An Exploratory Analysis Between Eating Habits and the Number of Confirmed Cases of COVID-19 and Death Cases of COVID-19 Across Different Countries in the World

Wendy Jiang, Bianca Xie, Sihao Feng

University of Washington

CSE 163 Section A

August 15th, 2022 Summer

Summary of questions and results

Q1: What are the top three factors contributing to people getting COVID-19 in terms of diet?

- The top three factors contributing to people getting COVID-19 in terms of diet are milk (excluding butter), animal products, and animal fats.

Q2: Does intake of more fat or intake of more protein help people get away from COVID-19?

- Intake of more fat helps people get away from COVID-19.

Q3: Is there a significant difference between undernourished countries with a higher recovery rate and countries with a lower recovery rate?

- There is a significant difference between undernourished countries with a higher recovery rate and countries with a lower recovery rate. This result is not surprising because much more animal fat does cause many health issues, hence animal fat intake definitely does not benefit the Covid. On the contrary, vegetable product intake is beneficial for human's body due to those vitamins which are essential and crucial for our daily life.

Q4: Does there appear to be a relationship between the number of deaths from COVID and the amount of intake from fat?

- We can find the most positive relationship of fat intake between the death rate from COVID is the animal fat. The most negative relationship of fat intake between death rate is the vegetable product. It is very surprising that Wendy owns a significantly higher recovery rate. Wendy's eating habits are healthy and benefit from COVID recovery.

Q5: Use multiple machine learning models to find the best tune model by comparing the predicted accuracy and mean squared error. Finding and predicting Wendy's diet habit's recovery rate.

- The tuned Decision tree model is the better predict model. The best model's hyperparameter is:

  {'max_depth': 5,

  'max_features': None,

  'max_leaf_nodes': 40,

  'min_samples_leaf': 5,

  'min_weight_fraction_leaf': 0.1,

  'splitter': 'random'}

  Wendy's recovery rate prediction from Wendy's diet is: 1.35789416%. So, we can conclude Wendy has a food intake habit that is beneficial for COVID recovery.

Motivation

At the end of 2019, COVID-19 swept across the world. People's lifestyles have been forced to change from outdoor to indoors. As a small portion of the victims of this pandemic, we have also recognized that our eating habits have been heavily influenced. Since we stay indoors more frequently and have less chance to exercise, it leads us to start to eat unhealthily and irregularly. And that will lead to an unhealthy body with a weak immune system. It might cause people to increase a high chance of getting COVID. To verify our hypothesis, we decided to research the relationship between the change in eating habits and the opportunities to get COVID-19 to encourage people to eat healthy food.

About Dataset

The dataset we found that could help to answer our question is from *Kaggle, "*Covid-19 healthy dataset*"*. Inside the dataset, we have a total of 5 CSV files. It includes data for food group supply quantities, nutrition values, obesity, and underweight percentages obtained from Food and Agriculture Organization. For comparison, the final columns also show the percentages of the population that are obese, undernourished, and COVID-19 confirmed/recovery/death cases.

A list of data set

- Fat_Supply_Quantity_Data
    - The proportion of fat consumed from various food kinds in multiple nations is included in this dataset.
- Food_Supply_kcal_Data
    - The proportion of food consumption (kg) in various nations is included in this dataset.
- Food_Supply_Quantity_kg_Data
    - This dataset covers various food kinds' energy consumption (kcal) in multiple nations.
- Protein_Supply_Quantity_Data
    - This dataset provides the proportion of protein consumed by various dietary groups in various nations.
- Supply_Food_Data_Descriptions
    - This dataset, received from FAO.org, displays the precise food items that fall under each category in the previous datasets.

Covid-19 healthy dataset link:

https://www.kaggle.com/datasets/mariaren/covid19-healthy-diet-dataset?resource=download

Method

- Q1: Use the Food_Supply_Quantity_kg_Data. Compute the correlation coefficient (from scipy.stats import pearsonr) between the Covid-19 confirmed rate, recovery rate, and deaths rate columns and types of take-in ingredients columns. Draw a heatmap to help us understand and visualize.
- Q2: First find out countries with top 5 recovery rates based on Food_Supply_Quantity_kg_Data dataset. Then use the Fat_Supply_Quantity_Data and Protein_Supply_Quantity_Data to illustrate the distribution of intake of fat and protein from different ingredients, trying to figure out the relationship between intake of fat or protein and the possibility of getting away from COVID-19.
- Q3: Employ Food_Supply_kcal_Data, combining with Supply_Food_Data_Descriptions and using 2 sample t test hypothesis testing to find the significant difference between unnourished rate and recovery rate. Use the seaborn to do visualization and comparisons.
- Q4: Find the country with the highest and lowest death rate. Possible seaborn functions we would like to use: Replot
    - Calculated the covariance matrix and correlation coefficient in Fat_Supply_Quantity_Data
- Q5: Split train and test data sets, train each model by using training sets, tune models by different parameters to find the best parameters of the model. Do predictions and plot the

actual data with predic data, calculate the accuracy and the mse. Conclude the best tuning model.

Results

Q1: What are the top three factors contributing to people getting COVID-19 in terms of diet?

To figure out the top three factors contributing to people getting covid-19, we use the correlation coefficient to represent how strong the connection between different diet ingredients and confirmed cases, recovery rate, and deaths Heatmap helps us to better illustrate the pairwise correlation coefficients. Based on the result, we find out that the top three factors are milk, animal products, and animal fats. It is surprising that milk would contribute to covid-19.

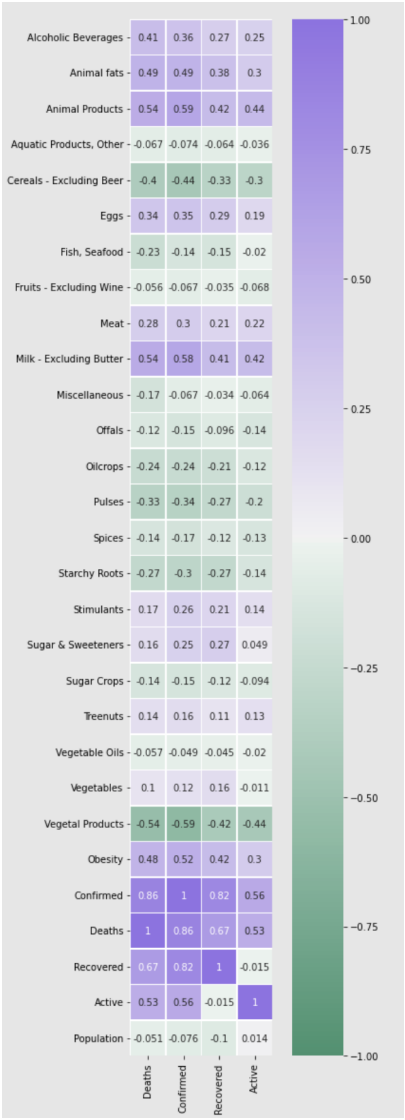


Figure. Correlation Coefficient Heatmap

With further research, there are some other paper indicating its association “higher intake of high-fat-dairy-product (OR: 1.40 CI: 1.09–1.92, p-trend = 0.03), high-fat milk (OR: 1.54 CI: 1.20–1.97, p-trend < 0.001), total yogurt (OR: 1.40 CI: 1.04–1.89, p-trend = 0.01), cheese (OR: 1.80 CI: 1.27–2.56, p-trend = 0.001), and butter (OR: 1.80 CI: 1.04–3.11, p-trend = 0.02) were related to increase the odds of COVID-19.” The paper indicates that different dairy products might lead to different degrees of getting COVID-19, and high-fat-dairy-products may increase

the probability of COVID-19 while moderate intake of total dairy and higher intake of low-fat milk had a protective effect on COVID-19. Since our data resource does not provide a specific description of the variety of dairy product, our result is actually reasonable though unexpected. Moreover, we also write test code to justify our research results. We randomly selected 80 percent of the data and repeated the procedure again. According to the plot we derive, we can conclude that the top three factors are still milk, animal product, and animal fats, indicating our conclusion is reasonable.

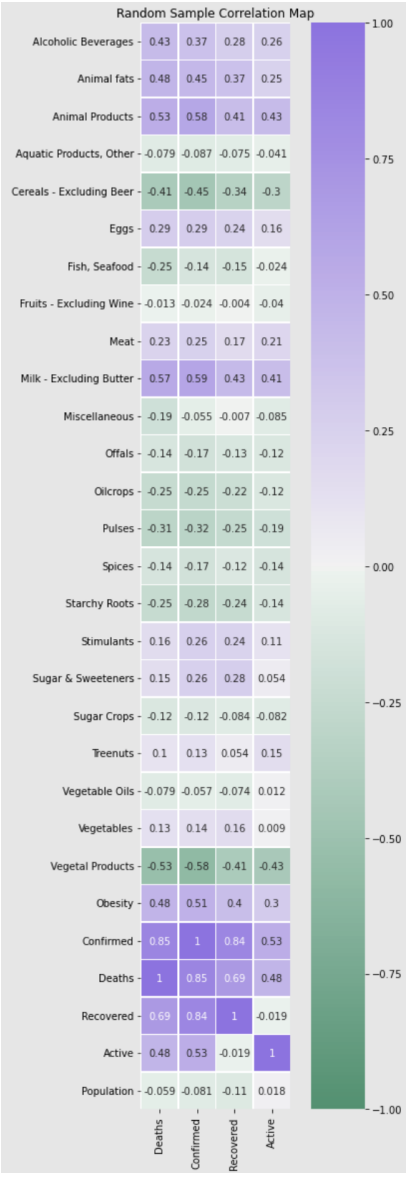


Figure. Random Sample Correlation Map

Q2: Does intake of more fat or intake of more protein help people get away from COVID-19?

From the data set Food_Supply_Quantity_kg_Data, we first find out the five countries with higher recovery rates.

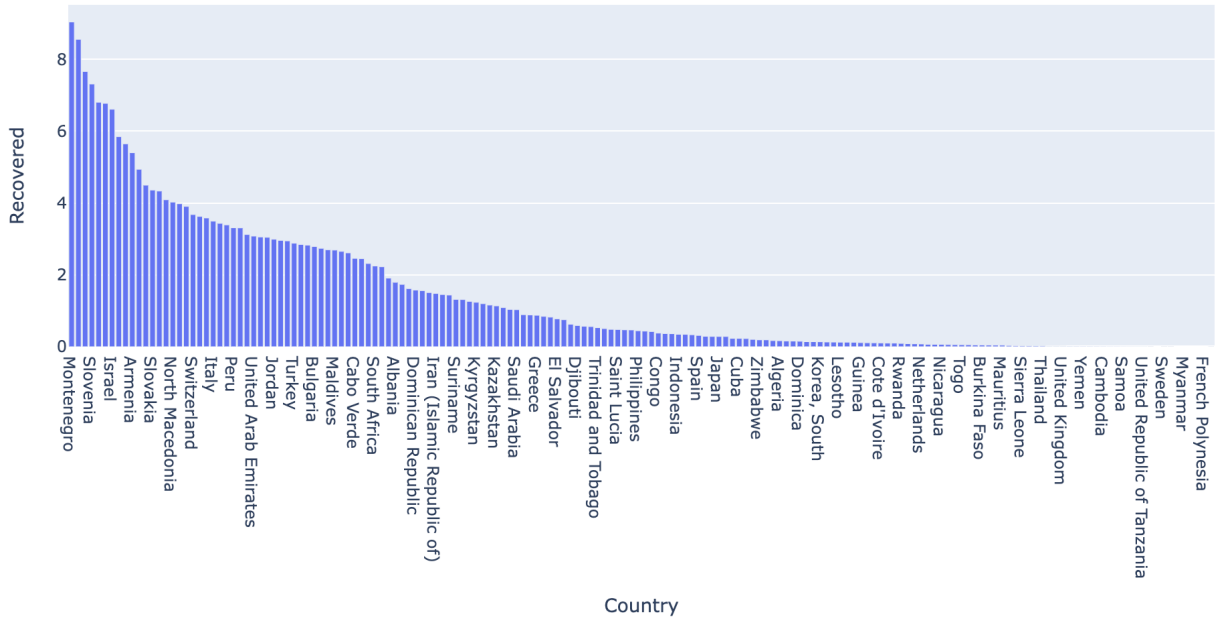


Figure. Recovered Rate of Every Country

Then for these five countries, we explore deeper into their fat and protein intake distribution with two given separate data sets Protein_Supply_Quantity_Data and Fat_Supply_Quantity_Data. Utilizing pie charts, we can clearly see the intake distribution of various food kinds. Combining the top three factors (milk, animal products, and animal fats) found in Q1 and the intake distribution plotted in Q2, we can know that all three factors contribute slightly more to fat intake than protein intake, which imply that fat seems to be more helpful to covid-19 recovery.

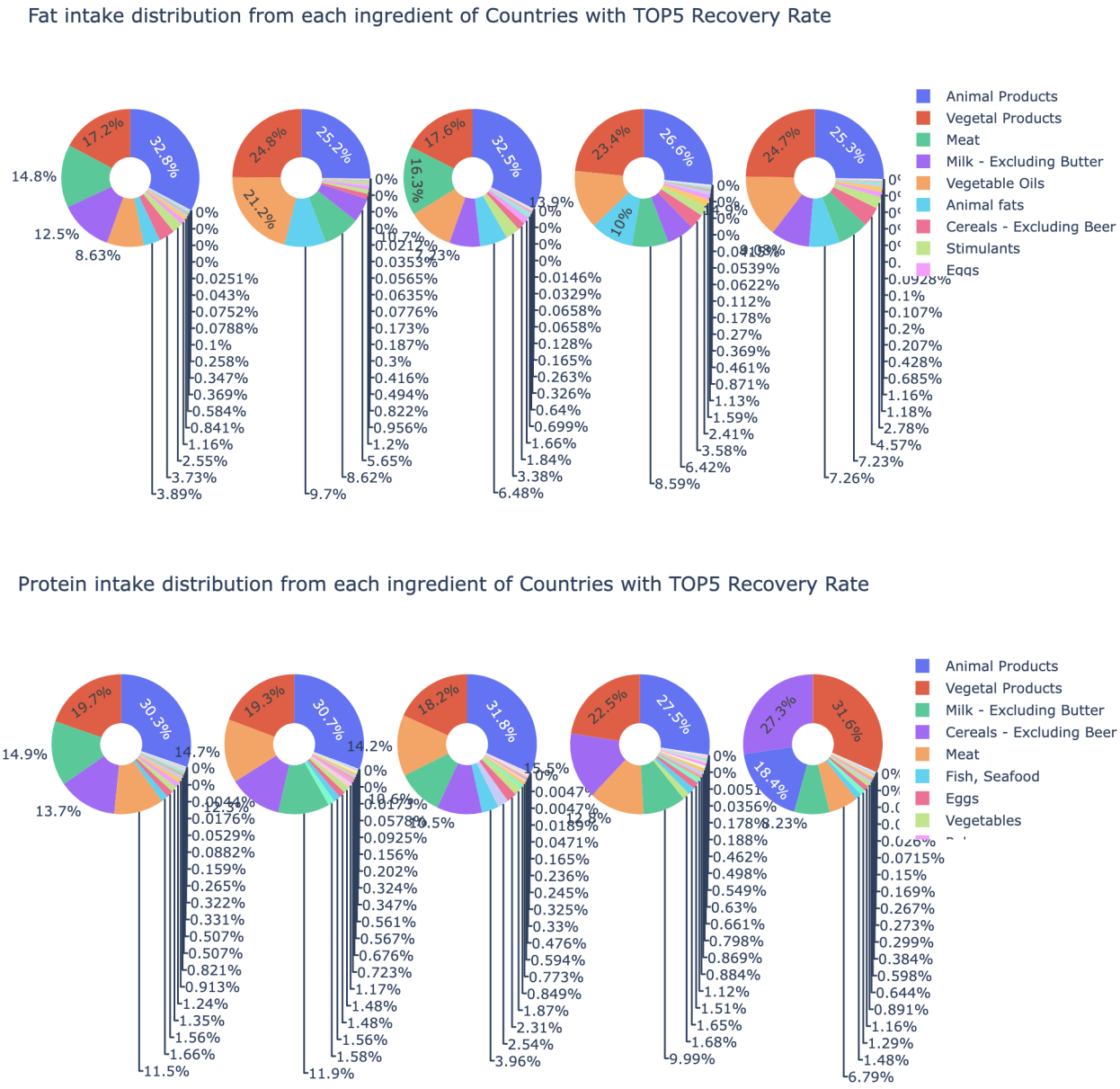


Figure. Pie Charts of fat and protein intake of top5 recovery rate countries

The result shocked us to some degree. Furthermore, we then find out the top 5 countries with higher deaths rate and explore their fat and protein intake distribution. We surprisingly found out the result is opposite to the previous result obtained from the top 5 recovery rate countries, which implies our previous conclusion was correct. Intake of fat helps people get away from covid-19.

Yet when we investigated more relationships between fat and COVID-19, we found that “Having obesity increases the risk of severe illness from COVID-19. People who are overweight may also be at increased risk”(“Obesity”, 2022) Which means that fat is not helping with the recovery but further deteriorates the disease. But based on our research result, we conclude that a small intake of fat that might help with COVID-19 recovery, but too much of fat intake will cause irretrievable deterioration to the disease.

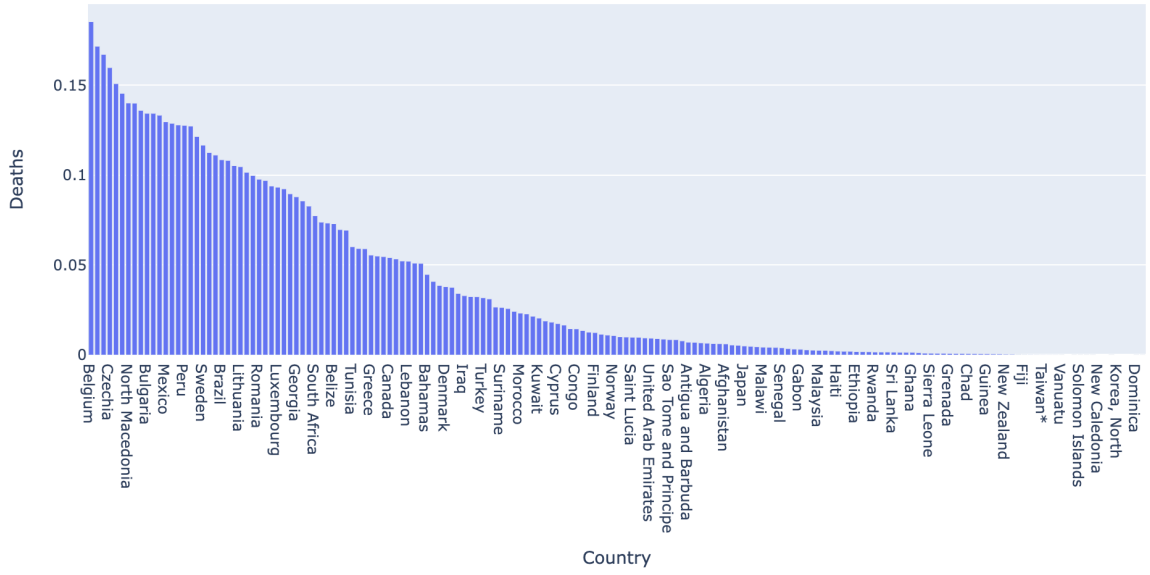


Figure. Death Rate for Every Country

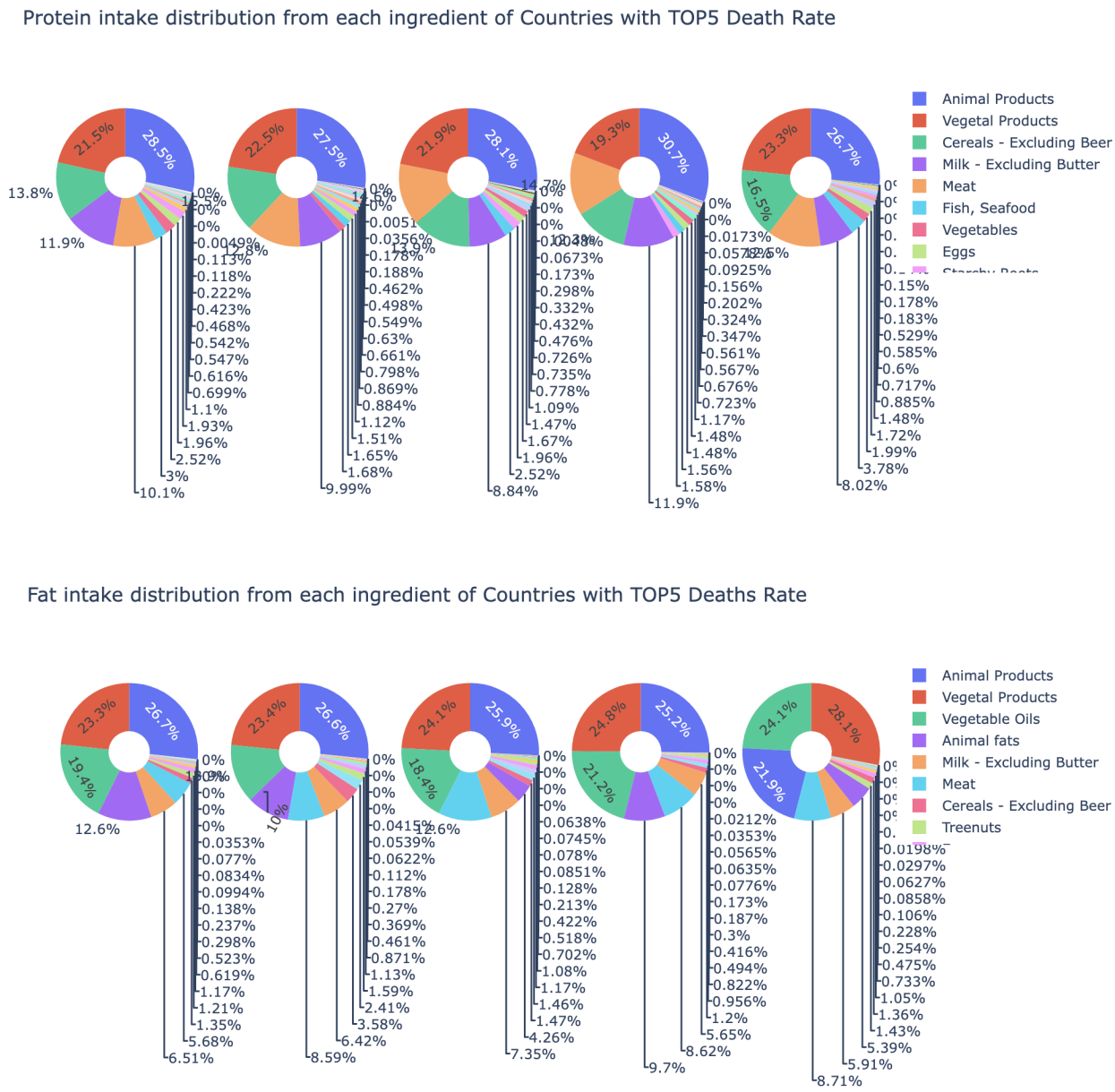


Figure. Pie Charts of fat and protein intake of top5 death rate countries

To further justify our result, we also write several test codes by randomly selecting 80 percent of the original data and taking the same operations. The result of sample test data is also consistent with previous results based on the original data set.

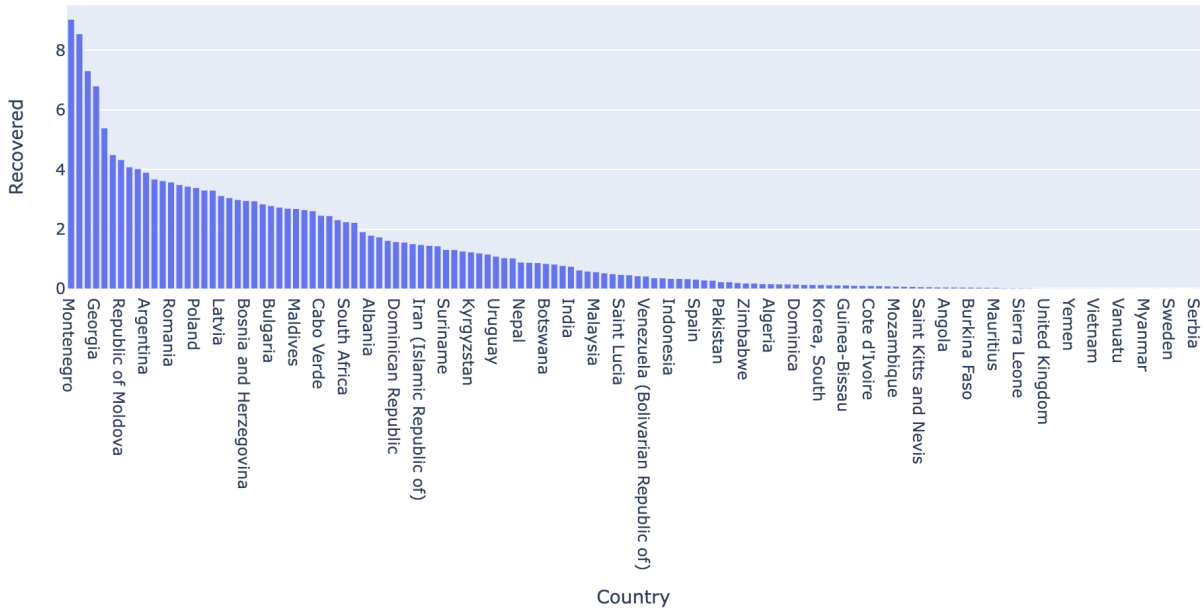
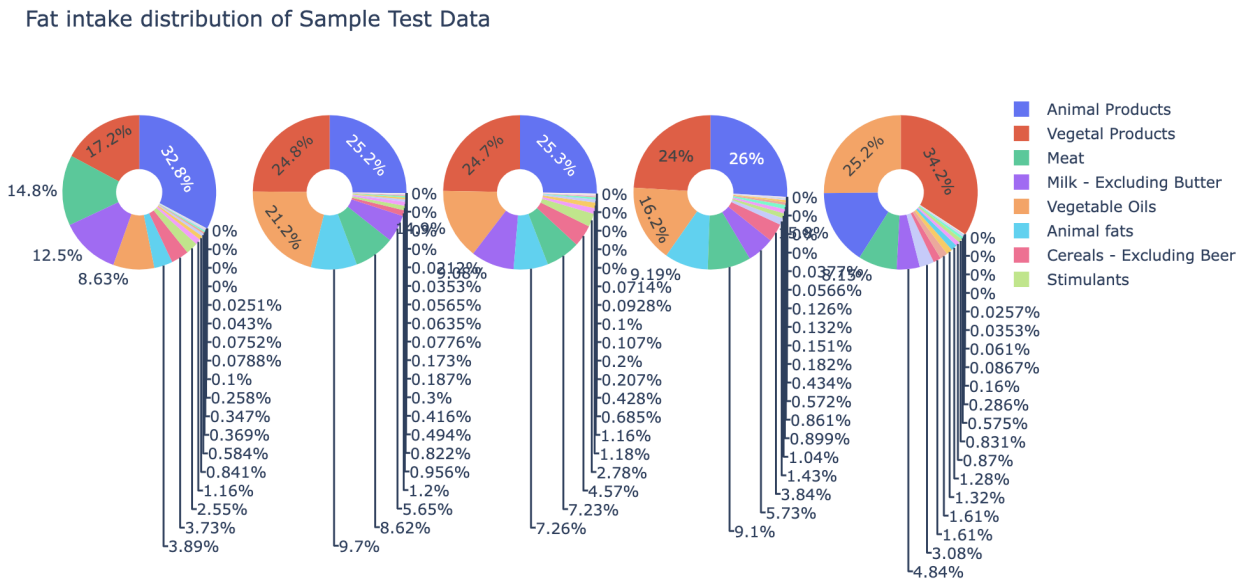


Figure. Recovery Rate for Every Country



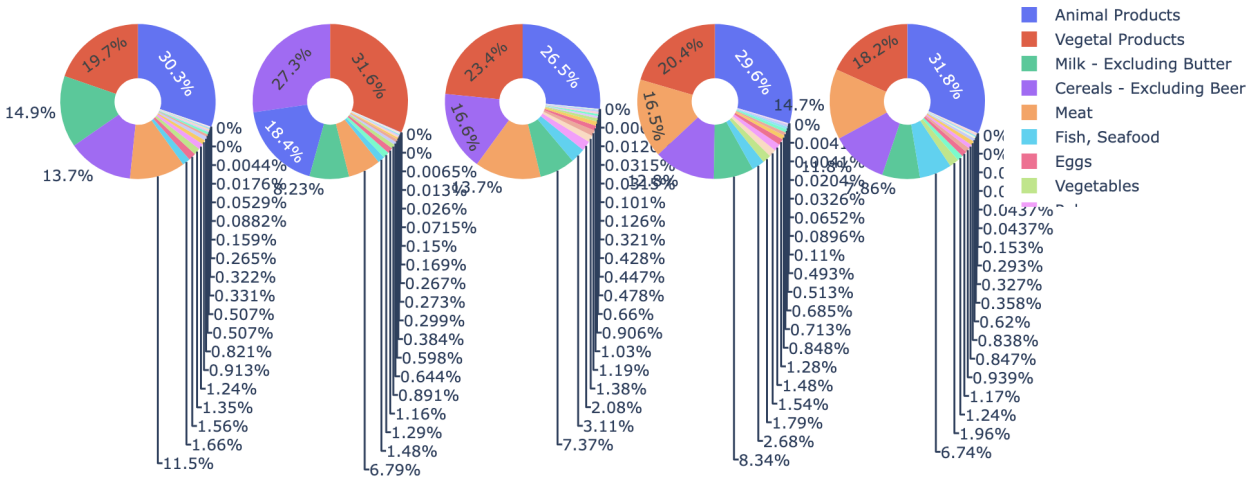


Figure. Pie Charts of fat and protein intake of top5 death rate countries for test

Q3: Is there a significant difference between the eating habits in countries with a higher recovery rate and countries with a lower recovery rate?

Hypothesis Statement:

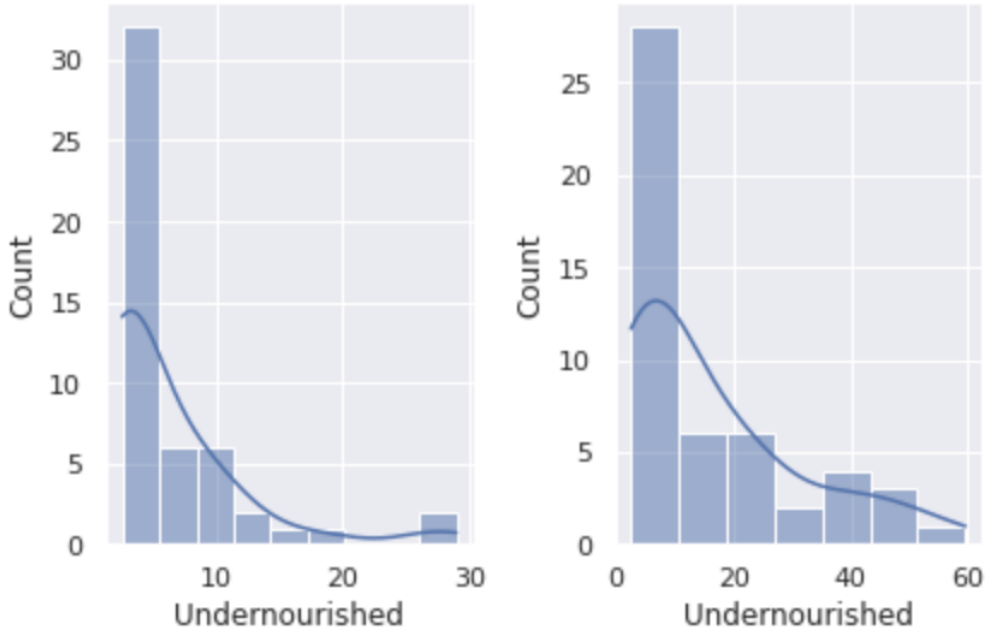
H0: μ1 = μ2 : there isn't a significant difference between the eating habits in countries with a higher recovery rate and countries with a lower recovery rate

H1: μ1 ≠ μ2: there is a significant difference between the eating habits in countries with a higher recovery rate and countries with a lower recovery rate

We select the Food_Supply_kcal_Data file as our input data. Based on the data in two numerical groups with independent random samples, we decided to use 2 sample 2 tests to conduct our hypothesis testing. And calculating the statistic of p-value and using 0.05 for our significance

level, we conclude that the p-value of 0.000363 is smaller than the significance level of 0.05, meaning we should reject the null hypothesis. We have enough evidence to support that there is a significant difference between the eating habits in countries with a higher recovery rate and countries with a lower recovery rate. This is not a surprising result since in common sense we know that people's intake of nutrition is related to human's health. And if we do not intake enough nutrition, it will lead to a bad health of our body. Which means that our immune system will be weak, and it will increase the chance of a person to get COVID-19 and increase the chance that they will have a long time to recover or die.

However, by choosing the data from Food_Supply_kcal_Data file and using a random generator to select 50 data, the condition check does not pass the normal condition test.



Where we should pass the normal condition test that the data should be normally distributed, the graph is actually skewed to the right. Yet the graph did make sense since the world is getting better so the undernourished countries are getting lesser. We still decided to conduct the 2 sample

t tests. This failed normal condition check might be one of the main leading reasons of causing p values too small that reject the null hypothesis.

Q4: Does there appear to be a relationship between the death rate from COVID and the amount of intake from fat?

To find the relationship between multiple variables in the dataset of foods' fat containing the death rate, using visualization is a good way to show our result. There are two main ways to help us consider the variable's relationship: correlation coefficient and covariance. The correlation coefficient is the specific measure that quantifies the strength of the linear relationship between two variables in a correlation analysis. Different from correlation, covariance is an indicator of the extent to which two random variables are dependent on each other. A higher covariance means a higher dependency.

Here is the full heatmap that recorded the variable's coefficient correlation across every variable:

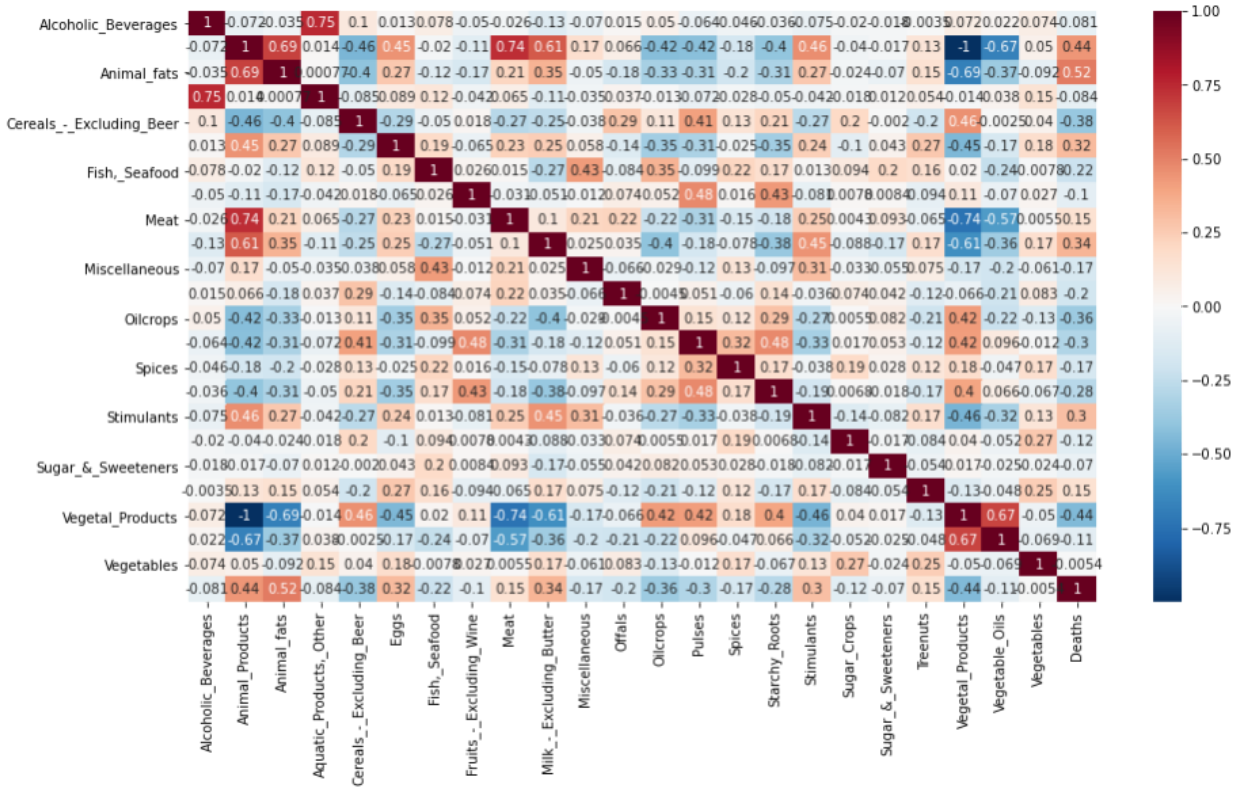


Figure. Heatmap of correlation of fat intake food

To get the more detailed relationship between variables with death rate, we can only focus on calculating the correlation coefficient of each type of food fat intake with the death rate, we can get the following result in this heatmap:

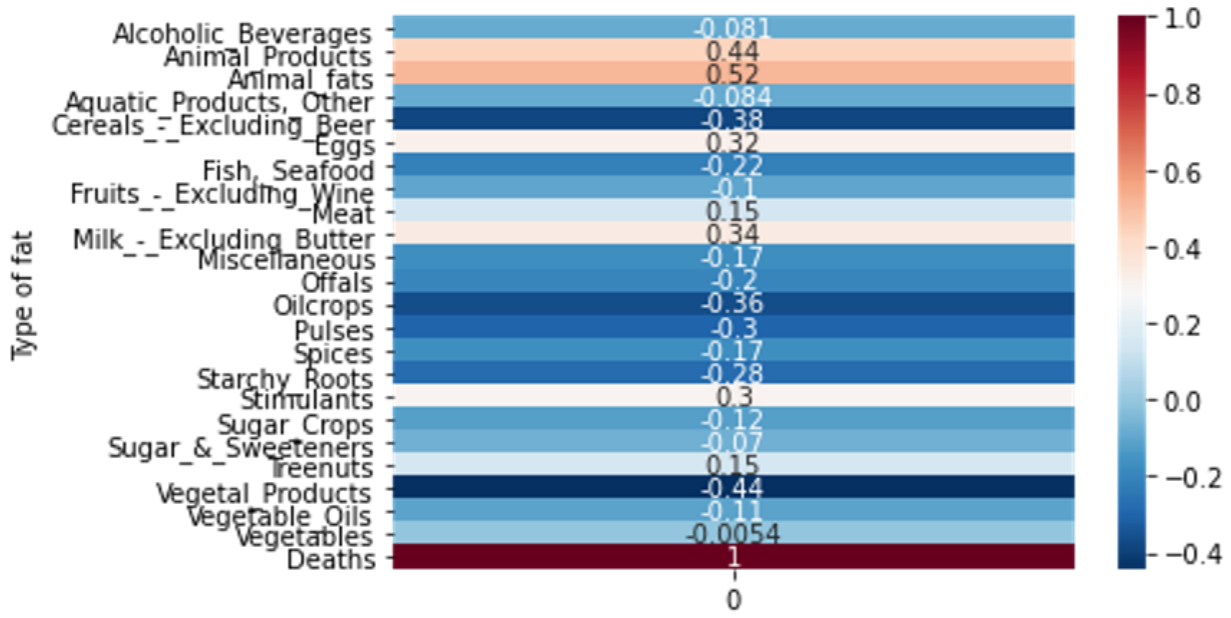


Figure. Heatmap of correlation of fat intake food with death rate

From the above heatmap figure, we can find the variable with the highest positive correlation between Death rates is Animal Fats, and the variable with the second highest correlation is Animal products. The most negative correlation variable with death rate is Vegetal Products. Since the heatmap can only show the general result, if we want to explore more detail, we use the scatter plot to show the trend between each variable with the death rate.

Here we use the seaborn.pairplot to show every column's data inside the scatterplot. Moreover, the diagonal in this pairplot will show the marginal distribution of the data in each column.

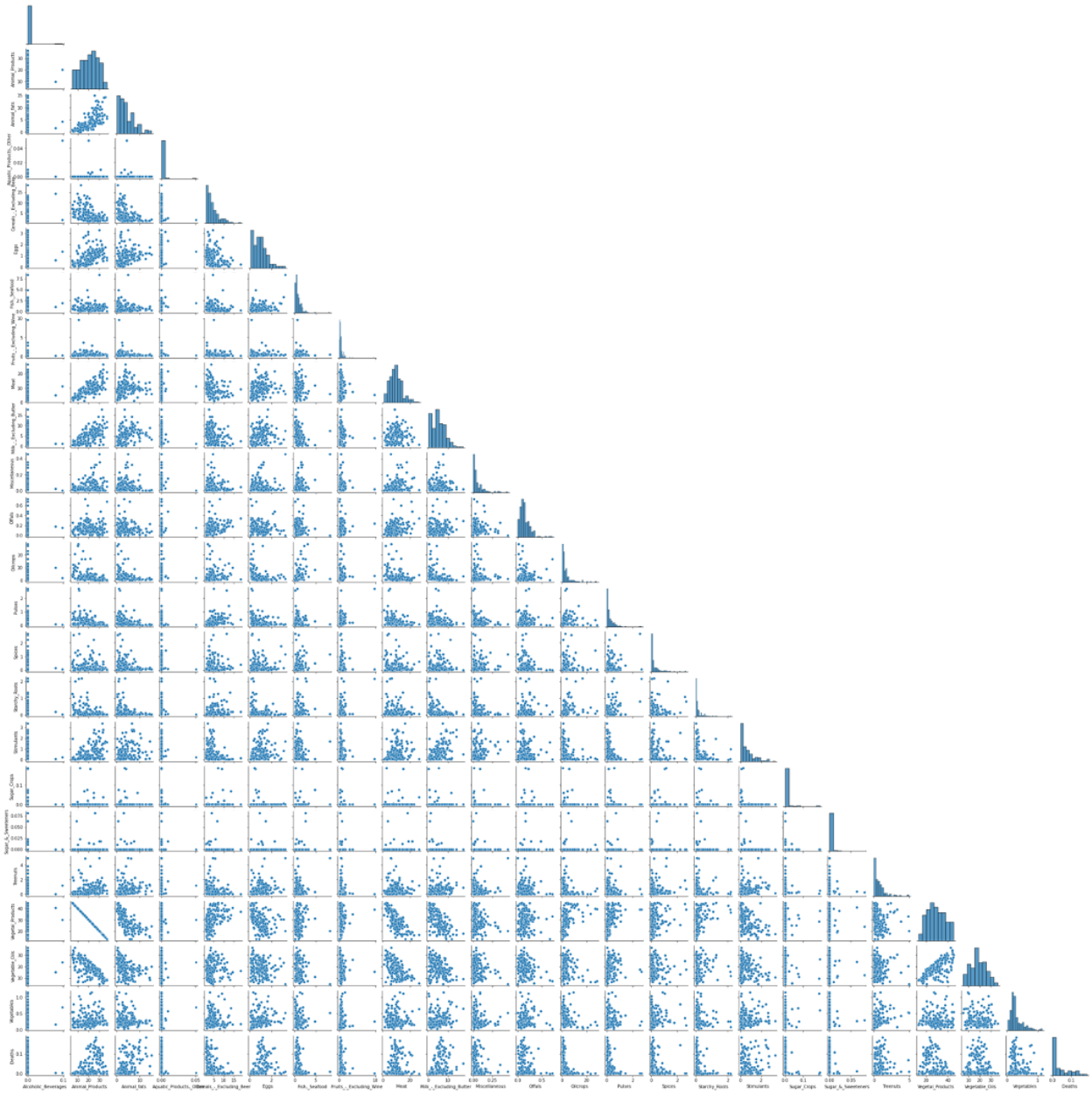


Figure. Pairplot of correlation of fat intake food

As we zoom in, we can see the Animal fat and Animal products, as well as Vegetal Products scatterplots with Death:

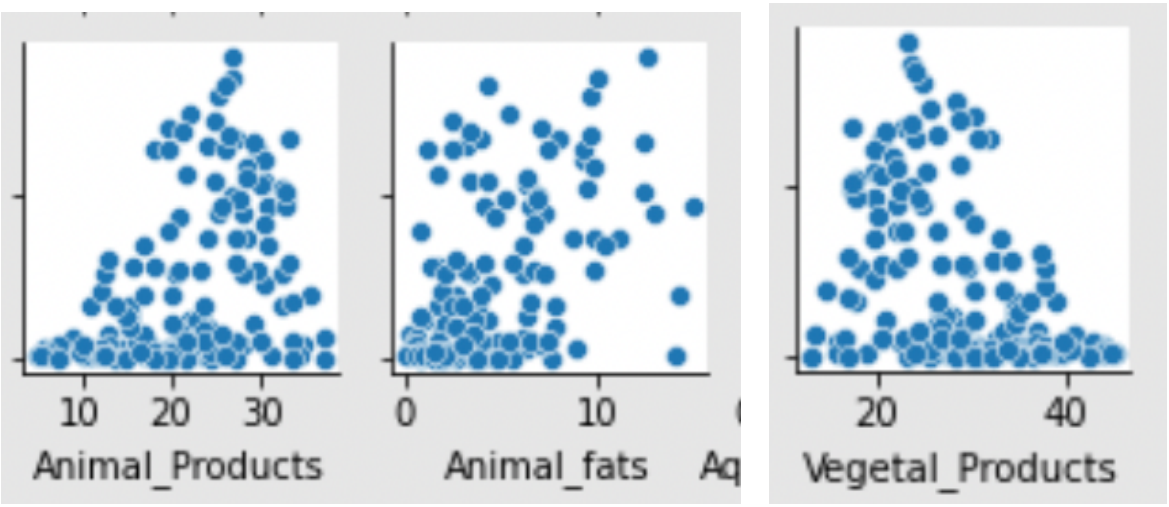


Figure. Zoom in scatterplots for animal products, animal fats, and vegetable products

As the above creates a correlation matrix between each variable, we can see that Animal fats have the strongest positive relationship with the death rates. Vegetable Products have the strongest negative relationship with the death rates.

Furthermore, we explored the variance between several types of fat intake with death rates. I selected animal fat, animal products, cereal, and vegetable products. Put all variables I selected with the death rate into an array. Then apply the covariance function. Here is the covariance matrix:

[[[1.09177015e+01 1.80362208e+01 -4.30264620e+00 -1.80352002e+01
8.33423992e-02]
[1.80362208e+01 6.36630471e+01 -1.18430937e+01 -6.36605504e+01
1.71100948e-01]
[-4.30264620e+00 -1.18430937e+01 1.03412043e+01 1.18425187e+01

-5.90334961e-02]

[-1.80352002e+01 -6.36605504e+01 1.18425187e+01 6.36580722e+01

-1.71109854e-01]

[8.33423992e-02 1.71100948e-01 -5.90334961e-02 -1.71109854e-01

2.35900530e-03]]

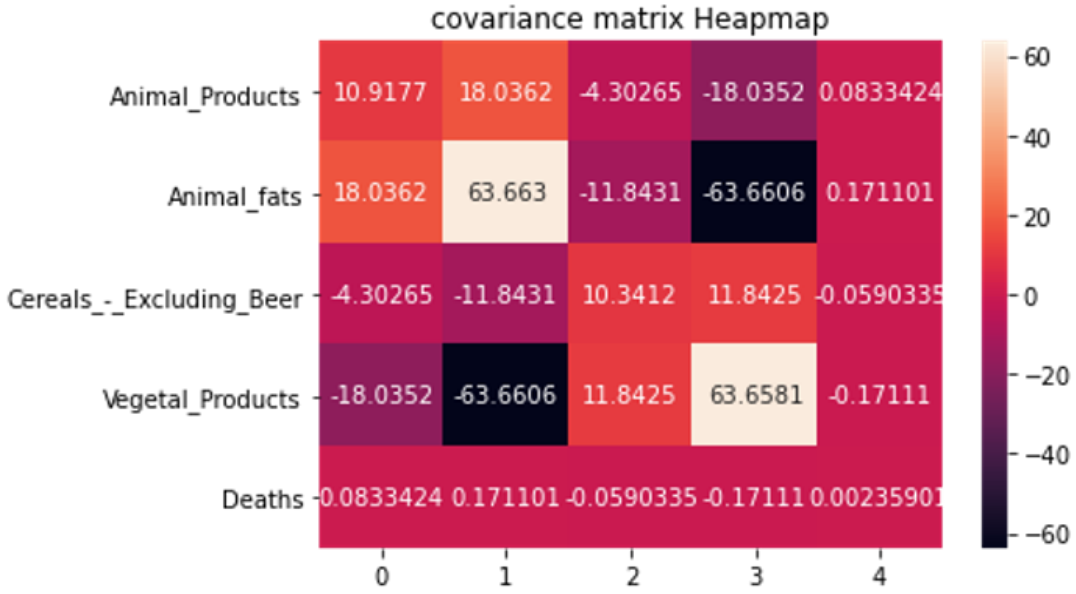


Figure. Heatmap for covariance matrix

As above, the most dependent variable with death rate is animal fats.

To see the trend of animal fat with death rate, we draw another scatterplot:

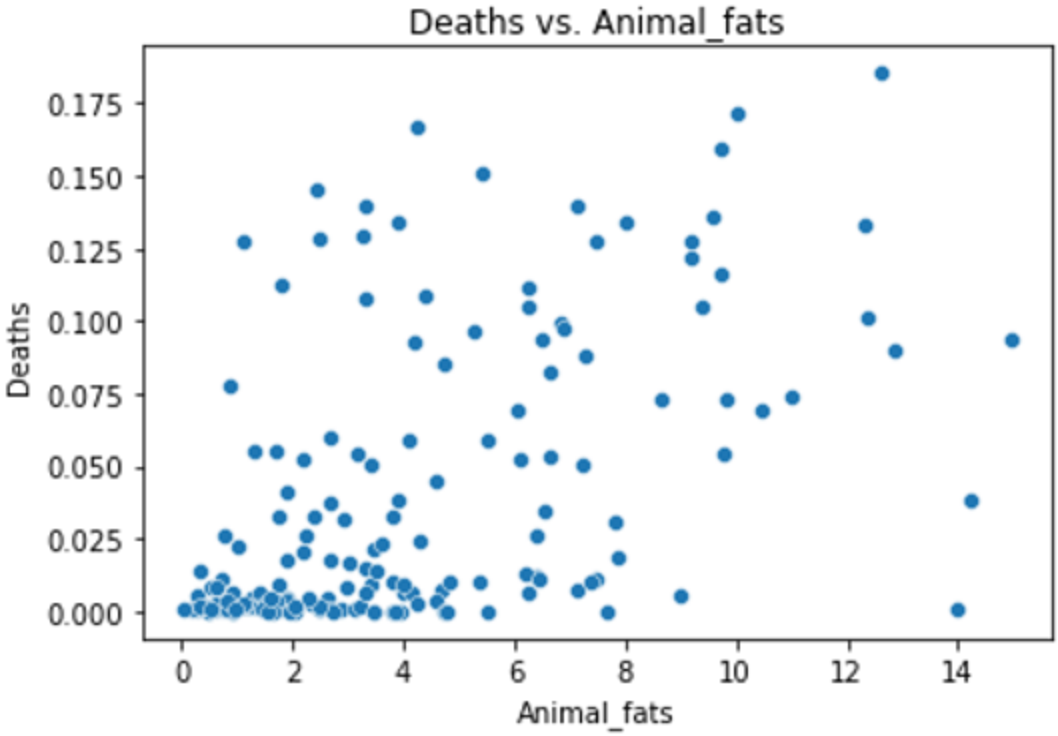


Figure. Deaths vs. Animal fats intake

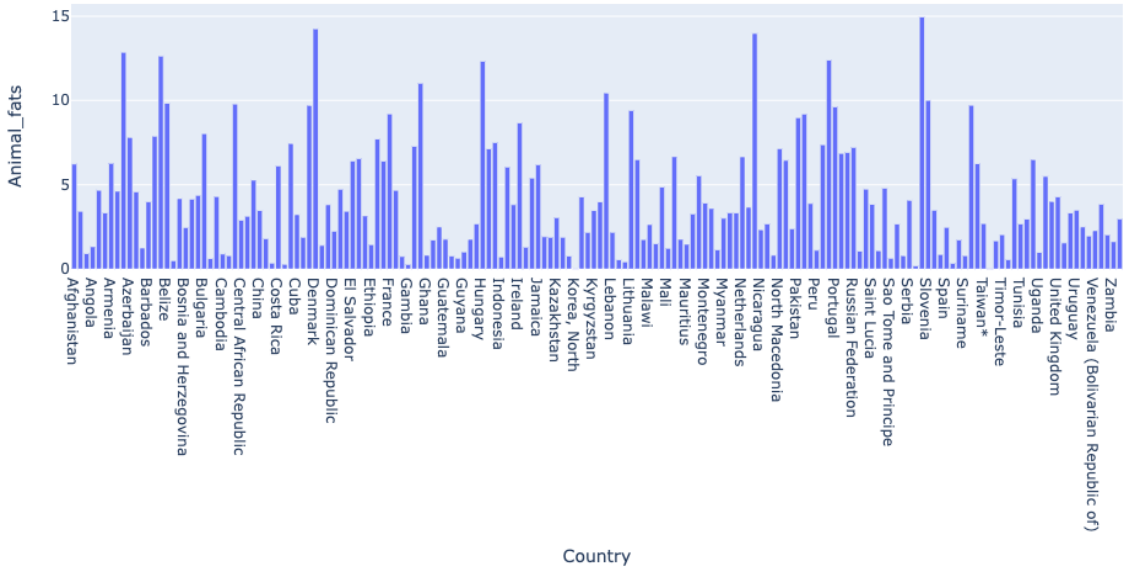


Figure. The animal fat intake in every country.

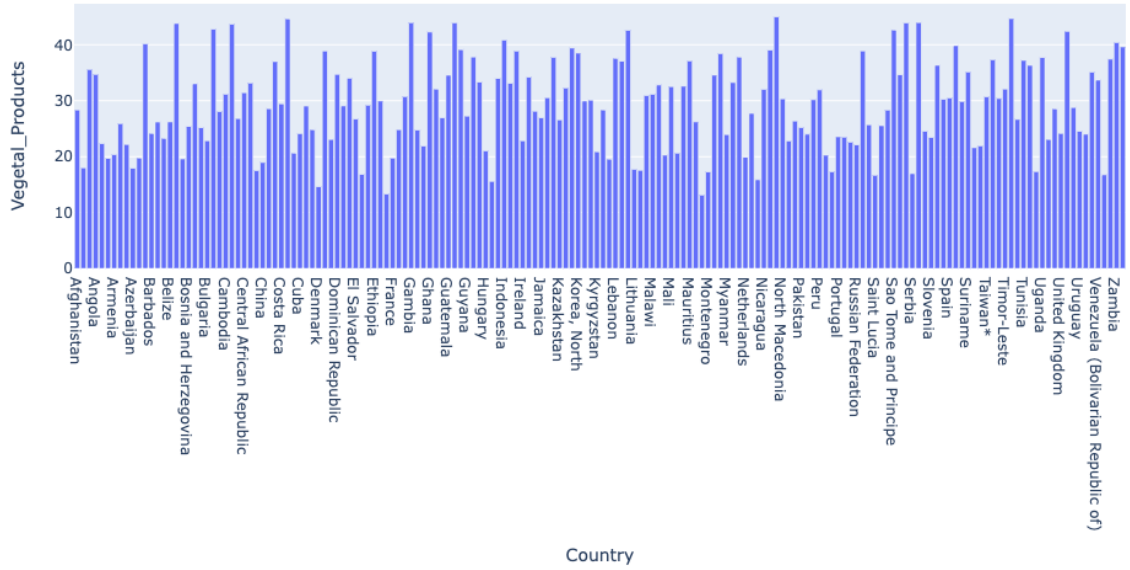


Figure. The vegetable product intake in every country

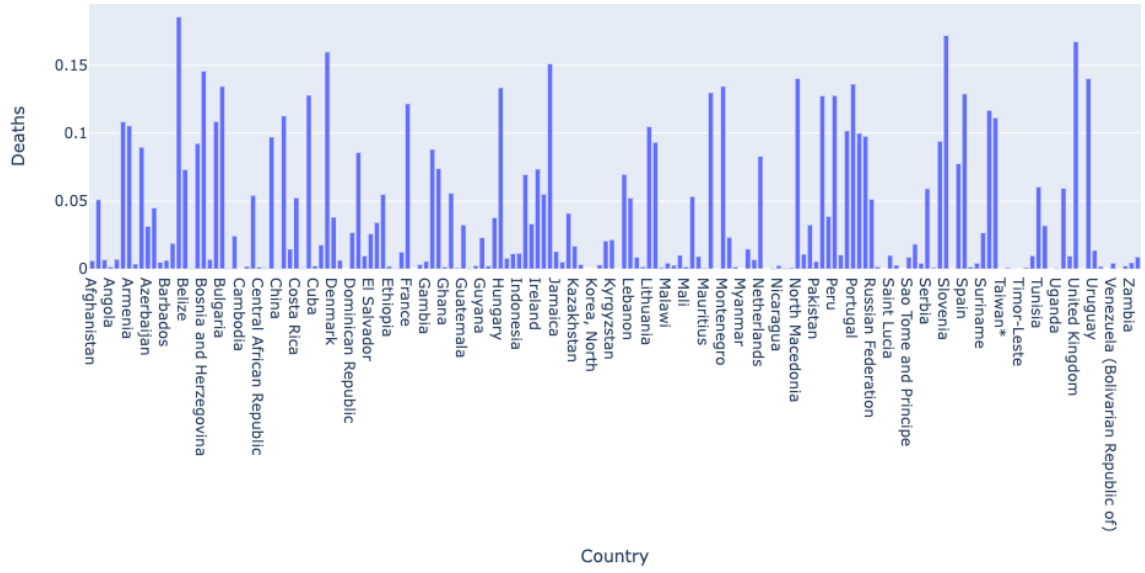


Figure. The death rate of every country

Based on the above two plots, we can see the country with higher animal fat intake owns a higher COVID death rate.

The regression relationship of Animal fats and Vegetable products in detial as below:

names	coef	se	T	pval	r2	adj_r2	CI[2.5%]	CI[97.5%]
Intercept	0.152237	0.029294	5.196850	6.475595e-07	0.331943	0.323094	0.094358	0.210116
Animal_fats	0.041266	0.015043	2.743264	6.819910e-03	0.331943	0.323094	0.011545	0.070988
Vegetal_Products	-0.003214	0.000708	-4.542058	1.132230e-05	0.331943	0.323094	-0.004612	-0.001816

Q5: Use multiple machine learning models to find the best tune model by comparing the predicted accuracy and mean squared error. Finding and predicting my diet habit's recovery rate.

Use Decision tree Regressor, Random Forest Regressor.

Since we want to build a model of predicted recovery rate, we set the label y is the "Deaths", and the features x is the data without "Deaths'. Since we do not have classified variables, we want to predict the death rate which is also the numeric variable. We definitely want to use Regression rather than Classification.

The first step is to drop all rows with NA value. Then, split our data into a training set and a test set. I set 30% to be a test set and 70% to be a training set. Our training set shape is (107, 23), and the test set shape is (47, 23).

The Decision tree Regressor

Build the decision tree and use the training set to fit the model. Checking the difference between the labeled y and predicted y without tuning the model:

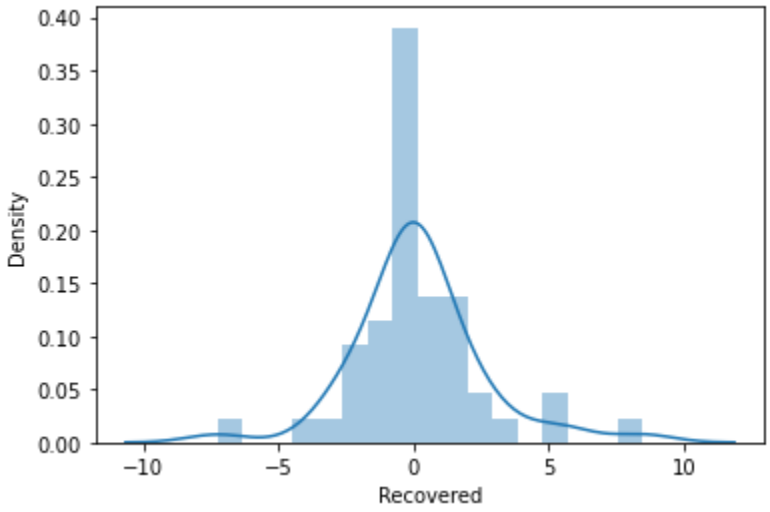


Figure. Range of prediction value and original range

The good bell curve only tells us the range of predicted values is within the same range as our original data range values are.

As we know, if we get a better prediction, the test result and prediction result should be close enough. Therefore, in a scatter plot, if points are almost uniformly distributed closely with the y=x line, the result would be good.

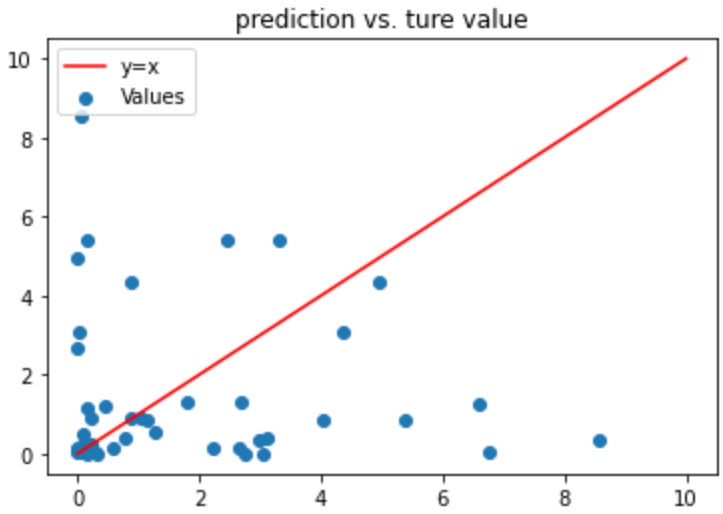


Figure. The scatter plot of prediction and true value before tuning Decision Tree Regressor

Before tuning, we got a 100% score on training data. On test data we got a -24.46% score. A negative score just means that the particular model is performing quite poorly. The reason for this is we did not provide any tuning parameters while initializing the tree as a result of which the algorithm split the training data till the leaf node. Due to this the depth of the tree increased and our model did the overfitting.

So, we use hyperparameter tuning by Gridsearch to find the best parameters for our decision tree model. Then we can get the best hyperparameters for our decision tree model are: {'max_depth': 5,

'max_features': None,

'max_leaf_nodes': 40,

'min_samples_leaf': 5,

'min_weight_fraction_leaf': 0.1,

'splitter': 'random'}

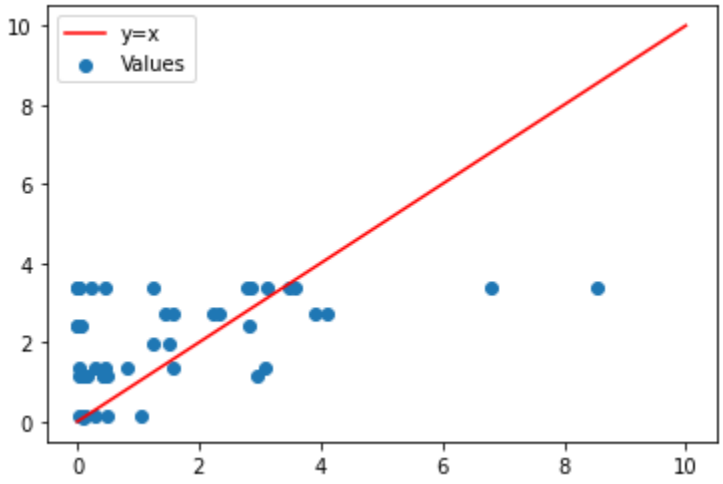


Figure. The scatter plot after hyperparameter tuning Decision Tree

Here we can see the above scatter plot looks a lot better.

We also can compare the Error rate of our model with the hyper tuning of parameters to our original model which is without the tuning of parameters.

Before tuning, we have:

MSE: 4.170436138299242

RMSE: 2.042164571796123

accuracy score: -0.24455965232483234

After tuning, we have:

MSE: 3.073229627619313

RMSE: 1.7530629274556326

accuracy score: 0.08287347653186317

So, the tuning model performs clearly better.

Random Forest Regressor

To find a better random forest predicting model, we also need to use hyperparameter selection. Here we use RandomizedSearchCV to select our hyperparameters. We select parameters of 'n_estimators', 'max_features', 'max_depth', 'min_samples_split', 'min_samples_leaf', and 'bootstrap'.

Our best parameters are:

{'n_estimators': 20, 'min_samples_split': 10, 'min_samples_leaf': 4, 'max_features': 'auto', 'max_depth': 40, 'bootstrap': True}

Here is our scatterplot of prediction value with true value

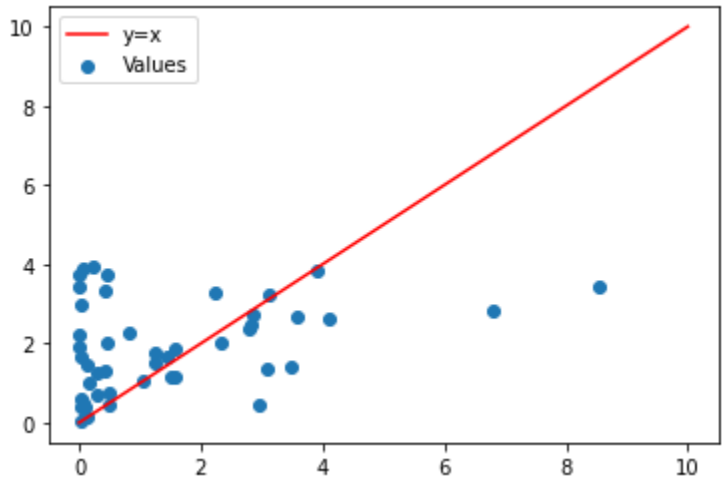


Figure. The scatter plot after hyperparameter tuning Random Forest

With this model, we can also estimate using mean squared error and mean squared error root:

MSE: 3.6061338216070937

RMSE: 1.8989823120837892

accuracy score: 0.23386092370459166

As above, the tuned Decision tree model is the better predict model. The best model's hyperparameter is:

{'max_depth': 5,

'max_features': None,

'max_leaf_nodes': 40,

'min_samples_leaf': 5,

'min_weight_fraction_leaf': 0.1,

'splitter': 'random'}

Then use the above tuned decision tree model predict Wendy:

	Wendy's
Alcoholic_Beverages	3.5000
Animal_fats	0.2212
Animal_Products	12.0000
Aquatic_Products,_Other	0.0000
Cereals_-_Excluding_Beer	8.0666
Eggs	0.7792
Fish,_Seafood	3.2750
Fruits_-_Excluding_Wine	5.8723
Meat	5.8477
Milk_-_Excluding_Butter	2.2041
Miscellaneous	3.3020
Offals	0.1704
Oilcrops	0.8734
Pulses	0.6391
Spices	0.1565
Starchy_Roots	4.0812
Stimulants	0.1721
Sugar_&_Sweeteners	5.3344
Sugar_Crops	0.0000
Treenuts	0.0852
Vegetable_Oils	0.8677
Vegetables	5.4725
Vegetal_Products	37.5167

As the decision tree model with hyperparameters prediction, Wendy got her recovery rate is 1.35789416%. Here may be a concern of the result, due to our dataset the recovery rate if for a whole country. So using the model to predict a single person may not be accurate. As the median recovery rate in all countries, the median recovery rate is 0.476994136691255% .

Hence, Wendy's recovery rate is significantly higher than the median recovery rate of all countries in our dataset. So, we can conclude Wendy has a food intake habit that is beneficial for COVID recovery.

Impact and Limitations:

The results of our research can be used to advise for food organization, or for specific reference when hospitals are matching meals to patients with COVID-19. It still implies people to intake food in an appropriate range, rather intake too much, too low, or only a few varieties in order to have a high recovery rate from COVID-19 and low possibility of death from COVID-19. However, this research has only some reference value, and it's result cannot compare with professional academic research results. Since this paper is based on Kaggle's COVID-19 Healthy Diet Dataset by Maria Ren and was updated 2 years ago, we suspect that the data may be a bit off from today's data. Although the authors state in their statement that the dataset is based on data from the Food and Agriculture Organization of the United Nations, Population Reference Bureau, and Johns Hopkins Center for Systems, we are not 100% sure that the authors did not omit or modify any data in the middle of the transformation. On the other hand, the data itself only contains a few variables for 170 countries, and each country only has one line of summarized data, which means the data is generalized rather than given in detail. We need further investigation for more accurate results from all the countries, even interviews with people who live in the country. Therefore, this research is not applicable to academic research, but only as a referencial resource used by such as non-profit organizations for daily food intake advice.

Challenge Goals

For implementation, using below challenge goal:

- Use different plot library(plotly.express and plotly.graph_objects)
- Scipy.stats to get covariance, correlation coefficient and 2 sample t test
- Tuning machine learning model
- Multiple dataset

Same as the challenge goals in our proposal, we did employ and explore multiple datasets to accomplish our research. To provide a clearer and more understandable visualization of the result, we not only utilize the plot functions we learned from courses, but also we import new plot libraries which are not covered in the lessons. To figure out the association between diet ingredients and COVID-19, we use different functions from various libraries to compute those related statistics. In addition, we utilize RandomizedSearchCV for Random Forest module tuning and Gridsearch for Decision Tree model tuning to find the best hyperparameters of each model. With the hyperparameter, we can improve our model's prediction accuracy.

Work Plan Evaluation

Following the work plan instructions in our proposal, we completed this final project step by step. We updated our understanding of the dataset and had much clearer methods for each research problem during preparation. We spent a lot of time working on processing data, coding, research and interpretation. As we started to interpret the result we derived, we also did some further research, such as reading other papers with relevant topics, to justify and support our result. We think our proposed work plan is fairly accurate and close to reality since we have reliable data resources, deliberate processing procedure, and credible academic research support.

Testing

For Q1, we randomly select 80 percent data from the original dataset and compute pairwise correlation coefficients of the selected dataset. Comparing the result of the original dataset and the sample dataset, we obtain the conclusion that our report is correct and we can trust the result. For Q2, we repeated the test procedure as for Q1. We also randomly selected part of the original data and repeated the data processing.

For Q5, split the test set from our original dataset by train_test_split with test_size=0.3, which is 30% of the whole dataframe. We did not use other data files as test files, our test set included inside the main dataset. We did not use assert statements either. We use a test set for machine learning models to do prediction and get the prediction accuracy and mean squared error to estimate our machine learning models. Moreover, to make sure our generated result is correct, we select using plot as evidence to support our claim and conclude results.

Collaboration State

This project is completed mainly by Sihao Feng, Wendy JIang, and Bianca Xie with help of University of Washington CSE 163 summer 2022 staff.

Research Citation:

Darand, M. (2022, April 29). *The association between dairy products and the risk of COVID-19*. Nature. https://www.nature.com/articles/s41430-022-01149-8?error=cookies_not_supported&code=be1d9392-606c-4f82-be18-c4e2eb6f94a0#:%7E:text=Our%20finding%20indicated%20that%20moderate,protective%20effect%20on%20COVID%2D19

Obesity, Race/Ethnicity, and COVID-19. (2022, May 20). Centers for Disease Control and Prevention. https://www.cdc.gov/obesity/data/obesity-and-covid-19.html#:%7E:text=Adults%20with%20excess%20weight%20are,COVID%2D19%20infection.

Code citation:

[Question 4 & 5]

Random Forest:

Arjunprasadsarkhel. (2021, August 18). *Simple random forest with hyperparameter tuning*. Kaggle. Retrieved August 15, 2022, from https://www.kaggle.com/code/arjunprasadsarkhel/simple-random-forest-with-hyperparameter-tuning

[Question 3 code]

How to conduct 2 sample t test by using python?

https://www.marsja.se/how-to-perform-a-two-sample-t-test-with-python-3-different-methods/

graphs:

https://stackoverflow.com/questions/6541123/improve-subplot-size-spacing-with-many-subplots-in-matplotlib

Random Sample:

https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.sample.html

[Question 1& 2 code]

https://seaborn.pydata.org/generated/seaborn.heatmap.html#seaborn.heatmap

https://plotly.github.io/plotly.py-docs/generated/plotly.graph_objects.Pie.html

https://plotly.com/python/creating-and-updating-figures/

https://seaborn.pydata.org/generated/seaborn.diverging_palette.html

With other in class knowledge taken with instructor Wen Qiu in CSE 163, Summer 2022.