**Comparison of Machine Learning Models to Predict COVID Outcomes using Food Consumption Data**

BIOS 635 - Introduction into Machine Learning

Alexis Payton, Eileen Yang, and Zhitong Yu

May 4, 2021

**Abstract**

**Purpose** - We compared the performance of different supervised machine learning methods on predicting cases and deaths in each country using diet data, and used unsupervised machine learning methods for the same data to explore further dietary indications of COVID-19 cases and deaths. Through the analysis, we hope to see whether diet can be a good predictor of COVID-19 cases and deaths.

**Methods** - We used supervised algorithms including k-nearest neighbors (KNN), linear regression, penalized regression, tree-based methods, and support vector regression (SVR) to predict continuous outcomes, confirmed cases per capita and deaths per capita. We also predict dichotomous outcomes, whether one country’s confirmed cases per capita are above the median cases per capita across all countries and whether one country’s deaths per capita are above median deaths per capita. Methods used for the classification include quadratic discriminant analysis (QDA), logistic regression in addition to penalized regression, tree-based methods, and SVR. As for unsupervised learning methods, k-means and principal component analysis (PCA) are used for both regression and classification. PCA and k-means, in particular, are used to improve the performance of KNN by reducing bias.

**Results** - Based on the 10-fold cross validation errors, lasso is the method with best performance across supervised regression and classification. Unsupervised machine learning tends to have higher errors than the supervised methods.

**Conclusion** - Based on our findings, diet may not be the best sole predictors to COVID-19 cases and deaths. Further analyses are needed to better understand factors contributing to COVID-19.

**Keywords**

Supervised machine learning, unsupervised machine learning, COVID-19 cases, COVID-19 deaths, nutrition, diet

**Introduction**

As of May 1, 2021, Severe Acute Respiratory Syndrome 2 (SARS-Cov-2) or Coronavirus disease 2019 (COVID-19) has been responsible for the deaths of over 3M people worldwide, half of million of which occured in the United States. Confirmed cases have topped more than 151M worldwide with approximately 32M occurring in the United States (*Johns Hopkins, 2021*). Unlike other contagious respiratory infections, COVID-19 can induce a wide degree of symptoms in its hosts varying from asymptomatic to paucisymptomatic responses, characterized by respiratory failure and multiple organ dysfunction syndromes (MODS) (*Cascella et. al*). Overall, its high level of transmission, wide variation in symptomatic manifestations, along with its long incubation period allowed the virus to successfully spread creating a global pandemic by March 2020. Even after recovery, there are reports of lingering side effects months after the initial diagnosis including shortness of breath, dizziness and confusion, and leg numbness in people now referred to as COVID-19 long haulers (*Harvard Medical School, 2021*).

As a novel disease, researchers have worked swiftly to uncover data on its pathology, transmissibility, and risk factors for severe immune response, since the virus first appeared in humans in December 2019. It has been discovered that age, sex, and other comorbidities impact a person’s risk for COVID-19 disease severity. Numerous studies have explored using supervised machine learning techniques to predict COVID-19 prognosis, mortality, and risk factors using clinical data (*Zimmerman et. al; Vaid et. al; Castelnuovo et. al*). However, an important research gap remains in regards to how diet can influence this risk of diagnosis or deaths. This current study seeks to address that data gap by evaluating various methods of supervised and unsupervised machine learning algorithms’ ability to predict COVID-19 cases and mortality based on food consumption categories delineated by country.

**Methods**

Data Description and Preprocessing

The project dataset was downloaded from Kaggle and contained information on the average food category breakdowns and COVID-19 cases and deaths of 170 countries *(Kaggle, 2021)*. Specifically, the dataset contained variables denoting what percentage of average food intake (in kilocalories) for a given country comes from a specific food category (ex. Vegetables or Animal Products), as well as variables that denoted what percentage of the population had confirmed cases of COVID-19 and had died of COVID-19 for the week of February 6, 2021. The data on food intake breakdowns was obtained from the Food and Agriculture Organization of the United Nations, and the COVID-19 data was obtained from the Johns Hopkins Center for Systems Science and Engineering website.

Initially, there were 24 food categories which served as our predictor variable set. The initial outcome variables were COVID-19 cases and deaths. Before analysis, data processing was completed to prepare the dataset for analysis. From the predictor variables, the variables for Miscellaneous and Spices were removed due to their ambiguous definitions. From correlation plots and the Spearman Correlation Coefficients, it was discovered that some predictor variables were highly correlated, so for most analyses these variables were removed from the predictors. For the outcome variables, population size, which was provided in the original dataset, was used to calculate COVID-19 cases and deaths per capita, which served as the continuous outcomes of interest. Two categorical outcomes for comparing the cases and deaths per capita to the median rate from all the countries (levels were “Above” and “Below” the median) were also created. All observations with missing values were removed before analysis.

R version 4.0.3 was used for supervised machine learning models and R version 4.0.5 was used for unsupervised machine learning models (*Comprehensive R Archive Network, 2021*).

Supervised Machine Learning Models

Supervised learning was used to predict the continuous outcomes of Cases per capita and Deaths per capita and categorical outcomes of confirmed cases vs median and deaths vs median. Furthermore, 10-fold cross validation (CV) was performed for each method, with the CV errors and CV error standard errors reported as measures of the models’ performance. Errors for continuous outcomes were defined as Mean Squared Errors, and errors for classification were defined as 1 - prediction accuracy. The specific supervised models used were K-nearest Neighbors, Regression methods like linear, logistic, lasso and ridge regression, QDA, tree-based methods like CART and Random Forest, and Support Vector Machine algorithms.

The K-Nearest Neighbors algorithm was implemented to predict both continuous and categorical outcomes. 10-fold cross validation was performed, with variables scaled and centered during preprocessing of each fold, and a TuneLength of 20 was used. The CV errors and the standard errors of these errors were then recorded.

Linear regression was used solely to predict continuous outcomes. Although assumptions of normality and linearity typically required for linear regression were not met by the data, this method was still included as to act as a comparison to some of the other methods used.

Logistic regression and QDA were used for classification only. QDA was used instead of LDA because the distribution of the data does not satisfy the multivariate normality and homogeneity of variance assumptions of LDA (**Figure S1** and **Figure S2** in the appendix). Some variables are not normal for each level of the grouping variable. Meantime, variances among group variables are not the same across levels of predictors.

Lasso and ridge penalized regression models were created to predict both continuous and categorical outcomes for our data set. For each of the continuous or categorical outcomes of interest, a lasso regression model was fit using both 10-fold CV and on the full dataset, with the lambda that minimized the mean squared error or classification error chosen as the optimal parameter. The features and respective coefficients from these lasso models were extracted and compared to the original set of predictors to determine which variables were being penalized and removed during the lasso regression.

Random Forest was used to predict both continuous and categorical outcomes. These models were tuned with a grid search to determine the optimal number of trees and number of predictors to consider at each split. Based on the results, the most important 10 predictors for each (categorical) outcome were extracted. Another tree-based method, single decision trees (CART), was also used to predict the continuous outcomes.10-fold cross validation was used for both the CART and random forest algorithms and CV errors/SEs were recorded. Both the full decision trees and the 10-fold CV were pruned by selecting the optimal complexity parameter, which minimized RMSE.

Support Vector Machine regression/modeling with linear, polynomial, and radial basis function (RBF) kernels was also performed, and hyperparameters for each respective model were tuned using grid search. Specifically, parameters for linear SVR for continuous outcomes were tuned based on a pre-selected grid search of cost = (1, 2, 3, 4, 5) and epsilon (0.1, 0.2, 0.3, 0.4), and parameters for linear SVM for classification were tuned using a grid search of cost = (1,2,3) and epsilon = (0.1, 0.5, 1). For the radial kernel, both continuous and categorical outcomes were tuned using a grid search of gamma = (0.001, 0.05, 0.1), cost = (1,2,3), and epsilon = (0.1, 0.5, 1). Finally, polynomial SVM for both continuous and categorical outcomes used a grid search of gamma = (0.001, 0.05, 0.1), cost = (1,2,3), and d = (2,3). CV errors and SEs following 10-fold cross validation were reported.

Unsupervised Machine Learning Models

Although unsupervised machine learning is used when the ground truth or outcome has not been established, it was used in this paper to reduce dimensionality with the aim of improving the bias-variance tradeoff in results from K Nearest Neighbor (KNN). To do so, Principal Component Analysis (PCA) was used to reduce the dimensionality of the 19 food categories that were scaled prior to analysis. PCA is a widely used computational technique that seeks to balance maintaining the variation in the original dataset, while reducing the number of predictors into principal components or eigenvectors. High dimensional datasets often are more difficult to visualize and require a longer amount of processing time, hence the need for PCA (*Jaadi, 2014*). The elbow method was used to determine the optimal number of principal components to use in KNN. Lastly, KNN was performed with the chosen eigenvectors on both outcomes (ie. cases and deaths) continuously and categorically using a 10-fold cross validation. The parameters of the KNN model (ie. TuneLength and pre-processing) mirrored that of the model used in the supervised machine learning models.

The performance of the algorithm was assessed using cross validation (CV) error and standard error. For continuous variables, average overall accuracy represented the CV error and the standard deviation of the overall accuracy represented the standard error from each of 10 folds. For categorical variables, average mean squared error (MSE) represented the CV error and the standard deviation of the MSE represented the standard error.

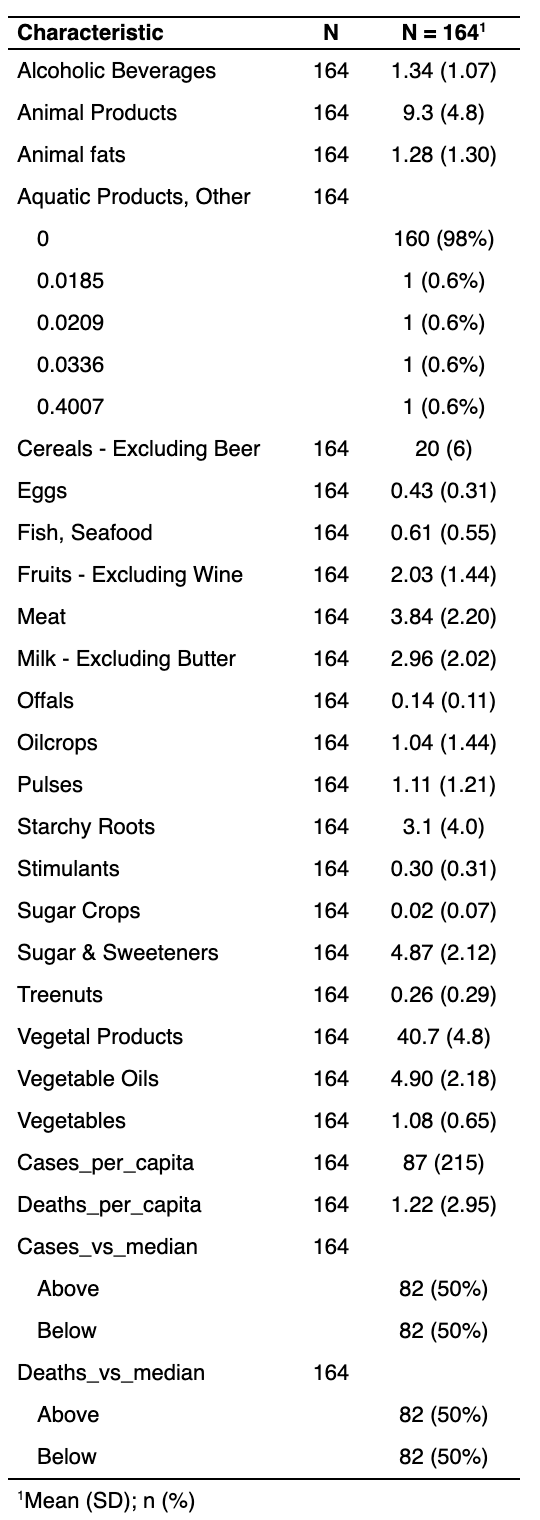
Another unsupervised machine learning approach, k-means, was used to investigate how well it would be able to cluster countries into the correct dichotomous outcome (ie. above or below the median cases or deaths). K-means works by minimizing the sum of squares or the within cluster variances and places variables in clusters that have the closest centroid or average. Prior to analysis, the variables were scaled and a various number of centers were then tested between 1 through 12. The Bayesian Information Criterion (BIC) was calculated for each model to identify which k-means model had the center with the smallest BIC, helping to assess how well the clusters fit the data.

**Results**

Summary Statistics

The summary statistics for all variables are provided in Table 1, with mean percentages of each food category provided in the rightmost column. Notice that Vegetal Products has the highest mean percentage by far of 40.7%. Animal Products was another relatively high-percentage category with a mean of 9.3%. However, as noted previously, these two variables were removed from the analyses due to high correlations with other variables, so most of our analyses were done using predictors with mean percentages less than 5%. Summary statistics for variables grouped by Cases vs Median and Deaths vs Median are provided in the appendix (**Tables S1 and S2 in appendix**).

**Table 1. Summary statistics for all variables.**



Correlations and Data Preprocessing

Prior to the implementation of machine learning algorithms, Spearman correlation coefficients between predictors to determine if any were highly correlated (⍴ > |0.7|). It was found that Animal Products was highly correlated with Meat, Animal Fats, Milk - Excluding Butter, and Eggs. Vegetal Products was another variable correlated with a plethora of predictors including Meat, Animal Fats, Milk - Excluding Butter, Eggs, and Animal Products (**Table 2**). For those reasons Animal and Vegetal Products were removed from further analyses.

**Table 2. Correlations between predictors**. This table includes Spearman correlation coefficients between variables that are highly correlated (⍴ > |0.7|).

|  |  |  |
| --- | --- | --- |
| **Variable 1** | **Variable 2** | **Spearman Coefficient** |
| Meat | Animal Products | 0.874 |
| Animal fats | Animal Products | 0.848 |
| Milk - Excluding Butter | Animal Products | 0.797 |
| Eggs | Animal Products | 0.723 |
| Vegetal Products | Eggs | -0.723 |
| Vegetal Products | Milk...Excluding.Butter | -0.797 |
| Vegetal Products | Animal fats | -0.848 |
| Vegetal Products | Meat | -0.874 |
| Vegetal Products | Animal Products | -1.000 |

Supervised Machine Learning Approaches

*Continuous Outcomes*

For Cases per capita, Lasso, Ridge regression, and linear regression performed the best with CV errors of 31800, 31900, and 32100, respectively. For deaths per capita, Lasso, Ridge regression, and linear SVR were the top three performing models with CV errors of 6.66, 6.81, and 6.77, respectively. In general, the Decision Tree and Random Forest algorithms had the weakest performance for predicting both continuous outcomes. Interestingly, however, for the Cases per capita outcome, the Decision Tree performed better than Random Forest, which is unusual.

For predicting the Cases per capita outcome with the full dataset, lasso selected 12 predictor variables to keep in the model: Animal Fats, Eggs, Fish/Seafood, Fruits - excluding Wine, Meat, Offals, Pulses, Starchy Roots, Stimulants, Sugar and Sweeteners, Vegetable Oils, and Vegetables (**Figure S3** in appendix). The variables which were most important for prediction, as noted by the magnitude of their coefficients, were Stimulants and Offals, which had coefficients of 389.77 and -239.49, respectively.

For predicting the Deaths per capita outcome with the full dataset, lasso selected 12 predictor variables to keep in the model: Alcoholic Beverages, Eggs, Fish/Seafood, Fruits - excluding Wine, Meat, Offals, Starchy Roots, Stimulants, Sugar Crops, Sugar and Sweeteners, Tree Nuts, and Vegetable Oils (**Figure S4** In appendix). Note that, similar to the lasso results from the Cases per capita model, the variable with the coefficient of the largest magnitude was Stimulants (4.75), followed by Offals (-2.30).

*Classification:*

When predicting confirmed cases vs. median, lasso and random forest have the best prediction performance across all the classification methods. When predicting deaths vs. median, random forest has the lowest cross validation error, followed by lasso with the second lowest error. So, lasso and random forest appear to be the top 2 methods across both the classification cases.

For cases vs median classification on the full dataset, lasso selected 7 predictors: Alcoholic beverages, cereals excluding beer, fruits - excluding wine, milk - excluding butter, stimulants, sugar crops and sugar & sweeteners (**Figure S5** in appendix). When fitting random forest on the full dataset, all variables were plotted according to their importance in a descending order (**Figure S7** in appendix). Their importance was measured by corresponding mean decrease in gini index or node impurity. A higher decrease in gini index indicates that the predictor contributes to the improvement of node purity more significantly. The top 5 most important variables in predicting cases vs median are Stimulants, milk excluding butter, sugar & sweeteners, eggs, and cereals - excluding beer. 4 out of 5 top predictors indicated by random forest were selected by lasso as well.

As for deaths vs median classification on the full dataset, the lasso model selected 5 predictors: alcoholic beverages, milk - excluding butter, stimulants, sugar crops, and sugar & sweeteners (**Figure S6** in appendix). Similar to random forest output in predicting cases vs median, all predictors are ranked in terms of their gini importance (**Figure S8** in appendix). The top 5 most important variables are stimulants, milk - excluding butter, sugar & sweeteners, eggs, and cereals excluding beer, and the first three of them were also selected in the lasso model.

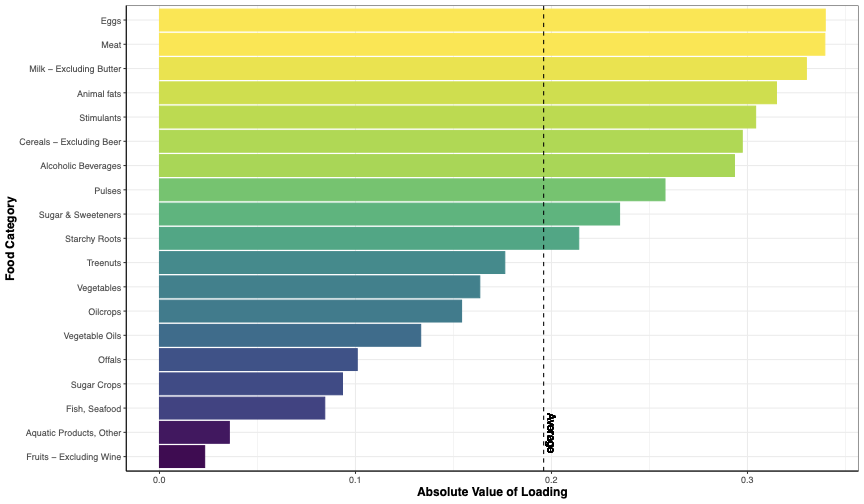
**Table 3.** **10-fold cross validation and standard errors across machine learning methods.** Table contains cross validation (CV) and standard (std) errors for all machine learning algorithms run.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Supervised Machine Learning** | | | | | | | |
| **Regression** | | | | **Classification** | | | |
| **Outcome** | **Method** | **CV Error** | **Std Error** | **Outcome** | **Method** | **CV Error** | **Std Error** |
| Confirmed cases per capita | KNN | 3.65E+04 | 4.48E+04 | Confirmed cases vs. median | KNN | 0.24 | 0.08 |
| Linear Regression | 3.21E+04 | 2.94E+04 | Logistic Regression | 0.21 | 0.12 |
| Lasso | 3.18E+04 | 3.11E+04 | Lasso | 0.20 | 0.09 |
| Ridge | 3.19E+04 | 3.28E+04 | QDA | 0.26 | 0.12 |
| Decision Tree | 4.14E+04 | 3.93E+04 | Random Forest | 0.20 | 0.10 |
| Random Forest | 6.27E+04 | 2.17E+04 | Ridge | 0.23 | 0.11 |
| Linear SVR | 3.50E+04 | 4.01E+04 | Linear SVM | 0.24 | 0.14 |
| Radial SVR | 3.84E+04 | 4.60E+04 | Radial SVM | 0.25 | 0.10 |
| Polynomial SVR | 3.59E+04 | 5.33E+04 | Polynomial SVM | 0.25 | 0.13 |
| Deaths per capita | KNN | 7.23 | 8.12 | Deaths vs. median | KNN | 0.25 | 0.09 |
| Linear Regression | 7.04 | 6.88 | Logistic Regression | 0.23 | 0.11 |
| Lasso | 6.66 | 7.14 | Lasso | 0.20 | 0.09 |
| Ridge | 6.81 | 7.48 | QDA | 0.23 | 0.10 |
| Decision Tree | 9.38 | 9.90 | Random Forest | 0.18 | 0.13 |
| Random Forest | 7.22 | 6.10 | Ridge | 0.21 | 0.09 |
| Linear SVR | 6.77 | 8.13 | Linear SVM | 0.23 | 0.10 |
|  | Radial SVR | 7.91 | 9.31 | Radial SVM | 0.24 | 0.12 |
|  | Polynomial SVR | 7.07 | 7.99 | Polynomial SVM | 0.29 | 0.12 |
| **Unsupervised Machine Learning** | | | | | | | |
| **Regression** | | | | **Classification** | | | |
| **Outcome** | **Method** | **CV Error** | **Std Error** | **Outcome** | **Method** | **CV Error** | **Std Error** |
| Confirmed cases per capita | KNN | 4.14E+04 | 5.93E+04 | Confirmed cases vs. median | KNN | 0.77 | 0.15 |
| Deaths per capita | KNN | 8.24 | 10.04 | Deaths vs. median | KNN | 0.74 | 0.15 |

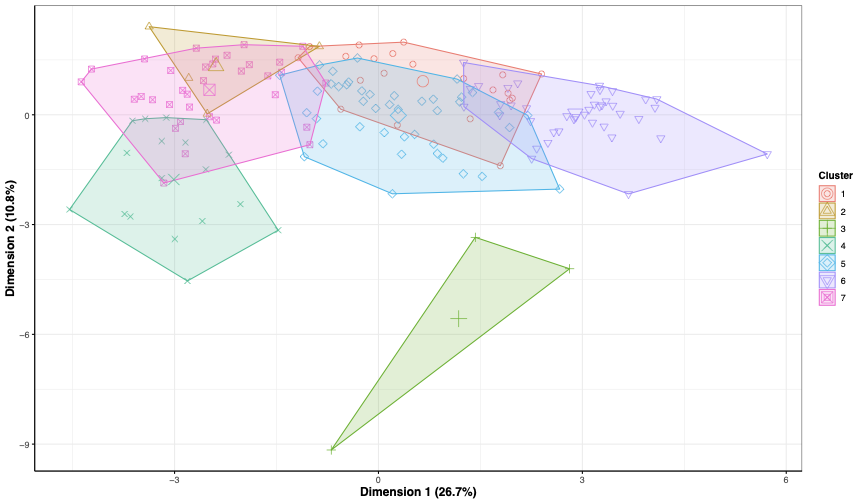
Unsupervised Machine Learning Approaches

As mentioned in the methods, PCA was run to reduce the dimensionality to obtain eigenvectors that still capture a majority of the variance in the original dataset. The top four predictors that contributed to the eigenvectors were all animal products and included eggs, meat, milk, and animal fats. Carbohydrate food categories tended to have moderate contributions, while plant based products contributed the least (**Figure 1**). The screeplot suggested retaining the first three principal components based on the elbow method to be used in the subsequent KNN analysis. These three eigenvectors captured 26.7%, 10.8%, and 8.3% of the variance, respectively, comprising 45.8% of the variance from the original dataset.

**Figure 1. PCA loadings of predictors.** The plot visualizes the magnitude or absolute value of each predictor’s loading on the first eigenvector. The dotted line represents the average magnitude of the loadings.

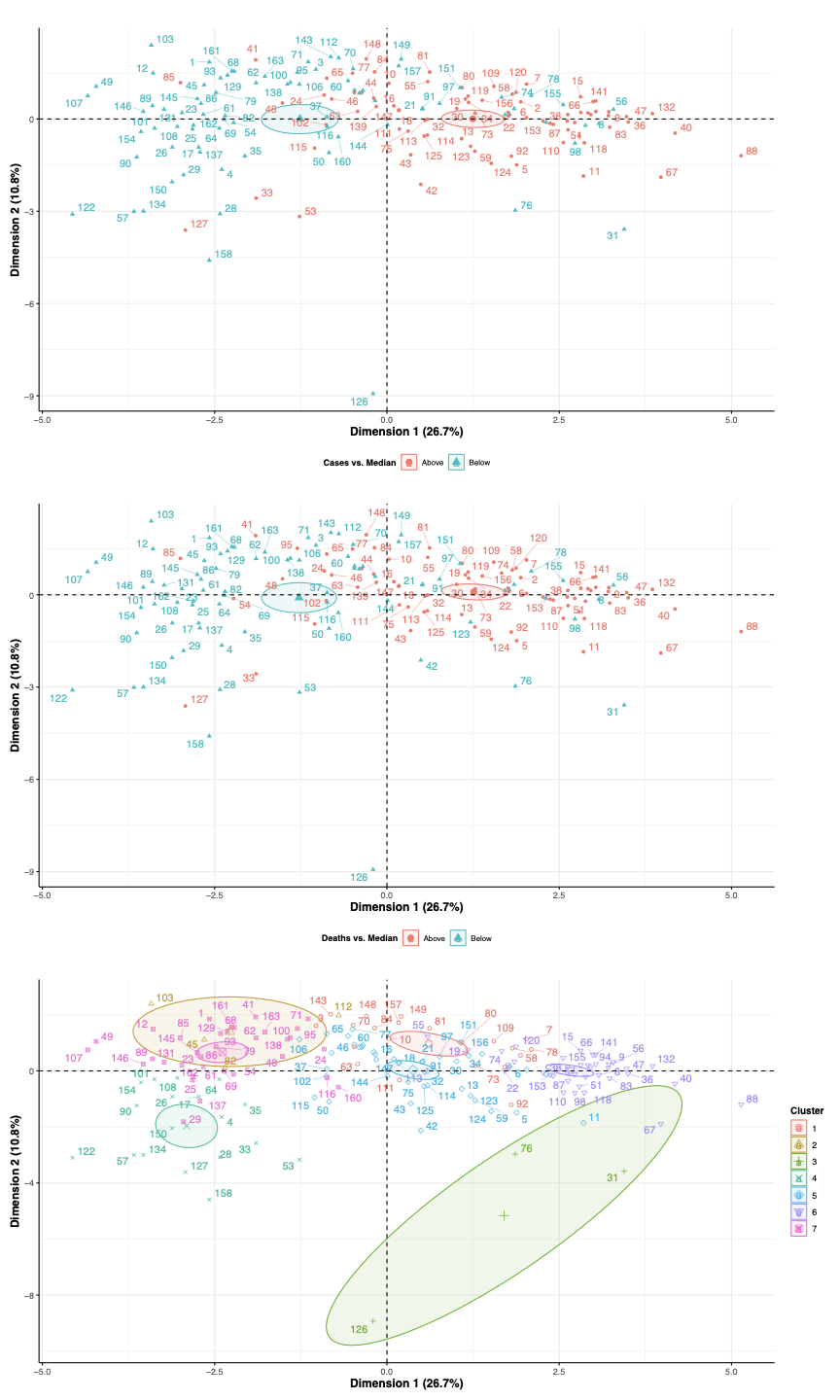
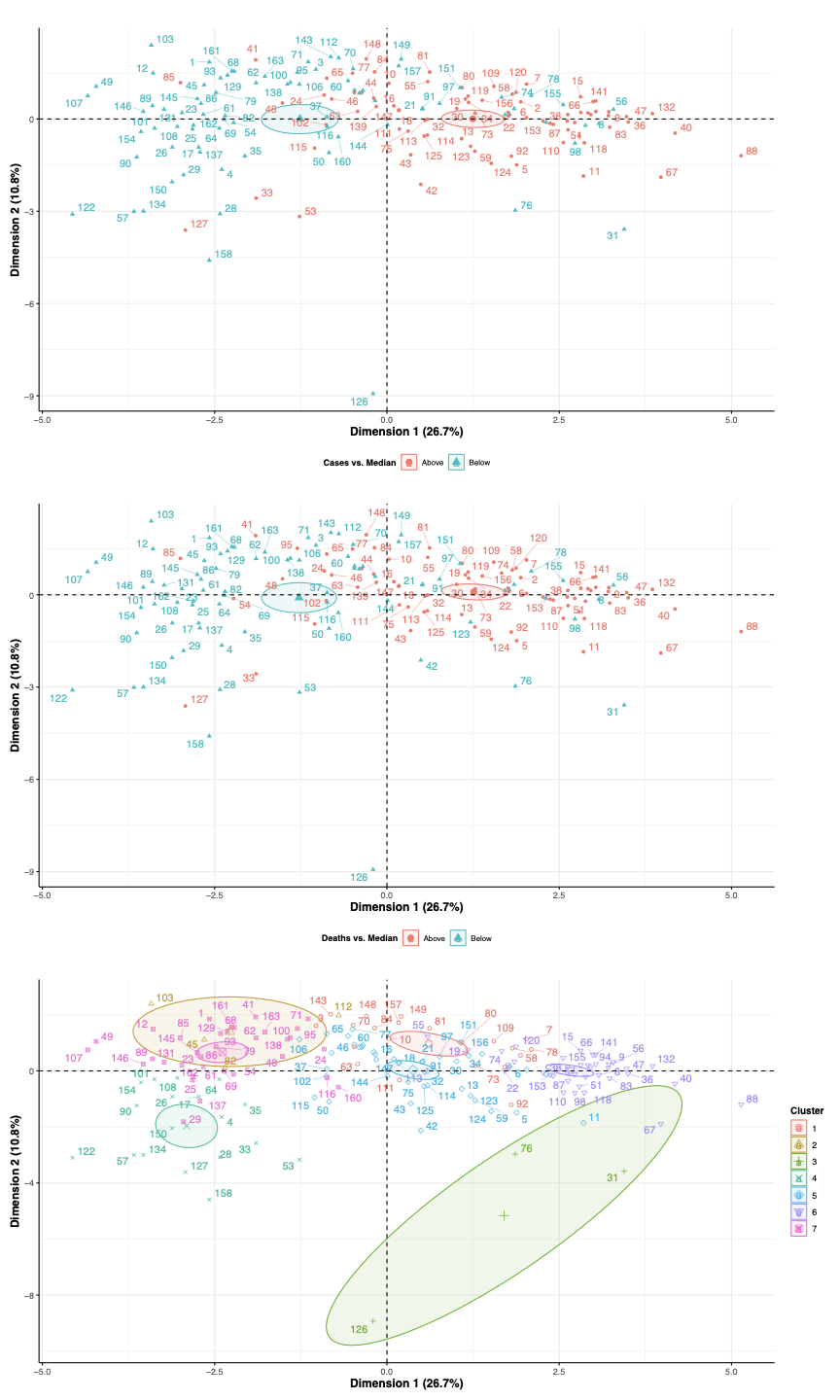
****

When the k-means algorithm was initially run, South Korea was an outlier in a cluster by itself; therefore, it was removed from the final dataset used in this analysis. This resulted in 7 clusters (k = 7) of countries being identified by k-means (**Figure 2**).

**Figure 2. K-means clusters of countries.** Seven clusters of countries identified by k-means clustering.  ****

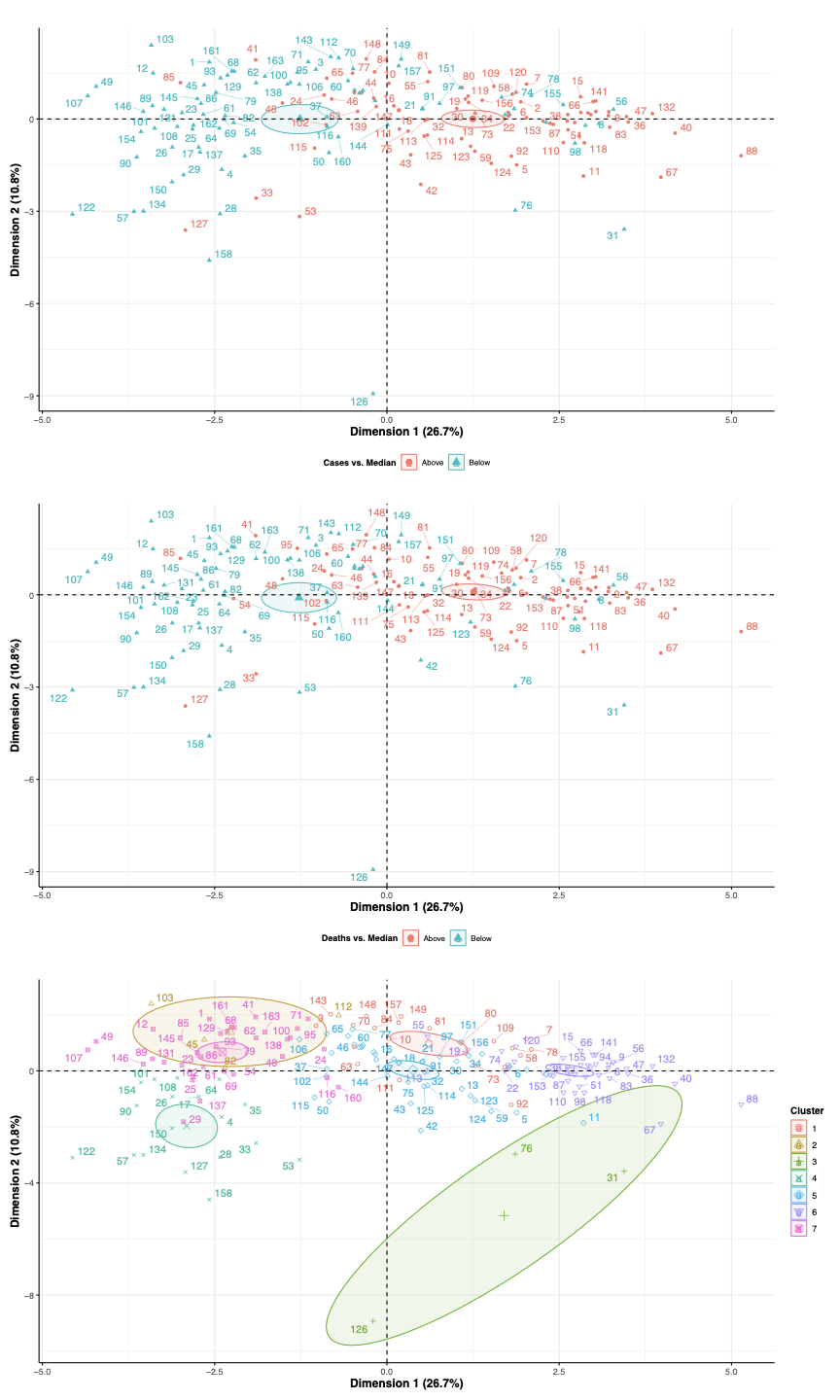
Next, clustering of the food category predictors was evaluated using k-means to determine if the algorithm would correctly suggest the number of clusters and resemble the grouping of the dichotomous outcomes. After running PCA, the compressed dataset was plotted with the first two principal components colored in three ways. The eigenvectors in plots A and B are colored by cases and deaths per capita, respectively either above or below the median. Plot C tags each country with one of the 7 k-means clusters (**Figure 3**). The eigenvectors shown in plots A and B appear almost identical, indicative of substantial similarities between these two outcomes. The k-means clusters in plot C poorly predict the correct number of clusters or outcomes with k = 7, rather than k = 2.

**Figure 3. Clustering of dichotomous outcomes and predicted k-means clusters.** A is colored by confirmed COVID-19 cases per capita above or below the median. B is colored by COVID-19 deaths per capita above or below the median. C is colored by the predicted number of k-means clusters (k = 7).



**B**

**A**



**C**

For the prediction results from the KNN regression analyses, cases per capita had a CV error of 4.14 x 10^4 and a standard error of 5.93 x 10^4, while deaths per capita had a CV error of 8.24 and a standard error of 10.04. For the prediction results from the classification analyses, confirmed cases vs. median had a CV error of 0.77 and a standard error of 0.15, while deaths vs. median had a CV error of 0.74 and a standard error of 0.15 (**Table 3**). Similar to the supervised machine learning results, the regression cases per capita results had the largest errors. All of the unsupervised machine learning KNN errors were higher than supervised machine learning approaches without PCA, suggesting that these models had both higher biases and variances.

**Discussion**

Interpretations of Models and Results

Among supervised machine learning methods, lasso has the best overall performance in both continuous and categorical outcome prediction. Lasso, as the only method used that conducts variable selection, improves the performance by eliminating unnecessary predictors. It also prevents overfitting with its penalty term that punishes complex models. WIth its advantages in variable selection and using a penalty term to prevent overfitting, it stands out from the rest. Additionally, according to the cross validation errors of SVM models with linear, radial, and polynomial kernels, linear kernel has the smallest errors in both continuous and categorical cases. It appears that a linear form is a better approximation than radial or polynomial forms. Lasso, similar to linear regression, follows a linear form. Therefore, another reason that lasso outperforms other models is that the linear form that lasso adopts is a better fit of the pattern of the data. However, one drawback of lasso is that it arbitrarily selects and keeps one feature from a group of correlated features. As a result, the variables the lasso selected in its final model may be the most important ones and instead are just a result of this random selection to reduce correlation.

Although random forest does not outperform other methods in predicting continuous outcomes, it did an excellent job in predicting categorical outcomes and achieved similar accuracy as lasso prediction. One explanation for this discrepancy may be because for this particular dataset with only dietary information given, it is much easier for random forest to predict whether cases or deaths are above or below the medians than a numeric value. For the classification, random forest does a great job in predicting cases vs median and deaths vs median. It does not do variable selection as lasso, but it can also effectively prevent overfitting issues by randomly selecting a subset of predictors to consider when tuning the best model. Another advantage of random forest is that it can rank the importance of all predictors in terms of gini index, while lasso may not be the best method indicating the most important predictors. However, one disadvantage of random forest is that because it is an aggregation of multiple trees, it is not easy to interpret or visualize.

Overall, unsupervised KNN models had higher biases and variances than all supervised machine learning models, including KNN. There were multiple factors that could further explain these results. Specifically, PCA works by performing a linear decomposition on the dataset and captures more of the variance when correlations between the predictors and the outcome are linear. The correlation matrix revealed non-linearity, which is a potential reason for the low degree of variance captured in the first eigenvectors. Another contributor could be the high variability of the dataset, making it difficult for PCA to capture a majority of the variance within the first few principal components. This dimension reduction method tends to work better with a “large” number of predictors as this was likely not met with 19 variables. Additionally, there was substantial overlap between the dichotomous outcomes of cases and deaths, suggesting there’s a lack of association between food categories and COVID-19 confirmed cases and deaths. All of these aforementioned factors are likely contributors to k-means predicting 7 outcomes or clusters rather than 2.

For the cases per capita results, all of the errors were extremely high and were not comparable to the errors from deaths per capita. It is unclear why that occurred given that the correlations between the predictors and outcomes were similar.

Significance of findings

The population of interest for this project was broadly defined as countries around the world, and the goal of the project was to identify how diet patterns played a role in predicting COVID-19 cases and deaths on a country level. This can potentially give implications to nutrition-related policies as well as individuals’ diet recommendations. There are many other factors related to COVID-19 cases such as mask availability and healthcare infrastructure and factors related to deaths such as healthcare resources and income level, but diet is a very important indicator. As a study suggests, diets such as high consumption of saturated fats, sugars, and refined carbohydrates contribute to obesity and type 2 diabetes, which are related to increased risk for severe COVID-19 pathology and mortality (*Butler and Barrientos, 2020*). So, it is important to explore how diet may help people to better protect themselves against COVID-19.

Based on the errors of both supervised and unsupervised methods, it does not appear that diet is a strong sole predictor of COVID-19 cases and deaths. However, diet does have some impact on COVID-19 cases and deaths. Stimulants, for instance, appears to be one of the contributors to confirmed cases as well as deaths. Similarly, sugar and sweeteners are also positively correlated to increased cases and deaths. These results are consistent with the Butler and Barrientos’ study findings that unhealthy diets such as high sugar uptake can lead to higher risk of health comorbidities of obesity/diabetes and COVID-19.

Limitations

One substantial limitation is that our study only focused on a single factor (ie. food consumption by category). The dataset did not incorporate other contributors known to have a large influence on health outcomes including, but not limited to universal healthcare, gross domestic product (GDP), messaging from government public health organizations, and health system infrastructure. It should also be mentioned that by only considering “big data'' with an aggregated dataset such as the one used in this paper, the “smaller” data that contributed to the aggregated values are overlooked, making it difficult to compare countries to each other. Other health determinants that are not captured at the country level, but have substantial influence at the individual level include socioeconomic status, education level, health insurance status, race, sex, public perception of COVID-19, location of residence, comorbidities, residing in multigenerational housing, and native language.

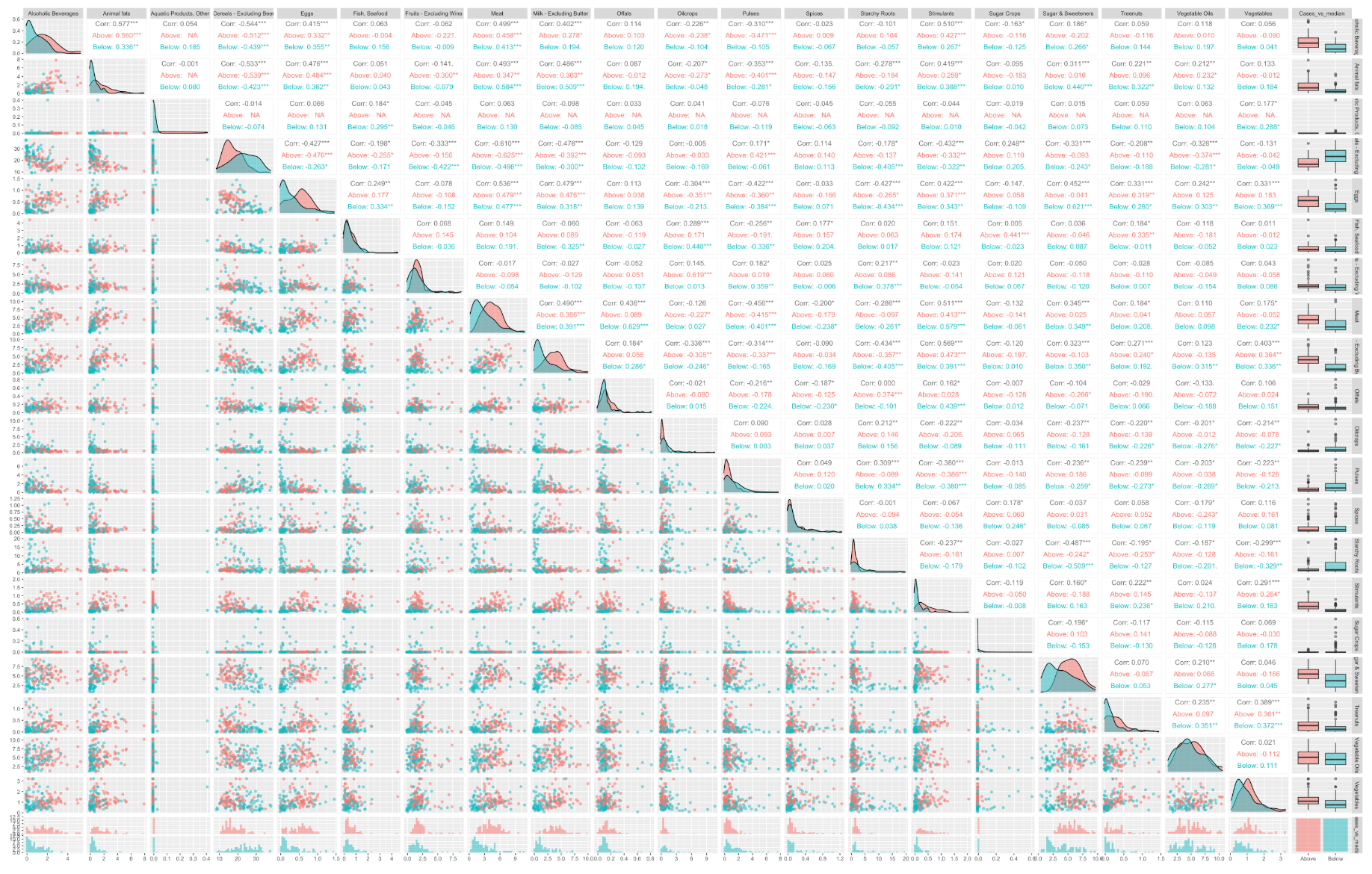
This dataset collected information on calorie percentage of energy intake for each country, meaning that higher calorie foods were weighted more in our analyses as shown in Figure 1. The models and results were unable to completely characterize the impact of food categories on health due to the lack of data on macronutrients, vitamins, and minerals that are more likely to have an effect on a person’s immune system and therefore response to COVID-19. Lastly, as this was a cross-sectional study food consumption and COVID prevalence were only from the week of February 6, 2021 and are not representative of more long-term trends.

Further Analyses

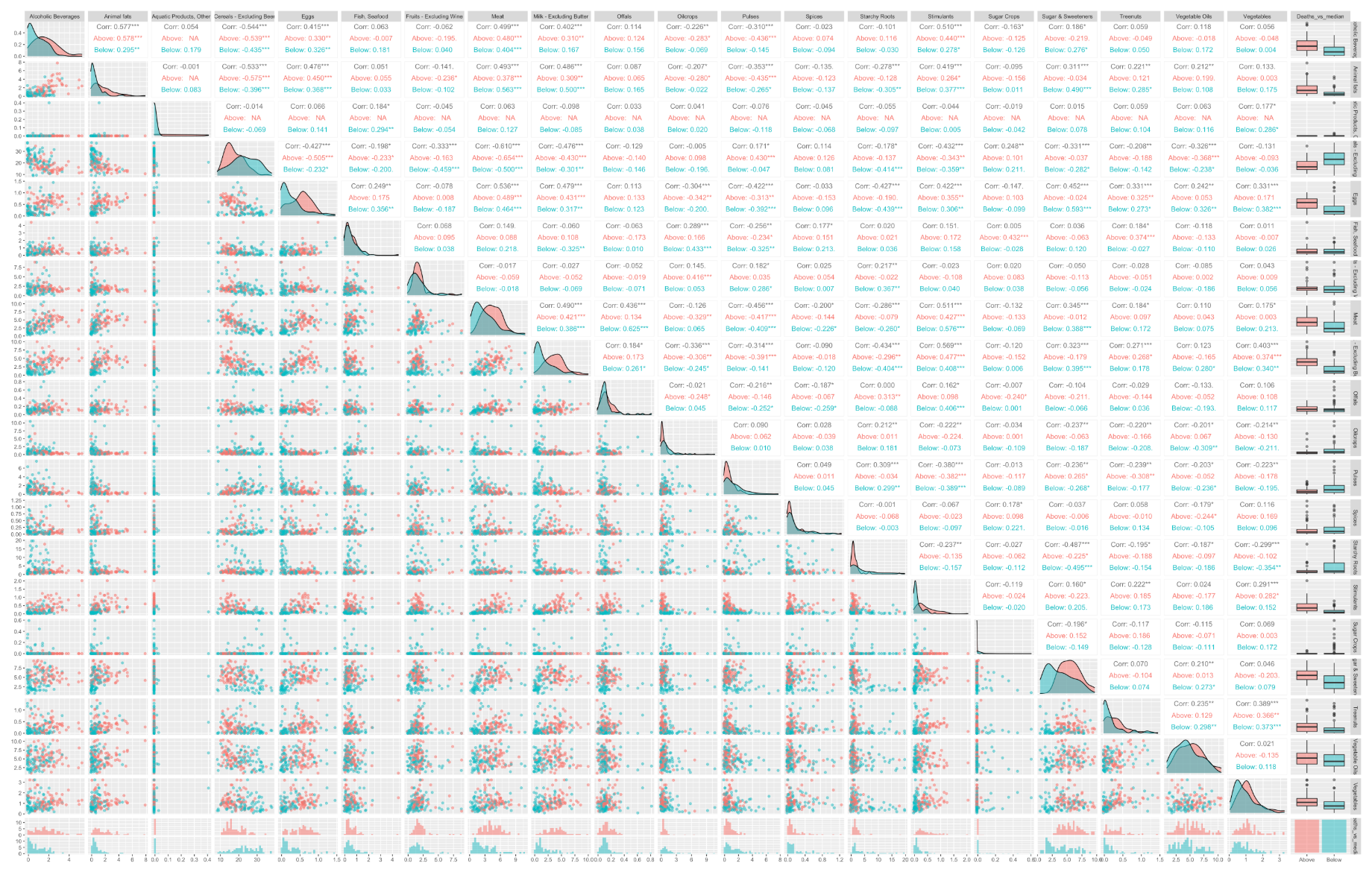
Given that our study was limited due to its focus on only one potential factor contributing to COVID cases and deaths, further studies that incorporate other potentially important factors, such as those mentioned in the Limitations section, would be helpful in understanding the true associated factors related to COVID-19 infection and mortality on a global level. An interesting further analysis from our dataset that could provide more information would be to isolate the countries that had the lowest case and death rates of COVID-19 to see if they shared any patterns in food category consumption and diet, as this could provide some insight into which diet patterns may be most effective in preventing COVID-19. Finally, another interesting further step to take would be to examine our research question on an individual level, to see if diet patterns could be used to predict COVID-19 cases and deaths among individuals. Overall, while our study provided a preliminary glimpse of the potential associations between diet patterns and COVID-19, further analyses are necessary to truly gain an understanding of COVID-19 spread at both the country and individual levels.

**Appendix**

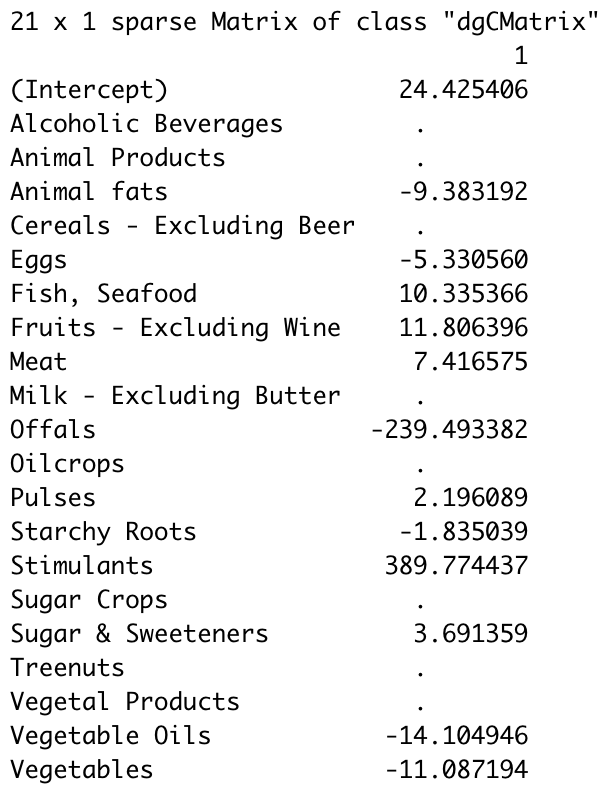
**Figure S1.** **Correlation Matrix and Scatterplots colored by “Cases\_vs\_median”**



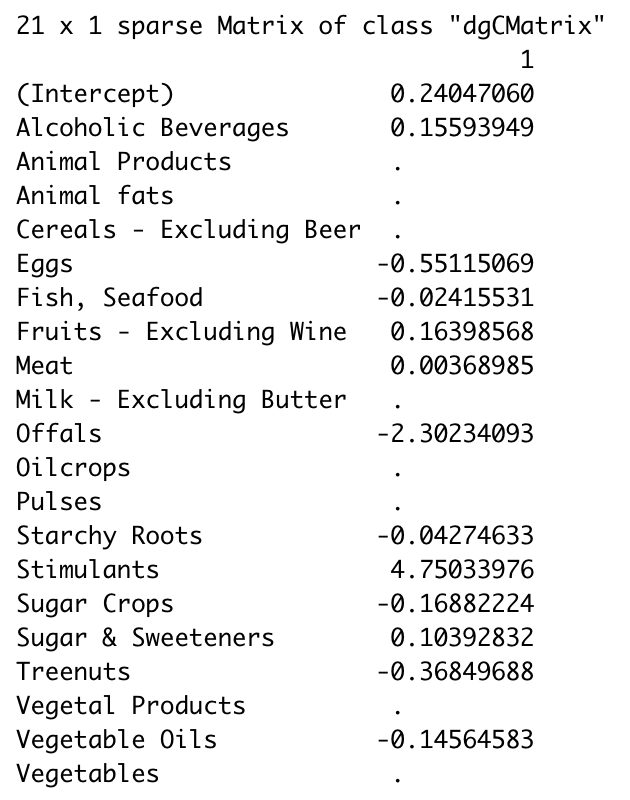
**Figure S2. Correlation Matrix and Scatterplots colored by “Deaths\_vs\_median”**



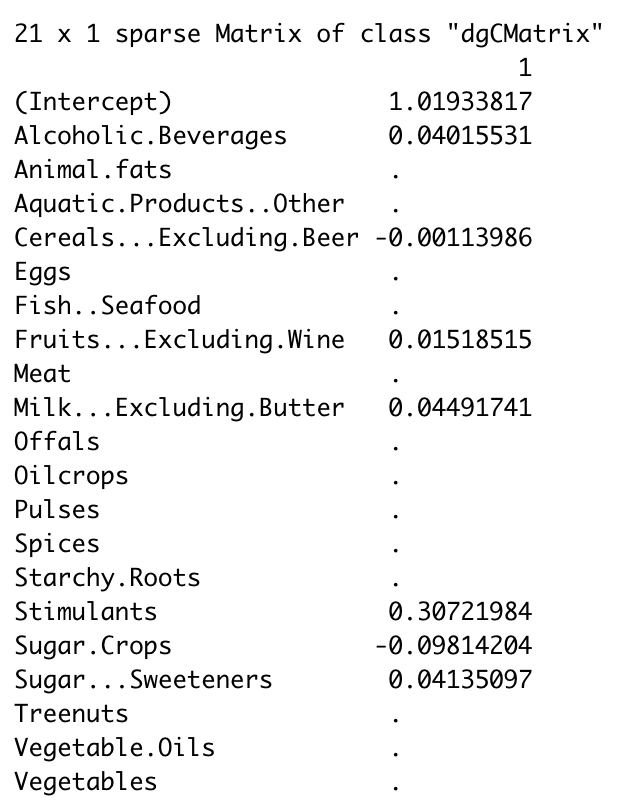
**Figure S3. Lasso Output Predicting Cases per capita on the Full Dataset:**



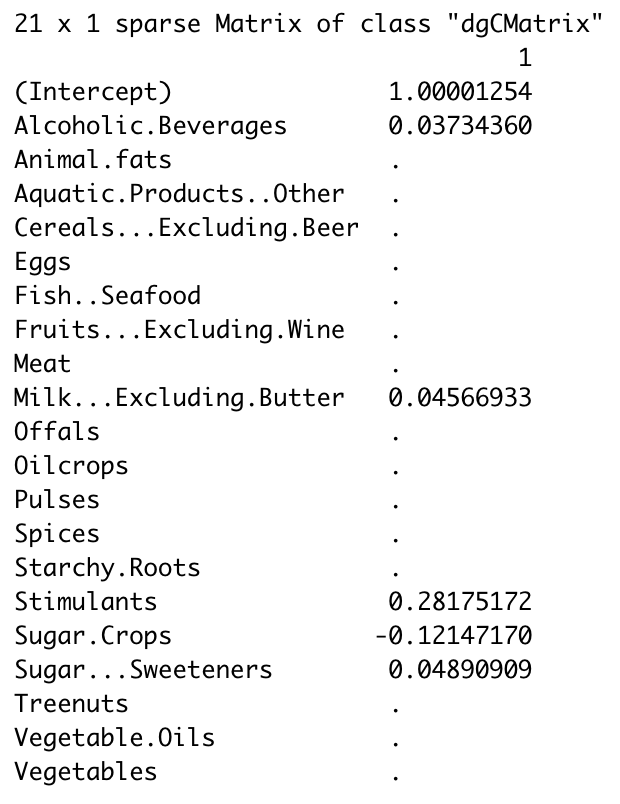
**Figure S4. Lasso Output Predicting Deaths per capita on the Full Dataset:**



**Figure S5. Lasso Output Predicting Cases vs Median on the Full Dataset:**



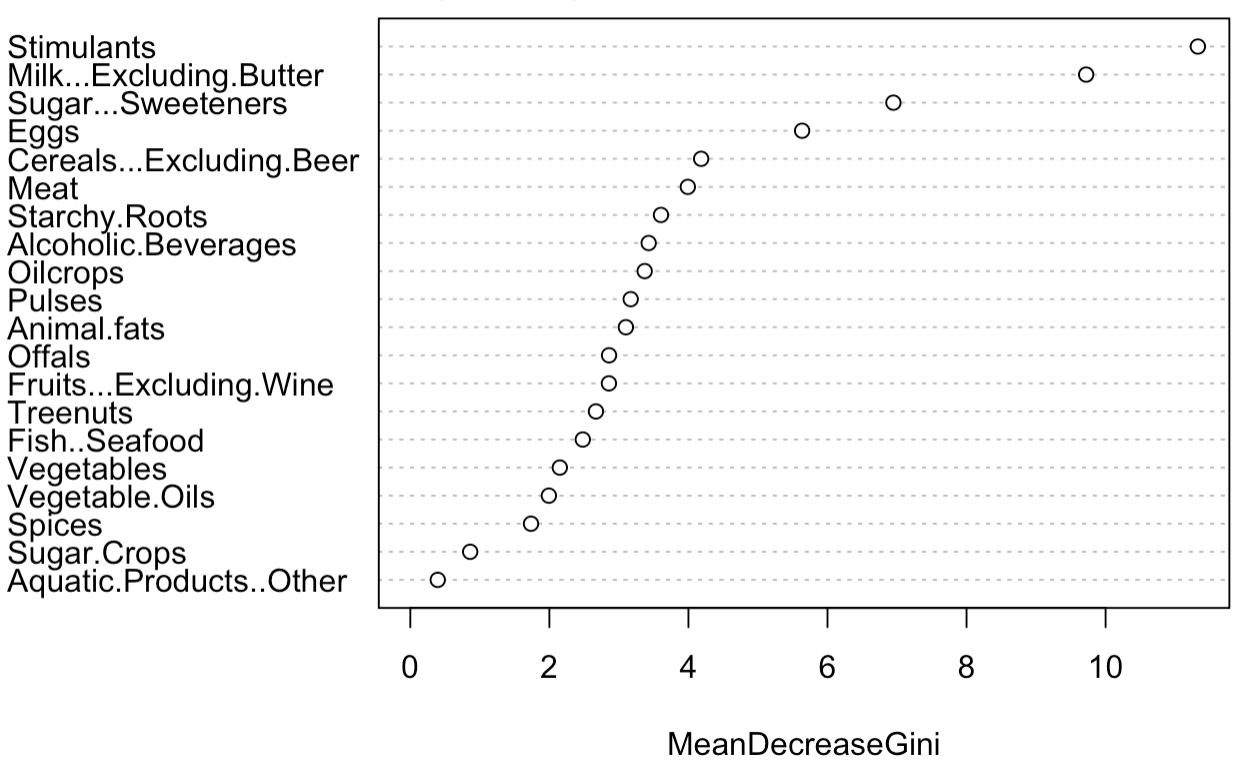
**Figure S6. Lasso Output Predicting Deaths vs Median on the Full Dataset:**



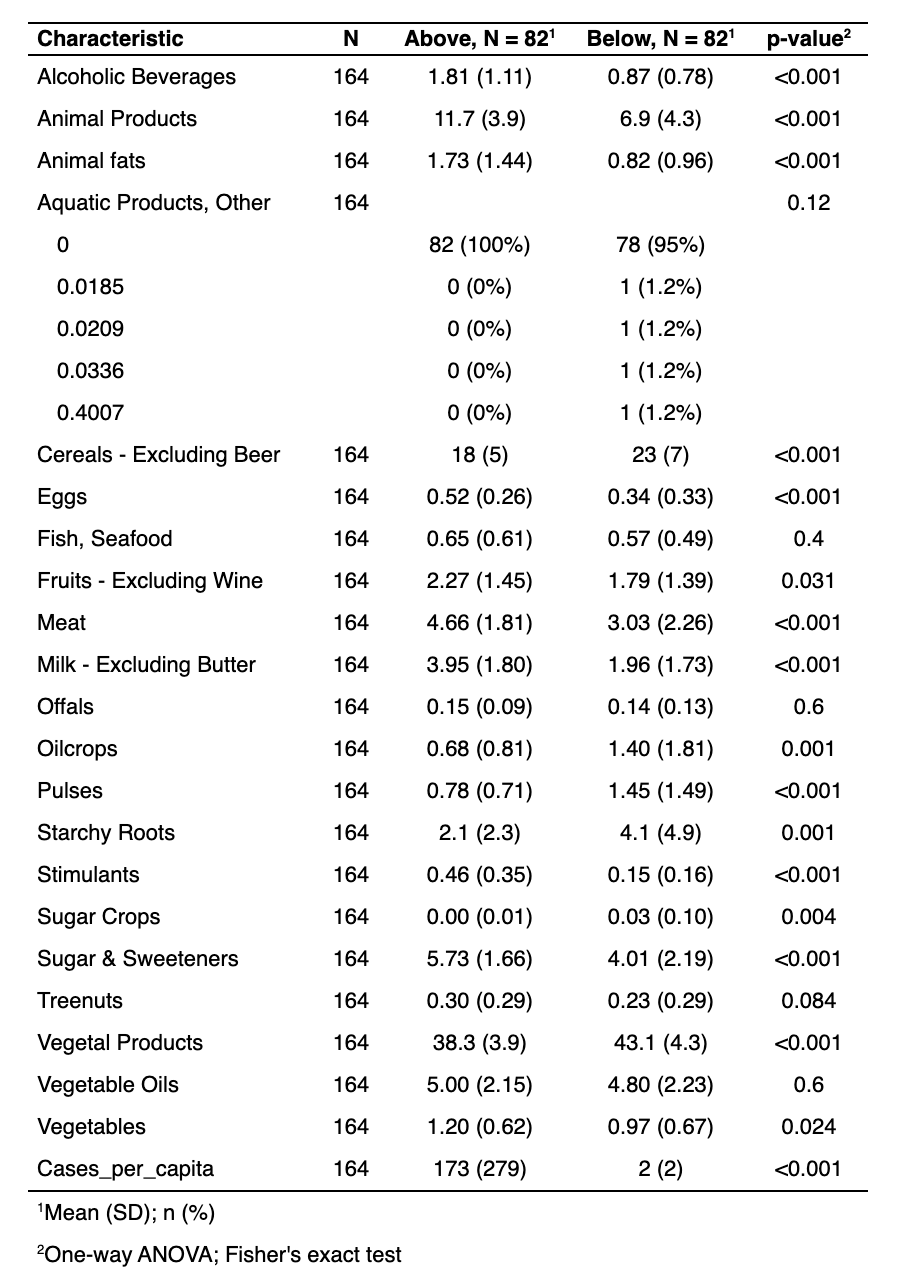
**Figure S7. Random Forest Predictor Importance in Predicting Cases vs Median on the full Dataset:**

****

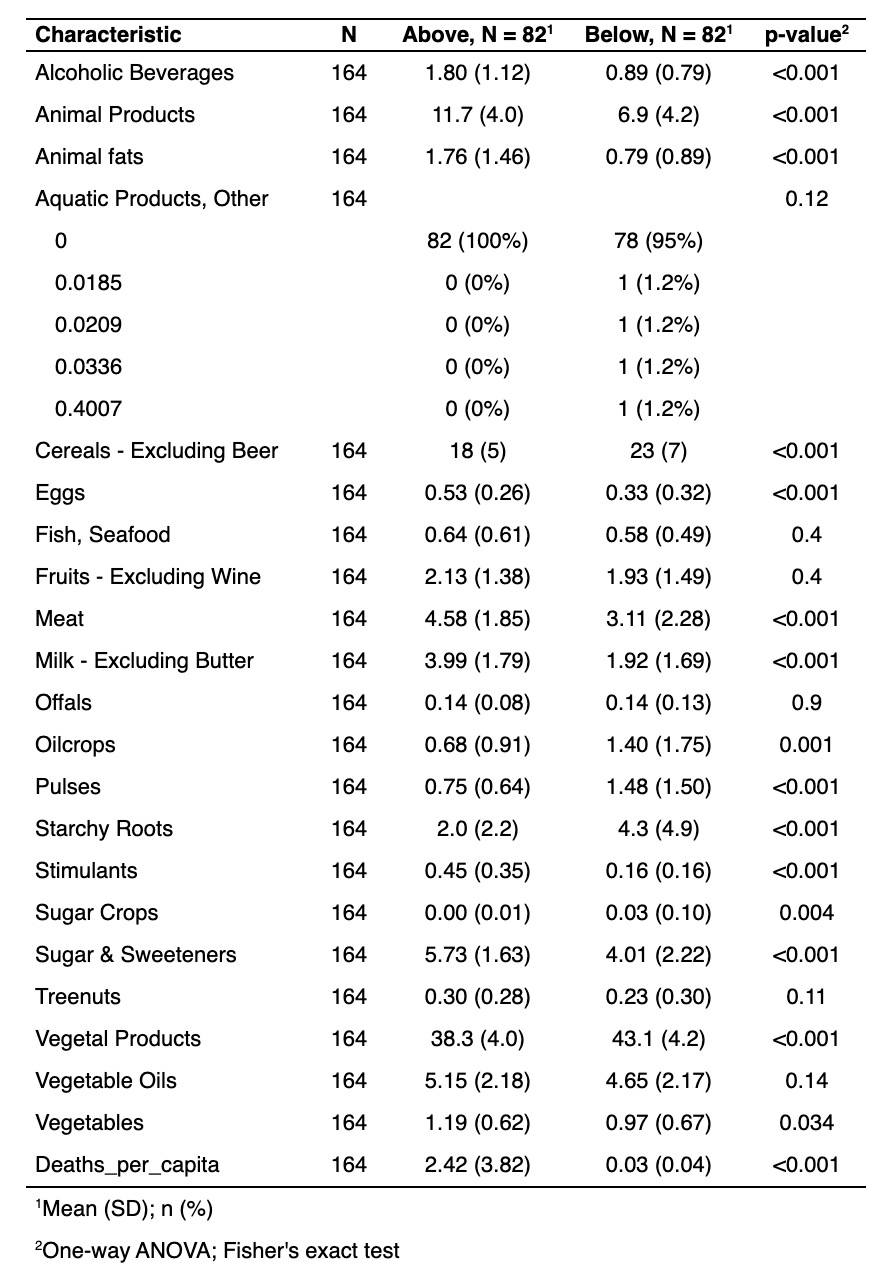
**Figure S8. Random Forest Predictor Importance in Predicting Deaths vs Median on the full Dataset:**



**Table S1. Summary Statistics Grouped by Cases vs. Median Outcome:**



**Table S2. Summary Statistics Grouped by Deaths vs. Median Outcome:**

****

**Bibliography**

The Comprehensive R Archive Network. (n.d.). https://cran.ism.ac.jp/.

Butler, M. J., & Barrientos, R. M. (2020). *The impact of nutrition on COVID-19 susceptibility and long-term consequences*. https://pubmed.ncbi.nlm.nih.gov/32311498/

*COVID-19 Map*. Johns Hopkins Coronavirus Resource Center. (n.d.). https://coronavirus.jhu.edu/map.html.

D;, Z. A. K. (n.d.). *Usefulness of machine learning in COVID-19 for the detection and prognosis of cardiovascular complications*. Reviews in cardiovascular medicine. https://pubmed.ncbi.nlm.nih.gov/33070540/.

D;, Z. A. K. (n.d.). *Usefulness of machine learning in COVID-19 for the detection and prognosis of cardiovascular complications*. Reviews in cardiovascular medicine. https://pubmed.ncbi.nlm.nih.gov/33070540/.

Jaadi, Z. (n.d.). *A Step-by-Step Explanation of Principal Component Analysis (PCA)*. Built In. https://builtin.com/data-science/step-step-explanation-principal-component-analysis.

Publishing, H. H. (n.d.). *If you've been exposed to the coronavirus*. Harvard Health. https://www.health.harvard.edu/diseases-and-conditions/if-youve-been-exposed-to-the-coronavirus.

Ren, M. (2021, February 7). *COVID-19 Healthy Diet Dataset*. Kaggle. https://www.kaggle.com/mariaren/covid19-healthy-diet-dataset?select=Food\_Supply\_Quantity\_kg\_Data.csv.