Homework 4 Topic: Clustering Analysis & Visualization on Fruit Consumption v.s. GDP per capita

Li, Peifeng
Nov 20th , 2021

Topic Background and Objectives:

Fruits, a kind of food have low in calories and fat and are a source of natural sugars, fibers and vitamins. According to USDA's food pyramid and nutrition guidelines, fruits is one of the basic four food groups that people should include in their daily meals.[1]

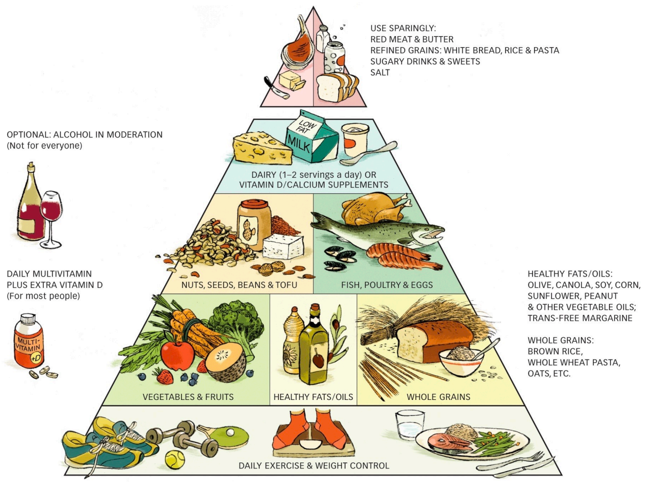
However, the unbalanced development of the world make some countries have the insufficient fruit supplement, which may lead to a serious of health problems in the residents like scurvy caused by lack of vitamin C or difficulty in maintaining a healthy body weight.[3]

As we all known, most of fruits basically grow in the places that are hot and with plenty of sunshine. But typically the countries in these area are not fully developed, which causes the unbalanced production of the fruit - which means some developed countries may pay more to get enough fruit supplement from other export country.

This topic is focus on analyzing the the relationship between fruits consumptions and the GDP per capita. By using high dimension reduction methods like MDS and tSNE to show the dissimilarities between entities and clustering methods like KMeans, Spectral Clustering and Agglomerative Clustering, we may find out countries that project to have similar characteristics. The insight may help the organizations to improve the healthy diet across the world.

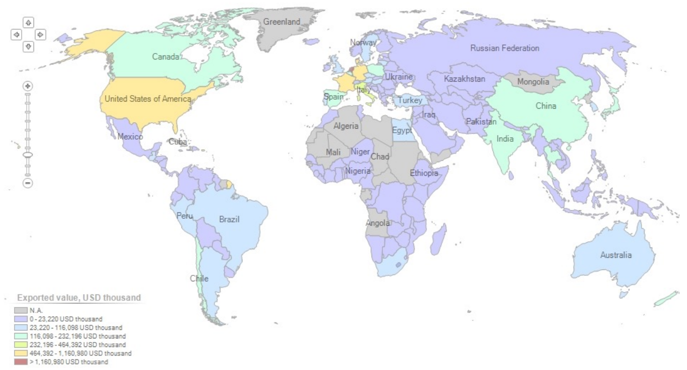
THE HEALTHY EATING PYRAMID

Department of Nutrition, Harvard School of Public Health



List of exporters for the selected product in 2020

Product : 1209 Seeds, fruits and spores, for sowing (excluding leguminous vegetables and sweetcorn, coffee, tea, maté and spices, cereals, oil seeds and oleaginous fruits, and seeds and fruit used primarily in perfumery, medicaments or for insecticidal, fungicidal o



Source:

[1] Wikipedia contributors. "Food pyramid (nutrition)." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia.

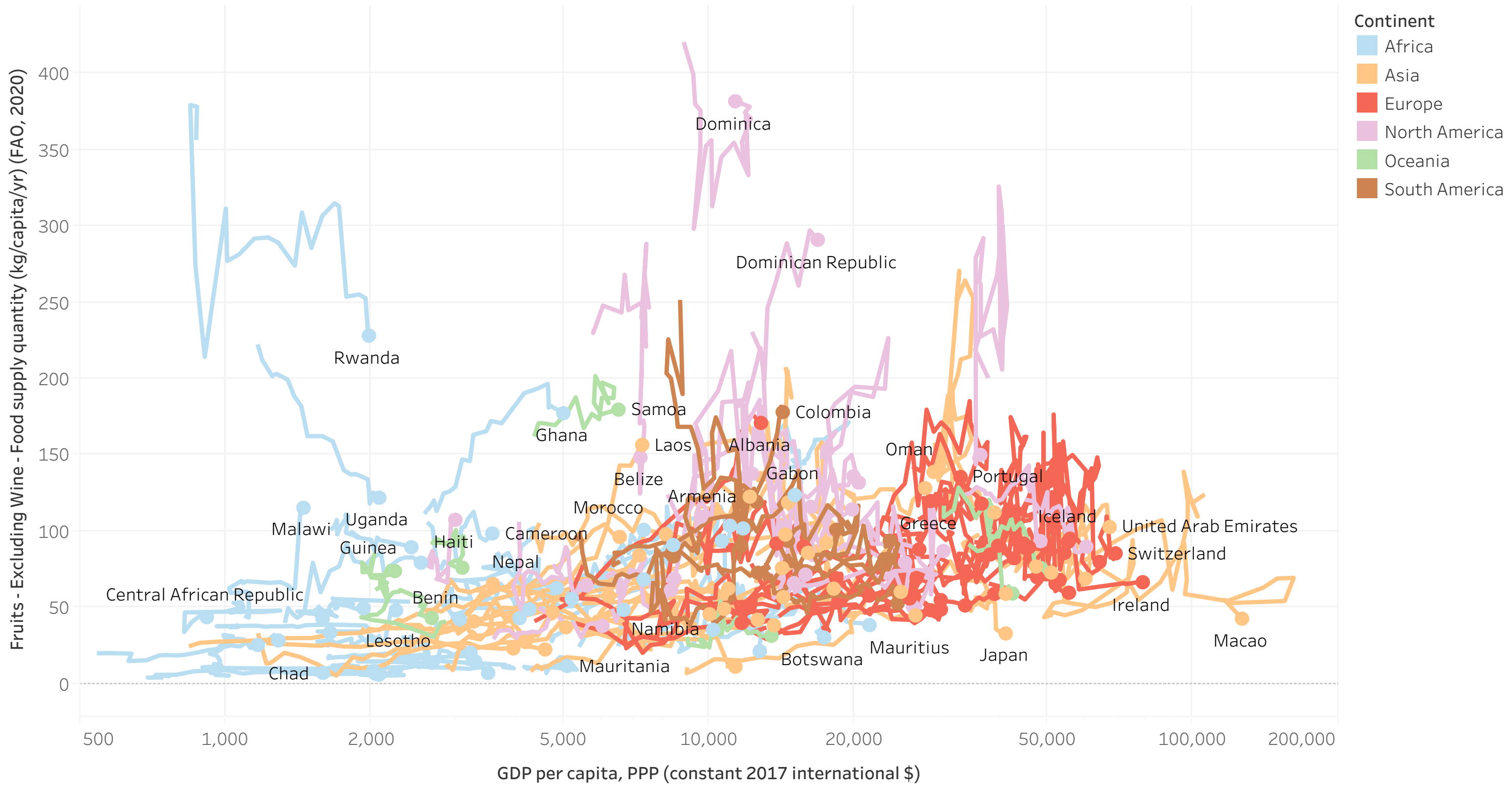
[2] Copyright © 2008. For more information about The Healthy Eating Pyramid, please see The Nutrition Source, Department of Nutrition, Harvard T.H. Chan School of Public Health, www.thenutritionsource.org, and **and** Eat, Drink, and Be Healthy, by Walter C. Willett, M.D., and Patrick J. Skerrett (2005), Free Press/Simon & Schuster Inc."

[3] Promoting fruit and vegetable consumption, World Health Organization

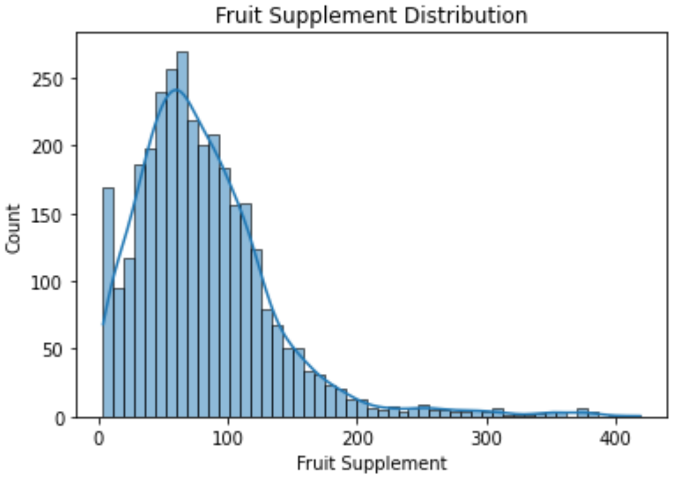
[4] 2020 World Fruit Export Map, Trade Map Org

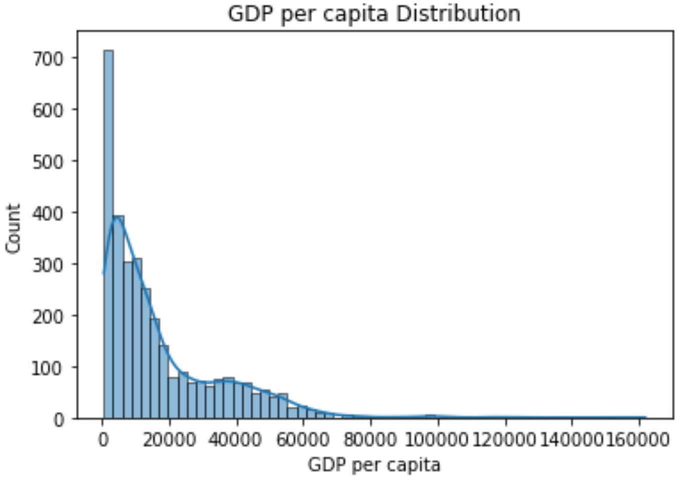
Overview:

Fruit Supplement v.s. GDP per capita, 1997 - 2017, shown by Continent



Sum of GDP per capita, PPP (constant 2017 international $) vs. sum of Fruits - Excluding Wine - Food supply quantity (kg/capita/yr) (FAO, 2020). Color shows details about Continent. The marks are labeled by Entity (fruit-consumption-vs-gdp-per-capita projections.csv). Details are shown for Entity.





The GDP per capita and the fruits supplement varies in each continents. Thus, we cannot suppose a sure conclusion by only looking at the geographical characteristics. It is interesting to see that the fruits supplement in most of entities has slight increases as the GDP per capita increases.

From the distribution graph we can see fruit supplement each year basically in the range of 50 - 100 kgs, While the GDP per capita is in the range of 4,000 – 20,000 dollars per year.

MDS and tSNE Dimension Reduction:

On the left panel we plot the entities after MDS with the bubble size as their GDP per capita (upper left) or Fruits supplement (bottom left). The entity bubbles basically falls in the upper right of the plot, with the bottom left empty.

The GDP per capita increases when we look from the right to the left. Fruit supplement, on the contrary, increases from top to the bottom.

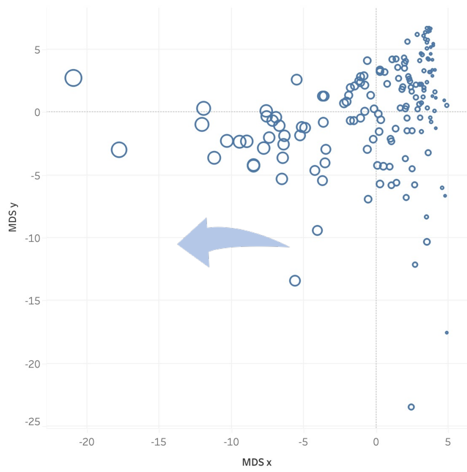
We may conclude that there may be three kind of countries. First, low GDP per capita and low fruit supplement – the top right corner of the MDS plot; Second, high GDP per capita but low fruit supplement – the top left corner of the MDS plot; Third, low GDP per capita but high fruit supplement – the bottom right corner of the MDS plot.

On the left panel we plot the entities after tSNE with the bubble size as their GDP per capita (upper right) or Fruits supplement (bottom right). The entity bubbles basically falls across the plot, with the bottom left empty.

The GDP per capita increases when we look from the top to the bottom. Fruit supplement, on the contrary, increases from left to the right.

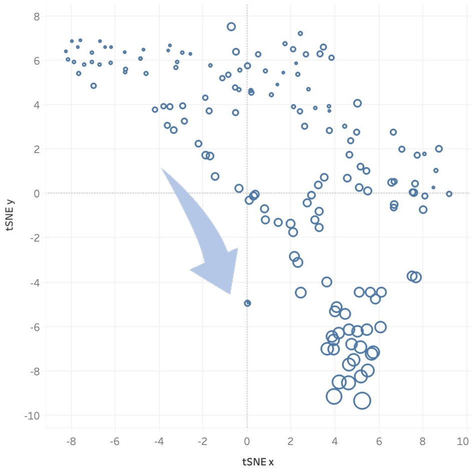
The three kind of countries we concluded on MDS could be founded as follows: First, low GDP per capita and low fruit supplement – the top right corner of the tSNE plot; Second, high GDP per capita but low fruit supplement – the bottom right corner of the tSNE plot; Third, low GDP per capita but high fruit supplement – the top right corner of the tSNE plot.

MDS - Continent



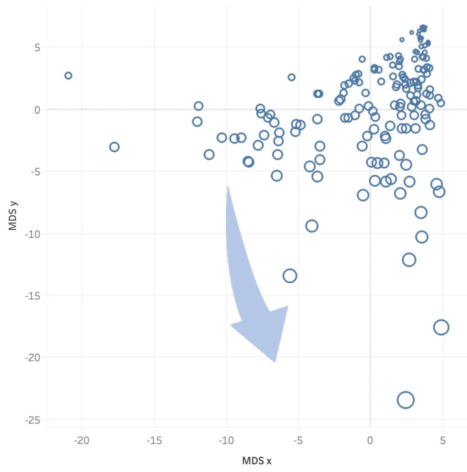
Sum of MDS x vs. sum of MDS y. Size shows sum of GDP per capita, PPP (constant 2017 international $). Details are shown for Entity.

tSNE - Continent



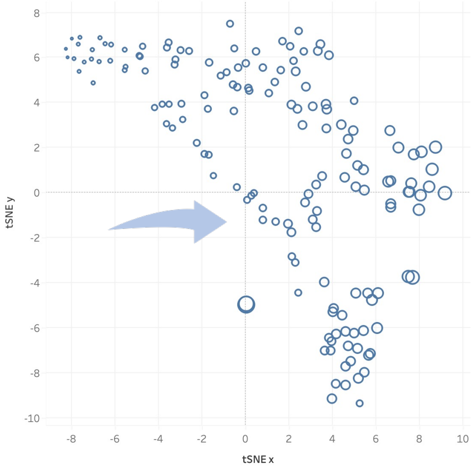
Sum of tSNE x vs. sum of tSNE y. Size shows sum of GDP per capita, PPP (constant 2017 international $). Details are shown for Entity.

MDS - Continent



Sum of MDS x vs. sum of MDS y. Size shows sum of Fruits - Excluding Wine - Food supply quantity (kg/capita/yr) (FAO, 2020). Details are shown for Entity.

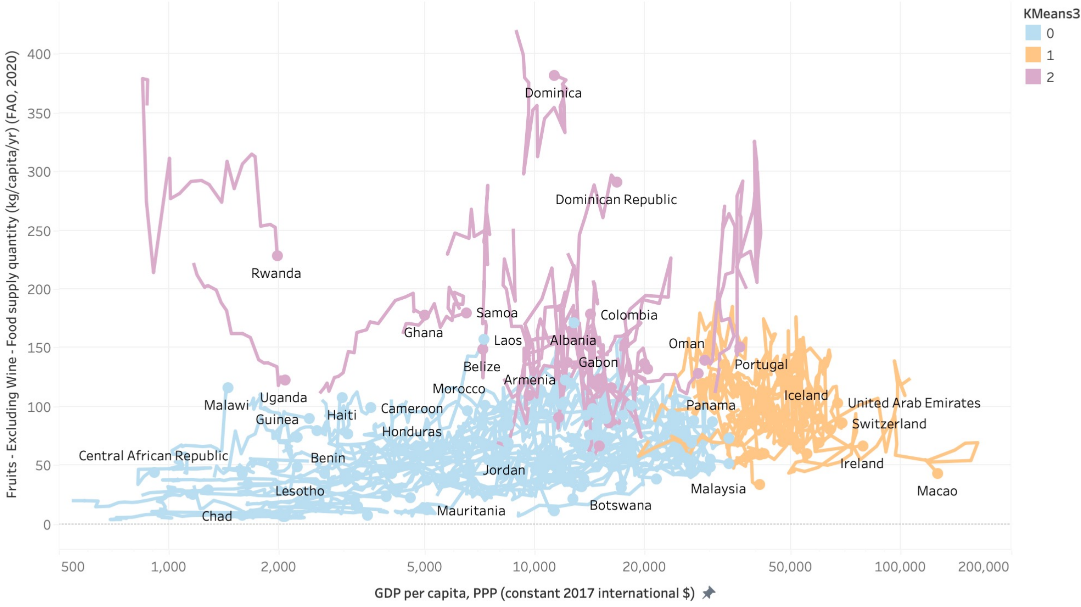
tSNE - Continent



Sum of tSNE x vs. sum of tSNE y. Size shows sum of Fruits - Excluding Wine - Food supply quantity (kg/capita/yr) (FAO, 2020). Details are shown for Entity.

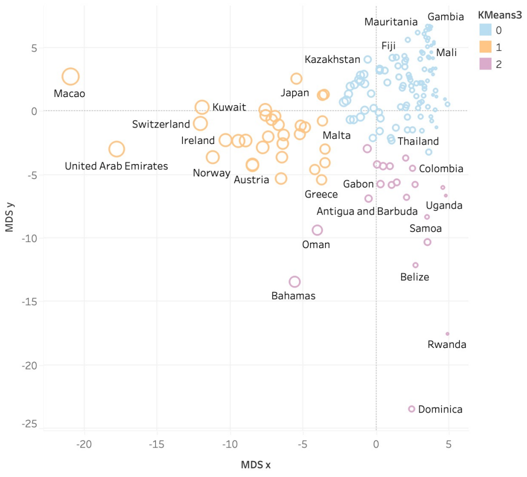
KMeans Projection:

Connected Scatter Plot - Kmeans 3 - 1997 - 2017



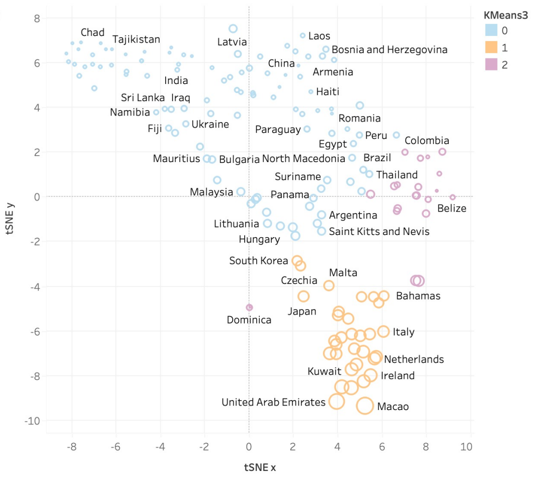
Sum of GDP per capita, PPP (constant 2017 international $) vs. sum of Fruits - Excluding Wine - Food supply quantity (kg/capita/yr) (FAO, 2020). Color shows details about KMeans3. The marks are labeled by Entity. Details are shown for Entity.

MDS - KMeans



Sum of MDS x vs. sum of MDS y. Color shows details about KMeans3. Size shows sum of GDP per capita, PPP (constant 2017 international $). The marks are labeled by Entity. Details are shown for Entity.

tSNE - KMeans



Sum of tSNE x vs. sum of tSNE y. Color shows details about KMeans3. Size shows sum of GDP per capita, PPP (constant 2017 international $). The marks are labeled by Entity. Details are shown for Entity.

The best KMeans model with K = 3 projects the entities into 3 major group, which is very closed to our conclusion in the MDS and tSNE projections. We can see the details in the MDS and tSNE map.

Yellow bubbles represent the high GDP per capita but low fruits supplement countries – e.g. Switzerland, UAE, Japan etc. These countries often are developed countries that have a impressive economy but due to very little land is used in growing the fruits, the fruits supplement rely on import. As shown in the connected scatter plot that GDP per capita of these entities are above 20,000 dollars, the fruit supplement basically in a range of 50 – 150 kgs.

Blue bubbles represent the low GDP per capita and low fruits supplement countries – e.g. China, Ukraine, Malaysia etc. These countries are basically developing countries or poor countries. People in these countries may not spend too much on fruits since fruits is a kind of luxury food than the basic grains.

Purple bubbles represent low GDP per capita but the high fruits supplement countries. Like Rwanda, Dominica, Colombia. These countries locates in tropics, which good for growing fruits. However, the tropic area is hard to develop, leads to a lower GDP than other countries.

Spectral Clustering Projection:

The best Spectral Clustering model with K = 4 projects the entities into 4 major group.The model further divide the the high GDP per capita but low fruits supplement countries in to two smaller group, representing as purple and blue objects in the plots.

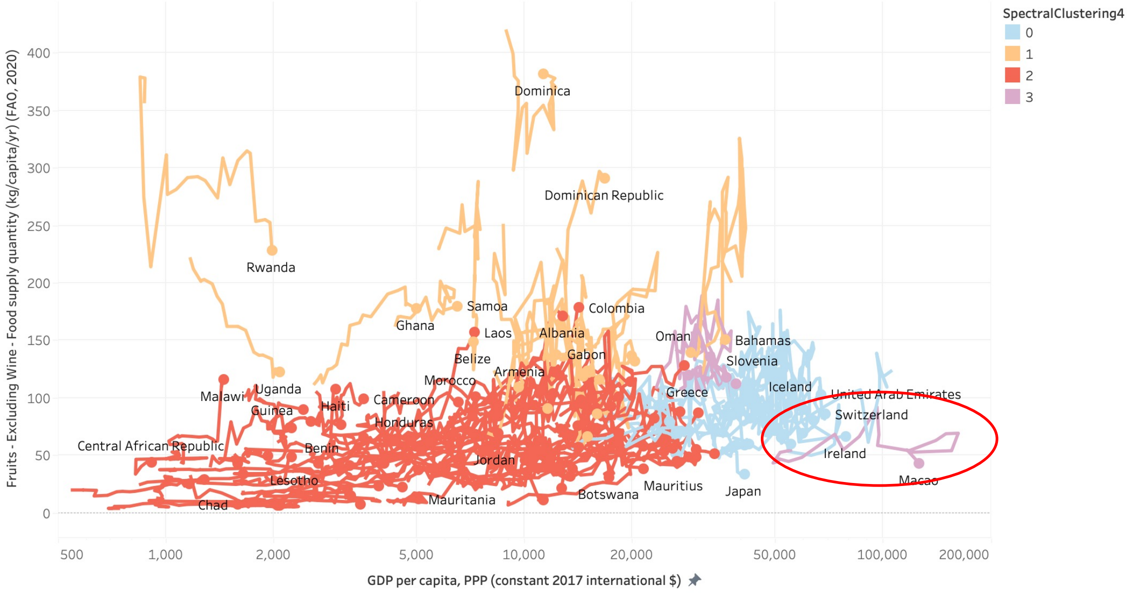
Since the other group are almost the same as the KMeans result, we are focusing on the difference between blue and purple group.

Both of blue and purple bubbles are representing the high GDP per capita but low fruits supplement countries. However, the entities with purple color – Macau, Malta, Greece tend to have a smaller land area than the blue ones.

Macau in the Purple bubbles seems like an outlier that it has a very large inertia within the cluster. Also we can see the Dominica in the tSNE plot is far from the orange group - low GDP per capita but the high fruits supplement countries

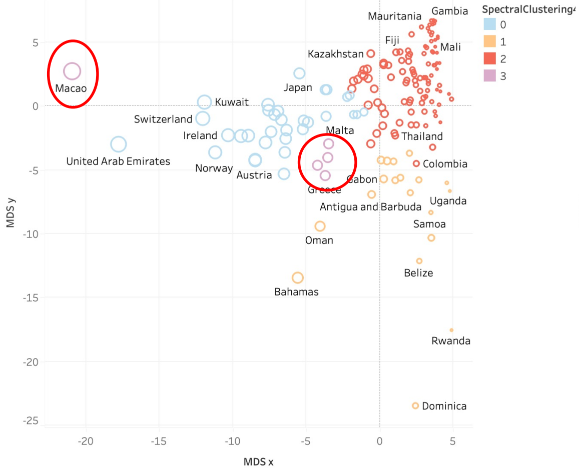
We may conclude that the spectral clustering might not work as good as the KMeans clustering model.

Connected Scatter Plot -SpectualClustering4 - 1997 - 2017



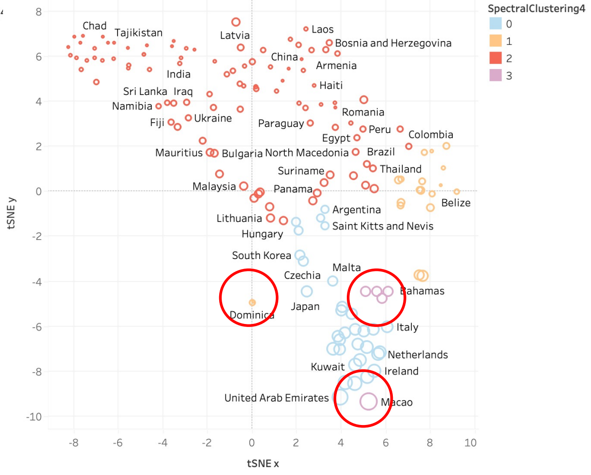
Sum of GDP per capita, PPP (constant 2017 international $) vs. sum of Fruits - Excluding Wine - Food supply quantity (kg/capita/yr) (FAO, 2020). Color shows details about SpectralClustering4. The marks are labeled by Entity. Details are shown for Entity.

MDS - SpectralClustering



Sum of MDS x vs. sum of MDS y. Color shows details about SpectralClustering4. Size shows sum of GDP per capita, PPP (constant 2017 international $). The marks are labeled by Entity. Details are shown for Entity.

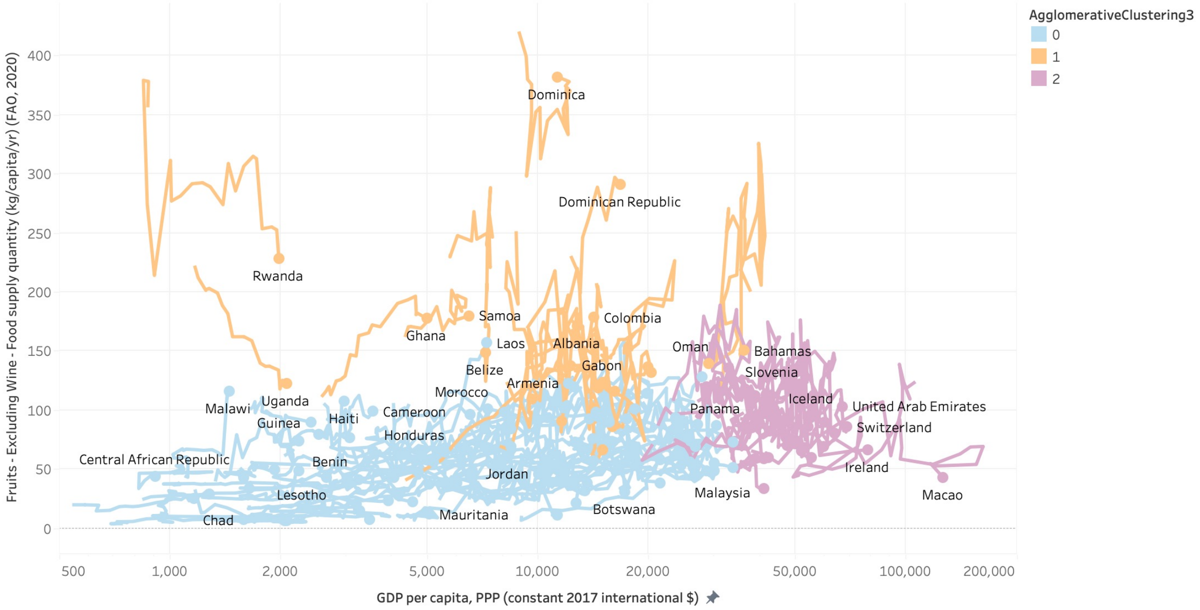
tSNE - Spectral Clustering



Sum of tSNE x vs. sum of tSNE y. Color shows details about SpectralClustering4. Size shows sum of GDP per capita, PPP (constant 2017 international $). The marks are labeled by Entity. Details are shown for Entity.

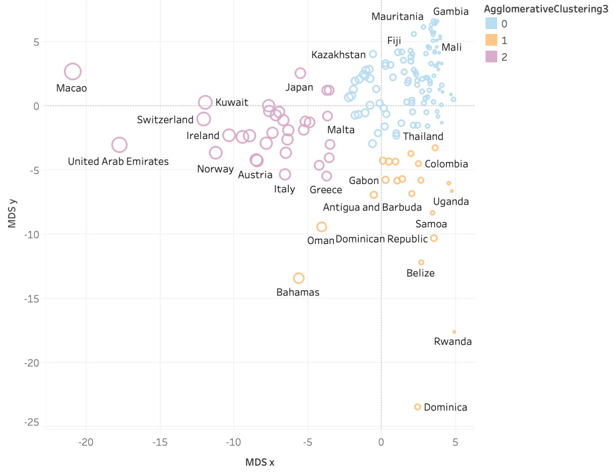
Agglomerative Clustering Projection:

Connected Scatter Plot - AgglomerativeClustering3 - 1997 - 2017



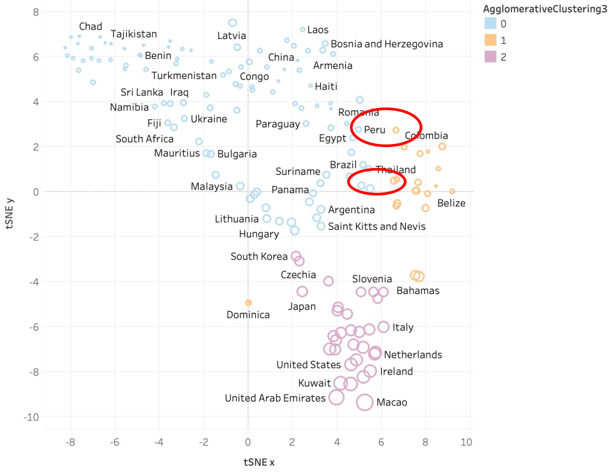
Sum of GDP per capita, PPP (constant 2017 international $) vs. sum of Fruits - Excluding Wine - Food supply quantity (kg/capita/yr) (FAO, 2020). Color shows details about AgglomerativeClustering3. The marks are labeled by Entity. Details are shown for Entity.

MDS - AgglomerativeClustering



Sum of MDS x vs. sum of MDS y. Color shows details about AgglomerativeClustering3. Size shows sum of GDP per capita, PPP (constant 2017 international $). The marks are labeled by Entity. Details are shown for Entity.

tSNE - Agglomerative Clustering



Sum of tSNE x vs. sum of tSNE y. Color shows details about AgglomerativeClustering3. Size shows sum of GDP per capita, PPP (constant 2017 international $). The marks are labeled by Entity. Details are shown for Entity.

The best Agglomerative Clustering model with K = 3 projects the entities into 3 major group, which is very closed to our conclusion in the MDS and tSNE projections. Also the KMeans model projections.

Interestingly, the difference between KMeans and Agglomerative Clustering projections is the boundary between two clusters.

As the red circles shown in the tSNE map, the different clustering algorithms assign the countries near the clustering boundary into different clusters.

It is because the points near the boundary is far from the clustering centroid. However, the idea of the clustering is to minimize the distance within the cluster and maximize the distance between different clusters. Thus, the different algorithms might have different results according to the boundary cases.

Except from that, the two algorithms are performs pretty good at divide the entities into three major clusters.

Conclusions:

- The major trend between the fruit supplements and GDP per capita is: as the GDP per capita growing, the fruit consumption growing in a slow fashion. But we also notice that fruits supplement in some high fruits supplement countries drops as their economy goes up.

- We can divide the entities into three major type of countries base on our high dimension reduction and clustering model:
    - a) high GDP per capita but low fruits supplement countries – e.g. Switzerland, UAE, Japan etc.
    - b) low GDP per capita and low fruits supplement countries – e.g. China, Ukraine, Malaysia etc.
    - c) low GDP per capita but high fruits supplement countries -e.g. Rwanda, Dominica, Colombia.

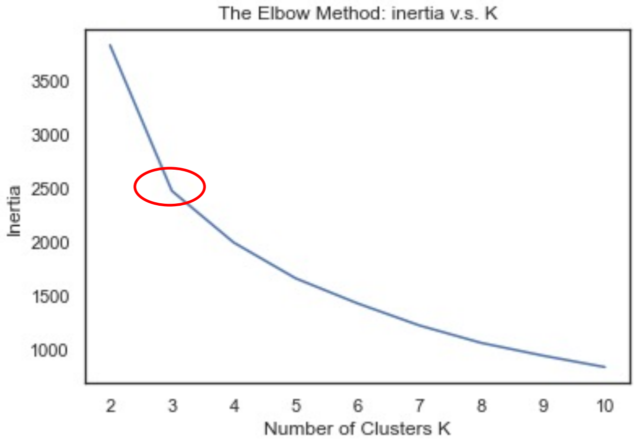
- Different clustering methods may have very similar results, basically the differences occurs near the cluster boundary.

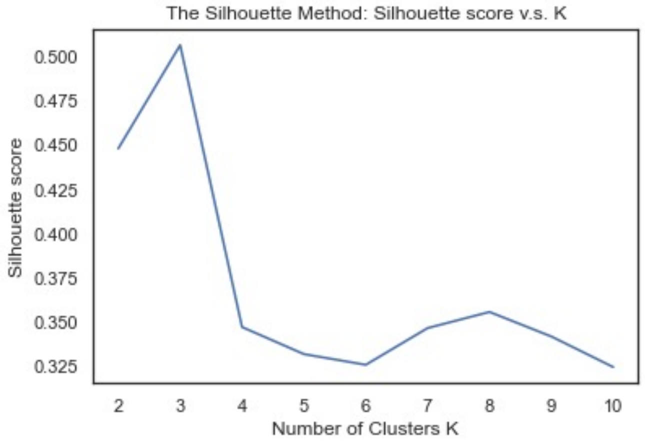
Appendix

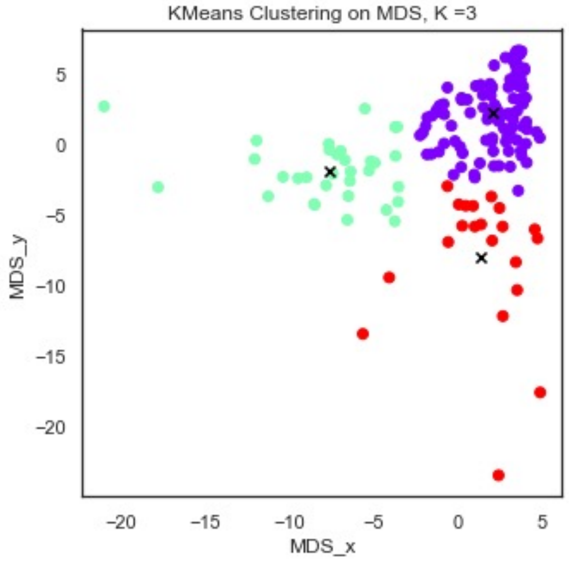
KMeans – Hyperparameter Tuning and Model Selection

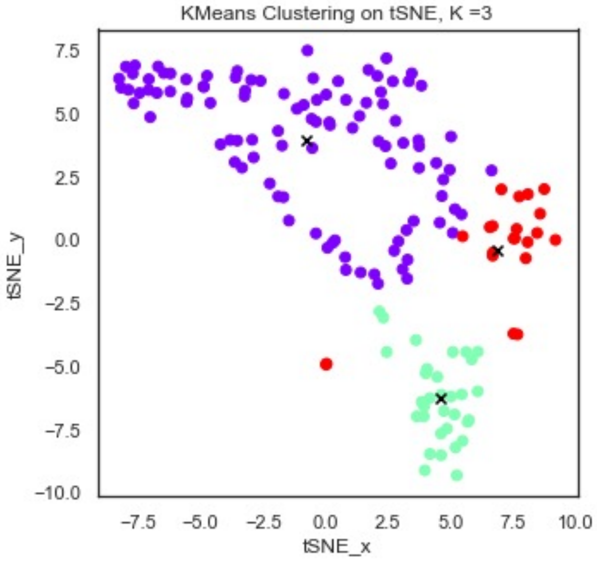
Elbow method is a naïve method to choose the best k. Its idea is to calculate the inertia(a.k.a. Within-Cluster-Sum of Squared Errors (WSS)) for different values of k, and choose the k for which inertia becomes first starts to diminish. In the plot of inertia v.s. k, this is visible as an elbow. Here the best k equals to 3.

The idea of the Silhuoette method is: The silhouette value measures how similar a point to its own cluster (cohesion) compared to other clusters. The larger Silhouette value, the better the model. Here the best k = 3, which is the same as the elbow method.









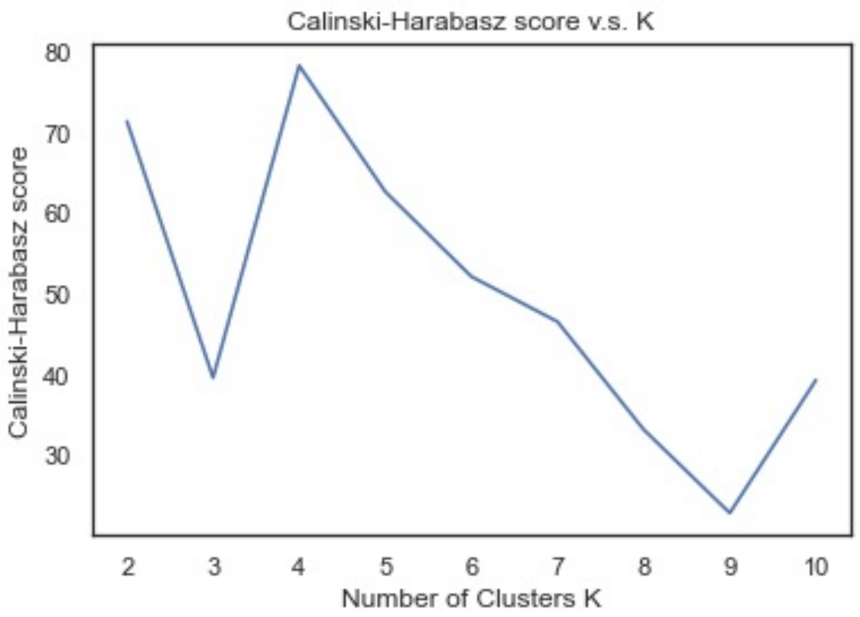
Spectral Clustering – Hyperparameter Tuning and Model Selection

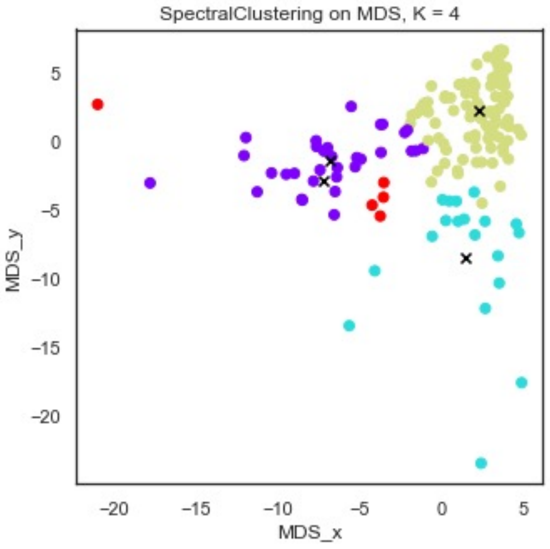
The general way to assign labels in SpectralClustering is 'KMeans', to show a difference with the KMeans methods, we try 'discretize' method here.

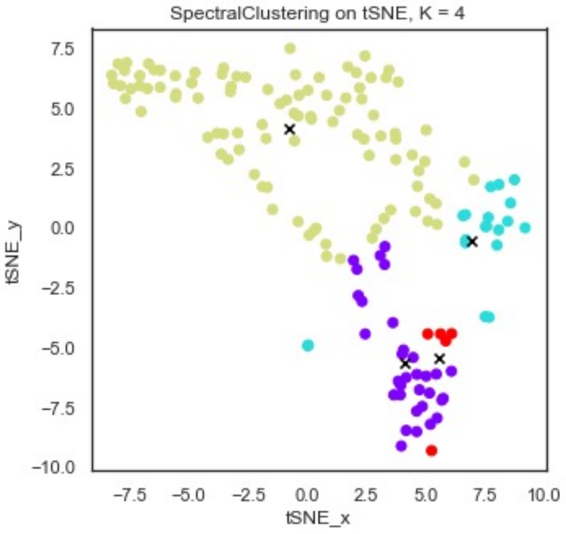
Note that Spectral Clustering performs a low-dimension embedding of the affinity matrix between samples. We usually use Calinski-Harabasz score to represent the Variance Ratio Criterion of the clustering model. The larger Calinski-Harabasz score, the better the clustering model is.

We can conclude the best k for spectral clustering is k equals to 4.

And we can see on the tSNE plot that the center of purple cluster is very closed to the center of red cluster, and there are also some points very far away from their cluster.





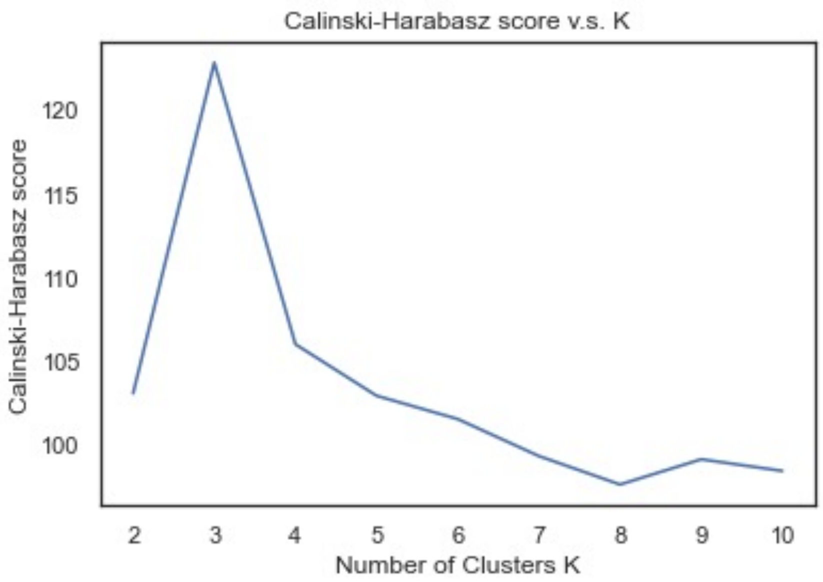


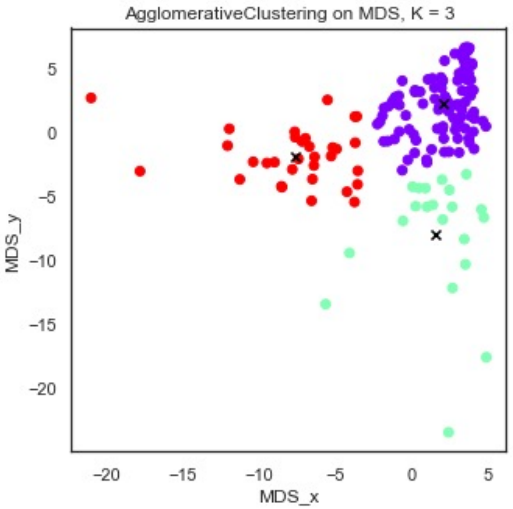
Agglomerative Clustering – Hyperparameter Tuning and Model Selection

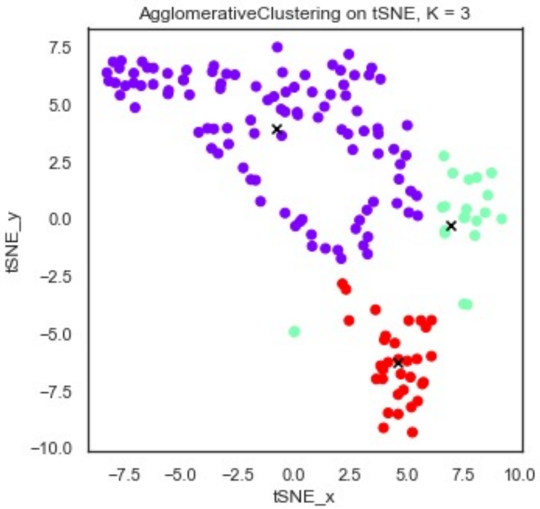
Agglomerative Clustering is a hierarchical clsuter algorithm that build nested clusters by bottom up merging or splitting them successively. The idea is each observation starts in its own cluster, and clusters are successively merged together.

As we all want to minimize the difference within clusters, we will try the 'Ward' linkage method - a variance-minimizing approach minimizes the sum of squared differences within all clusters. We also use Calinski-Harabasz score to represent the Variance Ratio Criterion of the clustering model. The larger Calinski-Harabasz score, the better the clustering model is.

We can conclude the best k for agglomerative clustering is k equals to 3.


Calinski-Harabasz score v.s. K
Calinski-Harabasz score
Number of Clusters K


AgglomerativeClustering on MDS, K = 3
MDS_y
MDS_x


AgglomerativeClustering on tSNE, K = 3
tSNE_y
tSNE_x