FINAL PROJECT REPORT

sxk190069@utdallas.edu

Contents

Executive Summary:

In the past couple months, we have witnessed doctors, nurses, paramedics and thousands of medical workers putting their lives on the frontline to save patients who are infected. And as the battle with COVID-19 continues, we should all ask ourselves – What should we do to help out? What can we do to protect our loved ones, those who sacrifice for us, and ourselves from this pandemic? What steps are important for us to take as a part of non-pharmaceutical intervention?

One simple answer to al these questions is that, we need to adapt healthy diet in order to protect our families and our own health.

To come up with insights I used an available and maintained Dataset- COVID-19 Healthy Diet dataset from Kaggle. This dataset contained relevant information which is been updated weekly with new versions of datasets *(Current version include COVID data from the week of 08/02/2020).*
I used a number of multivariate statistical techniques for analysis and concluded that lack of healthy diet might have impact on the health of families and own leading to the possibility of getting affected by the pandemic.

Introduction:

The USDA Center for Nutrition Policy and Promotion recommends a very simple daily diet intake guideline: 30% grains, 40% vegetables, 10% fruits, and 20% protein, but are we really eating in the healthy eating style recommended by these food divisions and balances?

The dataset consists of combined data of different types of food, world population obesity and

undernourished rate, and global COVID-19 cases count from around the world in order to learn more about how a healthy eating style could possibly help to combat the Corona Virus. And from the dataset, we can gather information regarding diet patterns from countries with lower COVID infection rate, and adjust our own diet accordingly.

I am interested in exploring these issues as we need to protect our families and our own health's by adapting to a healthy diet. In order to learn more about how a healthy eating style could help combat the Corona Virus, we can also gather information regarding diet patterns from countries with lower COVID infection rate, and adjust our own diet accordingly.

B. Data Description:

The entire dataset for this project consist of for different csv files.

The first csv file includes percentage of protein intake from different types of food in countries around the world. The last couple of columns also includes counts of obesity, undernourished, and COVID-19 cases as percentages of the total population for comparison purposes.

The second csv includes percentage of energy intake (kcal) from different types of food in countries around the world. The last couple of columns also includes counts of obesity, undernourished, and COVID-19 cases as percentages of the total population for comparison purposes.

The third csv file includes percentage of fat intake from different types of food in countries around the world. The last couple of columns also includes counts of obesity, undernourished, and COVID-19 cases as percentages of the total population for comparison purposes.

The fourth csv file includes percentage of protein intake from different types of food in countries around the world. The last couple of columns also includes counts of obesity, undernourished, and COVID-19 cases as percentages of the total population for comparison purposes.

All the above files have 170 observations and 32 variables.

C. Preprocessing Data:

Checking missing values, zeros, data type and unique values:

The first step to analyze a new dataset is to know if there are missing values and to understand datatypes.

I used the df_status function coming from the funModeling package which showed the numbers in relative and percentage values along with the infinite and zero statistics.

```
df_status(fat_data)
df_status(supply_kcal_data)
df_status(supply_kg_data)
df_status(protein_data)
```

data.frame 32 x 9 | data.frame 32 x 9 | data.frame 32 x 9 | data.frame 32 x 9

variable <chr>	q_zeros <int>	p_zeros <dbl>	q_na <int>	p_na <dbl>	q_inf <int>	p_inf <dbl>	type <chr>	unique <int>
Country	0	0.00	0	0.00	0	0	character	170
Alcoholic.Beverages	13	7.65	0	0.00	0	0	numeric	158
Animal.Products	0	0.00	0	0.00	0	0	numeric	170
Animal.fats	6	3.53	0	0.00	0	0	numeric	156
Aquatic.Products..Other	143	84.12	0	0.00	0	0	numeric	27
Cereals...Excluding.Beer	0	0.00	0	0.00	0	0	numeric	170
Eggs	0	0.00	0	0.00	0	0	numeric	170
Fish..Seafood	0	0.00	0	0.00	0	0	numeric	170
Fruits...Excluding.Wine	0	0.00	0	0.00	0	0	numeric	169
Meat	0	0.00	0	0.00	0	0	numeric	170

1-10 of 32 rows Previous 1 2 3 4 Next

- q_zeros: quantity of zeros (p_zeros: in percent)
- q_inf: quantity of infinite values (p_inf: in percent)
- q_na: quantity of NA (p_na: in percent)
- type: factor or numeric
- unique: quantity of unique values

1. **Handling Missing Values:**
   I used the function `df_status` that takes a data frame and returns a *status table* that can help us quickly remove features (or variables) based on all the metrics.

```
df_status(fat_data, print_results = FALSE) %>% select(variable, q_na, p_na) %>% arrange(-q_na)
df_fat_data<- fat_data
df_fat_data[is.na(df_fat_data)]=0


df_supply_kcal_data <- supply_kcal_data
df_supply_kcal_data[is.na(df_supply_kcal_data)]=0

df_supply_kg_data<- supply_kg_data
df_supply_kg_data[is.na(df_supply_kg_data)]=0

df_protein_data<- protein_data
df_protein_data[is.na(df_protein_data)]=0
```

To check the number of zeros in the four datasets:

```
df_status(df_fat_data, print_results = FALSE) %>% select(variable, q_zeros, p_zeros) %>%
arrange(-q_zeros)

df_status(df_supply_kcal_data, print_results = FALSE) %>% select(variable, q_zeros, p_zeros) %>%
arrange(-q_zeros)

df_status(df_supply_kg_data, print_results = FALSE) %>% select(variable, q_zeros, p_zeros) %>%
arrange(-q_zeros)

df_status(df_protein_data, print_results = FALSE) %>% select(variable, q_zeros, p_zeros) %>%
arrange(-q_zeros)
```

D. Exploratory Data Analysis:

1. **Getting common statistics:**
   For learning more about the common statistics like the total rows, total columns, column names, I used skim() function which gave detailed summary by providing larger sets of statistics.

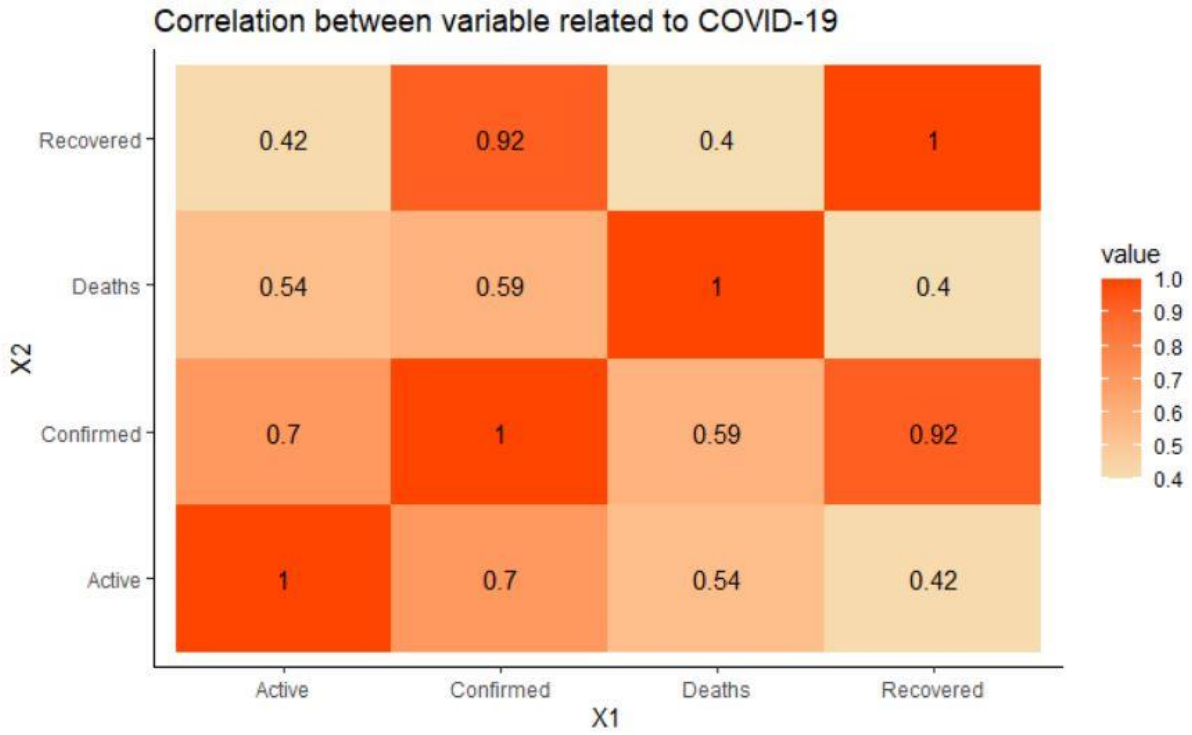
```
skim(fat_data)
skim(df_supply_kcal_data)
skim(df_supply_kg_data)
skim(df_protein_data)
```

```
-- Data Summary ------------------------
                           Values
Name                       df_fat_data
Number of rows             170
Number of columns          32
_____________________
Column type frequency:
  character                3
  numeric                  29
_____________________
Group variables            None

-- Variable type: character ------------------------------------------------------------------------
# A tibble: 3 x 8
  skim_variable               n_missing complete_rate   min   max empty n_unique whitespace
* <chr>                           <int>         <dbl> <int> <int> <int>    <int>      <int>
1 Country                             0             1     4    34     0      170          0
2 Undernourished                      0             1     1     4     0       99          0
3 Unit..all.except.Population.        0             1     1     1     0        1          0

-- Variable type: numeric --------------------------------------------------------------------------
# A tibble: 29 x 11
   skim_variable            n_missing complete_rate    mean      sd         p0     p25     p50
 * <chr>                        <int>         <dbl>   <dbl>   <dbl>      <dbl>   <dbl>   <dbl>
 1 Alcoholic.Beverages              0             1 1.03e-3 9.57e-3     0       0.      0.
 2 Animal.Products                  0             1 2.07e+1 8.00e+0     5.02    1.49e+1 2.09e+1
 3 Animal.fats                      0             1 4.14e+0 3.29e+0     0.0262  1.67e+0 3.31e+0
 4 Aquatic.Products..Other          0             1 4.50e-4 4.04e-3     0       0.      0.
```

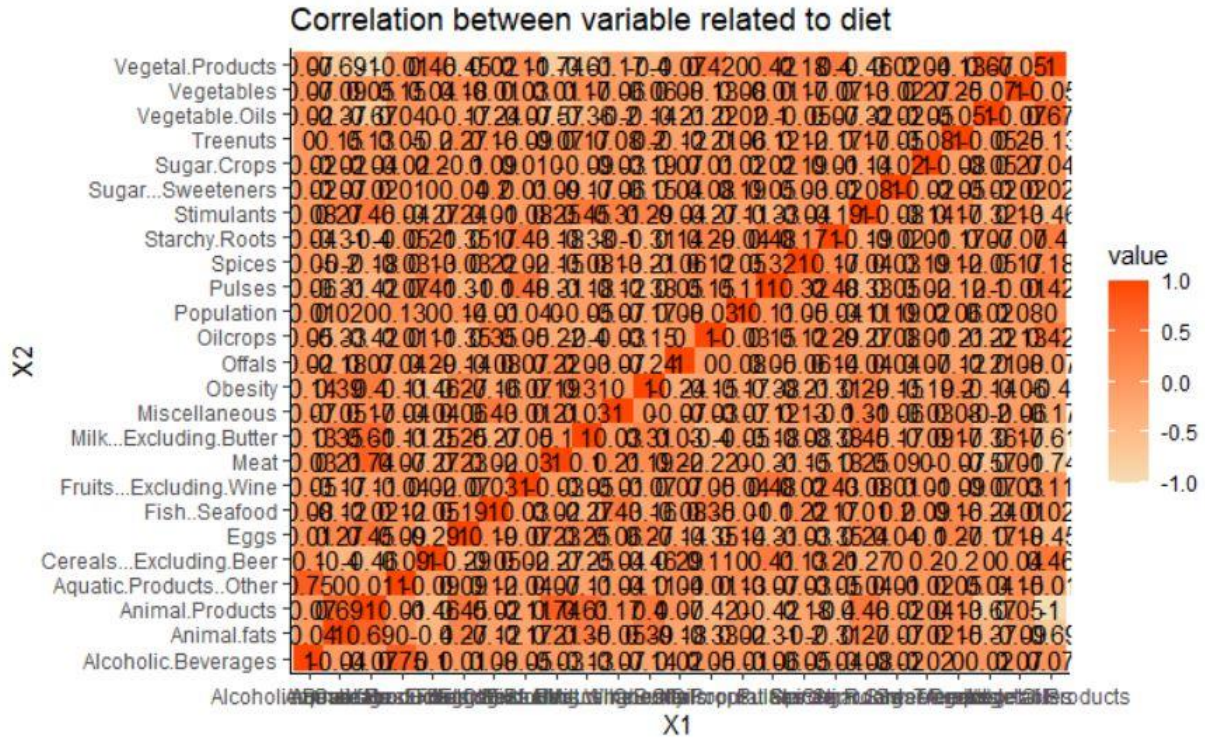
29	Population			0

	p75	p100	hist
*	<dbl>	<dbl>	<chr>
1	0.	9.76e-2	
2	2.69e+1	3.69e+1	
3	6.23e+0	1.49e+1	
4	0.	5.11e-2	
5	5.59e+0	1.84e+1	
6	1.28e+0	3.28e+0	
7	1.11e+0	8.41e+0	
8	5.78e-1	9.67e+0	
9	1.18e+1	2.64e+1	
10	7.63e-2	4.56e-1	
11	7.32e+0	1.78e+1	
12	1.89e-1	7.27e-1	
13	3.51e+0	2.86e+1	
14	3.42e-1	2.69e+0	
15	3.44e-1	2.69e+0	
16	2.12e-1	2.18e+0	

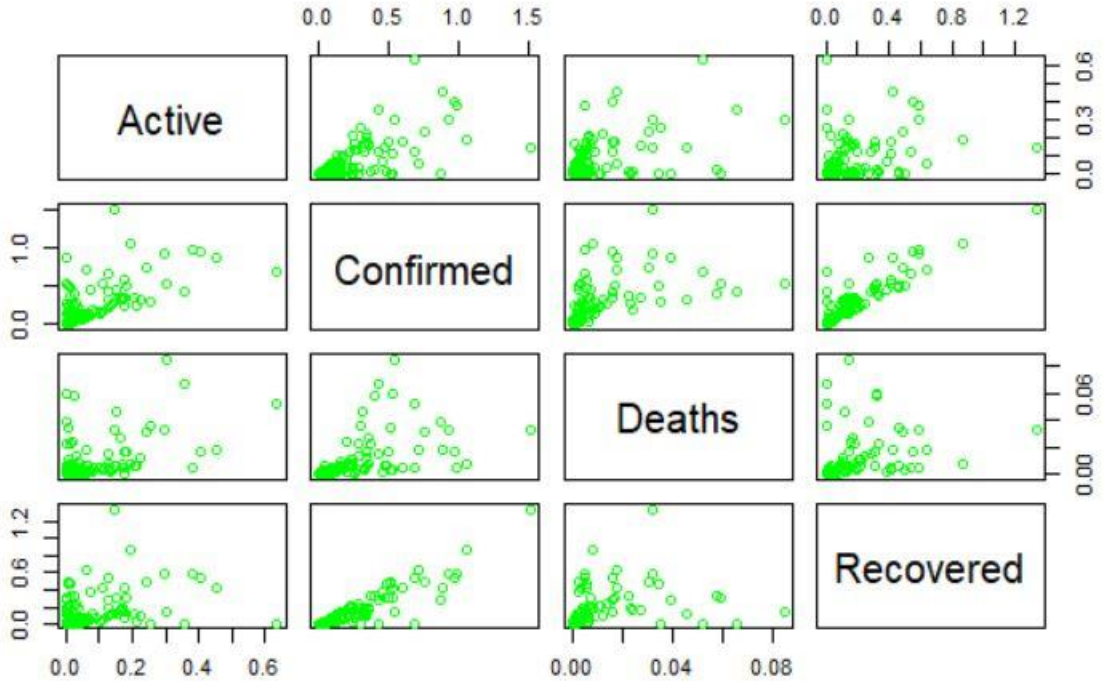
2. Correlation and relationship:



The above heat map shows the generalized correlation between the variable related to COVID-19 from which we can clearly see that there is a strong correlation between confirmed and recovered. Although, this is a generalized result, it might differ for various countries.



The above heat map shows the correlation between variables related to different types of food, world population obesity and undernourished rate.



The above scatter plot shows the spread of the variables related to COVID-19

E. Empirical Analysis:

I used Principal Component Analysis to better visualize the variation present in the dataset and in order to identify which variables are going to be useful in order to run model and perform further analysis by considering only those variables which gives 99% of cumulative proportion.

```
Importance of components:
                          PC1     PC2    PC3     PC4     PC5     PC6     PC7     PC8     PC9    PC10    PC11    PC12    PC13
Standard deviation     2.5583 1.62481 1.3910 1.30239 1.22296 1.19629 1.07943 1.05578 1.02713 0.9998 0.96331 0.90487 0.89469
Proportion of Variance 0.2337 0.09429 0.0691 0.06058 0.05342 0.05111 0.04161 0.03981 0.03768 0.0357 0.03314 0.02924 0.02859
Cumulative Proportion  0.2337 0.32802 0.3971 0.45770 0.51112 0.56223 0.60384 0.64365 0.68133 0.7170 0.75017 0.77942 0.80800
                          PC14    PC15    PC16    PC17    PC18    PC19    PC20   PC21    PC22    PC23    PC24    PC25
Standard deviation     0.82617 0.79929 0.78425 0.75506 0.73115 0.69702 0.67684 0.6438 0.61554 0.56035 0.51494 0.13514
Proportion of Variance 0.02438 0.02282 0.02197 0.02036 0.01909 0.01735 0.01636 0.0148 0.01353 0.01121 0.00947 0.00065
Cumulative Proportion  0.83238 0.85520 0.87716 0.89753 0.91662 0.93397 0.95033 0.9651 0.97866 0.98988 0.99935 1.00000
                           PC26       PC27         PC28
Standard deviation     0.0005876 0.0004224 0.000008216
Proportion of Variance 0.0000000 0.0000000 0.000000000
Cumulative Proportion  1.0000000 1.0000000 1.000000000
```

For example: The above figure shows the result of PCA on the file consisting of protein data.

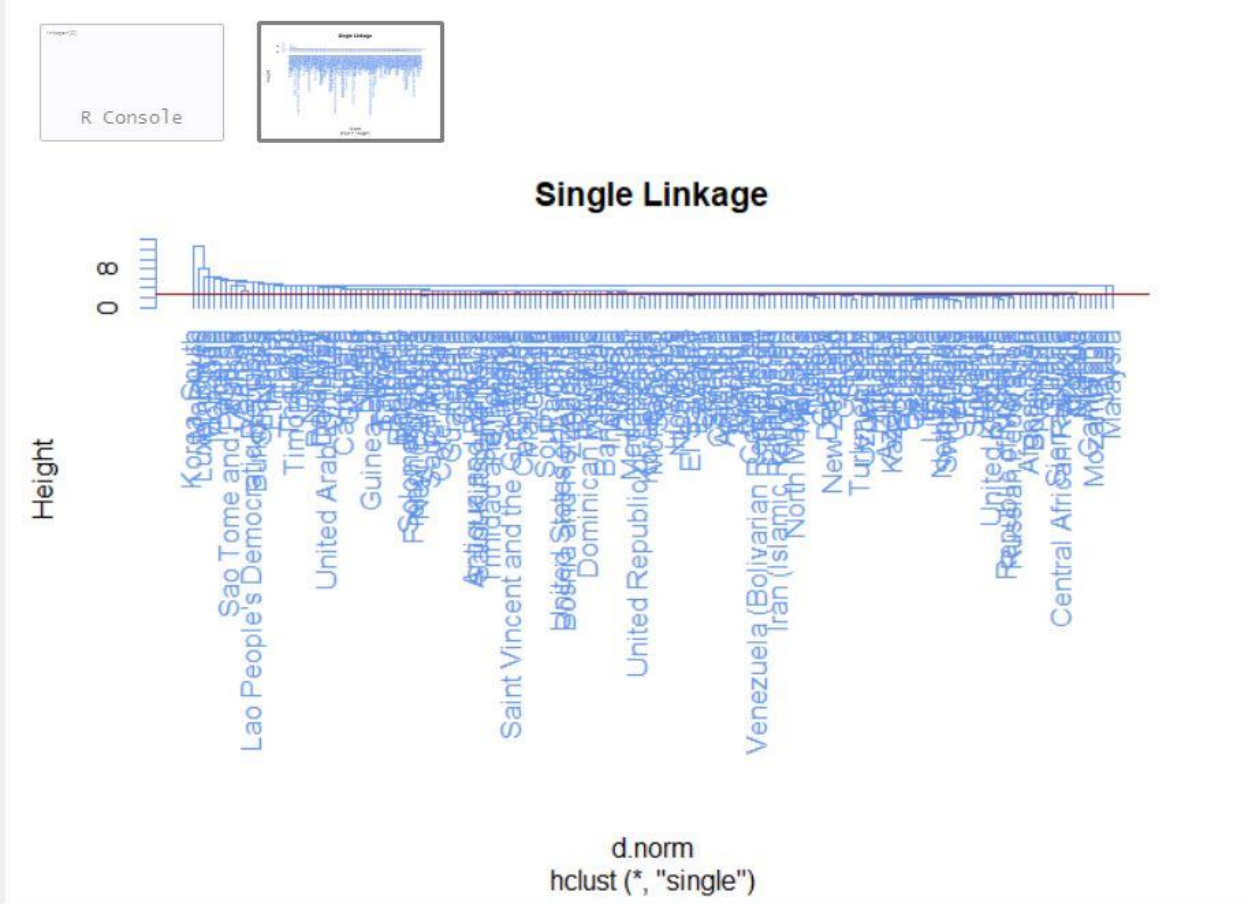
It shows that variables till PC24 gives 99% of cumulative proportion and hence I will be considering only those variables.

Data Partition:

I performed data partition by first randomly ordering all the files of the dataset and then splitting the data into 80% training and 20% test/ validation data.

Further I performed Clustering in order to get insights of which variable is prominent and which in not in all the clusters and get an idea about the spread of the variables.

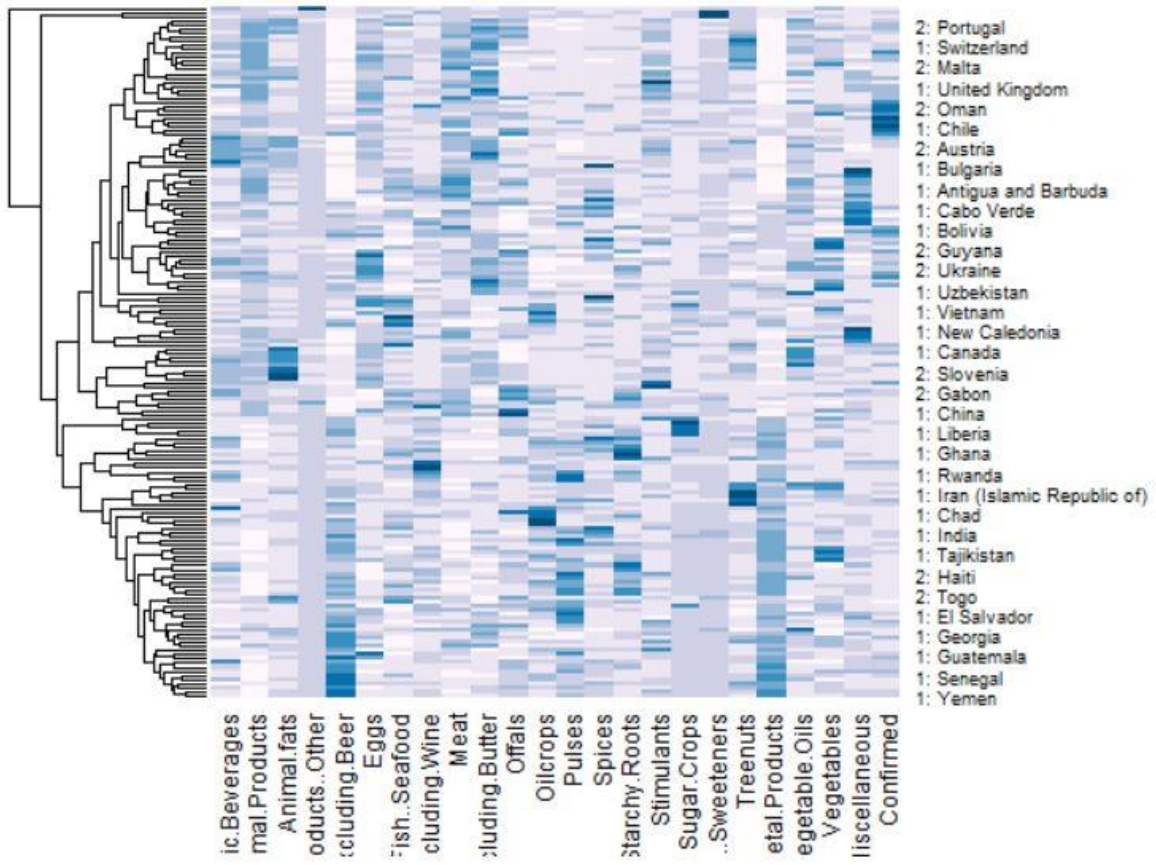
I choose hierarchical clustering technique to do the analysis further using dendrograms and membership in the clusters to perform in-depth analysis.



The above figure shows the dendrogram of protein content in the diet of people across countries.

Further I validated the clusters using heatmaps as follows:

i



Continued by assigning the test observations to the clusters built upon the training set and then using the subset function to build data frames which I used to perform regression.

I used Simple Linear Regression Method to perform regression on the clusters by keeping the Confirmed variable which tells us about the confirmed cases of covid-19 as the dependent variable.

I performed prediction based on the above regression and ended up evaluating the accuracy to compute the overall test-set accuracy of the cluster-then-predict approach, by combining all the test-set predictions into a single vector and all the true outcomes into a single vector and then evaluated the overall accuracy.

F. Conclusion:

Using the dataset I analyzed that which all countries have a certain eating habits leasing to the possibility of affecting their health and being prone to COVID-19.

I concluded that eating healthy and having a healthy diet does have a significant affect on the country for being more restraint against the pandemic.

Sources:

Website:

https://www.kaggle.com/mariaren/covid19-healthy-diet-dataset?select=Protein_Supply_Quantity_Data.csv

https://rstudio-pubs-static.s3.amazonaws.com/73672_c9cd25c8b1ab413490748e75a5aeba2b.html