# Food Recommendation System

Assisting food manufacturer to produce healthy food

# Project Objective

- To develop a service that helps food producers understand, which products should be produced in order to boost their customers' immune systems

- Dataset: **Kaggle- COVID-19 Healthy Diet Dataset**

  https://www.kaggle.com/mariaren/covid19-healthy-diet-dataset

# Dataset Description

Four Different types of nutritional data for 170 countries:
Percentage of consumption

- Protein
- Fat
- Amount(Kg)
- Energy (KCal)

Each of them contains

- 31 Identical Columns
- 24 different types of foods(eg. Animal product, vegetal product, fish, etc)
- Health condition and covid cases (Undernourished, Obesity, COVID death..)

# Dataset Overview

```
# Show a preview of data
Fat_Supply_DF.head()
```

| | Country | Alcoholic Beverages | Animal Products | Animal fats | Aquatic Products, Other | Cereals - Excluding Beer | Eggs | Fish, Seafood | Fruits - Excluding Wine | Meat | Miscellaneous | Milk - Excluding Butter | Offals | Oilcrops | Pulses |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Afghanistan | 0.0 | 21.6397 | 6.2224 | 0.0 | 8.0353 | 0.6859 | 0.0327 | 0.4246 | 6.1244 | 0.0163 | 8.2803 | 0.3103 | 1.0452 | 0.1960 |
| 1 | Albania | 0.0 | 32.0002 | 3.4172 | 0.0 | 2.6734 | 1.6448 | 0.1445 | 0.6418 | 8.7428 | 0.0170 | 17.7576 | 0.2933 | 3.1622 | 0.1148 |
| 2 | Algeria | 0.0 | 14.4175 | 0.8972 | 0.0 | 4.2035 | 1.2171 | 0.2008 | 0.5772 | 3.8961 | 0.0439 | 8.0934 | 0.1067 | 1.1983 | 0.2698 |
| 3 | Angola | 0.0 | 15.3041 | 1.3130 | 0.0 | 6.5545 | 0.1539 | 1.4155 | 0.3488 | 11.0268 | 0.0308 | 1.2309 | 0.1539 | 3.9902 | 0.3282 |
| 5 | Argentina | 0.0 | 30.3572 | 3.3076 | 0.0 | 1.3316 | 1.5706 | 0.1664 | 0.2091 | 19.2693 | 0.0000 | 5.8512 | 0.1878 | 0.0640 | 0.0213 |

```
Protein_Supply_DF.head()
```

| | Country | Alcoholic Beverages | Animal Products | Animal fats | Aquatic Products, Other | Cereals - Excluding Beer | Eggs | Fish, Seafood | Fruits - Excluding Wine | Meat | Milk - Excluding Butter | Offals | Oilcrops | Pulses | Spices | Starchy Roots |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Afghanistan | 0.0000 | 9.7523 | 0.0277 | 0.0 | 35.9771 | 0.4067 | 0.0647 | 0.5824 | 3.1337 | 5.5278 | 0.5916 | 0.2034 | 1.2479 | 0.1664 | 0.1941 |
| 1 | Albania | 0.1840 | 27.7469 | 0.0711 | 0.0 | 14.2331 | 1.8069 | 0.6274 | 1.2757 | 7.6582 | 16.4750 | 1.1084 | 0.3722 | 1.4555 | 0.0000 | 0.8867 |
| 2 | Algeria | 0.0323 | 13.8360 | 0.0054 | 0.0 | 26.5633 | 1.2916 | 0.6350 | 1.1624 | 3.5088 | 8.0616 | 0.3283 | 0.1830 | 2.5509 | 0.1776 | 1.4638 |
| 3 | Angola | 0.6285 | 15.2311 | 0.0277 | 0.0 | 20.3882 | 0.1756 | 5.4436 | 1.2754 | 7.6248 | 1.1460 | 0.8133 | 2.1534 | 4.0850 | 0.0000 | 5.1941 |
| 5 | Argentina | 0.1704 | 31.9799 | 0.0097 | 0.0 | 13.6702 | 2.0593 | 1.0223 | 0.5209 | 21.6250 | 5.8322 | 1.4313 | 0.0097 | 0.2434 | 0.0292 | 1.3096 |

# Dataset Overview

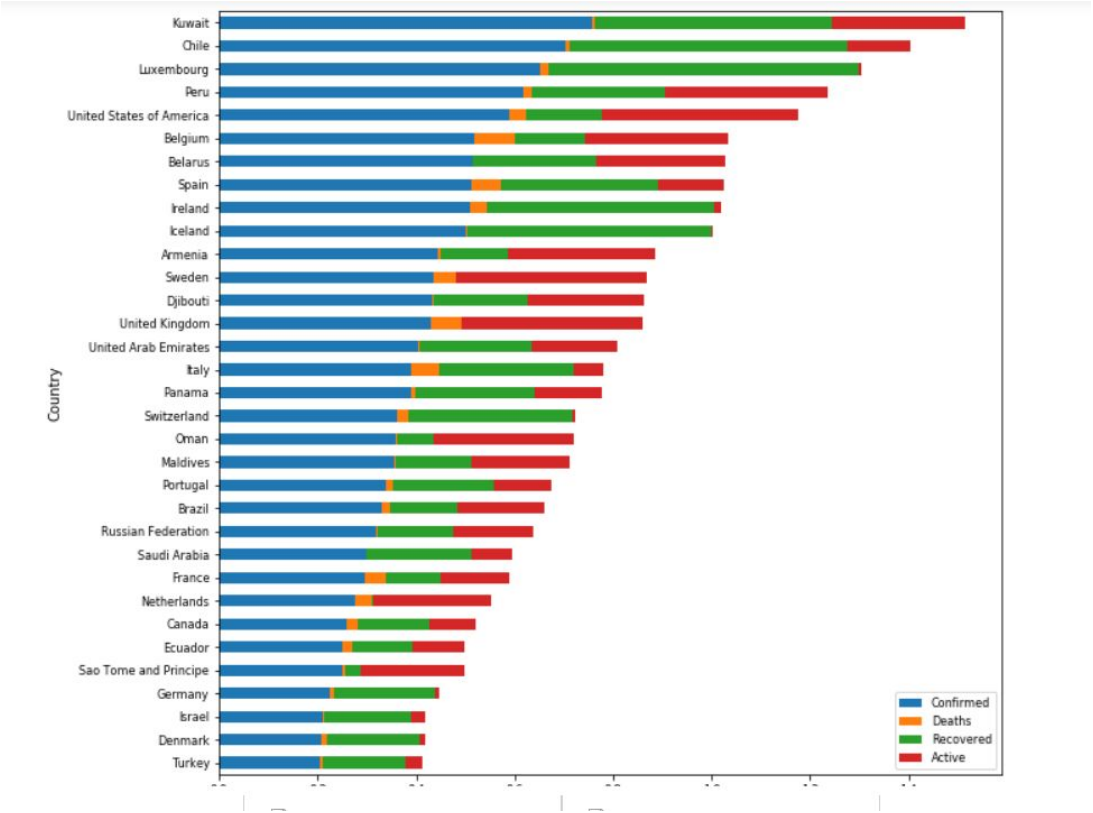
```
# Show a preview of data
Fat_Supply_DF.head()
```

| s | Sugar Crops | Sugar & Sweeteners | Treenuts | Vegetal Products | Vegetable Oils | Vegetables | Obesity | Undernourished | Confirmed | Deaths | Recovered | Active | Population | Unit (all except Population) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.0 | 0.0 | 0.7513 | 28.3684 | 17.0831 | 0.3593 | 4.5 | 29.8 | 0.053472 | 0.000938 | 0.004929 | 0.047605 | 38042000.0 | % |
| 0 | 0.0 | 0.0 | 0.9181 | 17.9998 | 9.2443 | 0.6503 | 22.3 | 6.2 | 0.043597 | 0.001190 | 0.032820 | 0.009587 | 2858000.0 | % |
| 6 | 0.0 | 0.0 | 0.8595 | 35.5857 | 27.3606 | 0.5145 | 26.6 | 3.9 | 0.023393 | 0.001629 | 0.015475 | 0.006289 | 43406000.0 | % |
| 8 | 0.0 | 0.0 | 0.0308 | 34.7010 | 22.4638 | 0.1231 | 6.8 | 25.0 | 0.000290 | 0.000013 | 0.000076 | 0.000200 | 31427000.0 | % |
| 5 | 0.0 | 0.0 | 0.1366 | 19.6449 | 17.3147 | 0.1878 | 28.5 | 4.6 | 0.050722 | 0.001478 | 0.015374 | 0.033870 | 44939000.0 | % |

```
Protein_Supply_DF.head()
```

| ugar & teners | Treenuts | Vegetal Products | Vegetable Oils | Vegetables | Miscellaneous | Obesity | Undernourished | Confirmed | Deaths | Recovered | Active | Population | Unit (all except Population) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.0000 | 0.1387 | 40.2477 | 0.0000 | 1.1370 | 0.0462 | 4.5 | 29.8 | 0.053472 | 0.000938 | 0.004929 | 0.047605 | 38042000.0 | % |
| 0.0042 | 0.2677 | 22.2552 | 0.0084 | 3.2456 | 0.0544 | 22.3 | 6.2 | 0.043597 | 0.001190 | 0.032820 | 0.009587 | 2858000.0 | % |
| 0.0000 | 0.2745 | 36.1694 | 0.0269 | 3.1267 | 0.1399 | 26.6 | 3.9 | 0.023393 | 0.001629 | 0.015475 | 0.006289 | 43406000.0 | % |
| 0.0092 | 0.0092 | 34.7782 | 0.0092 | 0.8133 | 0.0924 | 6.8 | 25.0 | 0.000290 | 0.000013 | 0.000076 | 0.000200 | 31427000.0 | % |
| 0.0049 | 0.0438 | 18.0176 | 0.0000 | 1.0516 | 0.0000 | 28.5 | 4.6 | 0.050722 | 0.001478 | 0.015374 | 0.033870 | 44939000.0 | % |

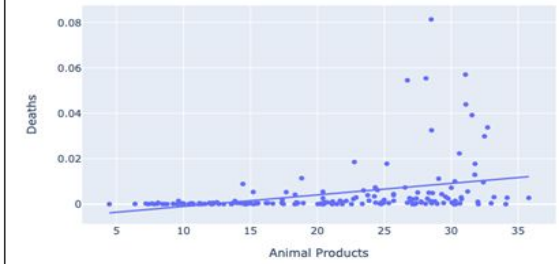
# COVID-19 cases, fatality and recovery in percent



# Data cleaning

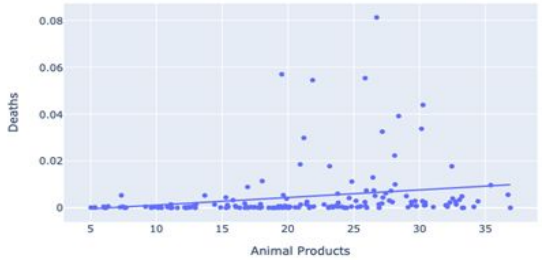
- Filling empty values:
  - Using most frequent value
- Removing Non Numeric Characters
  - (less than or greater than signs)
- Standard Scaler Transform
  - Standardizes a feature by subtracting the mean and then scaling to unit variance

# Data visualization: Positive correlation



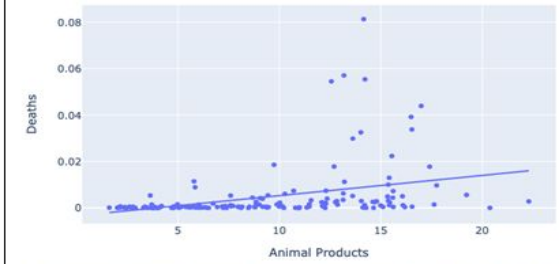

Correlation between Animal Products and Deaths is 0.34633459756709956

Protein data



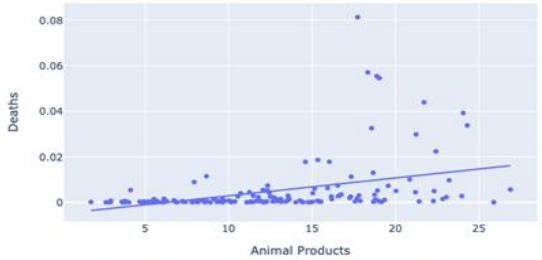

Correlation between Animal Products and Deaths is 0.22281945580963033

Fat data




Correlation between Animal Products and Deaths is 0.3562619529279708

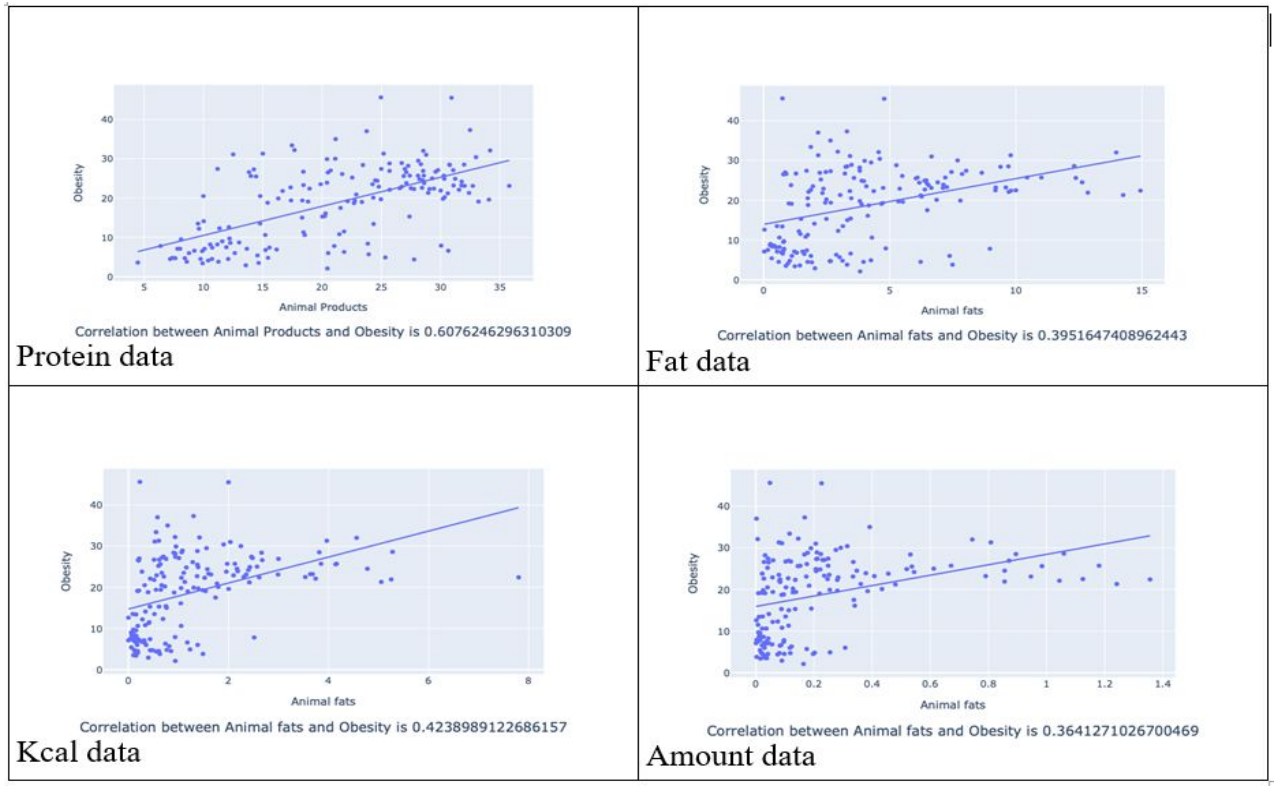
Kcal data



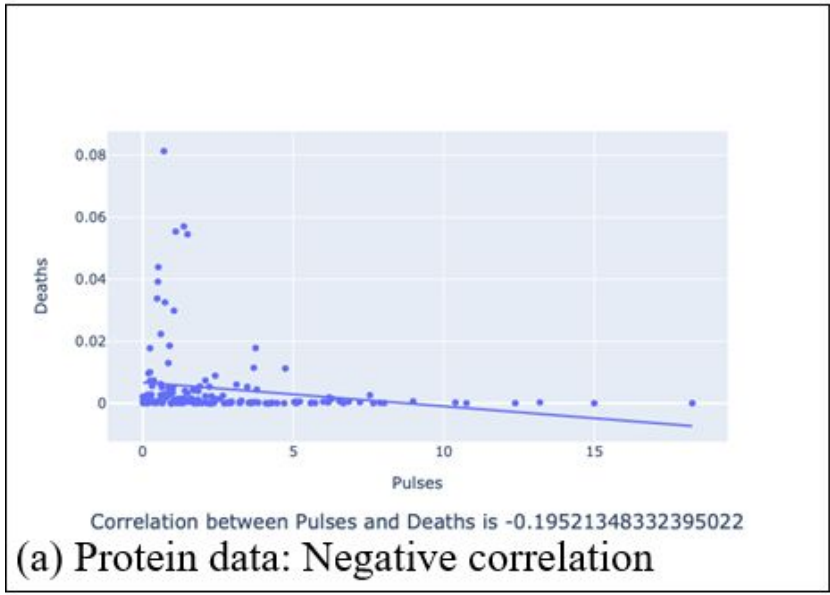

Correlation between Animal Products and Deaths is 0.3964601353235494

Amount data

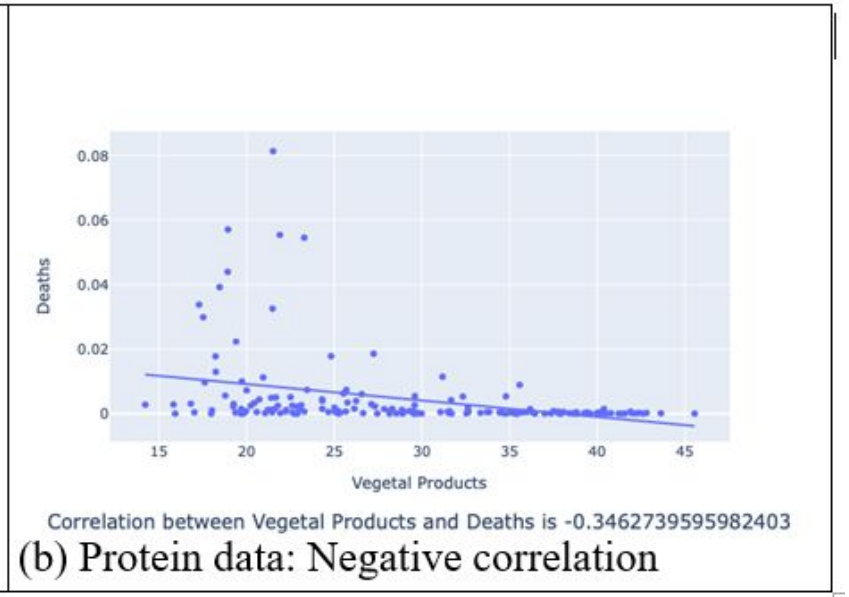
# Data visualization: Positive correlation



# Data Visualization: Negative correlation




(a) Protein data: Negative correlation




(b) Protein data: Negative correlation

# Input and Output of the model

- Input variables
  - Death, Recovery, Active and Infection rate of covid
  - Undernourished and Obesity

- Output variable
  - Different types of food percentage (Animal product, vegetal product etc.)

# Methodology

- Randomly shuffle
- Train data
  - 80%
- Hyper-parameter tuning
  - Random search
  - 5-fold cross validation
- Test data
  - 20%

Note: Only 170 samples, so we had less flexibility to increase the test data size

# Model Selection:

- Linear Regression
- LASSO
- Ridge Regression
- Support Vector Regression
- Decision Tree
- Random Forest
- Neural Network

# Model Selection

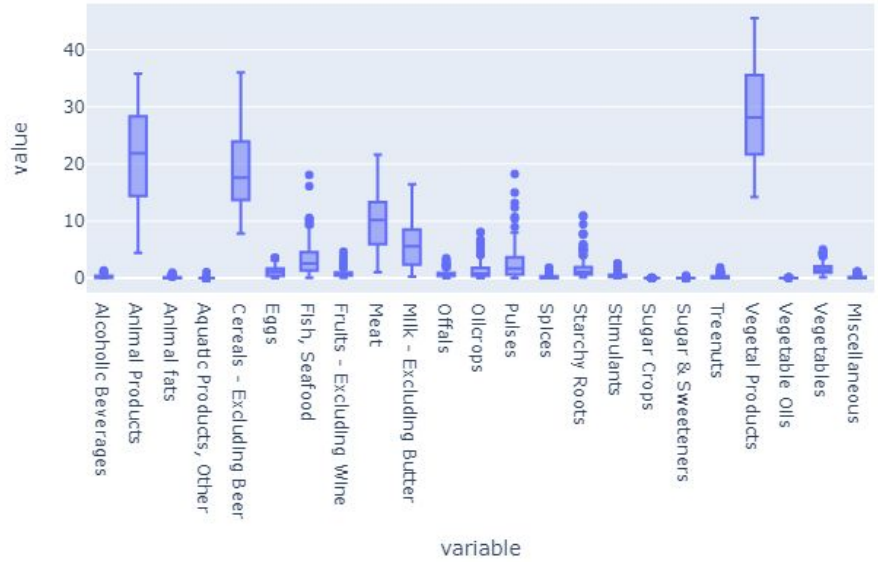
- 24 output variable

| Methods | R^2 Score(Fat) | R^2 Score(Protein) | R^2 Score(Kcal) | R^2 Score(Amount) |
|---|---|---|---|---|
| Linear Regression | -358.03 | -501.27 | -351.97 | -351.97 |
| Ridge | -0.08 | -0.08 | 0.07 | 0.07 |
| Lasso | -0.03 | -0.10 | 0.07 | 0.07 |
| SVR | -0.49 | -10.79 | 0.08 | 0.08 |
| Decision Tree | -0.01 | -0.10 | -0.03 | -0.03 |
| Random Forest | -0.06 | -0.98 | 0.08 | 0.08 |
| Neural Network | -0.13 | -0.19 | -0.30 | -0.23 |

*Table 1: $R^2$ accuracy after 5-fold cross validation for two nutrients consumption, calorie consumption and amount.*

# Distribution of data

## Protein dataset

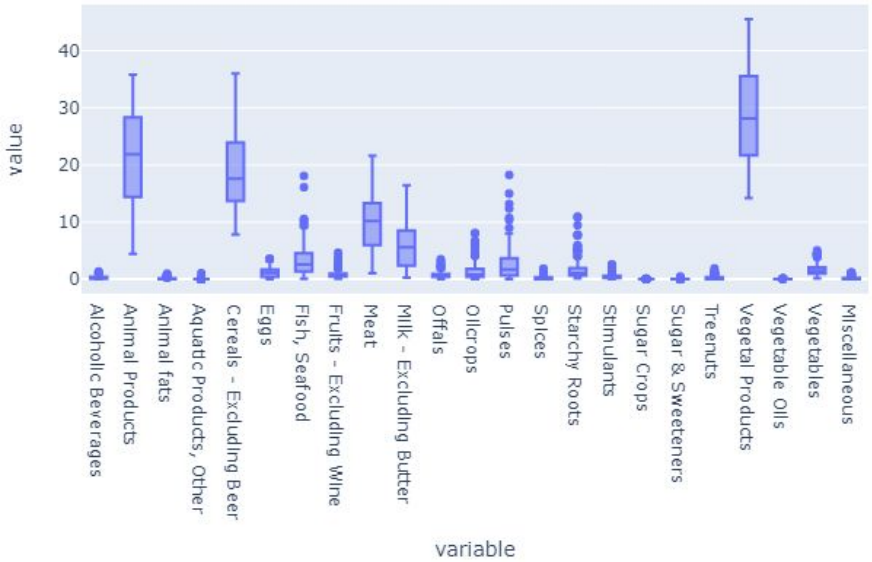



Top 8 using PCA

1. Vegetal Products
2. Fish
3. Seafood
4. Alcoholic Beverages
5. Sugar & Sweeteners
6. Fruits - Excluding Wine
7. Spices
8. Stimulants
9. Treenuts

# Distribution of data

## Fat dataset

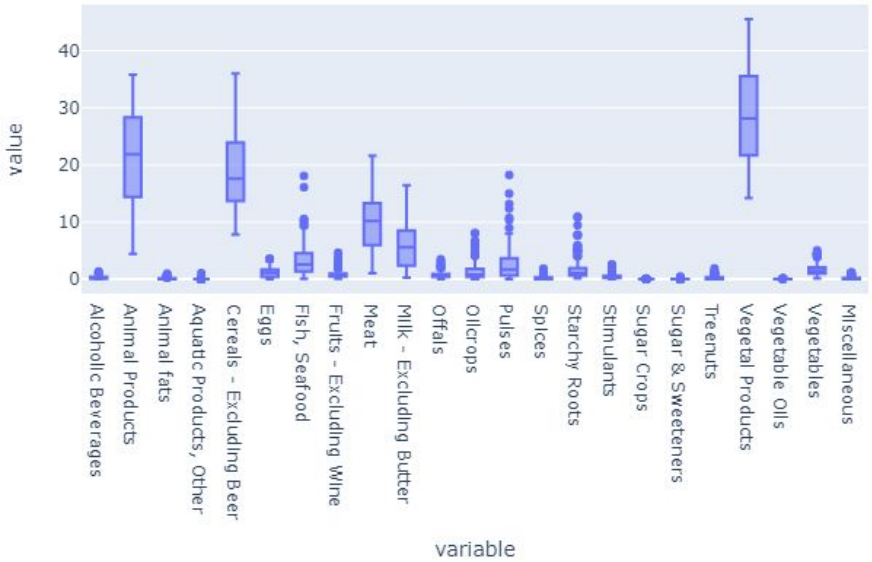



Top 8 using PCA

1. Animal Products
2. Vegetables
3. Miscellaneous
4. Fruits - Excluding Wine,
5. Milk - Excluding Butter
6. Starchy Roots
7. Animal fats
8. Fruits - Excluding Wine

# Distribution of data

## Kcal dataset

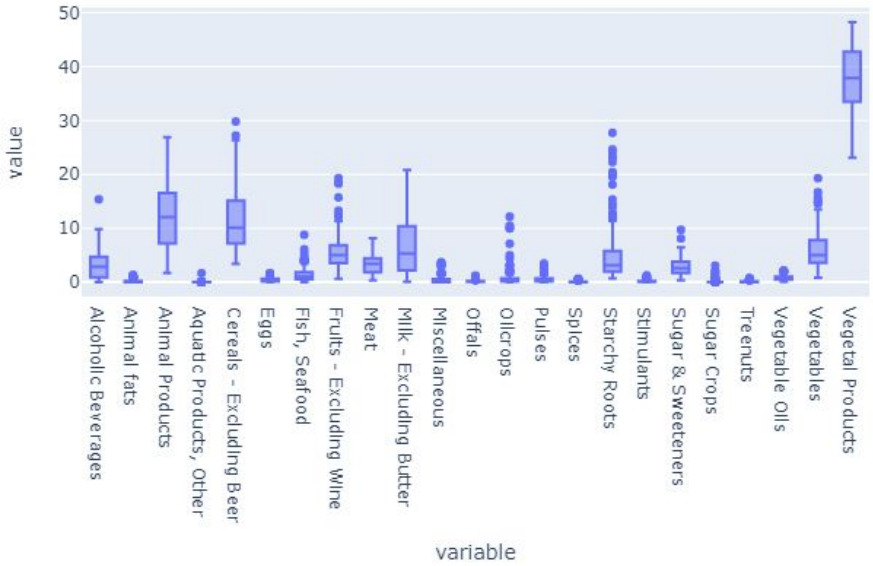



Top 8 using PCA

1. Animal Products
2. Offals
3. Sugar & Sweeteners
4. Miscellaneous
5. Animal fats
6. Vegetables
7. Starchy Roots
8. Sugar & Sweeteners

# Distribution of data

## Quantity dataset




Top 8 using PCA

1. Animal fats
2. Offals
3. Stimulants
4. Fruits - Excluding Wine
5. Vegetable Oils
6. Starchy Roots
7. Spices
8. Sugar Crops

# Model Tuning

- Training model on 8 selected output variable
- Animal Products, Animal fats, Meat, Cereals-Excluding Beer, Vegetables, Vegetal Products, Vegetable Oils, and Oil crops

| Methods | R^2 Score(Fat) | R^2 Score(Protein) | R^2 Score(Kcal) | R^2 Score(Amount) |
|---|---|---|---|---|
| Linear Regression | -258.79 | -81.34 | -450.49 | -1821.37 |
| Ridge | 0.09 | 0.19 | 0.15 | 0.07 |
| Lasso | 0.08 | 0.17 | 0.16 | 0.09 |
| SVR | 0.15 | 0.16 | 0.17 | 0.06 |
| Decision Tree | 0.06 | 0.15 | -0.06 | -0.11 |
| Random Forest | 0.19 | 0.23 | 0.27 | 0.06 |
| Neural Network | -0.64 | -0.69 | -0.81 | -0.68 |

*Table 2: Prediction on eight outputs for types of data.*

# Predicting recommended percentage of foods

- Best condition of a country in COVID-19 crisis
  - 0 death due to covid
  - 100% recovery rate from covid
  - 0% Undernourishment
  - 0% Obesity
- Usages of prediction
  - Based on above input we predict the output
  - Compare it with the current food habit of a country
  - Guide food manufacturers to produce food to increase immunity.

  Note: Detail description will be on the application presentation.