PREDICTING CHRONIC HUNGER

TABLE OF CONTENTS:

1. EXECUTIVE SUMMARY

The number of undernourished people in the world has been on the rise since 2014. Afther a decade long decline, the absolute number of undernourished people has increased to nearly 821 million in 2017. According to the Food and Agricultural Organization of the United Nations, in 2017. 7.5 percent of children under five were affected by wasting (low weight for height) consequently putting them at a higher risk of mortality.

THE GOAL OF THE ANALYSIS IS TO CONSIDER WHICH ECONOMIC, SOCIAL, AND POLITICAL FACTORS ARE INDICATIVE OF TRENDS IN CHRONIC HUNGER IN COUNTRIES AROUND THE WORLD AND PREDICT THE ANNUAL PREVALENCE OF UNDERNOURISHMENT.

Before we creat a predictive model whose goal is to predict the annual prevalence of undernourishment at the country level from other socioeconomic indicators, we first explore the data by calculating summary and descriptive statistics, clean, describe, and interpret the data set by creating visualizations of it.

After performing the analysis, we have come to conclusions:

Many factors can help indicate trends in chronic hunger, but the key features in this data analysis were:

- fertility_rate
- avg_supply_of_protein_of_animal_origin
- access_to_improved_sanitation
- access_to_improved_water_sources
- obesity_prevalence
- access_to_electricity

By observing the trend of these socioeconomic indicators, we can predict the prevalence of undernourishment.

2. DATA EXPLORATION

SCOPE OF THE RAW DATA

Data is compiled from the Food and Agricultural Organization of the United Nations as well as the World Bank, and it contains 1401 observations. Each row in the dataset represents a country in a given year. There are 45 variables provided in this dataset, and they are divided in ten categories:

ID

- country_code
- year

AGRICULTURE

- agricultural_land_area
- percentage_of_arable_land_equipped_for_irrigation
- cereal_yield
- droughts_floods_extreme_temps
- forest_area
- total_land_area
-

DEMOGRAPHICS

- fertility_rate
- life_expectancy
- rural_population
- total_population
- urban_population
- population_growth

ECONOMICS

- avg_value_of_food_production
- cereal_import_dependency_ratio
- food_imports_as_share_of_merch_exports
- gross_domestic_product_per_capita_ppp
- imports_of_goods_and_services
- inequality_index
- net_oda_received_percent_gni
- net_oda_received_per_capita tax_revenue_share_gdp
- trade_in_services
- per_capita_food_production_variability
- per_capita_food_supply_variability

EDUCATION

- adult_literacy_rate
- school_enrollment_rate_female
- school_enrollment_rate_total

FOOD SECURITY

- avg_supply_of_protein_of_animal_origin
- caloric_energy_from_cereals_roots_tubers

HEALTH

- access_to_improved_sanitation
- access_to_improved_water_sources
- anemia_prevalence
- obesity_prevalence
- open_defecation
- hiv_incidence

INFRASTRUCTURE
- rail_lines_density
- access_to_electricity
- co2_emissions

LABOR
- unemployment_rate
- total_labor_force

POLITICS
- military_expenditure_share_gdp
- proportion_of_seats_held_by_women_in_gov
- political_stability

LABEL
- prevalence_of_undernourishment

After performing the analysis of data, we found correlations between some features and our target variable prevalence_of_undernourishment. Variables that have some significant or less significant correlation are:

- life_expectancy
- avg_value_of_food_production
- gross_domestic_product_per_capita_ppp
- net_oda_received_percent_gni
- school_enrollment_rate_female
- school_enrollment_rate_total
- caloric_energy_from_ ereals_roots _tubers
- anemia_prevalence
- open_defecation
- political_stability

And the variables that have shown the most correlation with label are:
- fertility_rate
- avg_supply_of_protein_of_animal_origin
- access_to_improved_sanitation
- access_to_improved_water_sources
- obesity_prevalence
- access_to_electricity

3. ANALYSIS & METHODOLOGY

To inform ourselves with the data, we evaluate the key variables including the target variable. In the following section we explain the process and techniques used to analyze the data, including data cleansing, calculation of statistics, visualization and exploration.

DATA CLEANING

Some of these columns are unusable for one or more of the following reasons:

- Is a unique identifier for each row or country
- Is feature with mostly missing values

For these reasons, the columns have been ignored and filtered out of the analyzed data set (resulting in 40 columns).

DESCRIPTIVE STATISTICS

In the next selection we shown summary statistics for minimum, maximum, mean, median, standard deviation, and distinct count for some important variables. We also describe the data and shown the distribution for each key variable.

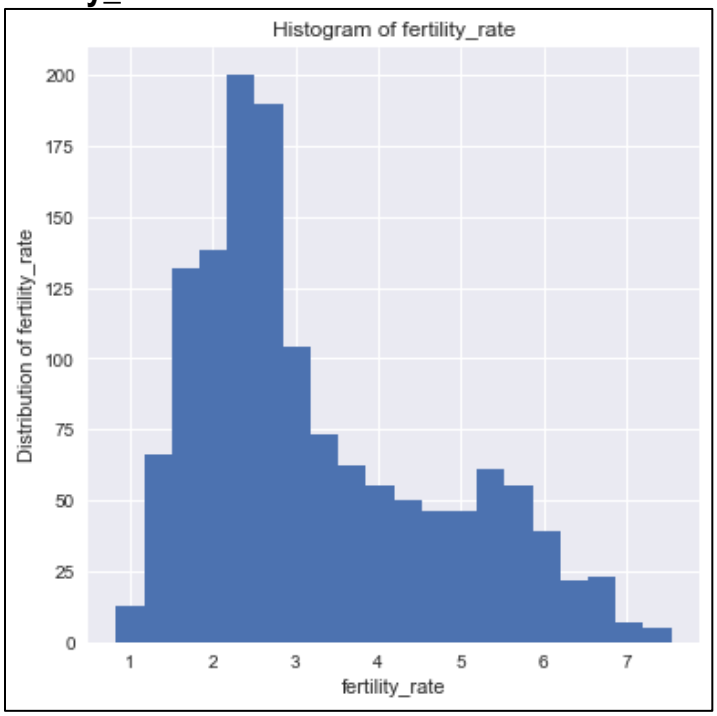
Summary statistics:

Column	Min	Max	Mean	Median	Std Dev	DCount
fertility_rate	0.83	7.54	3.25	2.75	1.47	1387
life_expectancy	38.20	84.77	67.11	69.85	8.78	1386
avg_value_of_food_ production	3.94	1042.48	229.47	205.28	149.05	1234
gross_domestic_product_ per_capita_ppp	573.16	137953.70	10843.43	6962.37	15275.31	1362
net_oda_received_percent_ gni	-0.66	189.13	6.10	2.16	12.02	1237
school_enrollment_rate_ female	35.62	101.61	88.67	93.56	12.86	795
school_enrollment_rate_total	35.33	101.77	90.25	94.64	11.16	897
avg_supply_of_protein_of_ animal _origin	2.95	83.21	27.96	25.14	15.98	1149
caloric_energy_from_ cereals_roots _tubers	22.58	84.38	50.88	50.30	13.92	1149
access_to_improved_ sanitation	10.33	101.74	65.05	73.46	28.42	1327
access_to_improved_water _sources	30.78	101.97	83.29	88.44	15.28	1339

anemia_prevalence	12.57	69.61	32.78	30.11	11.99	1321
obesity_prevalence	0.69	44.44	12.76	12.83	8.36	1244
open_defecation	0	66.68	11.70	4.77	15.13	1244
access_to_electricity	0.01	101.99	73.79	89.15	31.28	1397
political_stability	-2.78	1.37	-0.37	-0.28	0.85	1261
prevalence_of_ undernourishment	2.49	59.08	15.51	12.11	11.61	1401

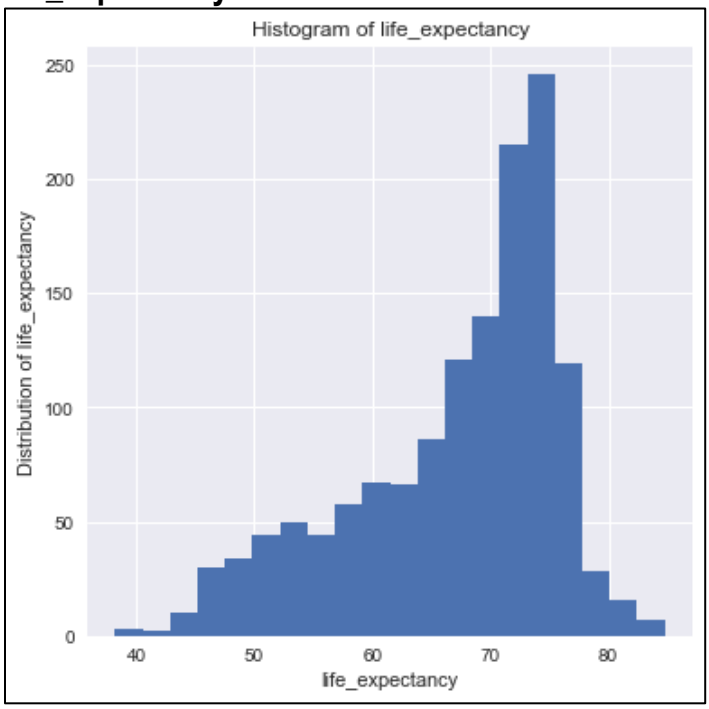
Distribution of the data for important variables:

fertility_rate



A fertility_rate is measured in births per woman. The average number of children per woman is 3.25.

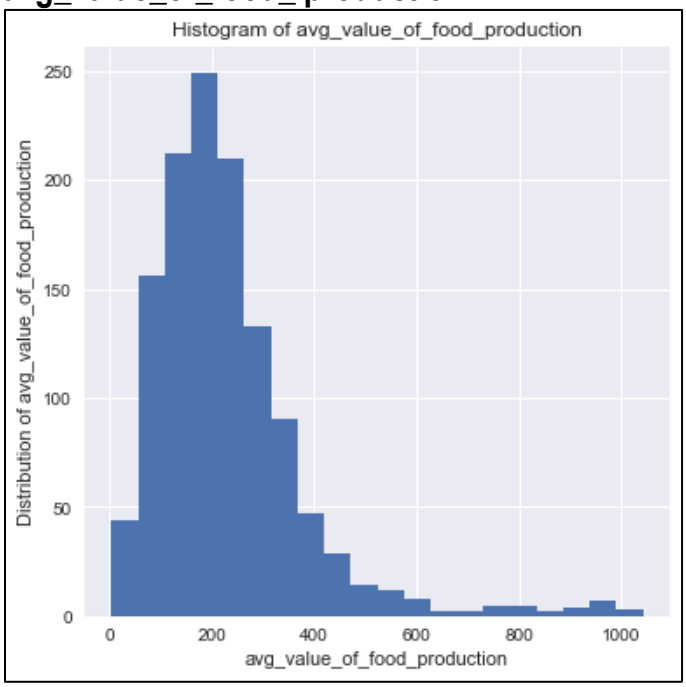
life_expectancy



Number of years a newborn infant would live if prevailing patterns of mortality at the time of its birth were to stay the same throughout its life.

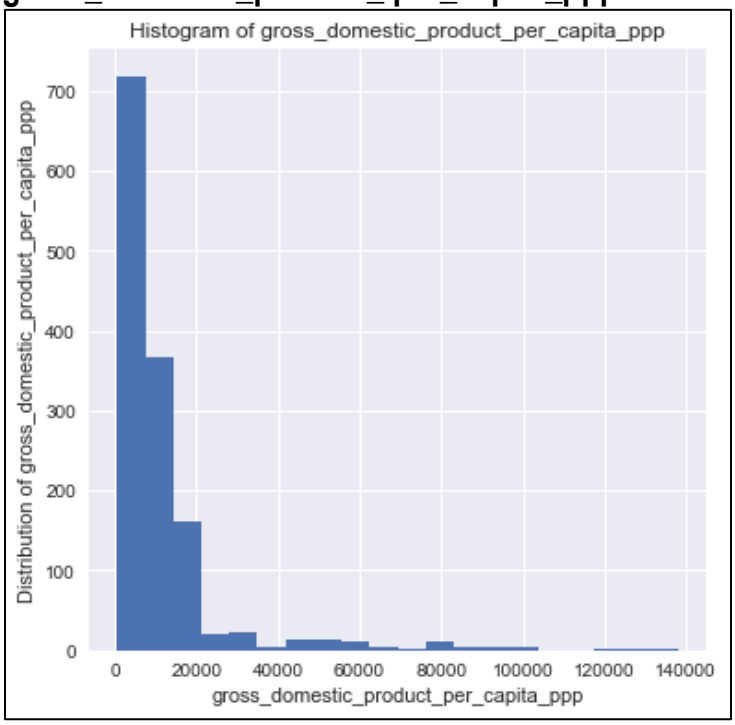
When reviewing the life_expectancy histogram plot, we notice that life expectancy data are left-skwed, with average lifespan of 67.11 years. Most of the data falling between 70 and 80 years.

avg_value_of_food_ production



Estimated food net production value of a country expressed in per capita terms. Values of avg_value_of_food_ production are right-skewed. More than 50% of food net production value is below 400 per person. In histogram we can see two little peaks around 800 and another around 1000 per person

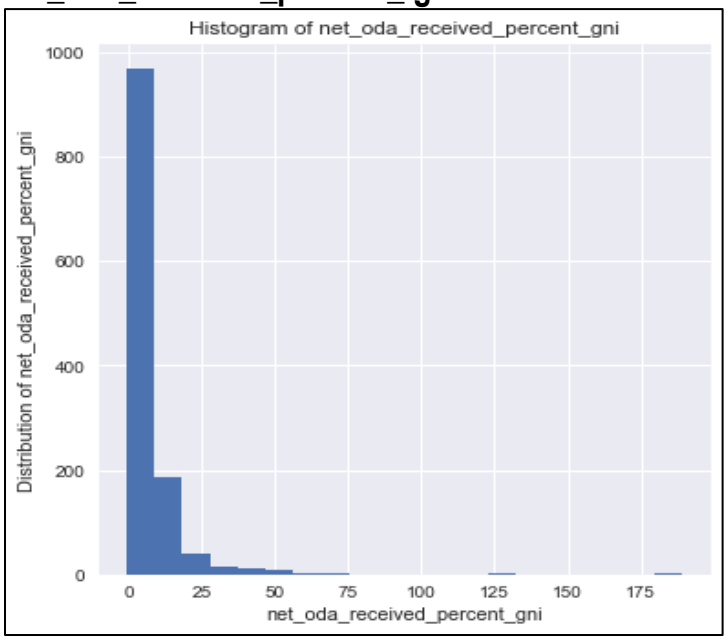
gross_domestic_product_ per_capita_ppp



The sum of gross value added by all resident producers in the economy plus any product taxes and minus any subsidies not included in the value of the products. It is divided by the total population to be expressed in per capita terms.

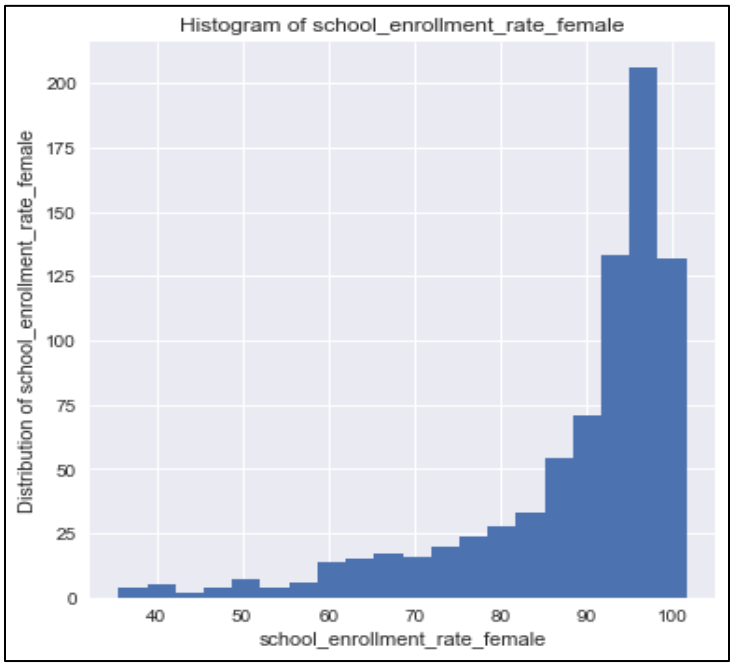
A histogram of the gross_domestic_product_ per_capita_ppp are extremely right-skewed. More than 90% of the gross_domestic_product_ per_capita_ppp data is below 20000.

net_oda_received_percent_ gni



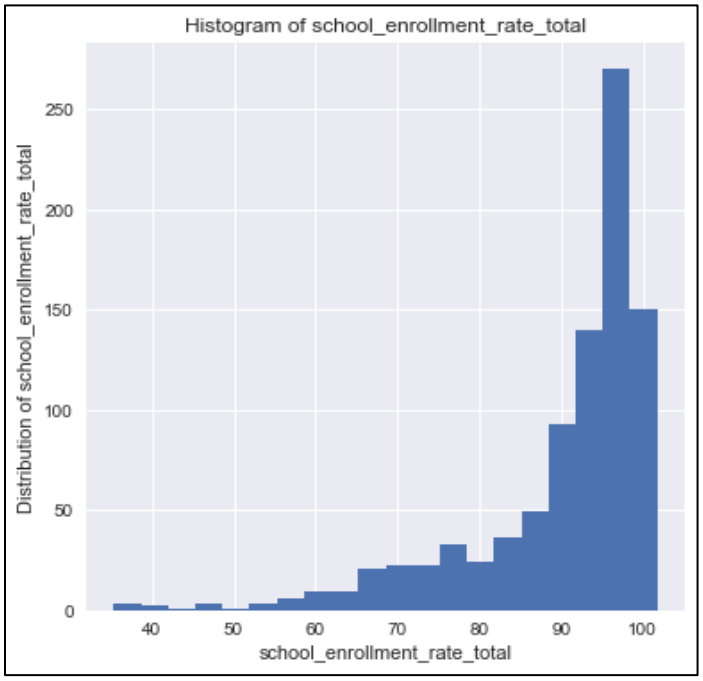
Net official development assistance received expressed as a share of gross national income (GNI). The ratio of aid to GNI provides a measure of recipient country's dependency on aid, where higher values indicate a greater dependency. The histogram for net_oda_received_percent_ gni is almost the same as for gross_domestic_product _per_capita_ppp, showing that countries with low gross domestic product, have lots of net official development assistance receive.

school_enrollment_rate_ female



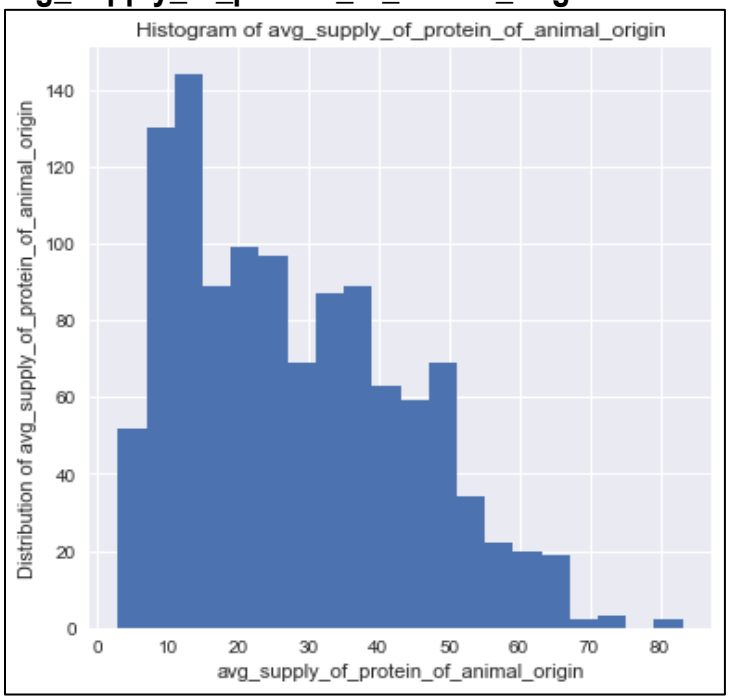
Percent of female primary education-aged children enrolled in school.
The distribution of school_enrollment_rate_female tells us that mostly of the data are between 90 and 100%, with average 88.67%.

school_enrollment_rate_total



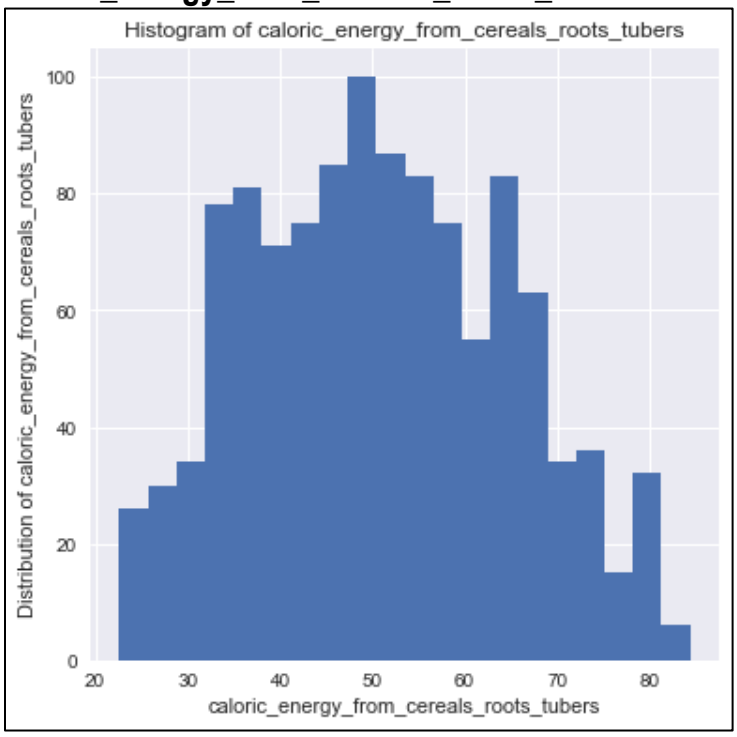
Percent of all primary education-aged children enrolled in school.
The histogram of school_enrollment_rate_total as well as for school_enrollment_rate_female shows that most percentages are between 90 and 100. That means, the more percent of all primary education-aged children are enrolled in school, the more percent of female primary education-aged children will be enrolled in school.

avg_supply_of_protein_of_animal_origin



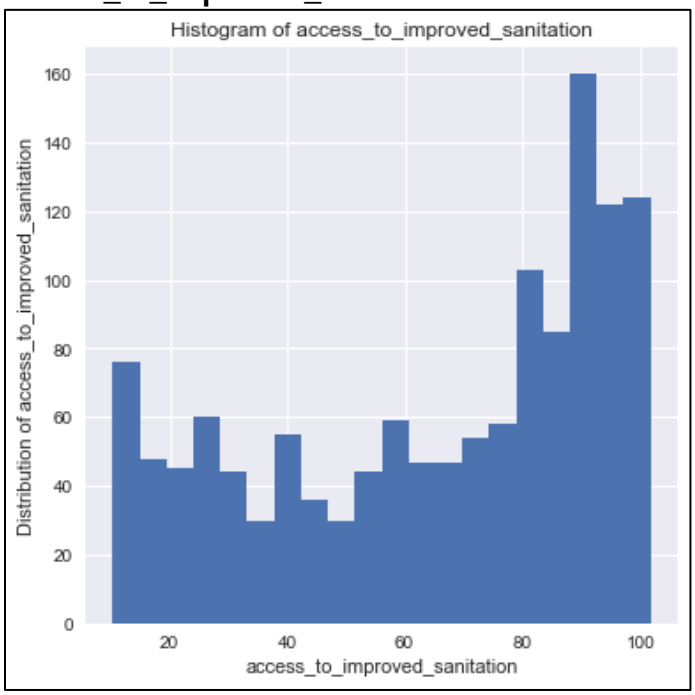
Average protein supply expressed in grams per capita per day. It includes protein from animal products.
The histogram shows more even distribution of avg_supply_of_protein_of_ animal _origin data, but still, it's lightly right-skewed. Minimum value is 2.95 grams per capita, and maximum value is 83.21 grams per capita.

caloric_energy_from_ cereals_roots _tubers



Percent of total dietary energy supply coming from cereals, roots and tubers.
The mean value for caloric_energy_from_ cereals_roots _tubers data are 50.88, and median 50.30. A small difference between these two values indicates that values of the data are evenly distributed, as we can see on histogram.

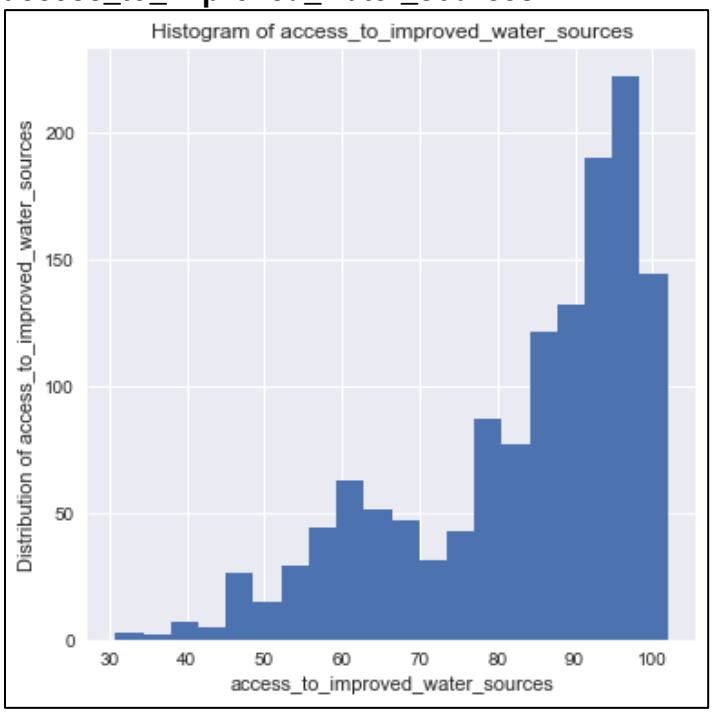
access_to_improved_sanitation



Percent of the population with at least adequate access to excreta disposal facilities that can effectively prevent human, animal, and insect contact with excreta.

The histogram of access_to_improved_sanitation shows us that the significant portion of access_to_improved_sanitation data are between 80 and 100%, but there is also lots of data between 10 and 30% access to improved sanitation.

access_to_improved_water_sources

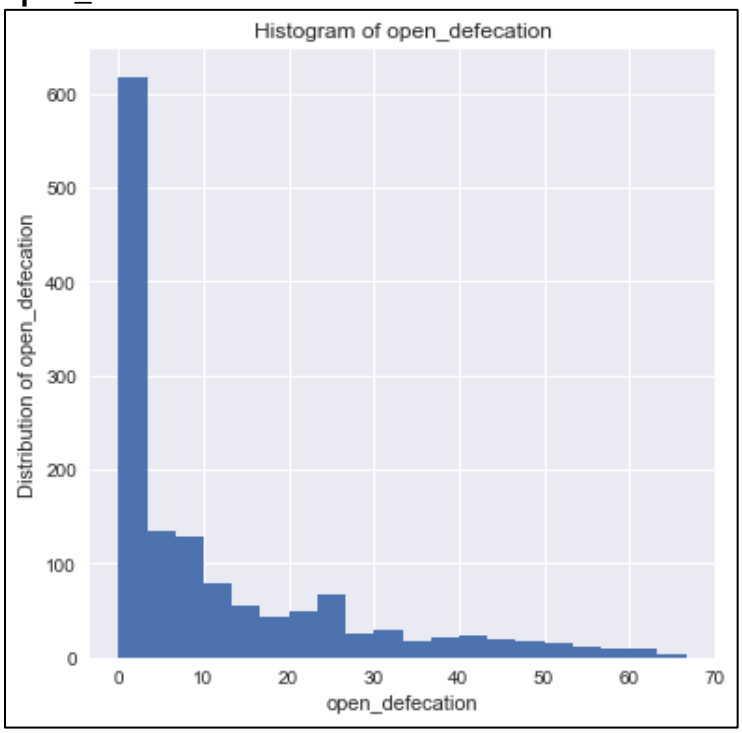


Percent of the population with reasonable access to an adequate amount of water from an improved source.

Percent of the population with reasonable access to an adequate amount of water from an improved source.

The histogram is left-skewed, with small peak around 60% The average value is 83.29%.

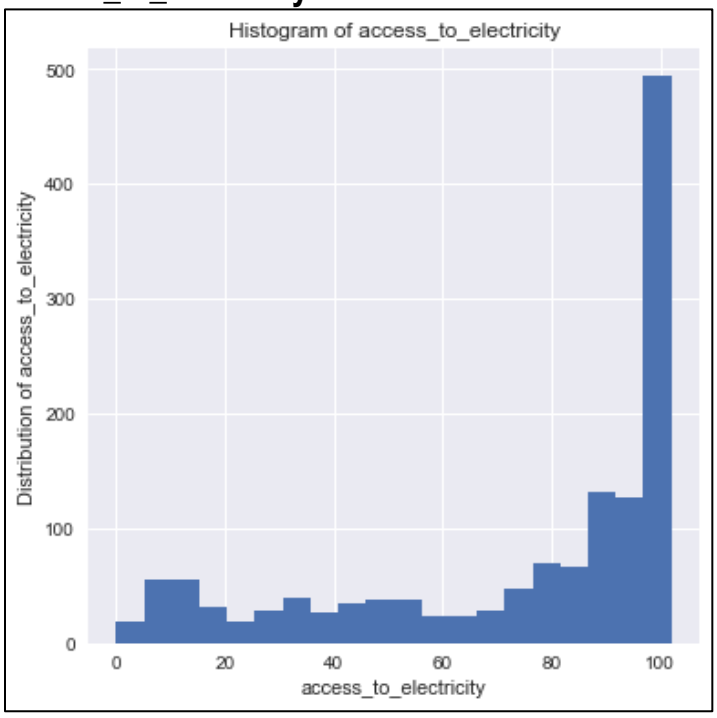
open_defecation



Percent of the population defecating in the open.

The open_defecation variable is also heavily right-skewed. The average percent of people having open defecation is 11.7%. The maximum percent is 66.68, but less than 5% of the open_defecation data are that high.

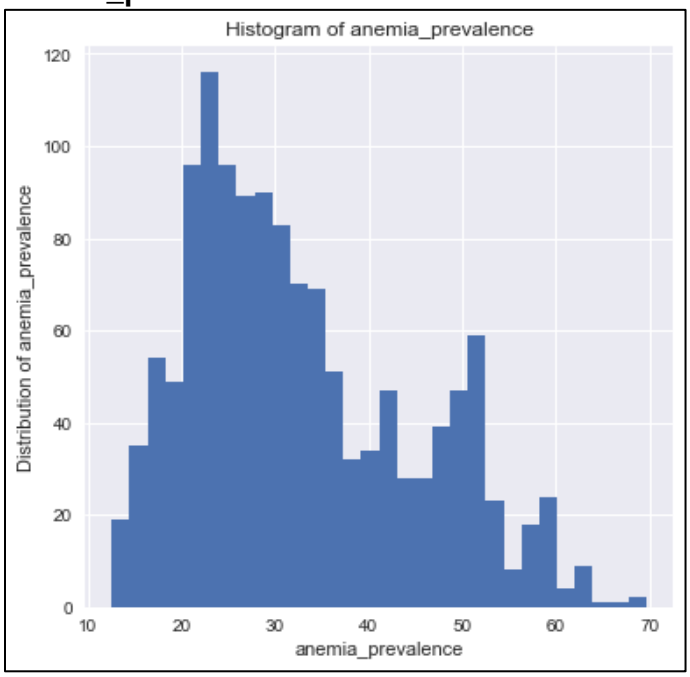
access_to_electricity



Percent of population with access to electricity.

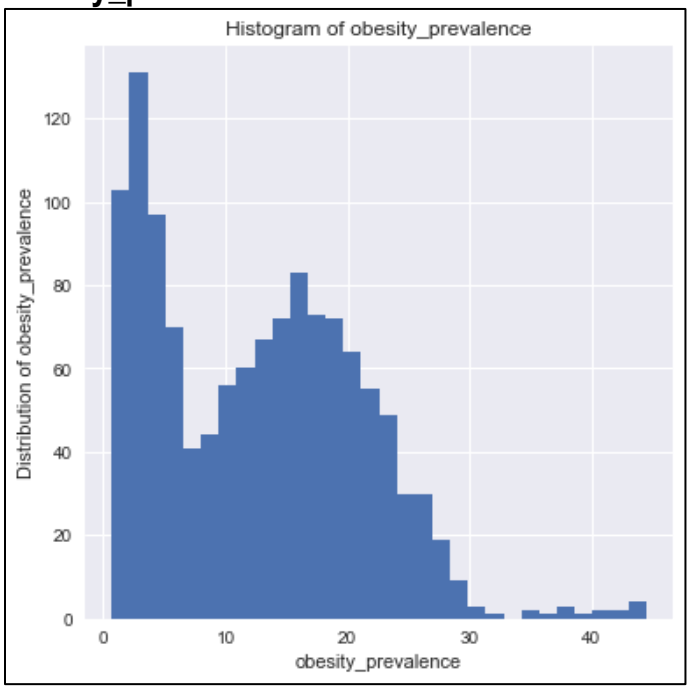
The histogram of access_to_electricity is left-skewed. Most of the data is between 80 and 100%.

anemia_prevalence



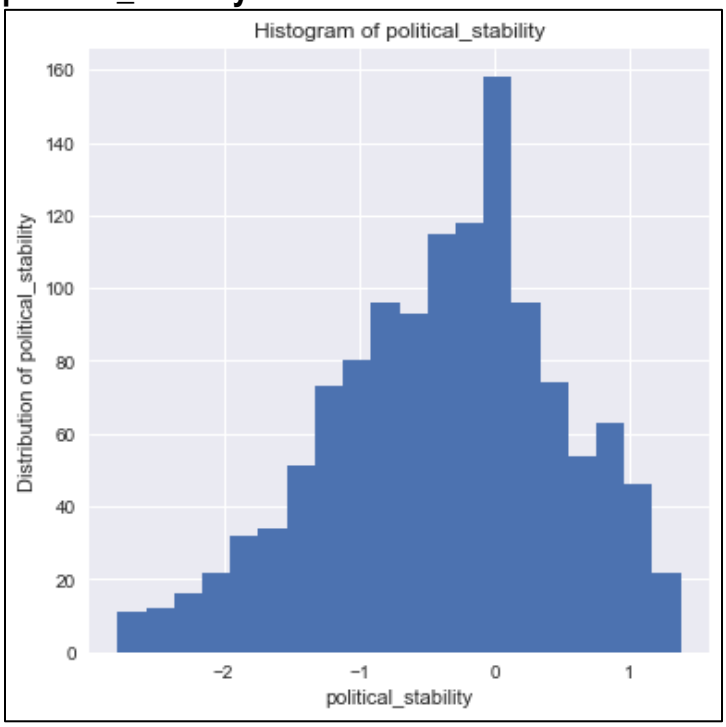
The anemia_prevalence is percent of women of reproductive age who meet the clinical definition of anemia. The histogram is right-skewed, with two peaks.

obesity_prevalence



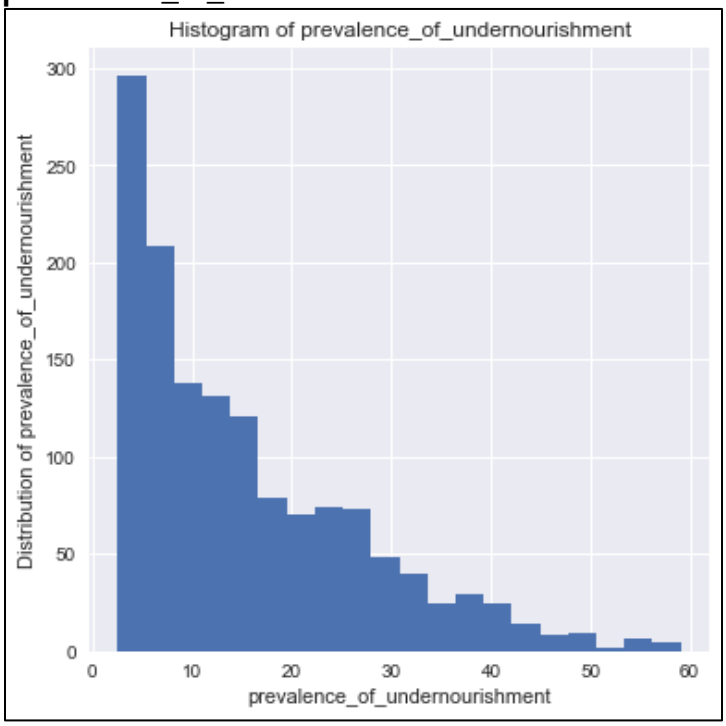
The obesity_prevalence represents the percent of adults ages 18 and over whose Body Mass Index is more than 30 kg/m2. The histogram shows two peaks, one around less than 5%, and other between 15 and 20%.

political_stability



The political_stability data are the index of the perceived likelihood that the government will be destabilized or overthrown by unconstitutional or violent means.

prevalence_of_undernourishment



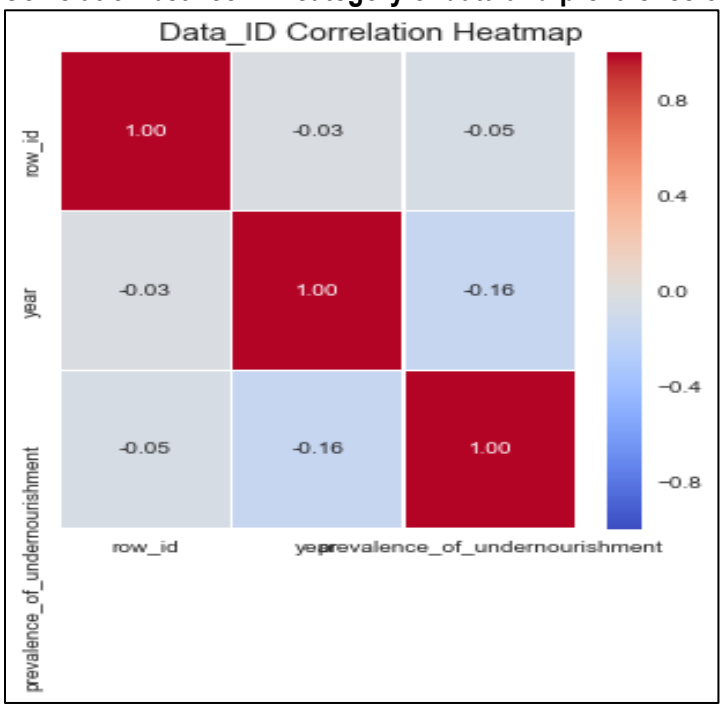
The probability that a randomly selected individual from the population consumes an amount of calories that is insufficient to cover her/his energy requirement for an active and healthy life

A histogram of the prevalence_of_undernourishment column shows that the prevalence _of_undernourishment values are right-skewed – in other words, most prevalence of undernourishment values are at the lower end of the range.

VARIABLE IMPACT ON PREVALENCE OF UNDERNOURISHMENT

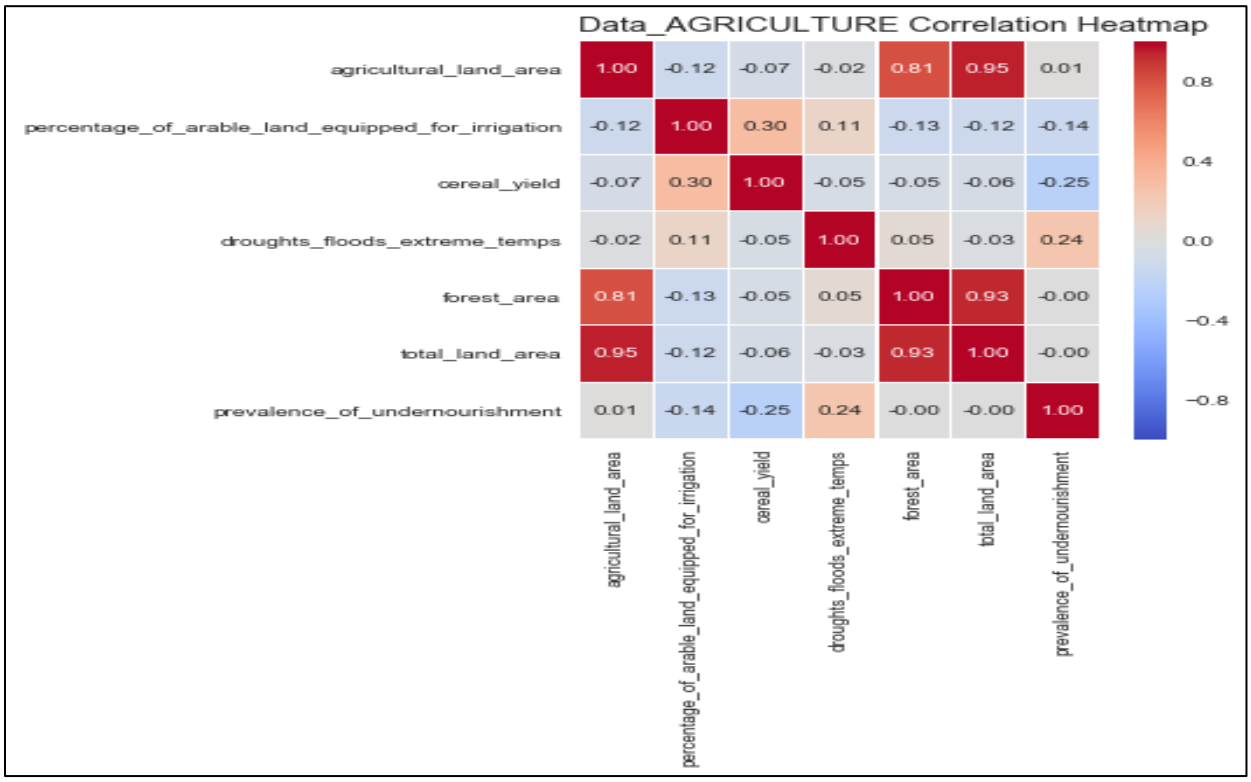
Now that we have described the data, we perform data attributes correlation heatmap to see the relationship between each variable and our outcome.

Correlation between ID category of data and prevalence of undernourishment



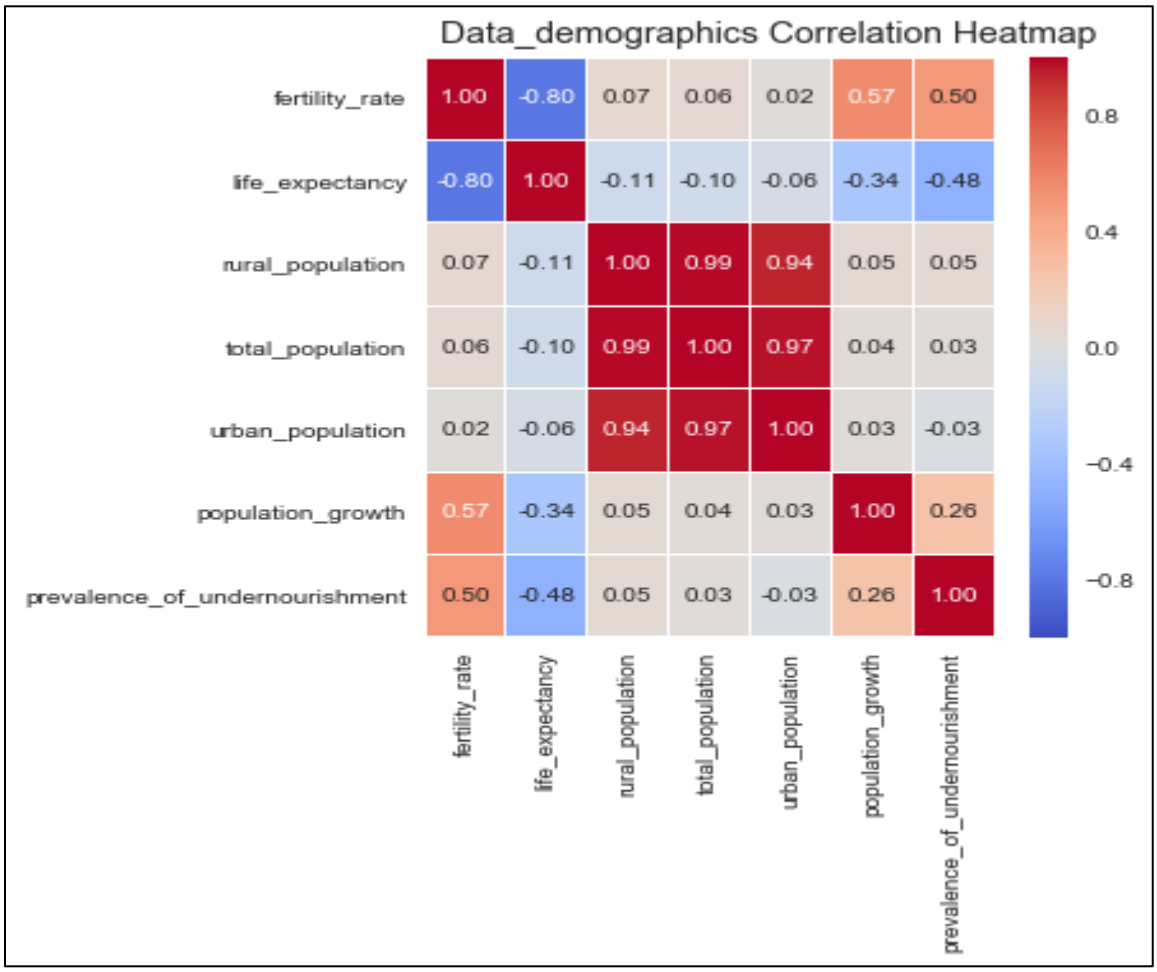
There is no signifficant correlation between year and prevalence of undernourishment.

Correlation between AGRICULTURE category of data and prevalence of undernourishment



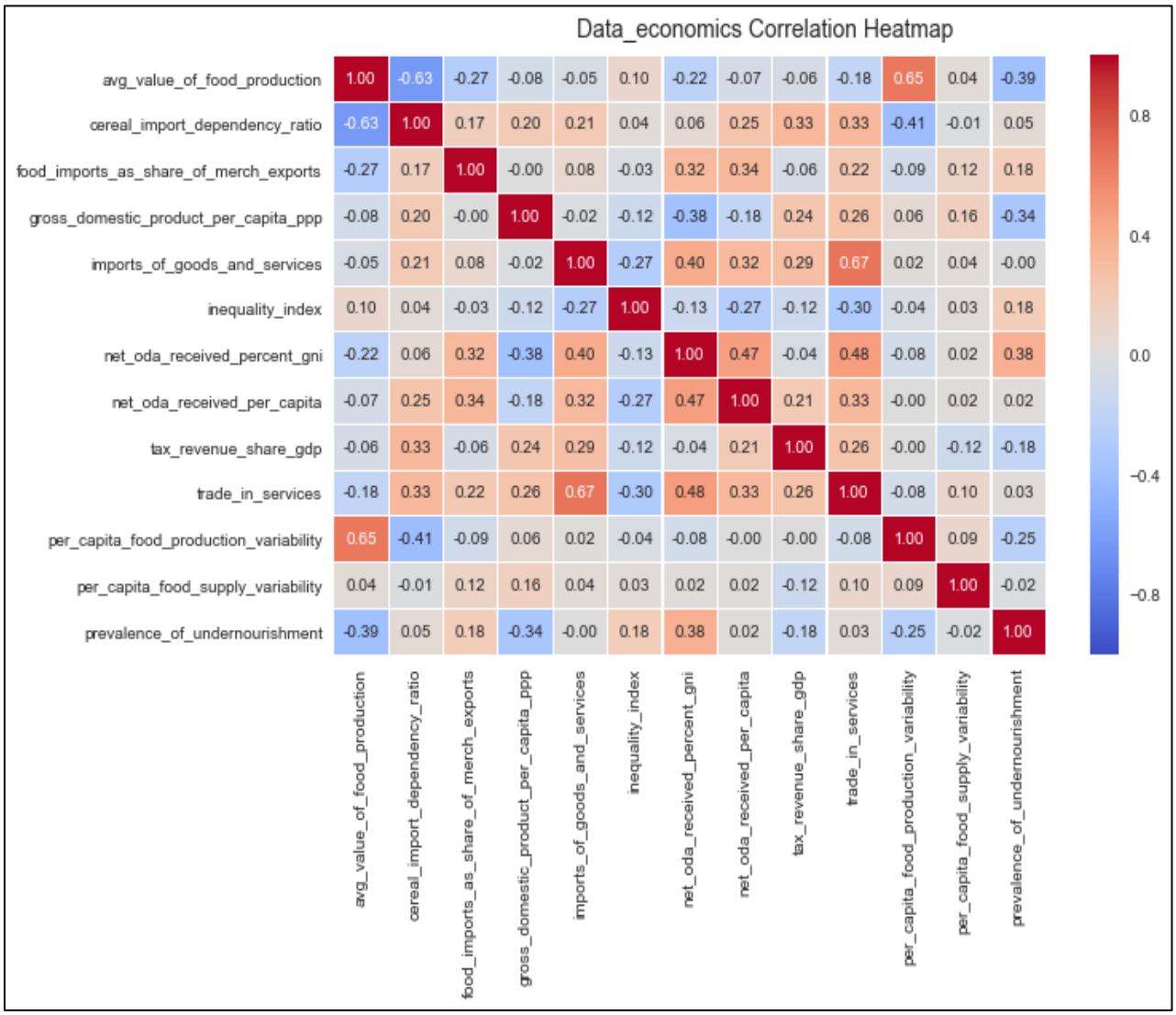
There is no signifficant correlation between columns in this category of the data and prevalence of undernourishment.

Correlation between DEMOGRAPHICS category of data and prevalence of undernourishment



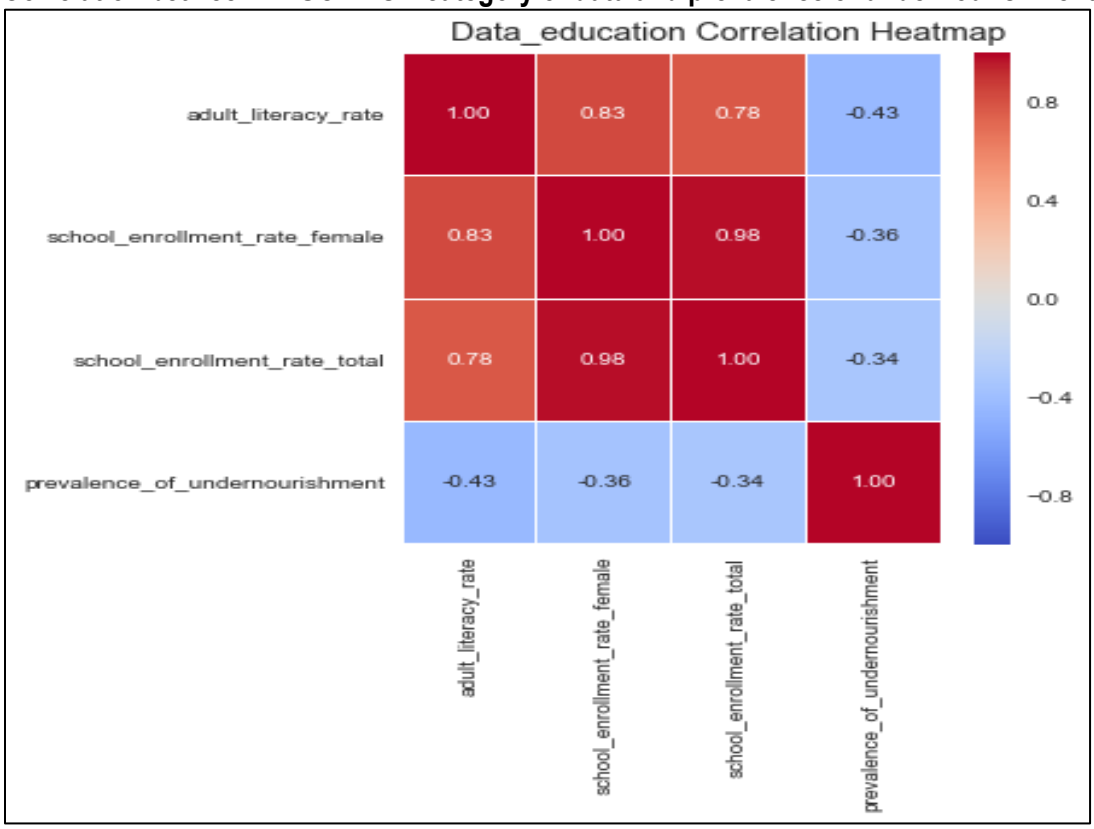
In this correlation heatmap we can see that ther is positive relationship between fertility_rate and prevalence of undernourishment, and negative relationship between life_expectancy and prevalence of undernourishment.

Correlation between ECONOMICS category of data and prevalence of undernourishment



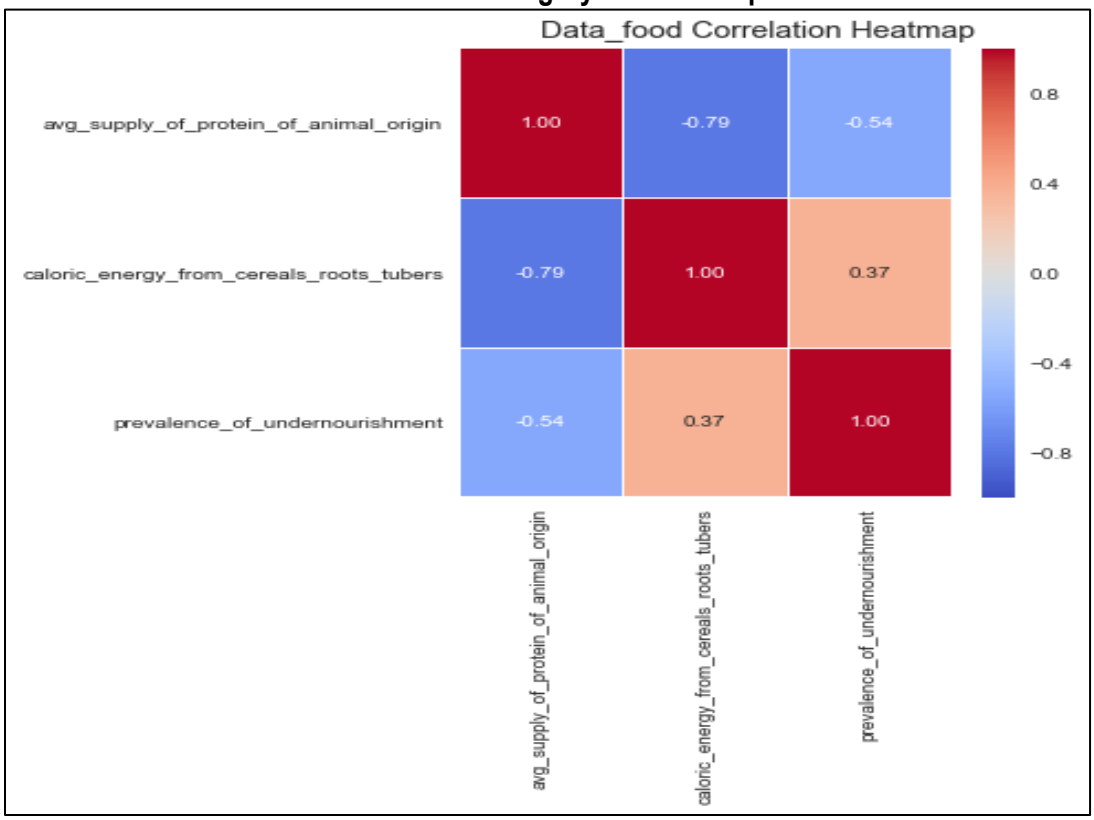
There is signifficant negative correlation between avg_value_of_food_production, gross_domestic_product _per_capita_ppp columns and prevalence of undernourishment, and signifficant positive correlation between net_oda_received_percent_gni and prevalence of undernourishment.

Correlation between EDUCATION category of data and prevalence of undernourishment



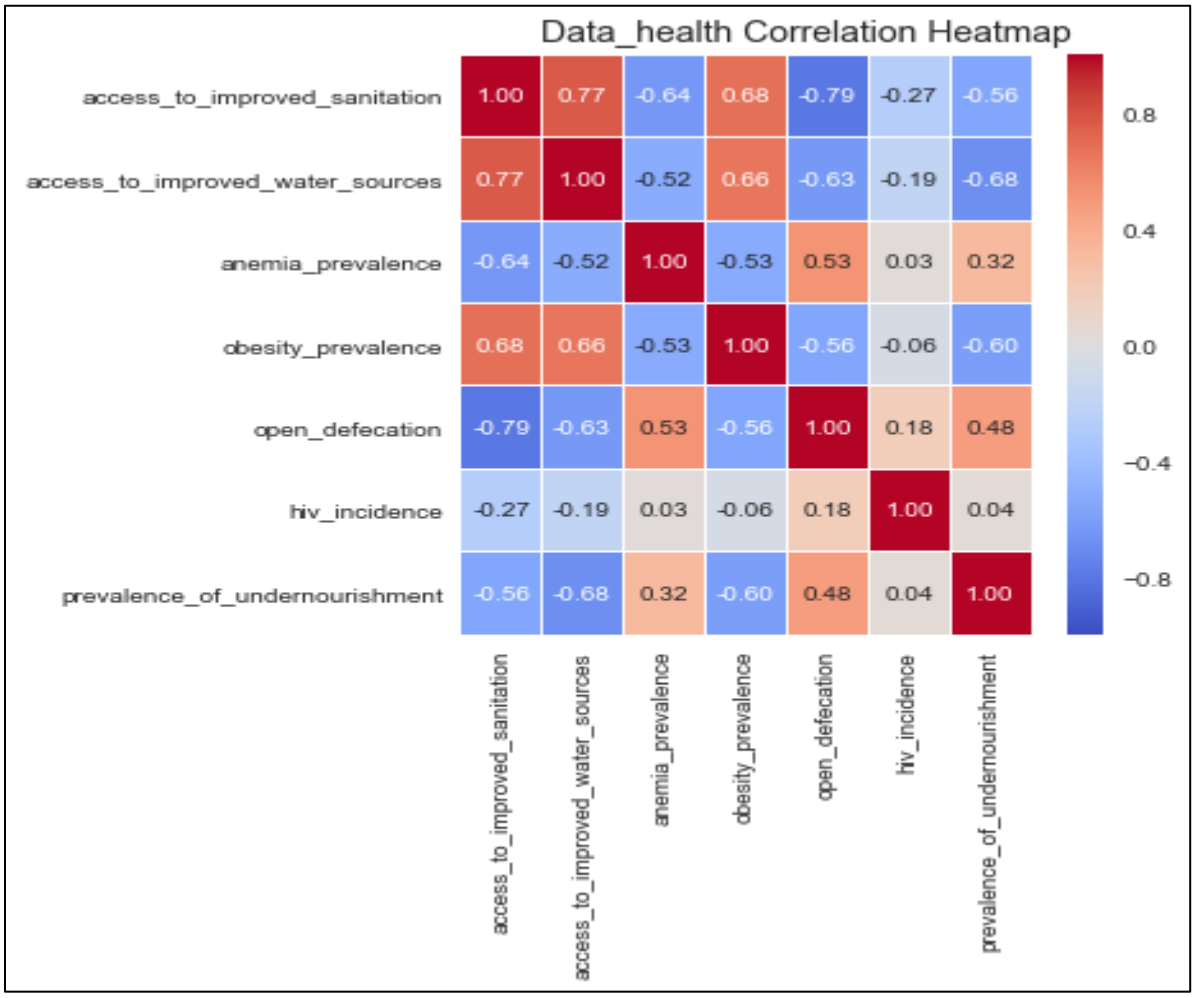
There is strong negative correlation between school_enrollment_rate_female, school_enrollment_rate_total columns and prevalence of undernourishment.

Correlation between FOOD SECURITY category of data and prevalence of undernourishment



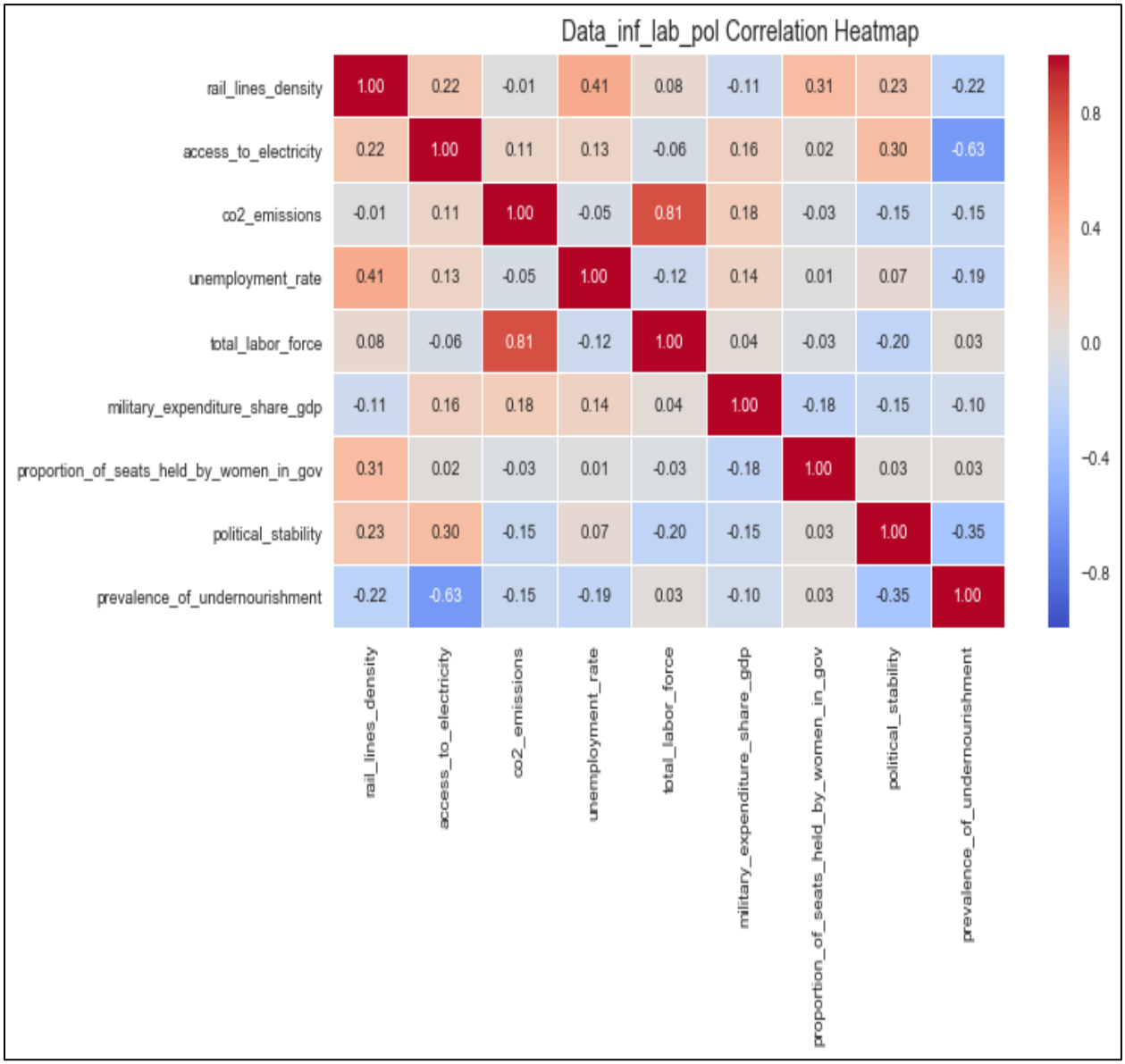
There is signifficant negative correlation between avg_supply_of_protein_of_animal_origin and prevalence of undernourishment, and positive correlation between caloric_energy_from_ cereals_roots _tubers column and prevalence of undernourishment.

Correlation between HEALTH category of data and prevalence of undernourishment



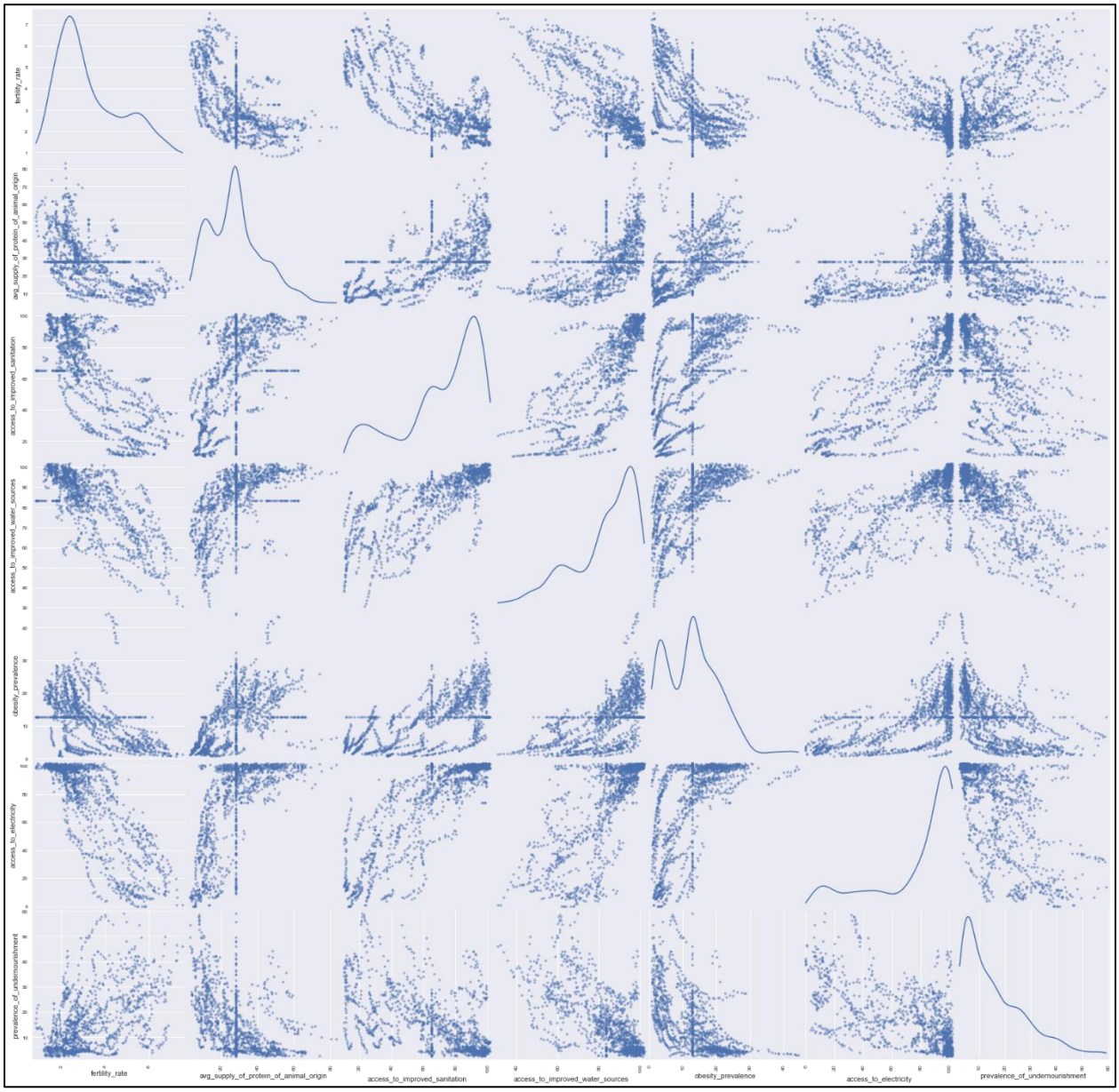
In this correlation heatmap we can see that ther is positive relationship between open_defecation, anemia_prevalence and prevalence_of_undernourishment, and negative relationship between access_to_improved_ sanitation, access_to_improved_water _sources, obesity_prevalence and prevalence of undernourishment.

Correlation between INFRASTRUCTURE, LABOR, POLITICS categories of data and prevalence of undernourishment



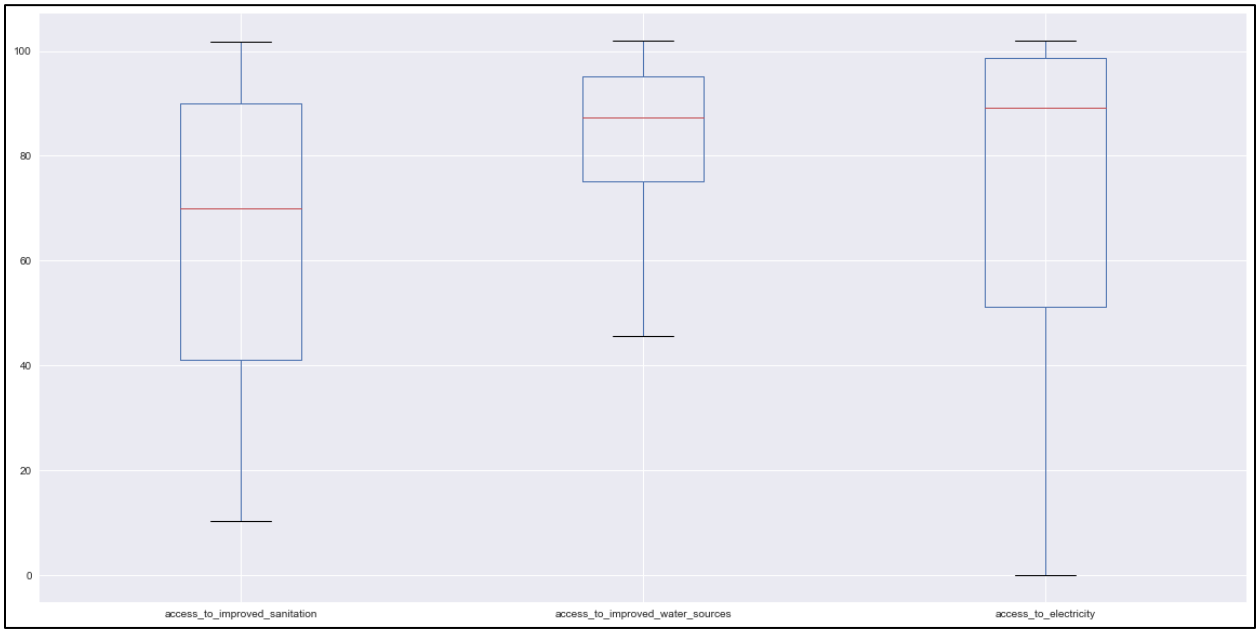
For the access_to_electricity and political_stability variables, we observe a high negative correlation with prevalence_of_undernourishment.

The Scatter Matrix Plot for six key features and prevalence_of_undernourishment reveals an expected relationship.



Viewing plots in the bottom row or the right-most column of this matrix shows an apparent relationship between prevalence_of_undernourishment and six key features.

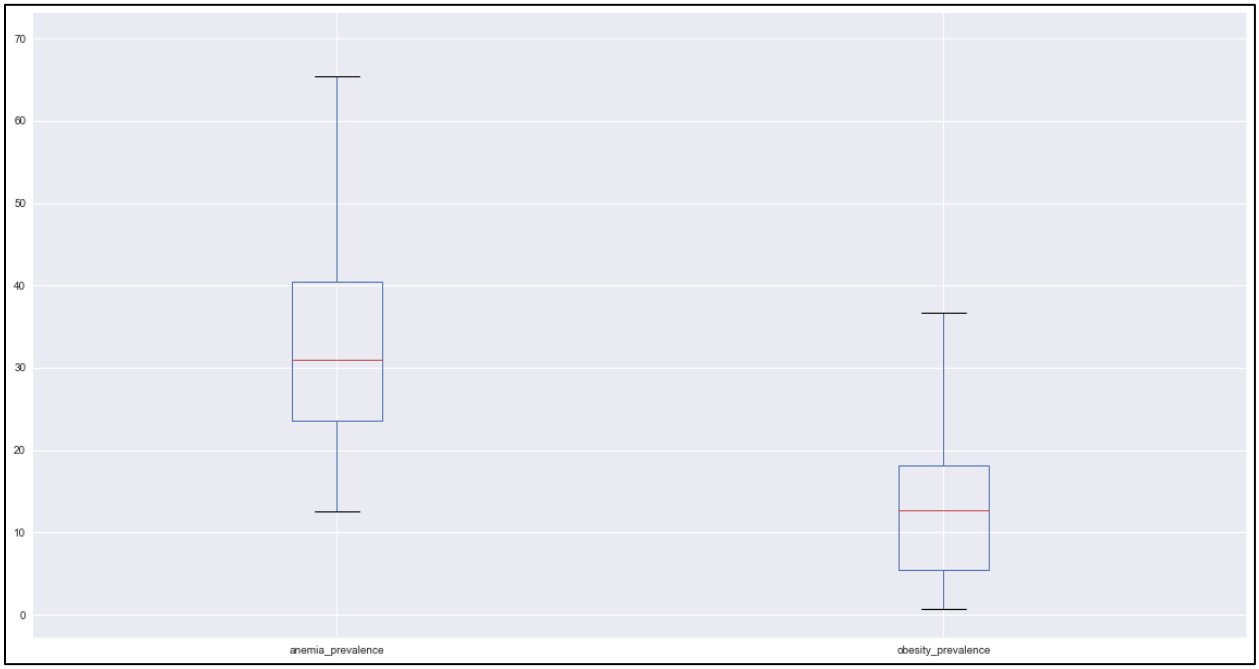
We used Boxplot to visualize the distribution of values of columns that represent living standard such as: access_to_improved_sanitation, access_to_improved_water_sources, access_to_electricity.



The box plots show some differences in terms of the median and distribution of values for these three features. For example:

- access_to_electricity have the largest range of values with some outliers around 0%, that indicates there is stll countries with population that have no access to electricity.
- access_to_improved_water_sources have the smallest range of values, with outliers around 45%.
- Median for access_to_improved_water_sources and access_to_electricity are almost the same.
- All of these three columns that represent living standard have most of their data above 40%.

We also used Boxplot to visualize the distribution of values of columns that represent health such as: anemia_prevalence and obesity_prevalence.

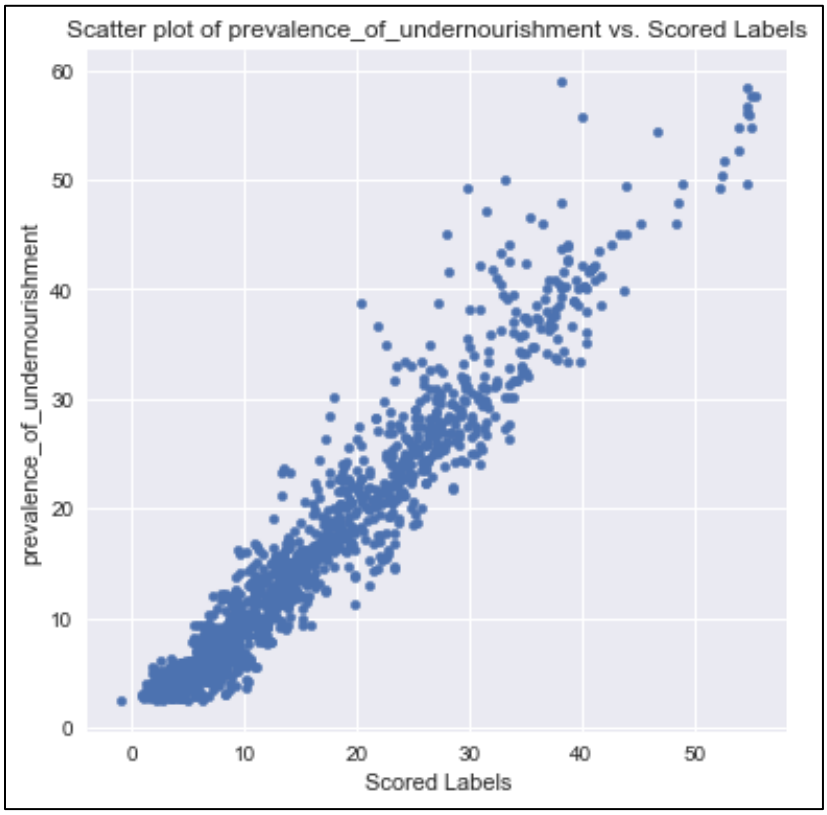


What we can see from these plot is that anemia_prevalence have higher values than obesity_prevalence.

REGRESSION

We have developed an understanding of the data set and select variables based on correlation heatmap with largest correlation index, either positive or negative.

Based on the analysis of the data, we created a regression model to predict the annual prevalence of undernourishment at the country level.
The model was created using the Linear Regression algorithm. The model was trained with 70% of the data and tested with the remaining 30%.
The Root Mean Square Error (RMSE) for the test results is 3.574332.
Here we can see a scatter plot that shows the predicted prevalence of undernourishment (Scored Labels) and the actual prevalence of undernourishment:



This plot shows a linear relationship between predicted and actual values in the test dataset.

4. FINDINGS & CONCLUSION

Our task was to create an regression model that will predict prevalence of undernourishment.
The raw data set contained 1401 observations and 45 columns. After cleaning and filtering the data, we ended up with a data set of 16 columns to develop our model.
We ended up choosing the Linear Regression algorithm.
The important variables for prevalence of undernourishment differ only slightly from the key variables for this label. The top variables that decreased the The Root Mean Square Error were:

- fertility_rate
- avg_supply_of_protein_of_animal_origin
- access_to_improved_sanitation
- access_to_improved_water_sources
- obesity_prevalence
- access_to_electricity

This analysis has shown that the prevalence of undernourishment can be confidently predicted from these top variables. The lower the value of avg_supply_of_protein_of_animal_origin, access_to_improved_sanitation, access_to_improved_water_sources, obesity_prevalence, access_to_electricity the higher is probabilities for undernourishment.