Agribusiness

Market Analysis.

1 INITIAL DATA ANALYSIS EXPLORATION.

1. Getting to know and understand the data numerically and visually
2. Statistical test.
3. Find the distribution of the dataset

Getting to know and understand the data numerically and visually.

	data_structure	type
Domain Code	object	categorical
Domain	object	categorical
Area Code (FAO)	int64	num discrete
Area	object	categorical
Element Code	int64	num discrete
...	...	...
Item	object	categorical
Year Code	int64	num discrete
Year	object	categorical
Unit	object	categorical
Value	object	categorical

# Column	# Dtype
0 domain_code_sicurity	object
1 domain_sicurity	object
2 country_code	int64
3 country	object
4 element_code_sicurity	int64
5 element_sicurity	object
6 item_code_sicurity	int64
7 item_sicurity	object
8 year_code_sicurity	int64
9 year_sicurity	object
10 unit_sicurity	object
11 value_sicurity	float64
12 domain_code_balance	object
13 domain_balance	object
14 element_code_balance	float64
15 element_balance	object
16 item_code_balance	float64
17 item_balance	object
18 year_code_balance	float64
19 year_balance	float64
20 unit_balance	object
21 value_balance	float64

Observing the datasets
The data was collected using the Fao data web app.

- Food Security, Food balance, Producer prices and Macro indicator.
- The columns and their observations.
- The Domain is what the dataset values describe, with the countries as Areas.
- The Element relates to the actual observation column called “value”.
- Item is the type of observation but also contains unit or aggregation information. Example value of food imports in total merchandise exports (percent) (3-year average) and Minimum dietary energy requirement (kcal/cap/day) one is an aggregation average by an amount of year, and the other is a minimum of a unit.
- A three-year column aggregated from 2017 to 2019 and a not aggregate from 2018.
- Every value is based on his unit, as it is visible in the Unit column, seven different ones.
- The flag columns and notes are metadata that may not be very important for our Analysis.

Data sample Macro Indicator.

Domain Code	Domain	Area Code (FAO)	Area	Element Code	Element	Item Code	Item	Year Code	Year	Unit	Value	Flag	Flag Description	Note
MK	Macro Indicators	2	Afghanistan	6109	Value Local Currency	22008	Gross Domestic Product	2018	2018	millions	1.296559e+06	X	International reliable sources	NaN
MK	Macro Indicators	2	Afghanistan	6110	Value US$	22008	Gross Domestic Product	2018	2018	millions	1.798697e+04	X	International reliable sources	NaN
MK	Macro Indicators	2	Afghanistan	6119	Value US$ per capita	22008	Gross Domestic Product	2018	2018	US$	4.838859e+02	X	International reliable sources	NaN
MK	Macro Indicators	2	Afghanistan	6183	Value Local Currency, 2015 prices	22008	Gross Domestic Product	2018	2018	millions	1.212817e+06	X	International reliable sources	NaN
MK	Macro Indicators	2	Afghanistan	6185	Value US$ per capita, 2015 prices	22008	Gross Domestic Product	2018	2018	US$	5.336178e+02	X	International reliable sources	NaN

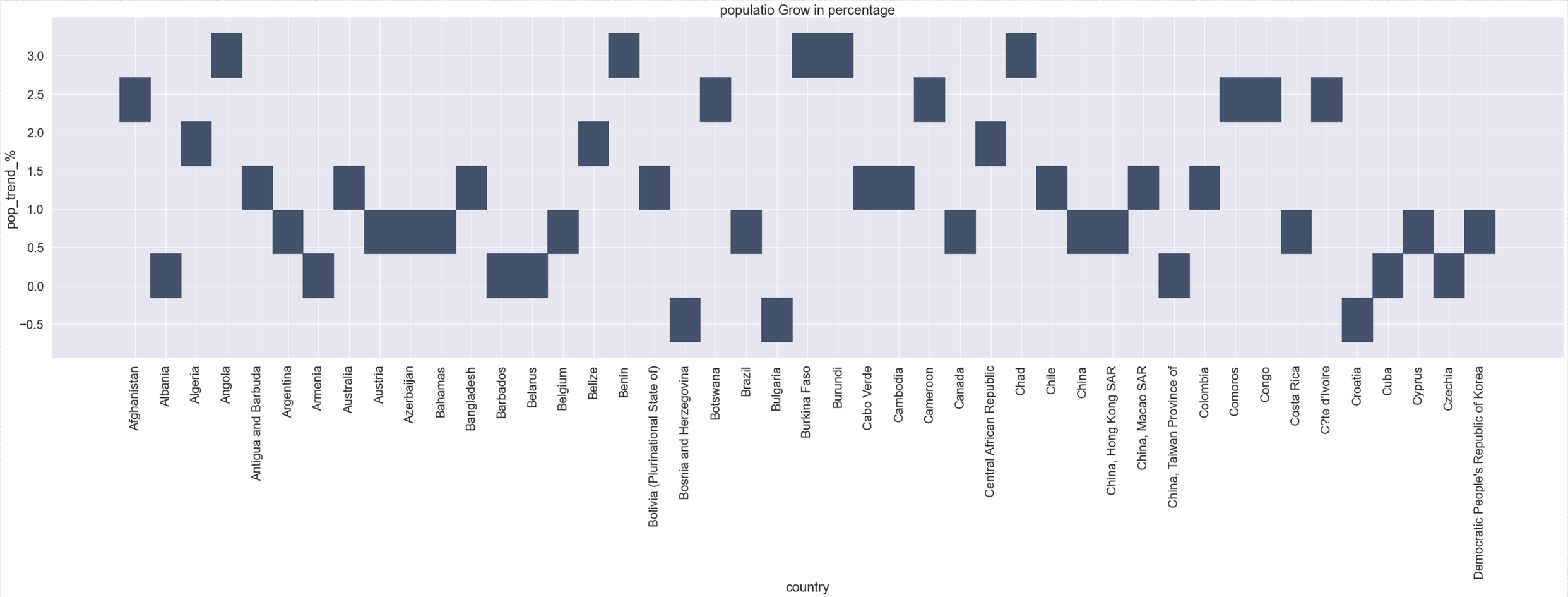
Statistical test.
Checking normality.

	variable	statistic	pvalue	hypothesis
2	value_import_quantity	0.5667678117752075	2.0747135565102326e-08	H1: value_import_quantity is not a Normal distribution!
3	value_domestic_supply_quantity	0.5613753795623779	1.7941514940389425e-08	H1: value_domestic_supply_quantity is not a Normal distribution!
4	value_protein_supply_quantity_(g/capita/day)	0.9496732950210571	0.15290087461471558	H0: value_protein_supply_quantity_(g/capita/day) is Normal distribution!
10	value_sanitation	0.7936556935310364	3.947396908188239e-05	H1: value_sanitation is not a Normal distribution!
11	value_automative_railway	0.43675392866134644	8.293304376572053e-10	H1: value_automative_railway is not a Normal distribution!
12	value_manufacture	0.2932596206665039	4.060358646529316e-11	H1: value_manufacture is not a Normal distribution!
13	value_legal_rank	0.9713712930679321	0.5575469136238098	H0: value_legal_rank is Normal distribution!
14	value_environmental	0.18972289562225342	5.942372895834103e-12	H1: value_environmental is not a Normal distribution!
15	value_eggs_hen_in_shell	0.18972289562225342	5.942372895834103e-12	H1: value_eggs_hen_in_shell is not a Normal distribution!

2 COUNTRIES ANALYSIS.

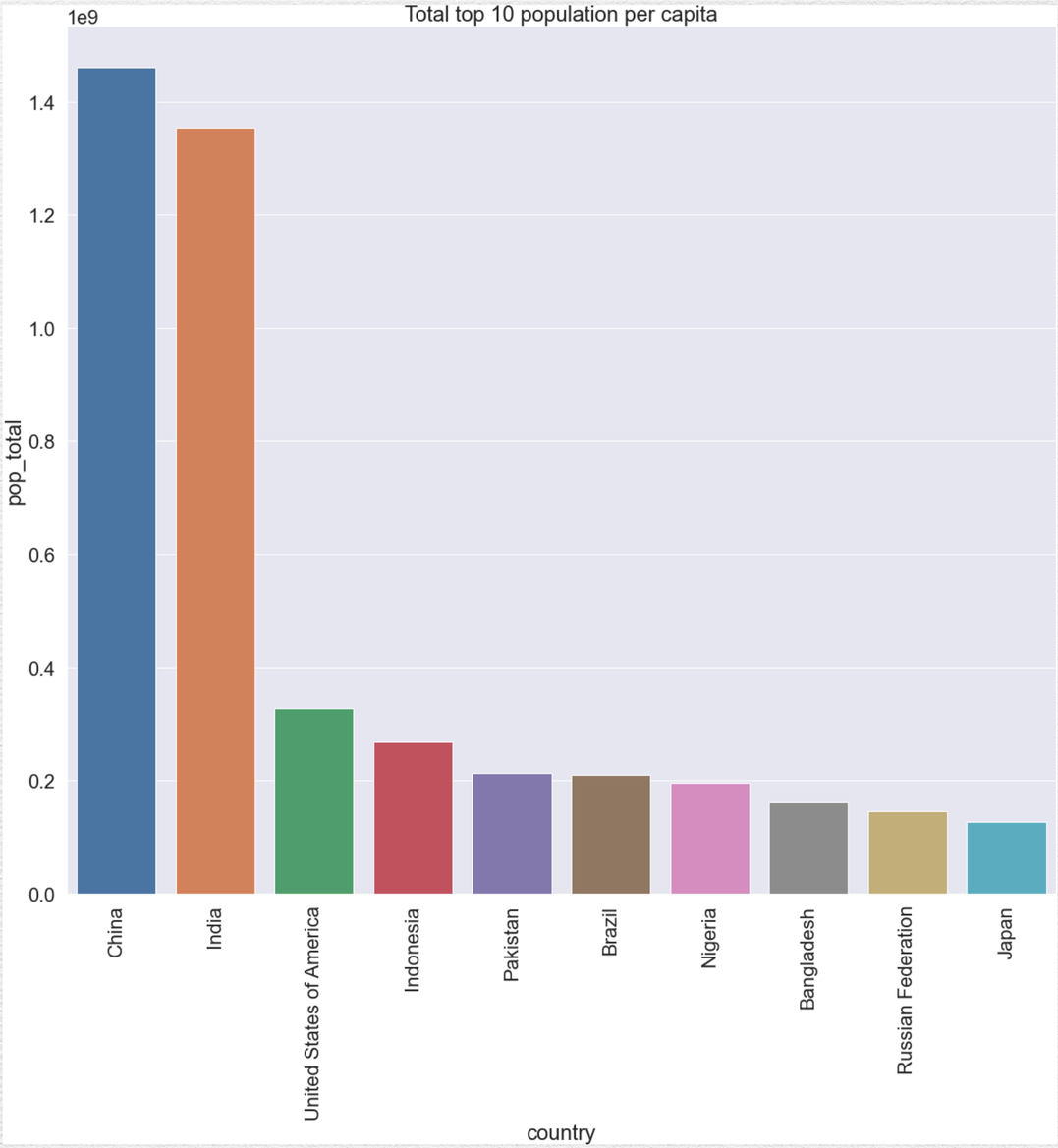
1. The difference in population between the previous year (elective) and the current year is expressed in percentage.
2. The proportion of animal proteins compared to the total amount of proteins available in the country.
3. Food availability in protein per capita.
4. Food availability in calories per capita.

The difference in population between the previous year (elective) and the current year is expressed as a percentage.



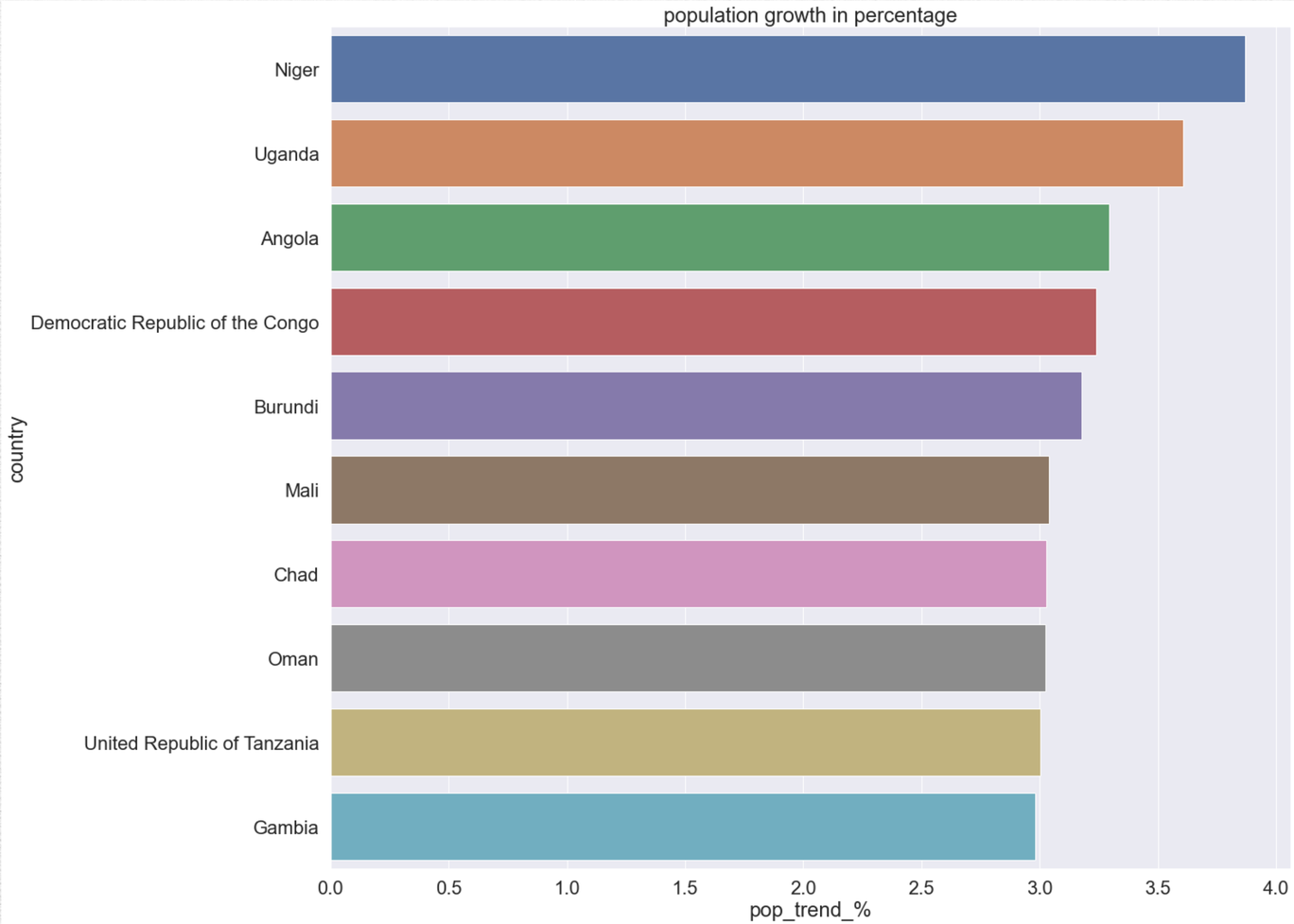
Top Populated Countries.

China has the highest population of all the countries, with a difference of more than 205 thousand per 1000 citizens compared to India in second place in this top 10; the USA here is in Third place here, and Brazil is part of the top 5.



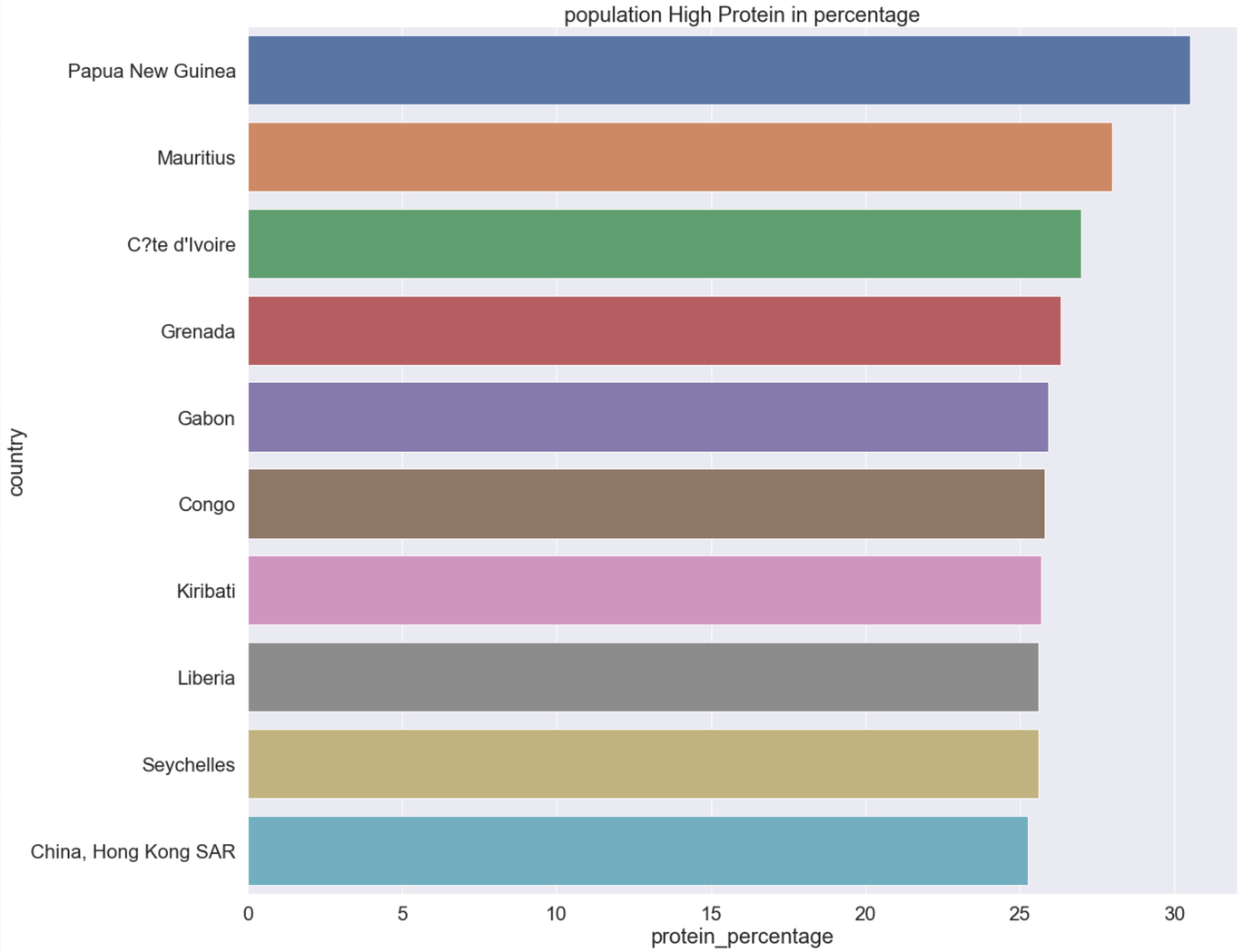
Top Populated Countries.

Now that we have our 10 top countries by population let us see how this population is growing! To our surprise, none of the top 10 countries per population make it to these new criteria pop_trend_% population growth in percentage.



Proportion of animal proteins compared to the total amount of proteins available in the country.

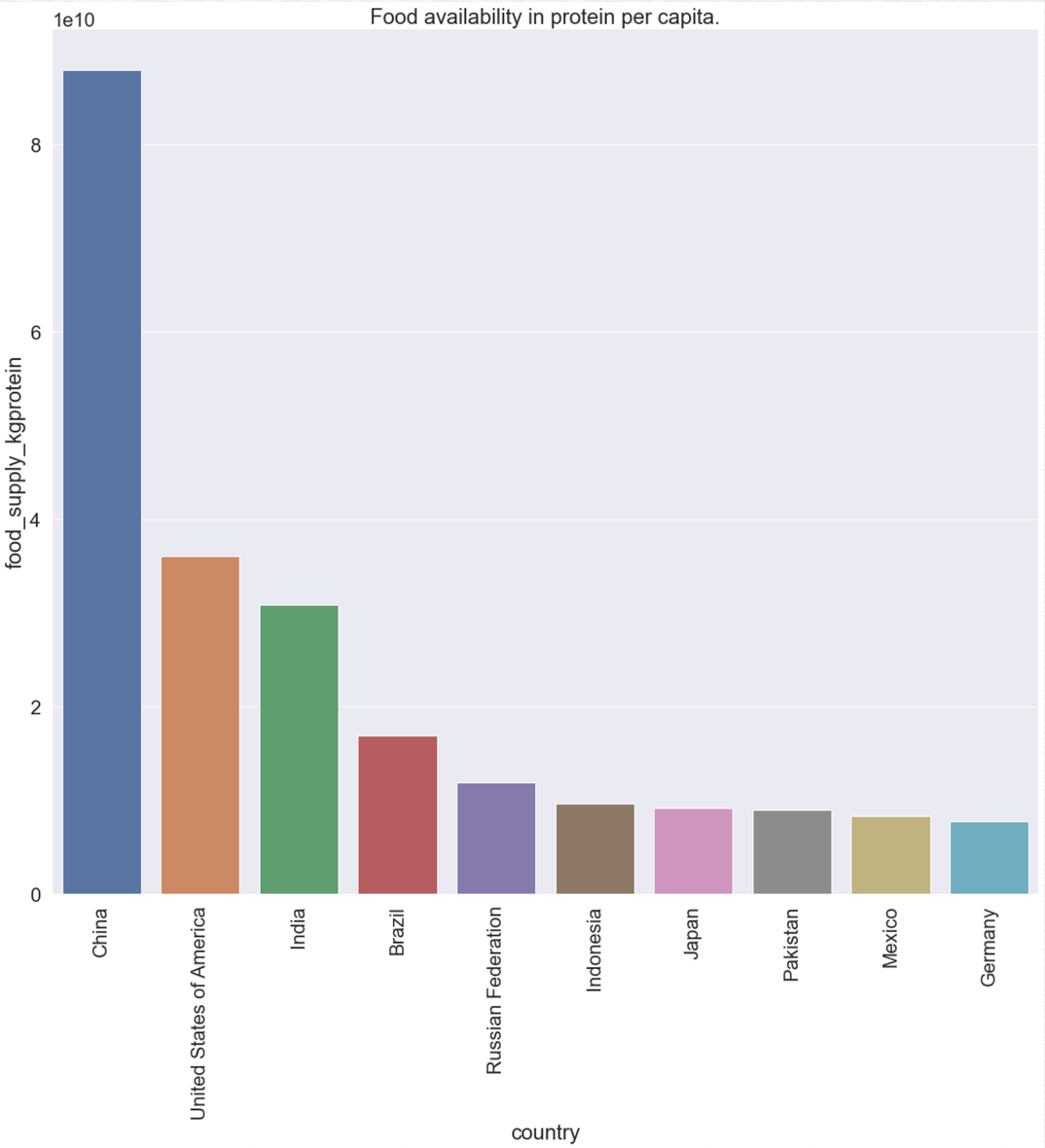
By Considering the protein consumption per country, we get a completely different figure here!



Food availability in protein per capita.

Here we get the expected figure, and these countries are highly or entirely industrialised, so as expected, they have a lot of protein supply.

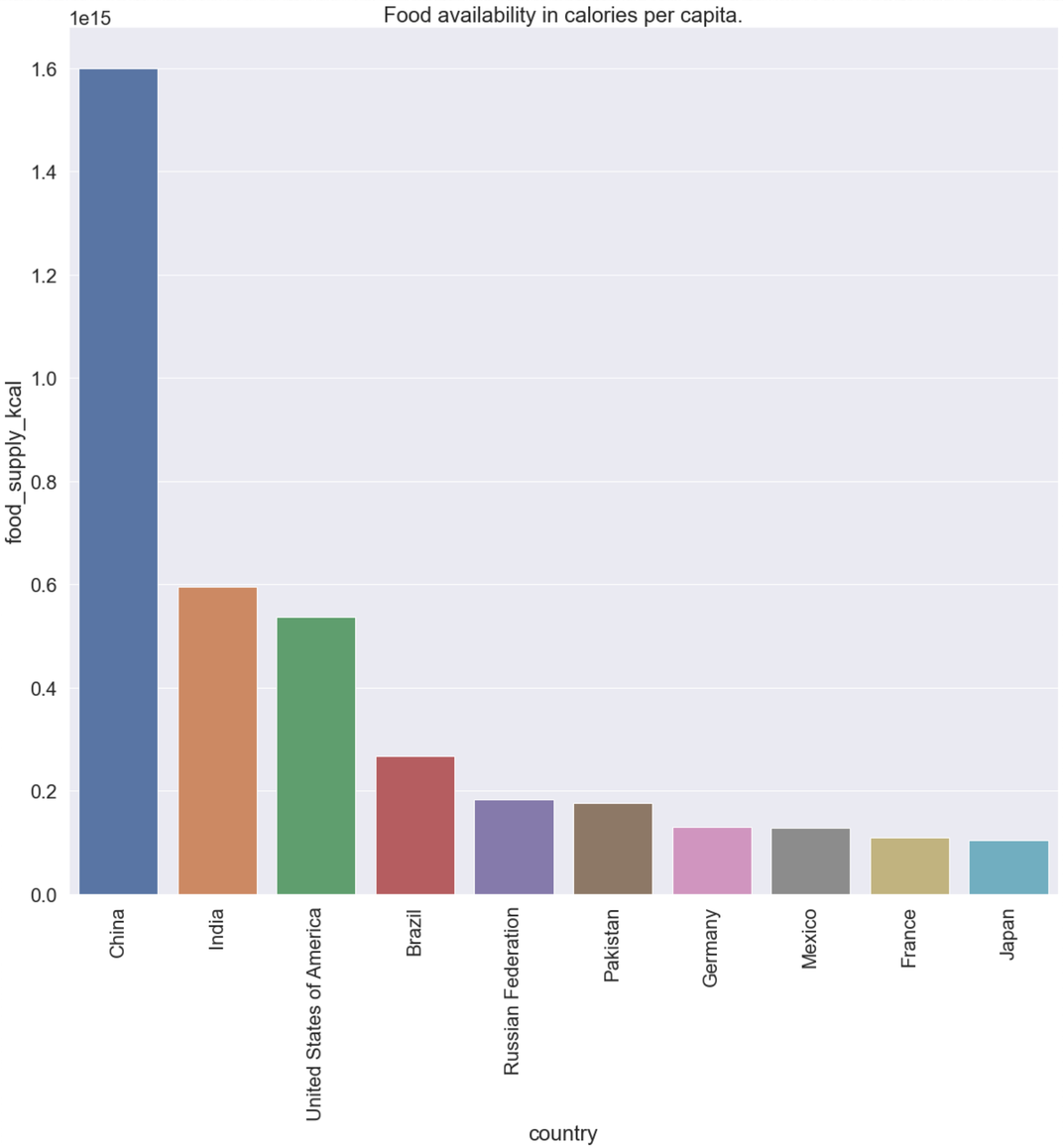
Because their population also influences this top 10 selection, we have China leading then the USA, India, Brazil, and Russian Federation.



Food availability in calories per capita.

This figure confirms the relation between population and food supply.

We have China as first in this selection, but India is before the USA, fallow by Brazil and Russian Federation.



3

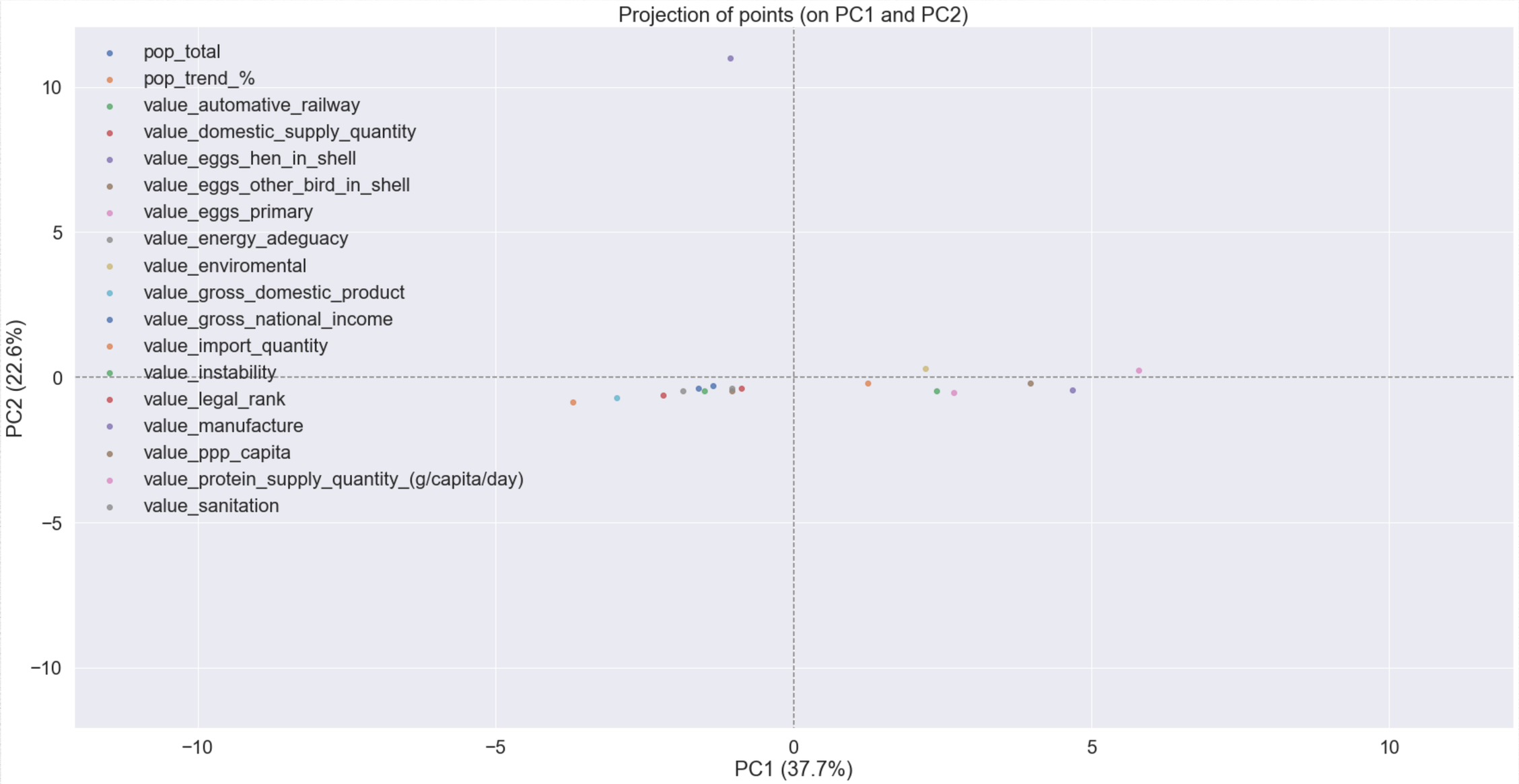
Variables and countries Conclusion.

- China is the country that most appears in these figures, having a considerable population that requires and consumes a lot of food and protein to maintain its workforce and continue to grow economically.
- Other persistent country here is USA and India, Brazil, and Russian Federation, perhaps because of their population.
- Brazil, India, and Russia are part of the so-called BRIC, a group of emergent strong economic countries that require a lot of food to keep growing.
- Germany and Japan are also present a few times more than other countries; they are economic powerhouses and have a significant amount of population and food, particularly protein.
- Worth to mention also is France, Mexico, Pakistan, and Indonesia are less present but are in some selections.
- From all the questions, we obtained ten countries with a total of 31 countries; they will be used in the next session of this Analysis, where we will apply more advanced algorithms that will reveal how these countries are related.

4 A dendrogram containing all the studied countries.

1. Performing PERSTL Analysis to identify critical countries in our data set.
2. Is egg's data values influential on the countries' Analysis?
3. PCA analysis, it is possible to reduce the number of dimensions to how many with?
4. Kmeas means Cluster analysis.
5. Hierarchical Clustering Dendrogram.

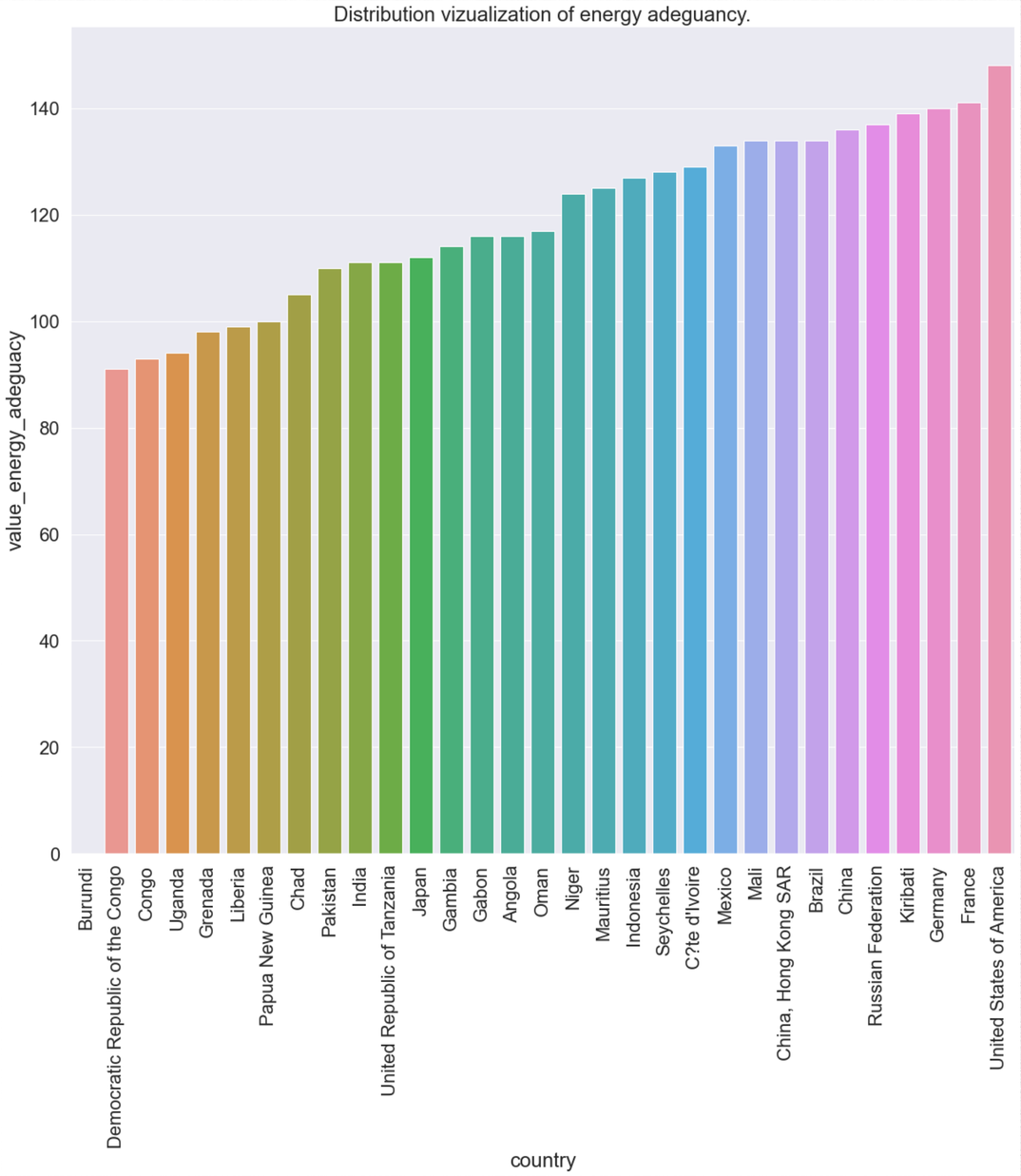
PERSTL Analisys and PCA.



Political.

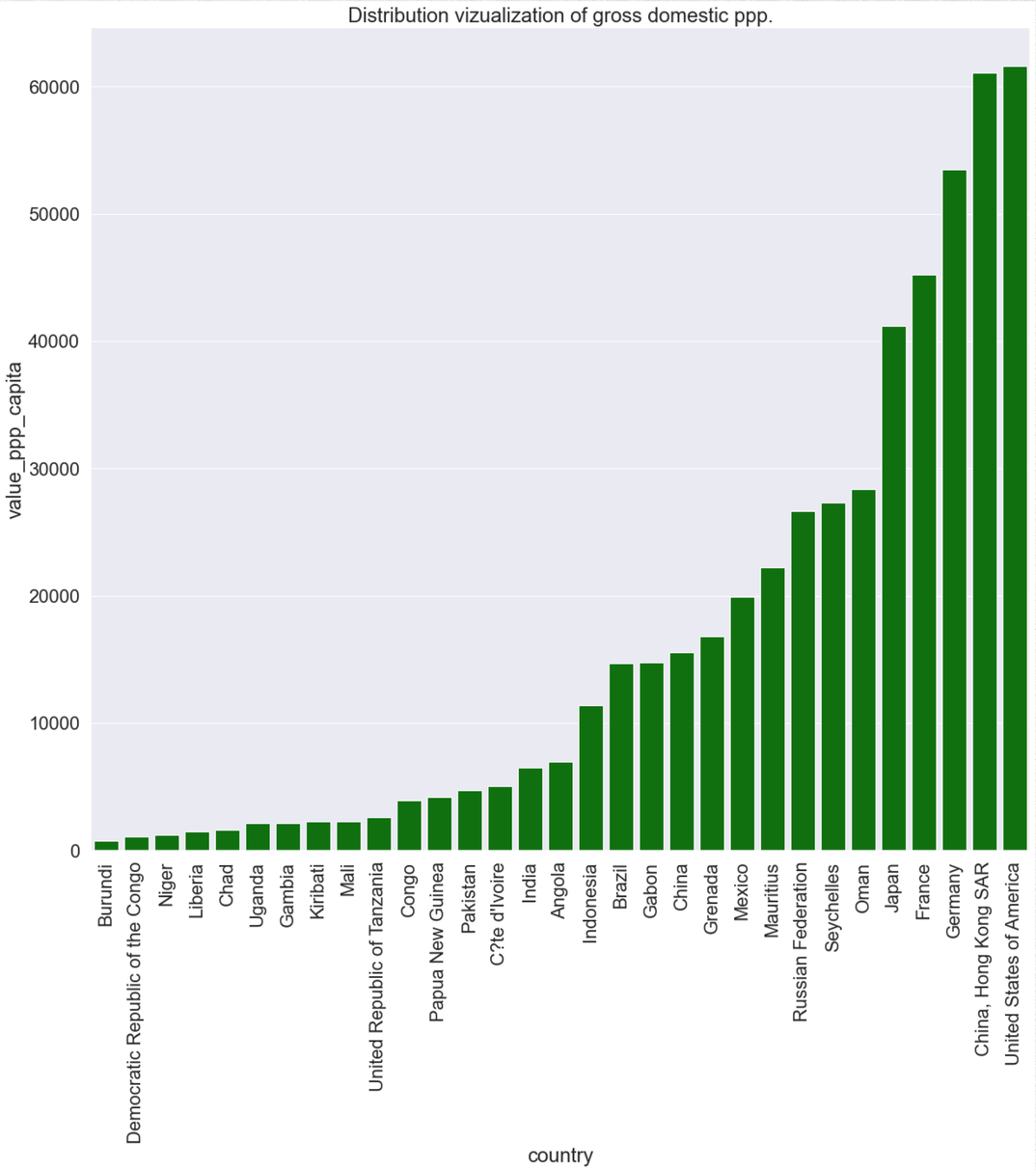
On Energy adequacy, we have values for all the countries apart from Burundi.

The most legal, stable countries have the highest values here!



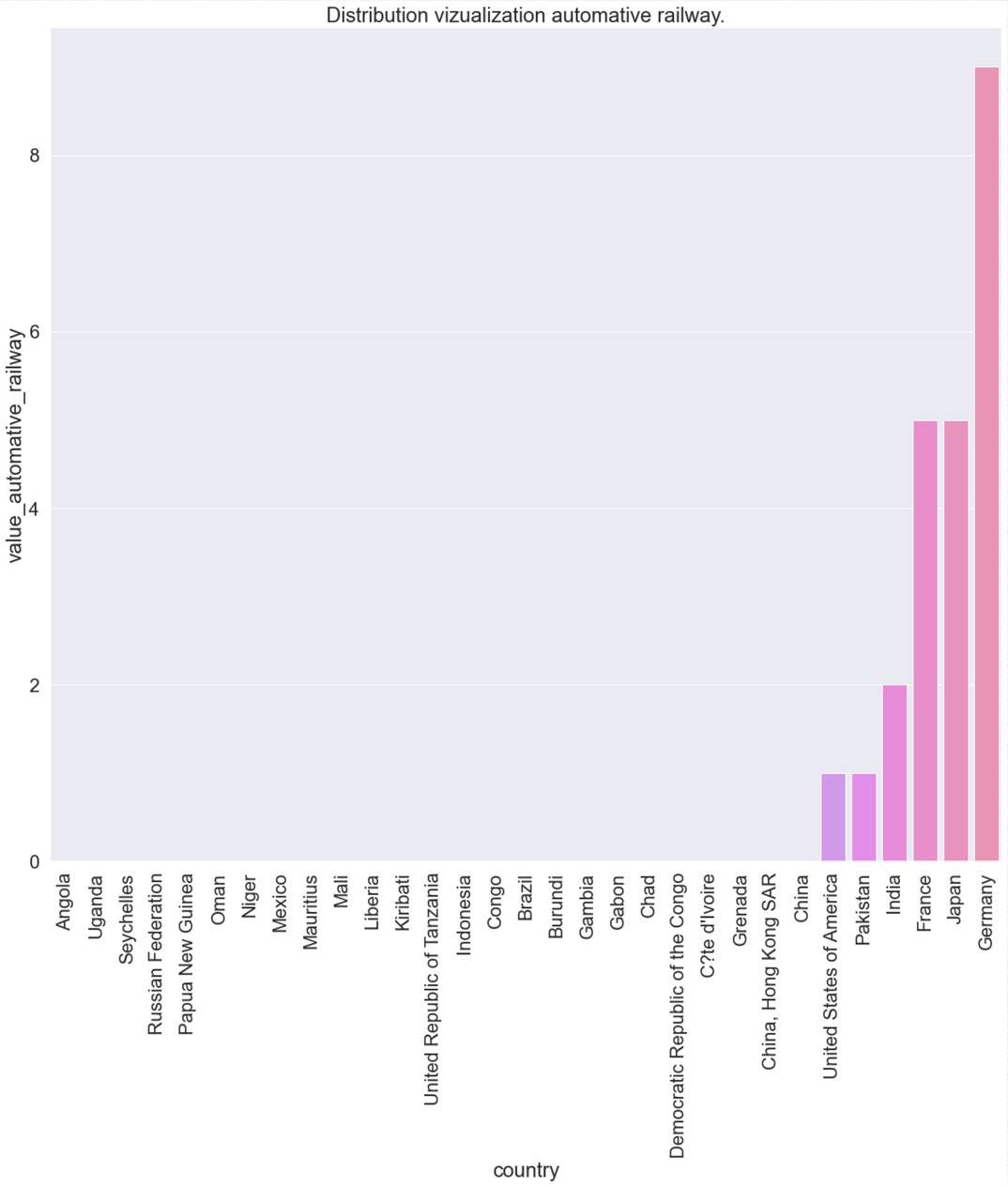
Economical.

At value_ppp_capita, as expected, small countries like Burundi have lower capital, and then we see the usual suspects, the USA, China, Germany, France, and Japan.



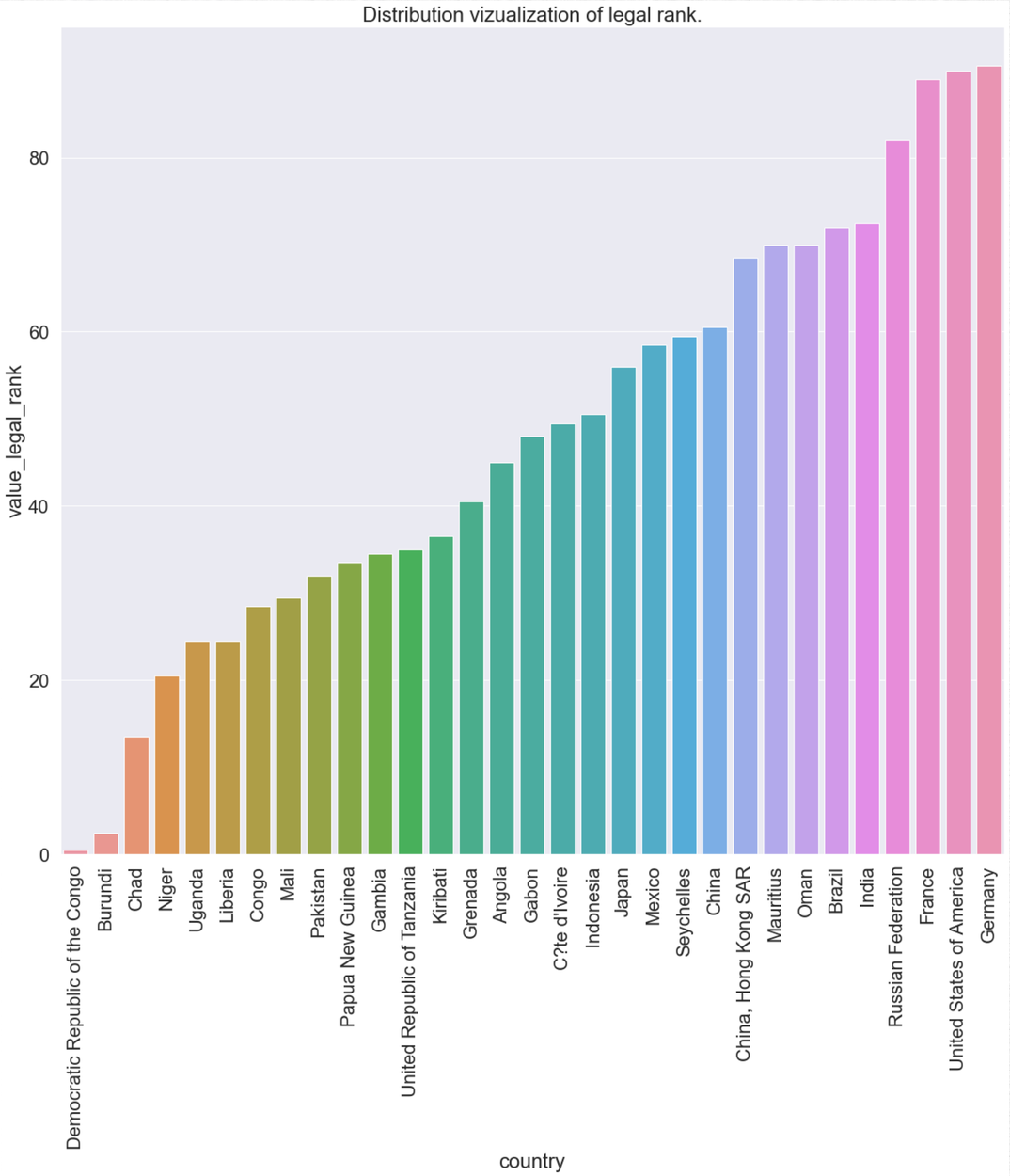
Technological.

value_automative_railway shows only value for highly populated countries: the USA, Pakistan, India, France, Japan, and Germany.



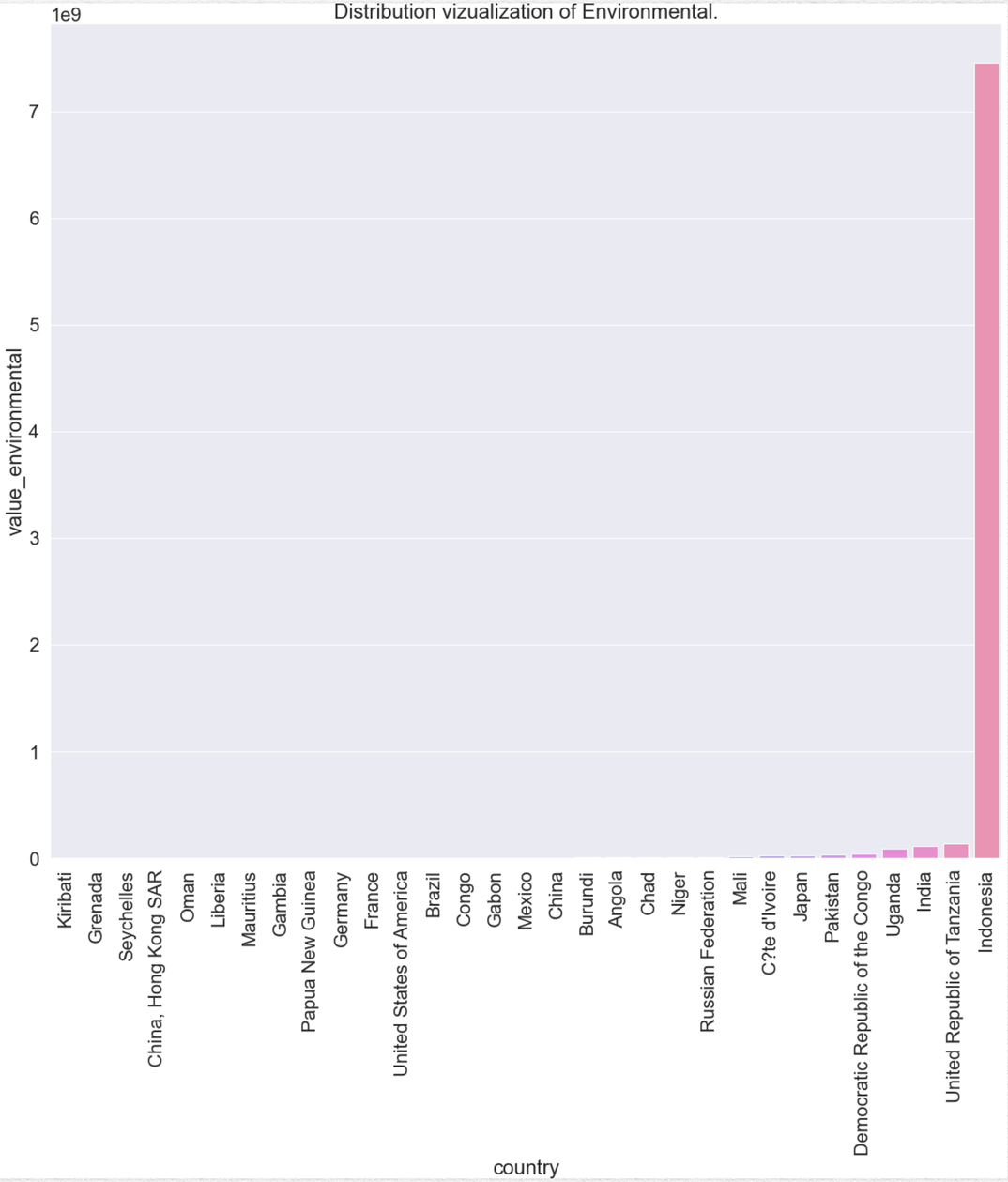
Legal.

value_legal_rank is calculated by the sum and difference related to a legal country state. Here we can see very low values countries', like DR Congo to the highest which is Germany.

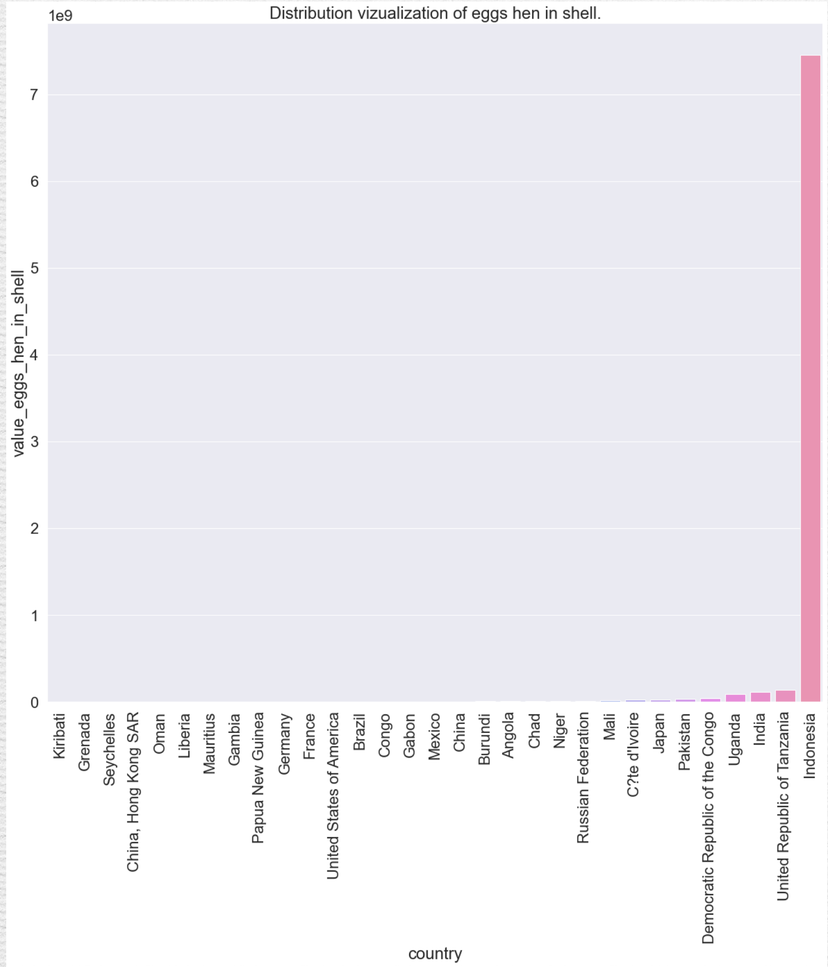


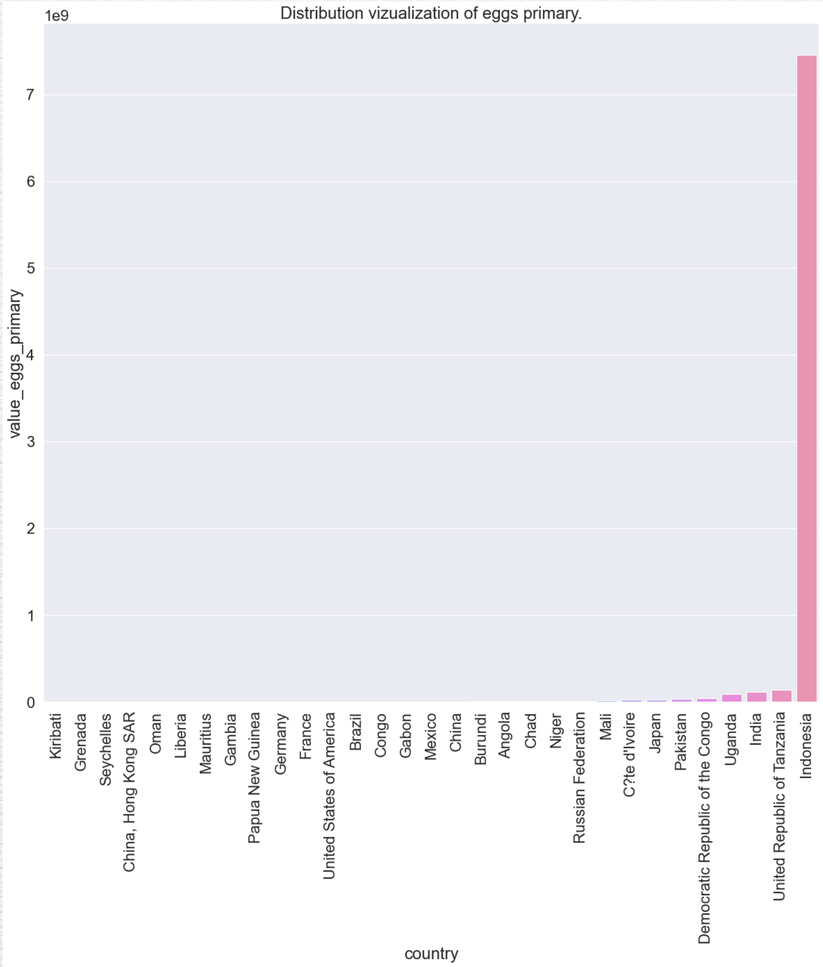
Environmental.

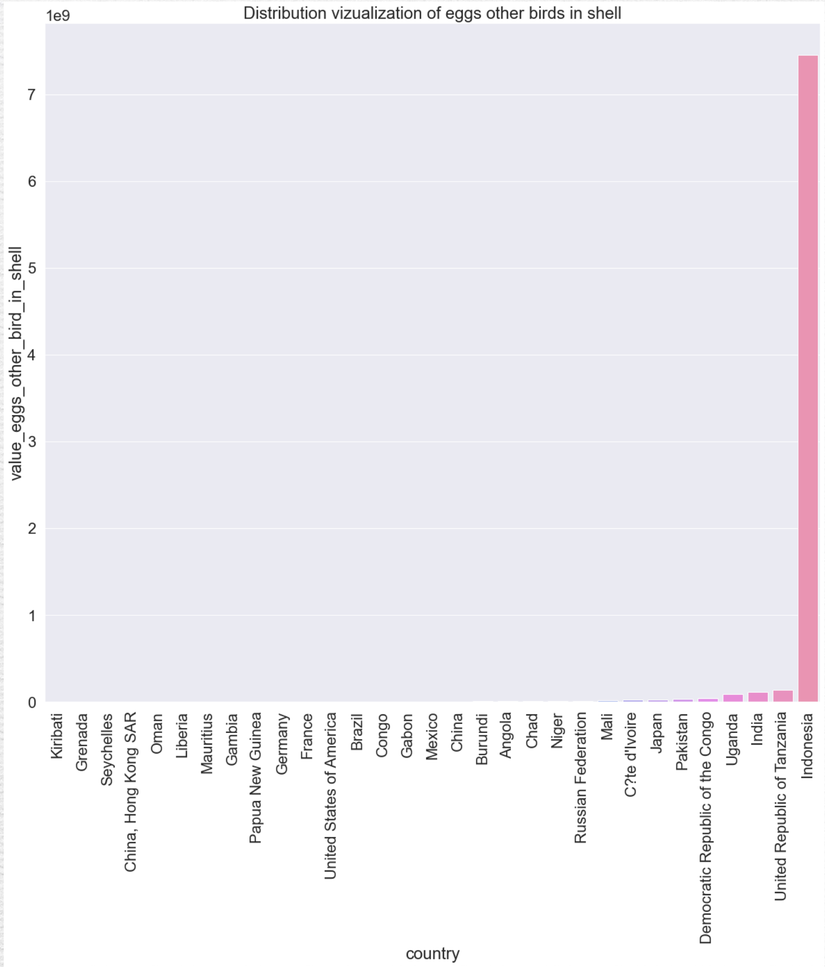
It was not possible to deduce more values from different data sets. There is a significant amount of countries that need more value here! It is still influential, as we will see.



Extra eggs variables and its distributions.





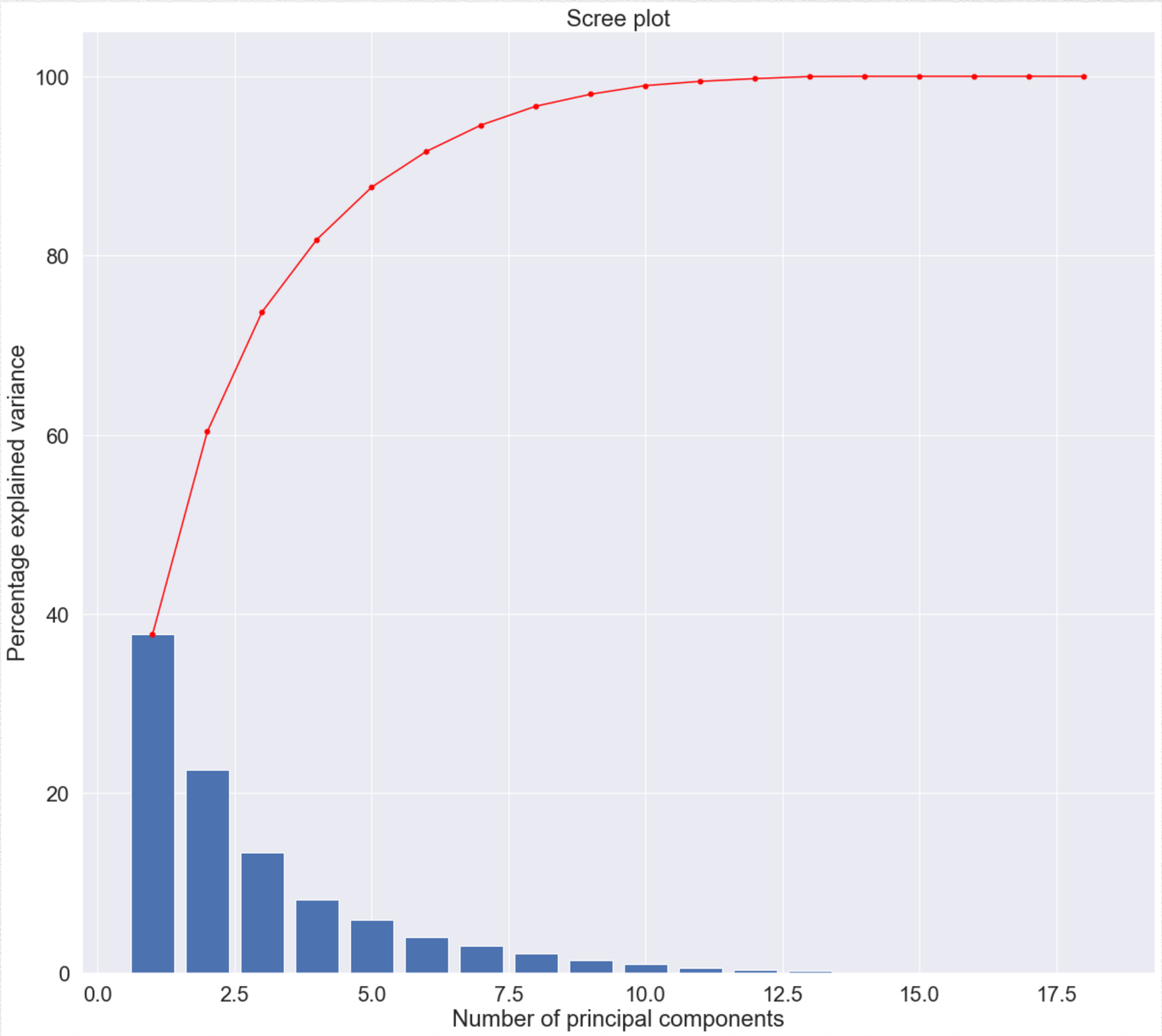


Here I add more relevant variables to the study, a data frame of egg values with their countries.

PCA scree plot.

What is the minimum amount of components that can explain most of our data set?

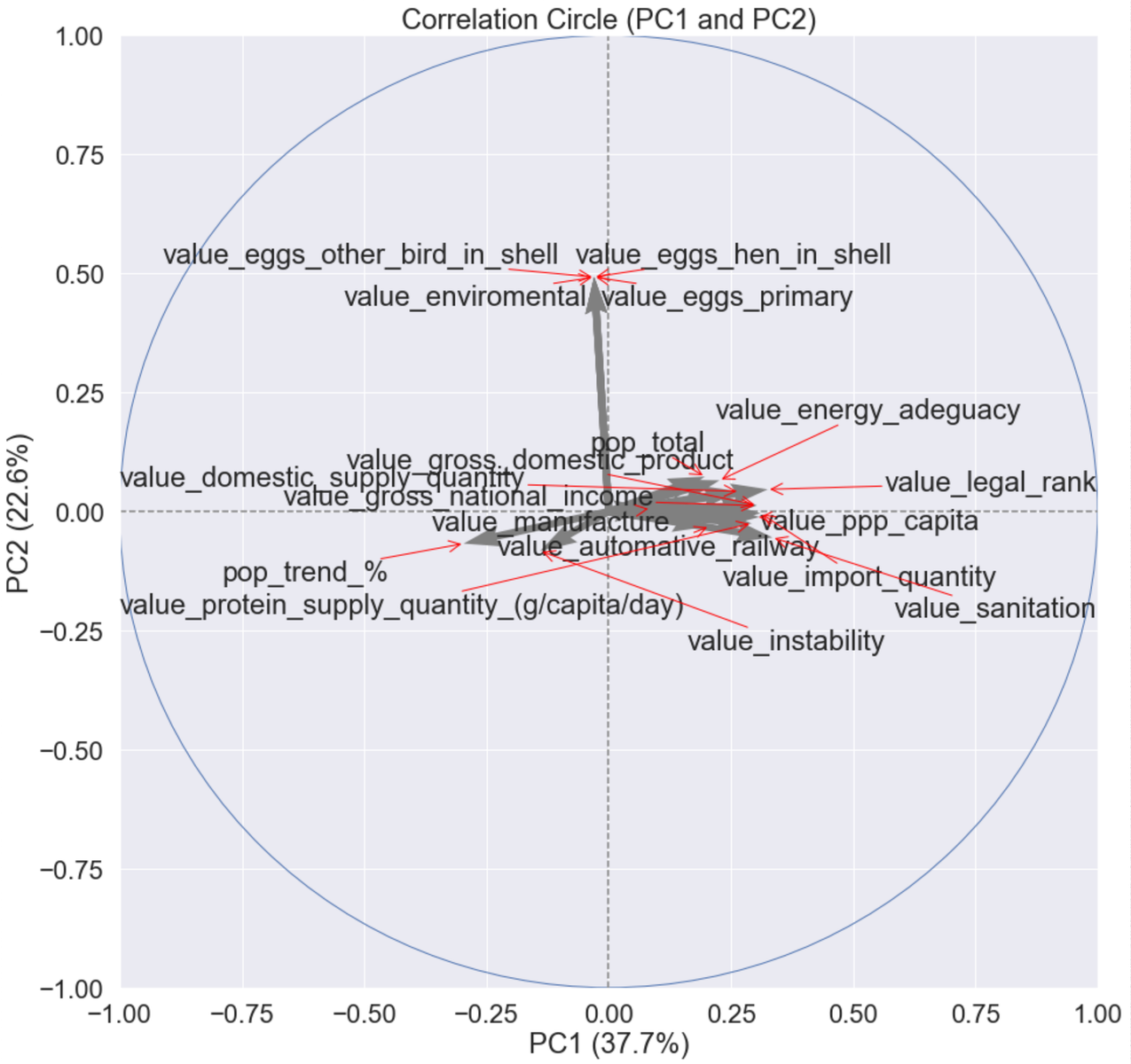
This question is a clear answer here in this scree plot. As you can see, with only 1 and 2 components, we can explain more than 50% of our data set.



PCA Correlation circle plot.

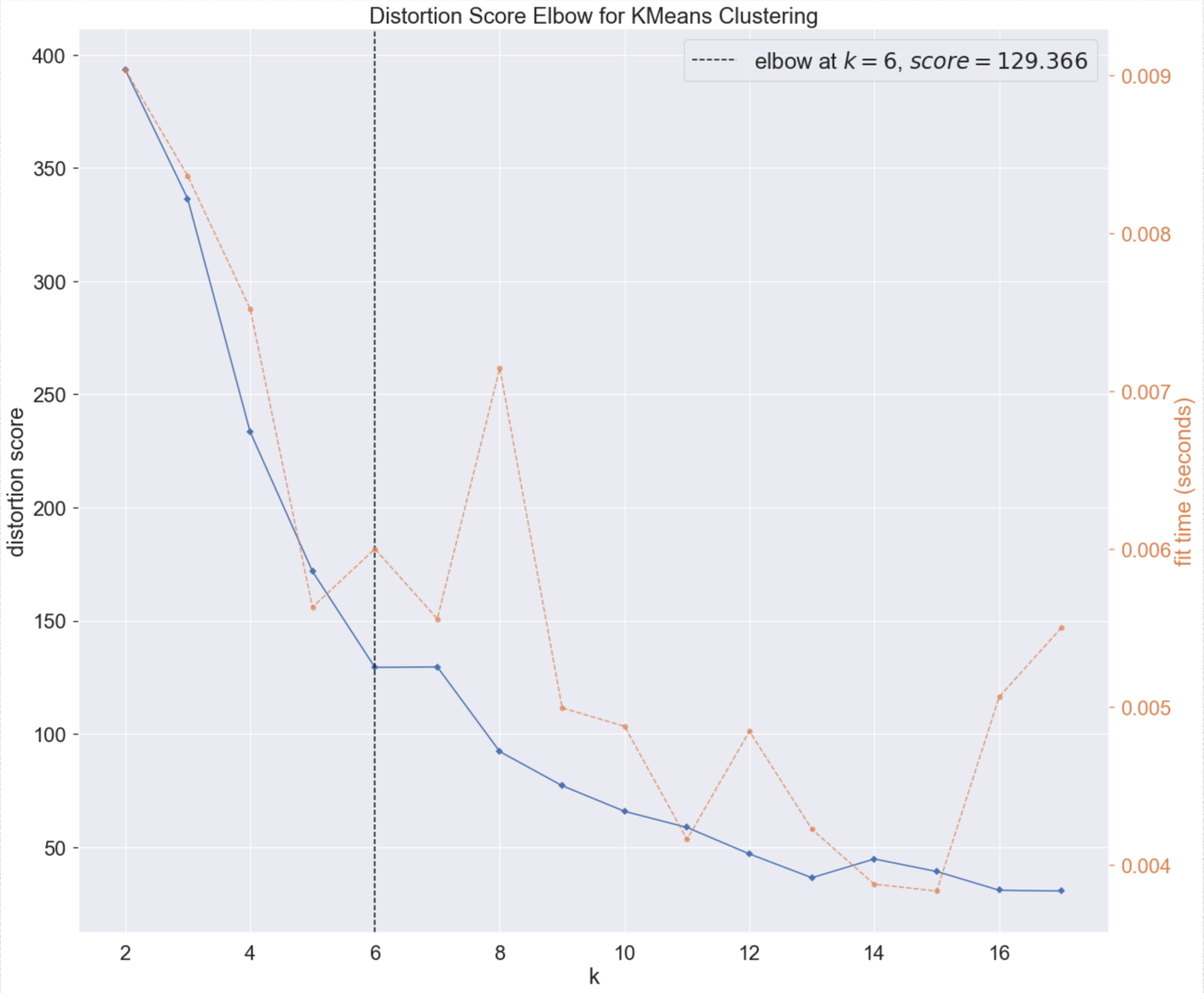
On this Circular factorial Map, we can visualize the influence of 2 components 1-2, and these components can explain 60,3 percent of the data set in consideration;

This is the highest value that can be plotted here according to the preview screed plot.



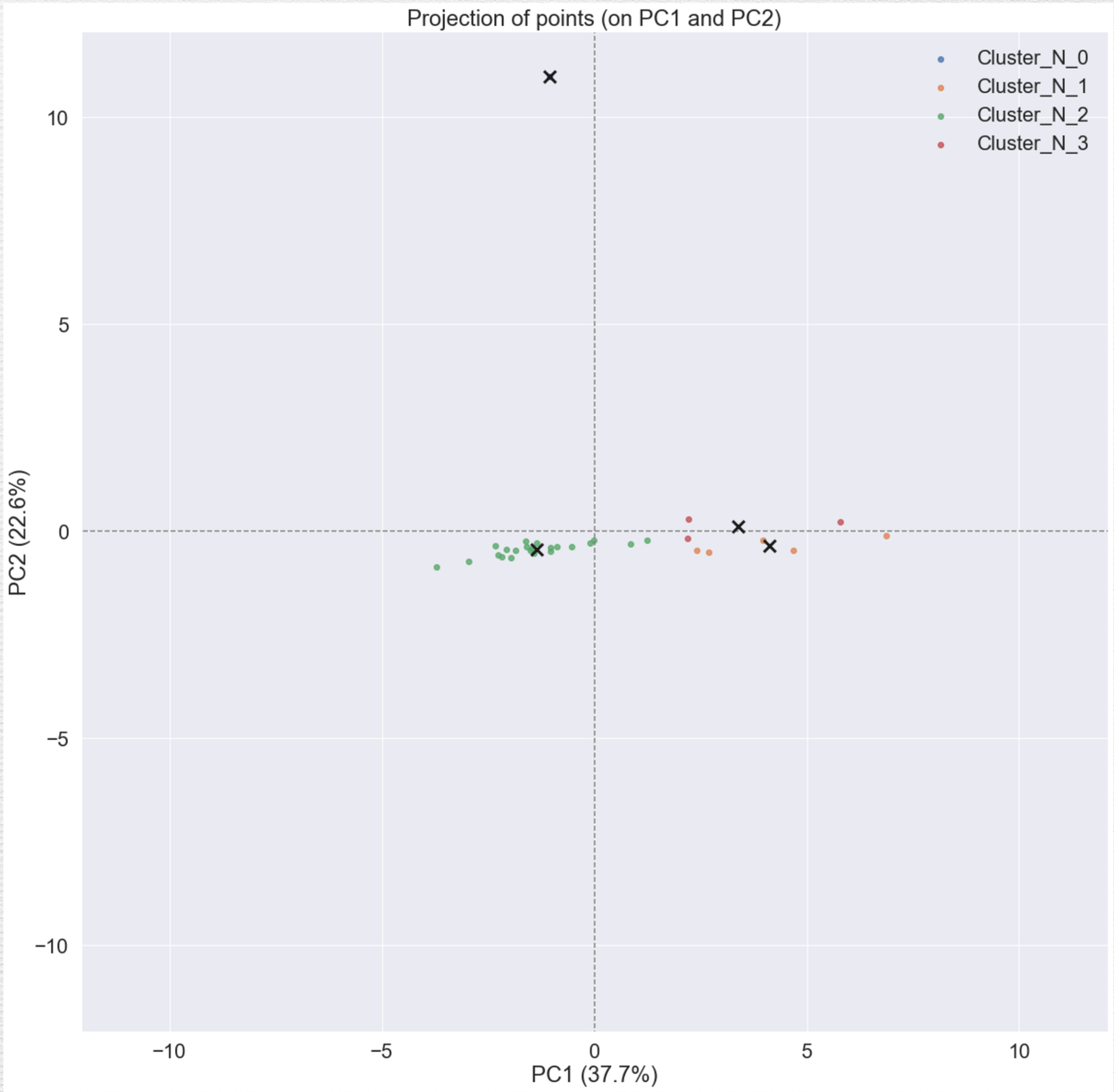
PCA Elbow plot.

I apply many different metrics in order to correctly validate the optimal amount of clusters.



PCA Kmeans the optimal number of cluster projection.

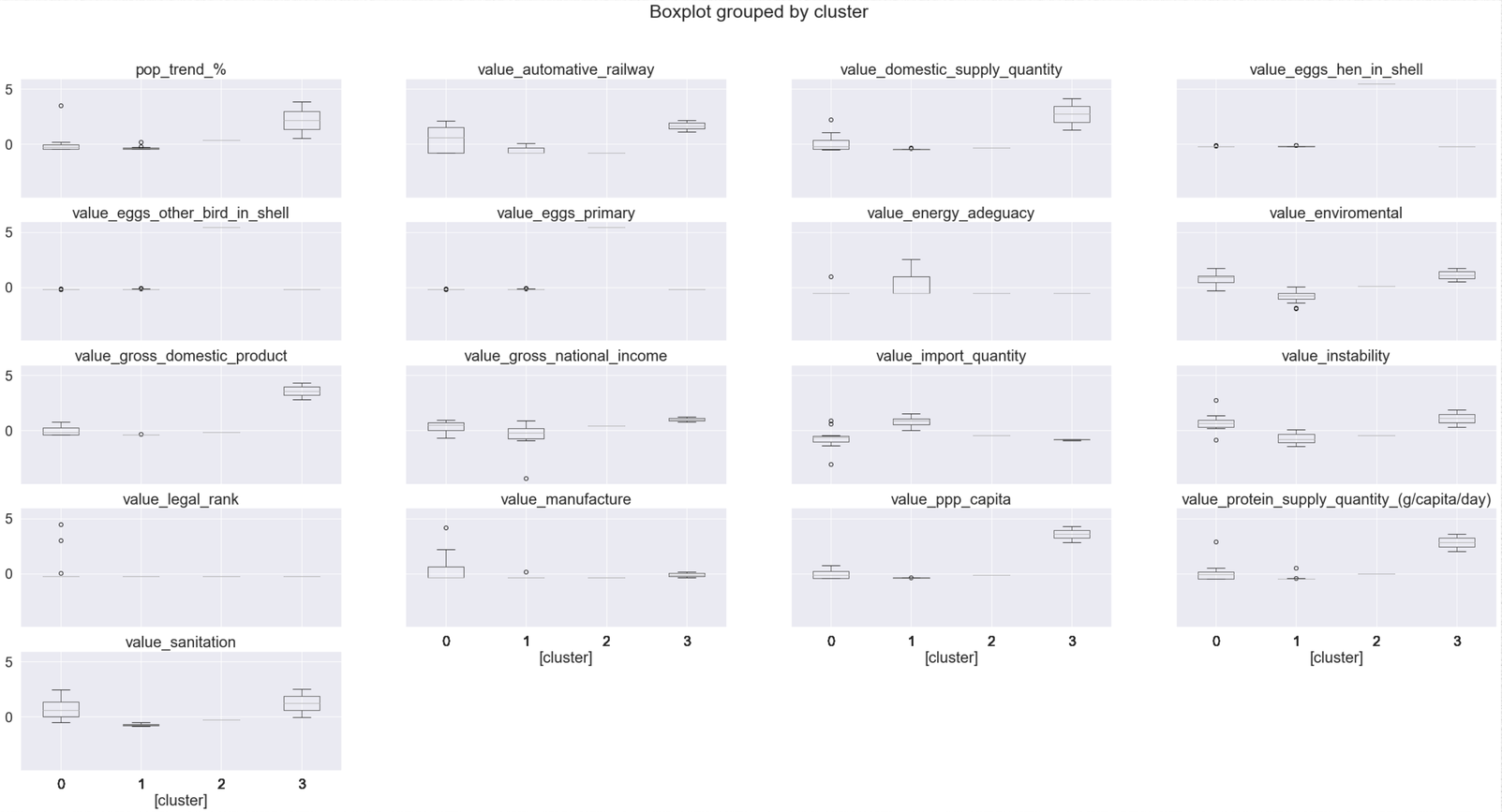
When applying the Kmeans algorithm to our PCA, four groups emerge from our optimal number of clusters. These groups will be visible on the dendrogram visualization revealing where each country belongs.



Cluster box plot groups.

Cluster 0, 1, and 3 variables are higher:

- Pop trend
- Value domestic supply
- Gross domestic product
- Instability
- PPP per capita
- Protein supply quantity
- Value sanitation

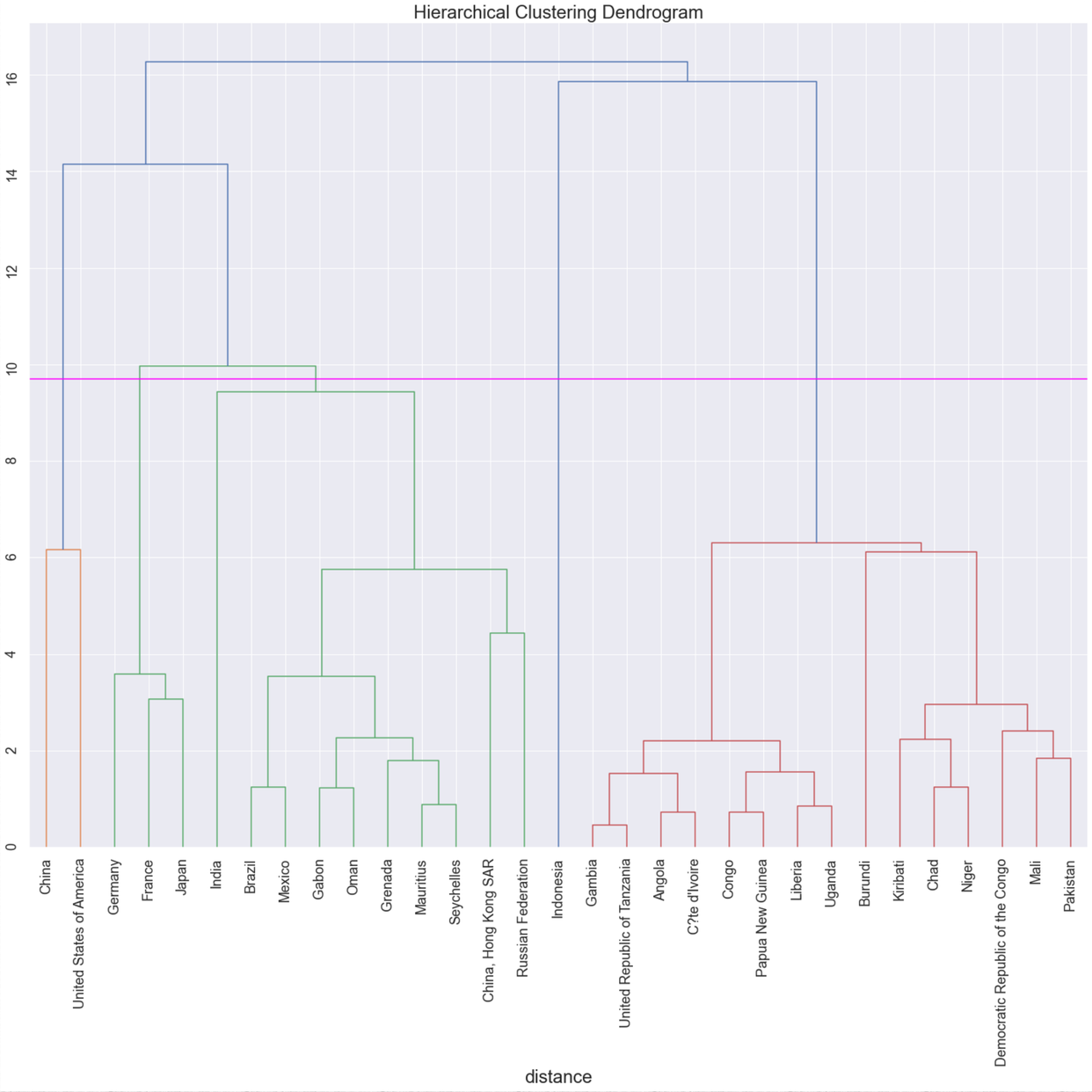


Dendrogram country hierarchical

How much does the dendrogram Explain the K-means clustering at this point of the Analysis?

We can see that almost all countries clustered are in the same branches, and it is more apparent now what variables have more influence on them from the Box plot, viz.

The dendrogram reflects the influence of the box plot cluster visualization.



5 Unsupervised ML analysis conclusion.

We came a long way here, accurately identifying the valuable data sets and ensuring this data is clean and consistent so that we can use it to make the right decision and that the company can do further Analysis!

We investigated which countries could become potential candidates for our recommendation by analyzing their distribution with existent variables.

After that, we calculated new variables, applied a PESTLE study analysis to understand countries and risk better, and used this insight on our ML algorithms; then, we gave a clear group of countries recommendations in a single cluster!

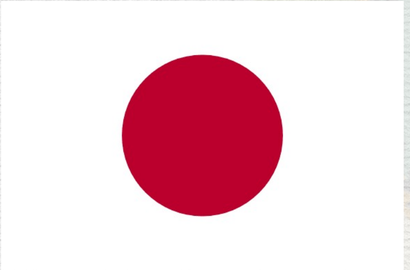
Next, I will use this cluster to present these selected countries.

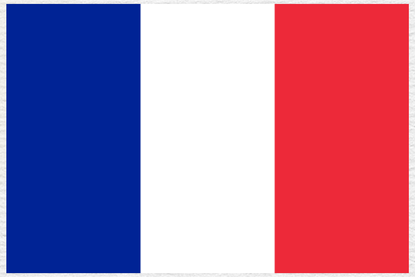
6 A short list (5-10) of potential countries.

China, France, Germany, India, Japan, United States of America

country	pop_total	pop_trend_%	value_import_quantity	value_domestic_supply_quantity	value_protein_supply_quantity_(g/capita/day)	value_instability	value_energy_adeguacy	value_gross_national_income	value_gross_domestic_product	value_ppp_capita	value_sanitation	value_automative_railway	value_manufacture	value_legal_rank	value_enviromental	value_eggs_hen_in_shell	value_eggs_primary	value_eggs_other_bird_in_shell	cluster
China	1460843634.0	0.42867577831103393	49754.169999999984	518735.94000000006	164.93999999999994	0.0	136.0	28992135.469826	29074191.961177	15525.0	64.0	0.0	0.0	60.5	4272693.62802	4272693.62802	4272693.62802	4272693.62802	3
France	66996130.0	-2.6908308556573184	14975.04	66415.44000000003	257.29	0.0	141.0	5621343.333865	5503382.233112	45251.0	78.0	5.0	0.0	89.0	313885.09373799997	313885.09373799997	313885.09373799997	313885.09373799997	3
Germany	83207517.0	0.47278764255810596	28773.419999999995	82743.0	254.14000000000004	0.0	140.0	8037330.011473	7824891.08429	53463.0	97.0	9.0	0.0	90.5	173732.81321300002	173732.81321300002	173732.81321300002	173732.81321300002	3
Japan	127328860.0	-0.2688636971116806	16828.460000000003	51205.40000000001	195.69	1.0	112.0	10401789.807165	10037272.430906	41182.0	80.0	5.0	0.0	56.0	24555730.580203008	24555730.580203008	24555730.580203008	24555730.580203008	3
United States of America	327425065.0	0.6019639494215766	19251.560000000005	318760.29999999993	301.29999999999995	0.0	148.0	42553872.557748005	42045087.0	61585.0	98.0	1.0	0.0	90.0	1489529.436458	1489529.436458	1489529.436458	1489529.436458	3

7 Conclusion and recommendation!





As a result of this agribusiness Market analysis, we have our five countries. China, account for a vast and unexplored Market. The Economically sthbilized countries, such As the United States of America, Japan, Germany, and France, Are a must-investment opportunity.