Khushi Jain
BSIS Practicum
12/14/2021

Trends in Diet and Covid

Introduction

This semester's research is a reflection of my interests in diet and the Covid-19 virus. Ever since the pandemic, I have been curious about the spread, recovery, and death rate of Covid in various countries around the globe. Some countries have recovered quickly from an epidemic, while others are struggling to control its spread. Key factors that are known to determine how well a country fights a pandemic are adhesion to quarantine regulations, sanitation, how many and how often people are wearing masks, and amount of medical resources. As a diet and fitness enthusiast, I wondered if health can also influence response to Covid. This curiosity led me into reading various journals/articles about Covid-19 and also searching for datasets. In the process of research, the Covid-19 Healthy Diet Dataset from Kaggle caught my attention. The "Food_Supply_kcal_Data.csv" dataset from this link is quite interesting, as it is able to summarize diet consumption for 170 countries. Although intriguing, I found this dataset to be questionable because I think it is quite difficult to determine average food consumption for an entire country. Every individual is very different, has a different lifestyle, and is bound to have very different diets. I believe it is impractical to represent one average diet for an entire country; especially for diverse countries like the US. Although I felt the dataset was a stretch, I was curious to see if it would be effective for analyzing trends between countries' average diets and their Covid infection, recovery, death, active, and total confirmed rates. The data about the diets comes from a reputable source: Food and Agriculture Organization of the United Nations (https://www.fao.org/faostat/en/#home).

Though there are too many extraneous factors to measure a direct correlation between diet and Covid, I felt it would be interesting to see if there is any sort of association. I was wondering if there is any way to predict a country's Covid rates based on their diet consumption from various food groups and rates of physical conditions, such as obesity and undernourishment. Although I wished to analyze data samples representing individuals/small groups rather than countries, I decided to proceed with this dataset, as I could not find a more

elaborate, conveniently formatted, and relevant dataset. Although I aimed to generalize my analysis, my formal research question was : “Is there an association between a country’s average diet and its Covid recovery rates?”

Dataset Overview

Columns:

1) *Country:* Countries around the world
2) *Alcoholic Beverages*: Percentage of energy intake (kcal) from alcoholic beverages
3) *Animal Products*: Percentage of energy intake (kcal) from animal products
4) *Animal fats*: Percentage of energy intake (kcal) from animal fats
5) *Aquatic Products, Other*: Percentage of energy intake (kcal) from aquatic product
6) *Cereals - Excluding Beer*: Percentage of energy intake (kcal) from cereal - excluding beer
7) *Eggs*: Percentage of energy intake (kcal) from eggs
8) *Fish, Seafood*: Percentage of energy intake (kcal) from fish, seafood
9) *Fruits - Excluding Wine*: Percentage of energy intake (kcal) from fruits - excluding wine
10) *Meat*: Percentage of energy intake (kcal) from meat
11) *Milk - Excluding Butte*r: Percentage of energy intake (kcal) from milk - excluding butter
12) *Miscellaneous*: Percentage of energy intake (kcal) from miscellaneous
13) *Offals*: Percentage of energy intake (kcal) from offals
14) *Oil Crops*: Percentage of energy intake (kcal) from oil crops
15) *Pulses*: Percentage of energy intake (kcal) from pulses
16) *Spices*: Percentage of energy intake (kcal) from spices
17) *Starchy Roots*: Percentage of energy intake (kcal) from starchy roots
18) *Stimulants*: Percentage of energy intake (kcal) from stimulants
19) *Sugar Crops*: Percentage of energy intake (kcal) from sugar crops
20) *Sugar & Sweeteners*: Percentage of energy intake (kcal) from sugar and sweeteners
21) *Tree Nuts*: Percentage of energy intake (kcal) from treenuts
22) *Vegetal Products*: Percentage of energy intake (kcal) from vegetal products
23) *Vegetable Oils*: Percentage of energy intake (kcal) from vegetable oils
24) *Vegetables*: Percentage of energy intake (kcal) from vegetables
25) *Obesity*: Obesity rate (%)
26) *Undernourished*: Undernourished rate (%)
27) *Confirmed*: Percentage of confirmed COVID-19 cases
28) *Deaths*: Percentage of confirmed COVID-19 cases
29) *Recovered*: Percentage of COVID-19 recovered
30) *Active*: Percentage of COVID-19 active cases
31) *Population*: Population count

Methods and Analysis

Predicting levels of Recovery% to Death% ratio

With this project, I essentially covered a data pipeline. In the first half of the semester, I spent time planning and implementing basic data processing steps: reading data into a pandas dataframe and removing null values, data manipulation: selecting columns of interest and

inputting them into an X and Y dataframe, and running predictive analytics: K-means clustering. The X dataframe contained the average food consumption in kilocalories from each of the food groups and national obesity and undernourished percentages. The Y array contains levels of percentage of cases of Recovery: high, medium, and low (ranges defined by Recovery% quantiles) .

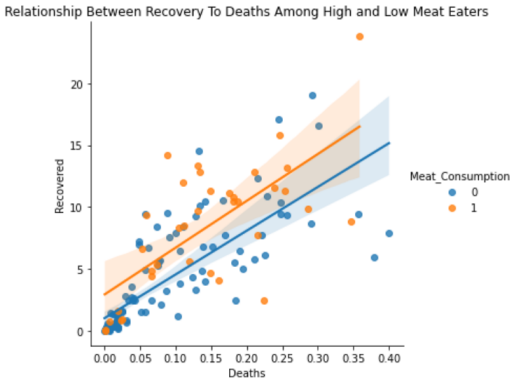
The K-means clustering model was fit to the data in the X dataframe. K-means is an unsupervised classification algorithm that splits inputted data points into K (programmer-specified number) clusters, such that data from each of those clusters is as far apart from each other as possible . The centroids/averages of each cluster are selected to minimize the sum of variances (between each datapoint in a cluster and its respective centroid) for each cluster. New data points are added to the clusters containing the centroid they are closest(distance) to. The points in each cluster are then re-averaged and new centroids are selected. The purpose of this algorithm is to check if the mathematically (model is not trained) created clusters each correspond to a Y classification/label. If the majority of the data points from each cluster have a particular Y classification/label, then the pattern exists. In other words, there is likely a relationship between the selected X variables and Y categorical variable. For my scenario, I am testing to see if my model would naturally split into 3 clusters of countries: low, medium, and high Covid recovery rate.

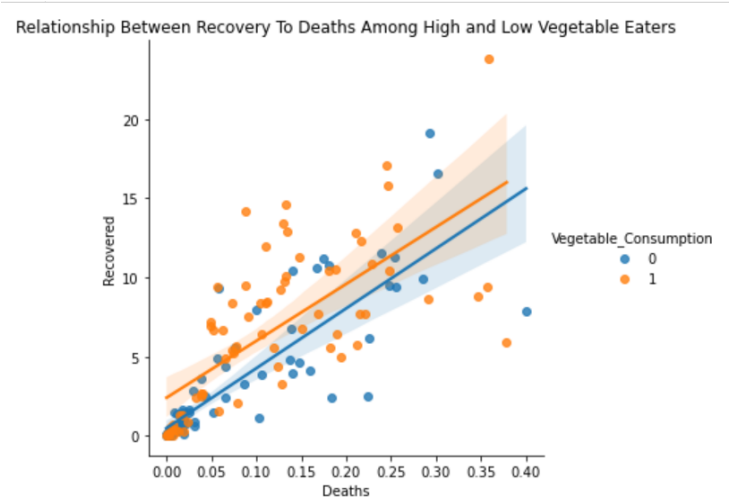
In the second half of the semester, when printing a correlation table, I noticed that Covid recovery and death rates had the same-sign correlation coefficients with most of the food groups. This is problematic for visualizations because this means that food groups causing more deaths were also causing more recoveries. I was initially confused about this, but realized this was because deaths and recoveries were dependent on confirmed cases: deaths + recoveries + active= confirmed. To address this, instead of looking at death and confirmed cases individually, I decided to observe their ratio - Recovery% : Death%. This way the metric of measurement is not dependent on Confirmed cases. So, I added a column calculating Recovery% : Death% to the dataframe.

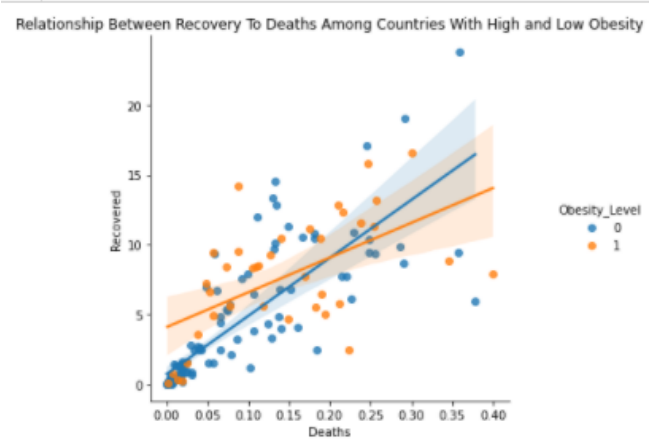
While calling the *pandas "describe"* function on the dataframe, I found out that the summary statistics for the Recovery%: Death% had "inf" and "nan" values. The reason for this was because multiple countries had a 0% death rate or a 0% recovery rate. Even countries like the United States had a 0% recovery rate, which was definitely inaccurate. So, I spent time

searching for datasets that contain updated Covid rates and population values, and came across this Kaggle dataset: Covid19 in World Countries-Latest Data. I had to read-in this dataset and add columns for converting Confirmed, Recovery, Death, and Active cases to a percentage. I also had to account for inconsistent country naming standards in the new dataframe and change various country names to match those in the original dataframe. This is important because I could not just merge both data frames, as they had different columns. So, I had to write a function that updates the Confirmed%, Recovery%, Death%, Active%, and Population columns in the original dataframe whenever a traversed country matches the country name in the updated data frame.

After doing all the required data processing, I printed out some graphs: displayed below. I added columns in my dataframe that classify countries' meat consumption levels, vegetable consumption levels, and population Obesity percentages as low (0) or high(1), based on quantiles. I created 3 Covid Recovery% vs Covid Death% graphs each colored by the columns described in the previous sentence. I made the graphs using the *lmplot* function from the *seaborn* library



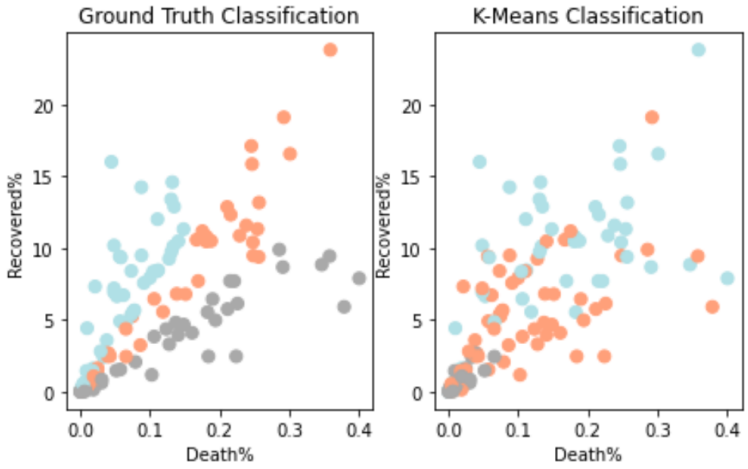




These graphs suggest that the ratio of recovered% : death% does not differ much among points of varying levels of vegetable and meat consumption. However, countries with lower obesity levels are likely to have a greater Recovery%: Death% ratio for countries with a high Confirmed% of cases.

Next, I proceeded to build a clustering model, in which cluster classifications are compared to actual classifications (Y) of Recovery%:Death% ratio: low (0), medium(1), high(2). I leveraged the *KMeans* function from *sklearn.cluster* library to generate the model. The ranges for each classification level were selected based on quantiles. The X_input variables are: 'Alcoholic Beverages', 'Animal Products', 'Animal fats', 'Aquatic Products, Other', 'Cereals - Excluding Beer', 'Eggs', 'Fish, Seafood', 'Fruits - Excluding Wine', 'Meat', 'Milk - Excluding Butter', 'Miscellaneous', 'Offals', 'Oilcrops', 'Pulses', 'Spices', 'Starchy Roots', 'Stimulants', 'Sugar Crops', 'Sugar & Sweeteners', 'Treenuts', 'Vegetal Products', 'Vegetable Oils', 'Vegetables'. The results are displayed below.

K-Means Clustering (Ground Truth Classification: Y = low (gray), med(orange), or high(blue) Recovery%: Death% ratio)



Classification Report For Predicting Recovery% : Death% level: low (0), med(1), or high(2)

Overall Accuracy: 38%

Precision = TruePositives / (TruePositives + FalsePositives)

Recall = TruePositives / (TruePositives + FalseNegatives)

*F1-score = (2 * Precision * Recall) / (Precision + Recall)*

	precision	recall	f1-score	support
0	0.44	0.38	0.41	50
1	0.28	0.34	0.31	50
2	0.45	0.42	0.43	50
accuracy			0.38	150
macro avg	0.39	0.38	0.38	150
weighted avg	0.39	0.38	0.38	150

The results show that the model does a poor job categorizing data according to Recovery%: Death% ratio level. The model only demonstrates a 38% overall accuracy and the clusters in the graphs do not match up at all.

I decided to try a Random Forest Classifier (RFC) Model to see if I can obtain a higher prediction accuracy. A RFC is a supervised learning model that generates a programmer-specified number of decision trees and selects the output based on a voter-based system, where the output for majority of the decision trees becomes the final output. I used the

RandomForestClassifier function from the *Sklearn* library to implement the model. I used a 80-20% training to test ratio to set up X_train, X_test, y_train, y_test. I also ran the model 30 times to compute average accuracy of the RFC model for predicting level of Recovery%: Death% ratio. I used the same X inputs as the previous model. The results are displayed below.

Actual vs Predicted Results for Predicting Recovery% : Death% level using RFC

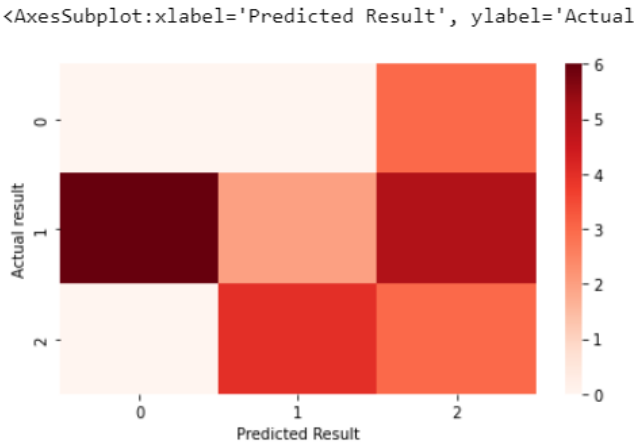
Xaxis- Predicted Result, Yaxis- Actual Result: low (0), med(1), or high(2)

Overall Accuracy: 34.9%

```
0.3492753623188406
```

:

Predicted Result	0	1	2
Actual result			
0	0	0	3
1	6	2	5
2	0	4	3



RFC was also not successful, as it was only able to demonstrate an overall 34.9% testing accuracy in predicting Recovery%: Death% ratio.

I decided to try one other model: Linear Regression (Ordinary Least Squares). I used the OLS function from the statsmodels.formula.api library to implement the model. This time, instead of using levels, I used the Recovered%:Death% ratio as Y directly. This is because the model is a form of regression rather than classification.

<u>Ordinary Least Squares Summary Output Table for Predicting Recovered% : Death% Ratio</u>

R-Squared = 0.472, AIC = 1295, BIC = 1368

OLS Regression Results

Dep. Variable:	Q("Ratio of Recovery to Death")	R-squared:	0.472
Model:	OLS	Adj. R-squared:	0.332
Method:	Least Squares	F-statistic:	3.362
Date:	Fri, 17 Dec 2021	Prob (F-statistic):	1.12e-05
Time:	18:34:12	Log-Likelihood:	-621.67
No. Observations:	120	AIC:	1295.
Df Residuals:	94	BIC:	1368.
Df Model:	25		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	3.402e+06	4.04e+06	0.841	0.402	-4.63e+06	1.14e+07
Q("Alcoholic Beverages")	-3.433e+04	4.05e+04	-0.847	0.399	-1.15e+05	4.61e+04
Q("Animal Products")	-3.396e+04	4.04e+04	-0.840	0.403	-1.14e+05	4.63e+04
Q("Animal fats")	-3.409e+04	4.05e+04	-0.842	0.402	-1.14e+05	4.63e+04
Q("Aquatic Products, Other")	-3.405e+04	4.05e+04	-0.841	0.402	-1.14e+05	4.63e+04
Q("Cereals - Excluding Beer")	-3.432e+04	4.05e+04	-0.847	0.399	-1.15e+05	4.61e+04
Q("Eggs")	-3.408e+04	4.05e+04	-0.842	0.402	-1.14e+05	4.63e+04
Q("Fish, Seafood")	-3.404e+04	4.05e+04	-0.841	0.402	-1.14e+05	4.63e+04
Q("Fruits - Excluding Wine")	-3.432e+04	4.05e+04	-0.847	0.399	-1.15e+05	4.61e+04
Q("Meat")	-3.409e+04	4.05e+04	-0.842	0.402	-1.14e+05	4.63e+04
Q("Milk - Excluding Butter")	-3.408e+04	4.05e+04	-0.842	0.402	-1.14e+05	4.63e+04
Q("Miscellaneous")	-3.433e+04	4.05e+04	-0.847	0.399	-1.15e+05	4.61e+04
Q("Offals")	-3.405e+04	4.05e+04	-0.841	0.402	-1.14e+05	4.63e+04
Q("Oilcrops")	-3.432e+04	4.05e+04	-0.847	0.399	-1.15e+05	4.61e+04
Q("Pulses")	-3.432e+04	4.05e+04	-0.847	0.399	-1.15e+05	4.61e+04
Q("Spices")	-3.428e+04	4.05e+04	-0.846	0.400	-1.15e+05	4.62e+04
Q("Starchy Roots")	-3.432e+04	4.05e+04	-0.847	0.399	-1.15e+05	4.61e+04
Q("Stimulants")	-3.428e+04	4.05e+04	-0.846	0.400	-1.15e+05	4.62e+04
Q("Sugar Crops")	-3.435e+04	4.05e+04	-0.848	0.399	-1.15e+05	4.61e+04
Q("Sugar & Sweeteners")	-3.433e+04	4.05e+04	-0.847	0.399	-1.15e+05	4.61e+04
Q("Treenuts")	-3.43e+04	4.05e+04	-0.846	0.400	-1.15e+05	4.62e+04
Q("Vegetal Products")	-3.373e+04	4.04e+04	-0.835	0.406	-1.14e+05	4.64e+04
Q("Vegetable Oils")	-3.432e+04	4.05e+04	-0.847	0.399	-1.15e+05	4.61e+04
Q("Vegetables")	-3.434e+04	4.05e+04	-0.847	0.399	-1.15e+05	4.61e+04
Q("Obesity")	1.6203	1.020	1.589	0.115	-0.404	3.645
Q("Undernourished")	0.0812	0.583	0.139	0.889	-1.076	1.238

Omnibus:	55.548	Durbin-Watson:	1.915
Prob(Omnibus):	0.000	Jarque-Bera (JB):	237.065
Skew:	1.560	Prob(JB):	3.33e-52
Kurtosis:	9.138	Cond. No.	4.79e+07

The R squared value is 0.472, which means that the model explains 47.2% of the variation in Ratio of Recovery% to Death%. R-squared indicates the predictive power of the linear model, 0

being lowest and 1 being the highest. This model is performing better than the previous ones as they were only able to give an overall accuracy in the 30s%.

Predicting Confirmed% levels

While trying out different models, I discovered that the same models used above are performing much better in predicting levels of infection (Confirmed%) rather than levels of the Ratio% : Death% ratio. I found that very interesting, but I am not sure why this might be happening. This would imply that there is an association between diet and infection rates, which does not biologically make much sense. Results for visualizations/calculations are below.

Ordinary Least Squares Summary Output Table for Predicting Confirmed Covid Case%

R-Squared = 0.643, AIC = 676.6, BIC = 749.1

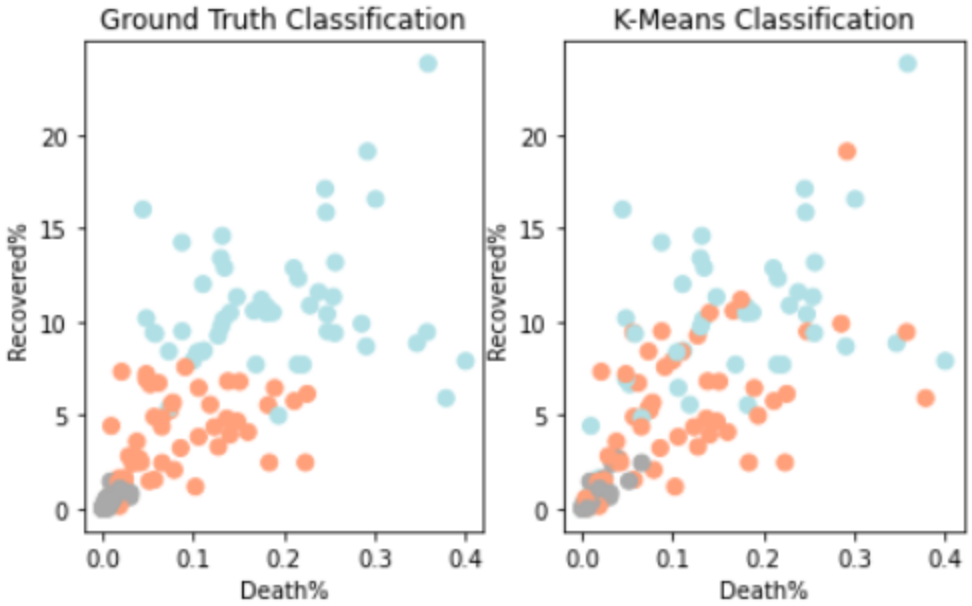
OLS Regression Results

Dep. Variable:	Q("Confirmed")	R-squared:	0.643
Model:	OLS	Adj. R-squared:	0.549
Method:	Least Squares	F-statistic:	6.783
Date:	Fri, 17 Dec 2021	Prob (F-statistic):	3.30e-12
Time:	19:42:20	Log-Likelihood:	-312.31
No. Observations:	120	AIC:	676.6
Df Residuals:	94	BIC:	749.1
Df Model:	25		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.623e+05	3.07e+05	0.529	0.598	-4.47e+05	7.72e+05
Q("Alcoholic Beverages")	-1625.0758	3076.609	-0.528	0.599	-7733.755	4483.603
Q("Animal Products")	-1610.6702	3069.935	-0.525	0.601	-7706.098	4484.758
Q("Animal fats")	-1635.8327	3072.787	-0.532	0.596	-7736.923	4465.258
Q("Aquatic Products, Other")	-1643.1995	3072.809	-0.535	0.594	-7744.335	4457.936
Q("Cereals - Excluding Beer")	-1625.5155	3076.605	-0.528	0.599	-7734.186	4483.155
Q("Eggs")	-1634.1383	3072.481	-0.532	0.596	-7734.622	4466.345
Q("Fish, Seafood")	-1636.5471	3072.718	-0.533	0.596	-7737.501	4464.406
Q("Fruits - Excluding Wine")	-1625.5162	3076.635	-0.528	0.599	-7734.248	4483.215
Q("Meat")	-1635.9709	3072.766	-0.532	0.596	-7737.019	4465.078
Q("Milk - Excluding Butter")	-1635.9082	3072.759	-0.532	0.596	-7736.943	4465.127
Q("Miscellaneous")	-1628.0775	3076.619	-0.529	0.598	-7736.776	4480.621
Q("Offals")	-1642.9820	3072.219	-0.535	0.594	-7742.945	4456.981
Q("Oilcrops")	-1625.5091	3076.711	-0.528	0.599	-7734.391	4483.373
Q("Pulses")	-1625.4441	3076.581	-0.528	0.599	-7734.068	4483.180
Q("Spices")	-1626.1221	3076.713	-0.529	0.598	-7735.007	4482.763
Q("Starchy Roots")	-1625.5212	3076.591	-0.528	0.599	-7734.164	4483.122
Q("Stimulants")	-1619.3500	3076.563	-0.526	0.600	-7727.939	4489.239
Q("Sugar Crops")	-1622.1910	3076.587	-0.527	0.599	-7730.826	4486.444
Q("Sugar & Sweeteners")	-1625.3197	3076.610	-0.528	0.599	-7734.002	4483.363
Q("Treenuts")	-1622.7921	3077.084	-0.527	0.599	-7732.415	4486.831
Q("Vegetal Products")	-1621.3122	3066.012	-0.529	0.598	-7708.951	4466.327
Q("Vegetable Oils")	-1625.6734	3076.610	-0.528	0.598	-7734.354	4483.007
Q("Vegetables")	-1626.2691	3076.546	-0.529	0.598	-7734.823	4482.285
Q("Obesity")	0.1227	0.077	1.585	0.116	-0.031	0.276
Q("Undernourished")	-0.0235	0.044	-0.530	0.597	-0.111	0.064

Omnibus:	4.274	Durbin-Watson:	2.166
Prob(Omnibus):	0.118	Jarque-Bera (JB):	5.234
Skew:	0.104	Prob(JB):	0.0730
Kurtosis:	4.002	Cond. No.	4.79e+07

K-Means Clustering (Ground Truth Classification: Y = low (gray), med(orange), or high(blue) Confirmed%)



Classification Report For Predicting Level of Confirmed Covid Case%: low (0), med(1), or high(2)

Overall Accuracy: 72%

Precision = TruePositives / (TruePositives + FalsePositives)

Recall = TruePositives / (TruePositives + FalseNegatives)

*F1-score = (2 * Precision * Recall) / (Precision + Recall)*

	precision	recall	f1-score	support
0	0.86	0.74	0.80	50
1	0.58	0.70	0.64	50
2	0.72	0.68	0.70	50
accuracy			0.71	150
macro avg	0.72	0.71	0.71	150
weighted avg	0.72	0.71	0.71	150

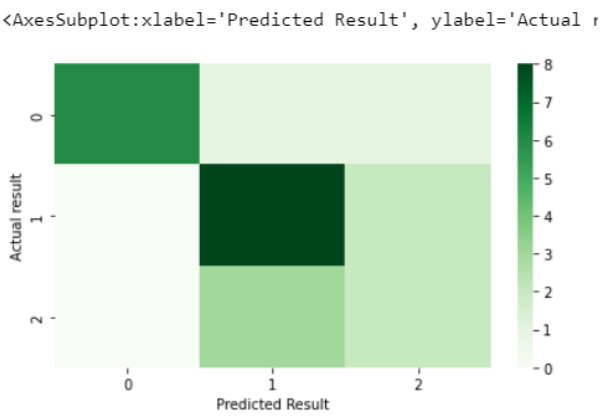
Actual vs Predicted Results for Predicting Level of Confirmed Covid Case% using RFC

Xaxis- Predicted Result, Yaxis- Actual Result: low (0), med(1), or high(2)

Overall Accuracy: 71.74%

0.717391304347826

Predicted Result	0	1	2
Actual result			
0	6	1	1
1	0	8	2
2	0	3	2



The clustering model for predicting Confirmed% levels demonstrates double the accuracy of the Recovery%: Death% model. The R-squared for predicting Confirmed% is also higher, which means the linear regression model has greater predictive power. This linear regression model also has a lower AIC and BIC score, indicating a more parsimonious model. In addition, the RFC model also demonstrates about double the testing accuracy.

Conclusion and Future Work

To conclude, for this dataset, Rain Forest Classifier and K-means Clustering are not the best algorithms for classifying levels of Recovery%: Death%. As predicted earlier, it is very difficult to generalize a diet for an entire country. Furthermore, there are so many additional factors that play a role in recovery, such as medical resources, genes, and rest. Also, if a country has more old people, there will probably be more deaths due to Covid, regardless of the average diet.

Linear Regression (with Ordinary least squares loss function) is still somewhat reliable for predicting Recover%:Death%. However, it may be useful to do hypothesis testing to find out which input-variable coefficients are non-zero, or actually influence the target variable: ratio of Recover%:Death%. Also, it may be beneficial to do recursive feature elimination or some sort of backwards/forward elimination to get a smaller number of input variables and potentially create a more parsimonious model.

This dataset and all 3 models work unexpectedly well for predicting Confirmed% of Covid cases in a country. I suspect that this is most likely a coincidence because I am not aware of any biological reasons why diet might contribute to reducing infection rates. If not a coincidence, healthier diets are probably resulting in weaker symptoms, resulting in not feeling the need to get tested. If fewer people test, there would be fewer confirmed cases.

In the future, I would like to do more in-depth research on machine learning algorithms to figure out any additional important steps and optimal predictive analytics methods for case-to-case scenarios .Type and size of dataset and what is being predicted determine what kind of ML is best suited for the problem. I can still try methods such as PCA analysis and K-nearest neighbors for predictions as well.

I also realized that I didn't observe a correlation table nor a pair plot to find any collinear input variables. Collinear pairs need to be eliminated to prevent exaggerated results and to produce an overall parsimonious model.

I also wanted to explore if any particular food groups are effective for predicting Covid infection rates Confirmed%). So, I ran a RFC model with just 'Animal Products', 'Animal fats', 'Meat' for X, and achieved 64.5% testing accuracy. I also ran a model with just 'Excluding Wine', 'Pulses', 'Spices', 'Vegetables', 'Treenuts', 'Vegetal Products', and 'Starchy Roots' for X and achieved a 62.5% testing accuracy. I also ran a RFC with just physical condition variables: "Obesity" and "Undernourished" and obtained a 61.7% testing accuracy.

In the future, I would like to observe dietary data for individuals infected with Covid rather than an entire country. This would make the data a lot more objective. I would like to analyze how their dietary patterns impact recovery time. I also would like to leverage Mechanical Turk crowdsourcing to get input on the choice of ML algorithms I have leveraged.