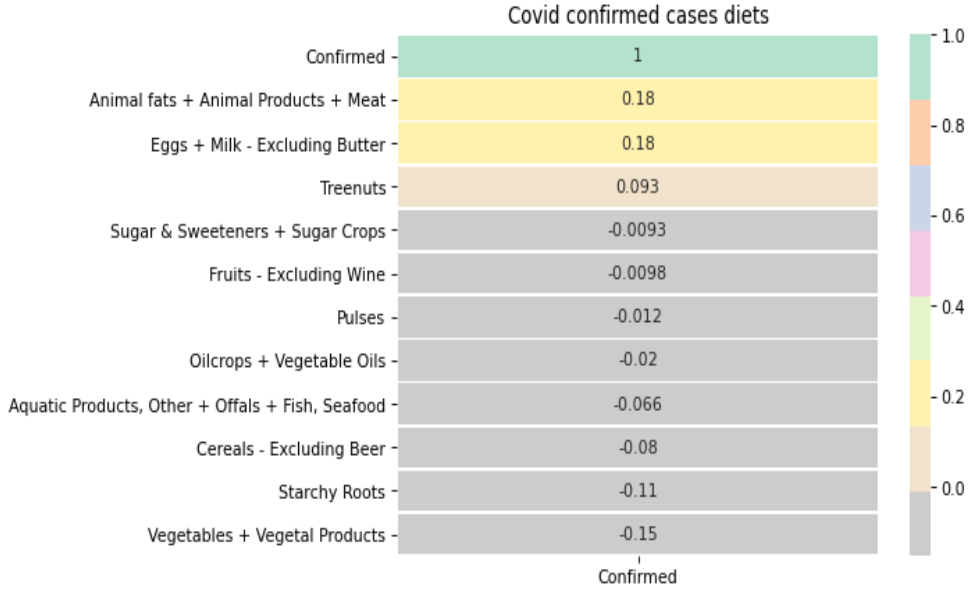
  I consider the CRF of ***Yemen*** (19.22%) as an outlier and remove it in the following association analysis.

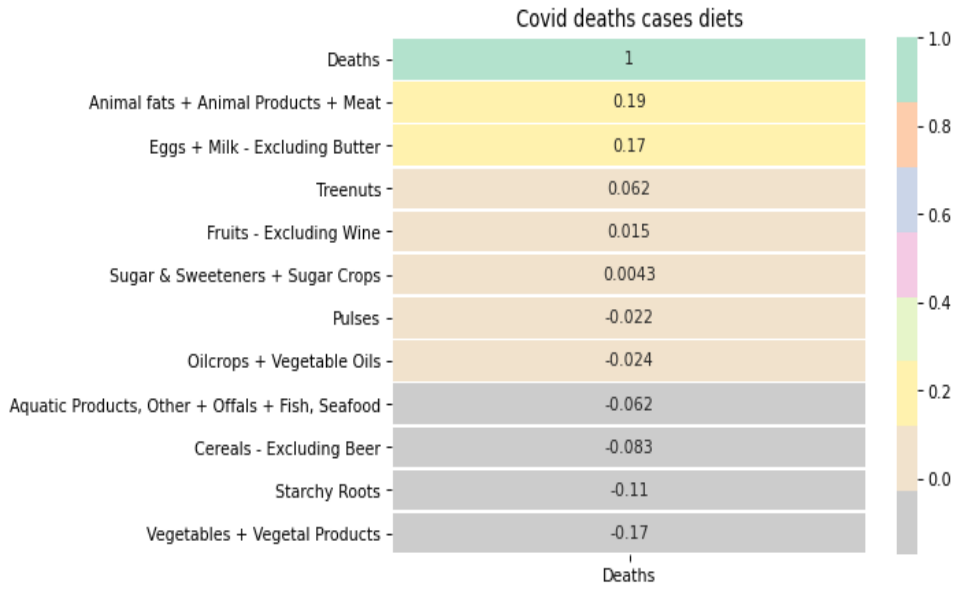


# EDA

- **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.

# EDA

- **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)*.

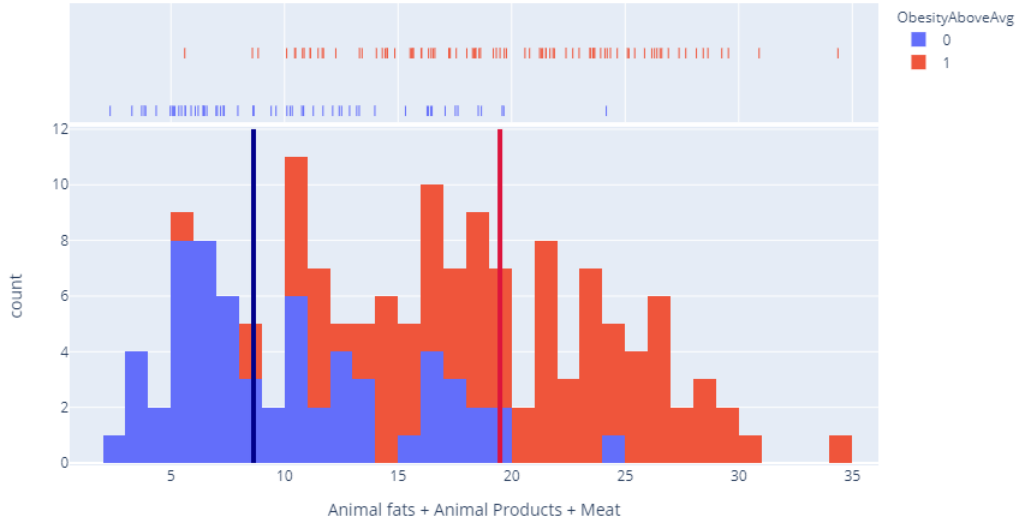
# EDA

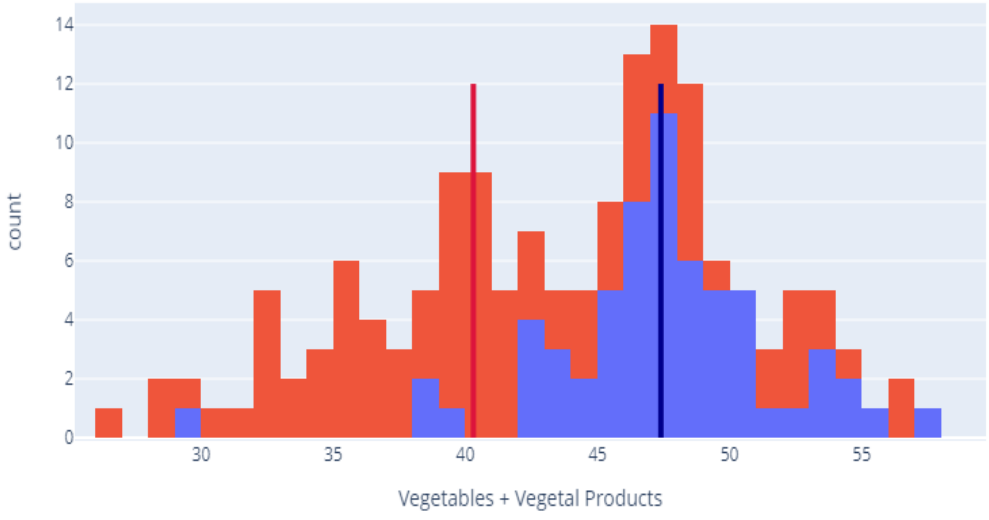
- **Food Consumption & Malnutrition**

  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).



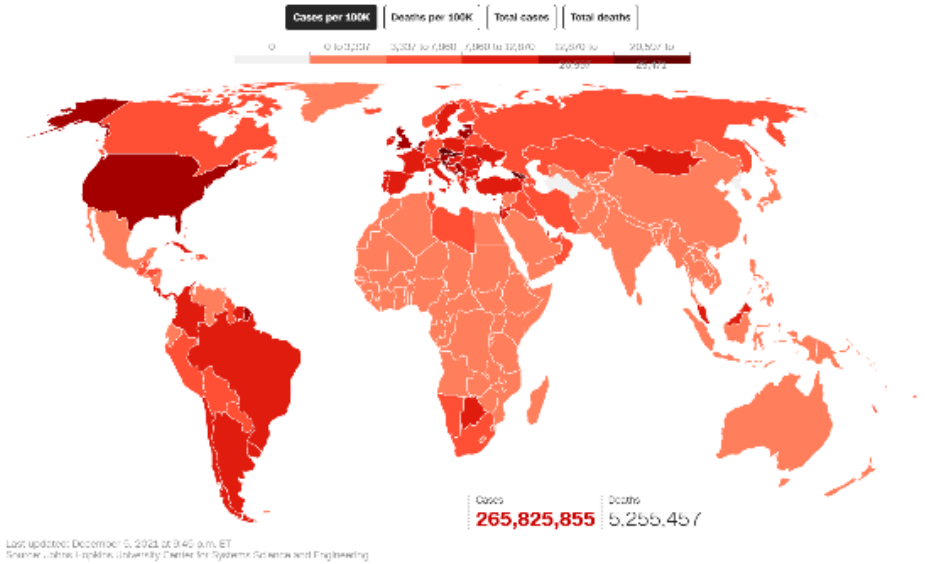



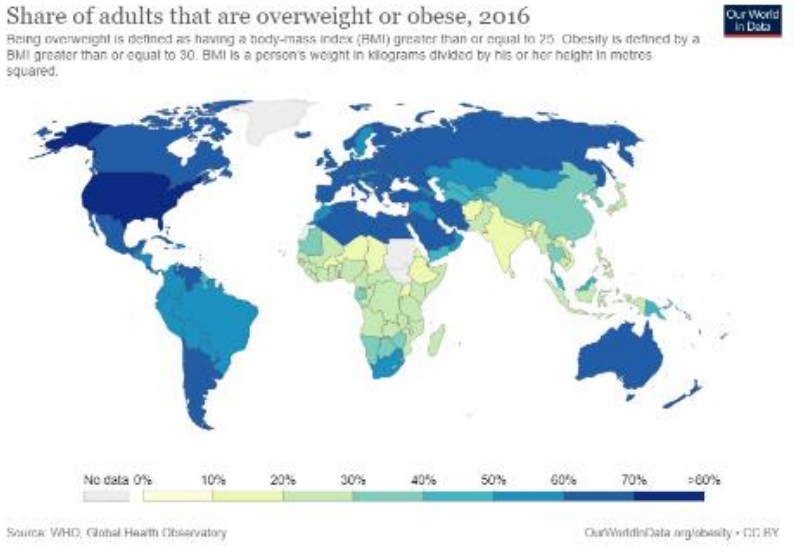
# EDA

- **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.

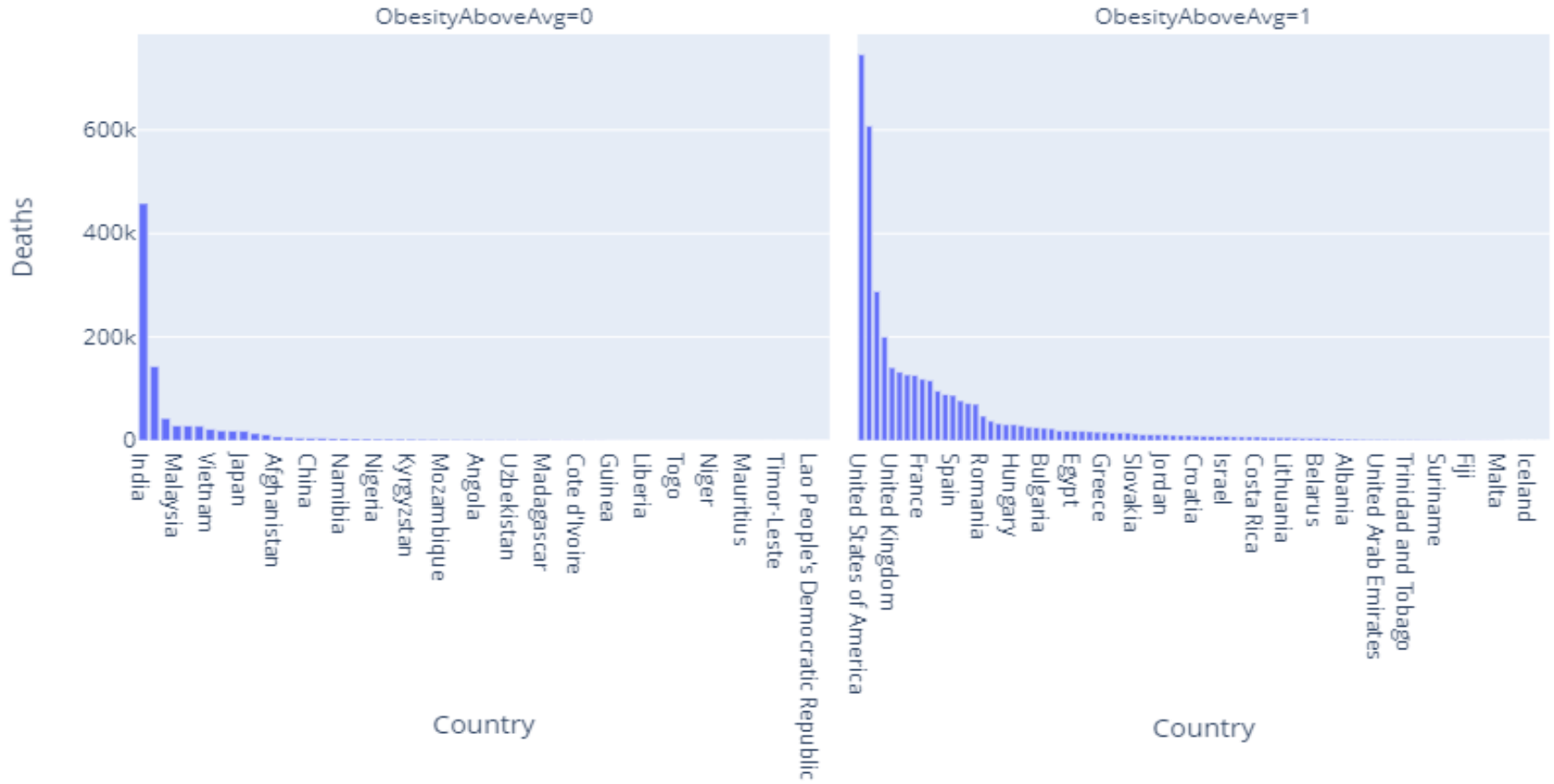


# EDA

- **Obesity & COVID-19**

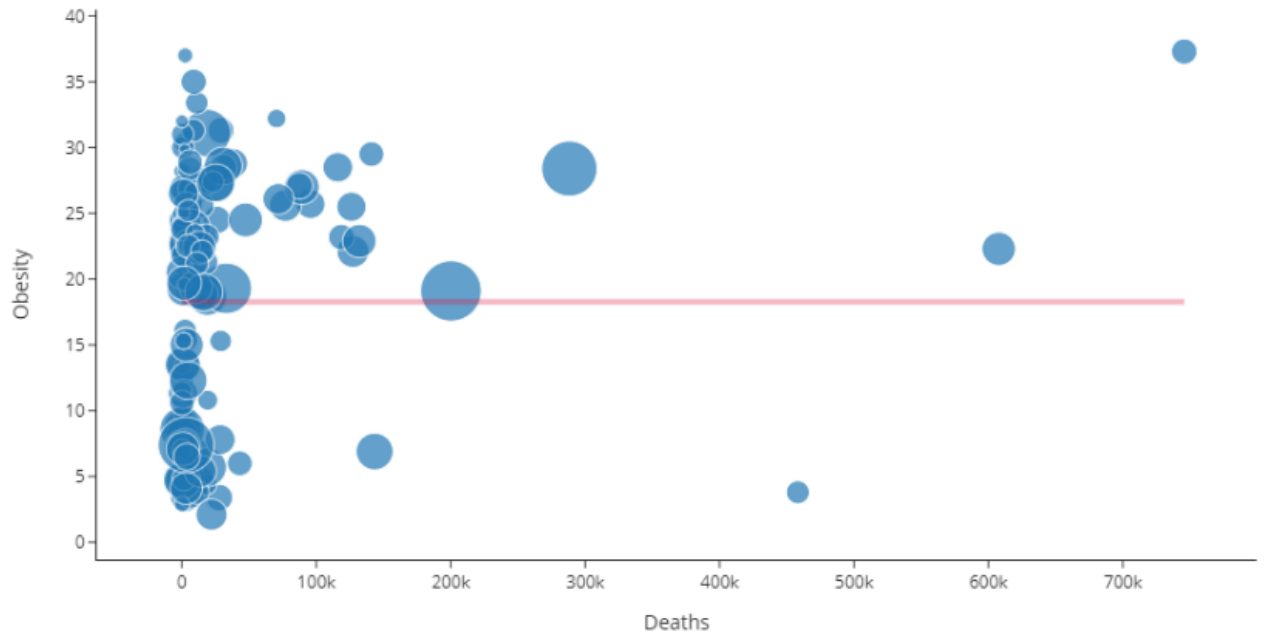
  Secondly, *HOC* have more COVID-19 deaths cases.



# EDA

- **Obesity & COVID-19**

  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.

# EDA Summary

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

- Yemen's CRF (19.22%) is an obvious outlier.
- Association between diet and COVID-19 is not strong (highest average correlation coefficient is around 0.18).
- 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).
- High Obesity Countries (HOC) have more COVID-19 deaths cases and higher Case Fatality Rate (CFR).

# Proposed modeling technique

In this project, methods used and the corresponding targets could be concluded as below.

| Main part | Description of Methods Used | Targets |
|---|---|---|
| *Data Preprocessing* | Conduct general exploratory data analysis & data visualization by using Excel/Tableau | To organize our own datasets preliminary (reduce redundant data, update data and handle missing data) |
| | Deal with personalized categories of foods based on NIH research by using Excel | To further categorize food data into *healthy* and *unhealthy* food groups |
| *Data Mining* | Figure out basic data features by using Excel/Python | To observe and report general distribution of each variable |
| | Detect further association by using Python | To analyze possible relationships among each two variables |
| | Explore classification & regression analysis by Python, mainly constructing and comparing models based on Ridge regressor, SVR, Random Forest, and XGBoost, as well as trying some hyperparameter fine-tuning with a simple Grid Search | To model and check the associations found before, and further answer our direct problem: whether countries with healthier eating habits can be less impacted by COVID-19? |

# Proposed modeling technique

The next modeling part is going to **explore classification & regression analysis by Python, mainly constructing and comparing models based on Ridge regressor, SVR, Random Forest, and XGBoost, as well as trying some hyperparameter fine-tuning with a simple Grid Search.**

| Main part | Description of Methods Used | Targets |
|---|---|---|
| *Data Preprocessing* | Conduct general exploratory data analysis & data visualization by using Excel/Tableau | To organize our own datasets preliminary (reduce redundant data, update data and handle missing data) |
| | Deal with personalized categories of foods based on NIH research by using Excel | To further categorize food data into *healthy* and *unhealthy* food groups |
| *Data Mining* | Figure out basic data features by using Excel/Python | To observe and report general distribution of each variable |
| | Detect further association by using Python | To analyze possible relationships among each two variables |
| | Explore classification & regression analysis by Python, mainly constructing and comparing models based on Ridge regressor, SVR, Random Forest, and XGBoost, as well as trying some hyperparameter fine-tuning with a simple Grid Search | To model and check the associations found before, and further answer our direct problem: whether countries with healthier eating habits can be less impacted by COVID-19? |

# Thank You