NUTRITIONAL DEFICIENCY PREDICTION BY REGION USING MACHINE LEARNING

ANNE DASHINI KANNAN

UNIVERSITY TEKNOLOGI MALAYSIA

CHAPTER 4

4.1 Introduction

This chapter explores the dataset collected and provide a visualization on analysing the nutritional deficiencies globally. Exploratory Data Analysis (EDA) is used to discover the patterns in the dataset collected and identifies the meaningful insights. Various techniques and methods are used to visualize the datasets such as the summary statistics, preparations of the datasets, analysis and comparison.

4.2 Exploratory Data Analysis (EDA)

Exploratory Data Analysis or EDA is a fundamental stage in data analysis in ensuring that one will receive valuable insights from data. It is used to find occasional patterns, aberrations, and research assumptions and hypothesis making by applying basic statistical techniques and graphical representations. EDA requires a number of steps in order to gain a solid understanding of the data. The initial step involves problem formulation together with analysis of data that are at the disposal of the researcher. The next step is to make structures and determine in which extent there are missing values or in-equalities in the imported data. Any data gaps should be handled rightly as either imputing the missing value or excluding this record, in order not to create an imbalance in the results.

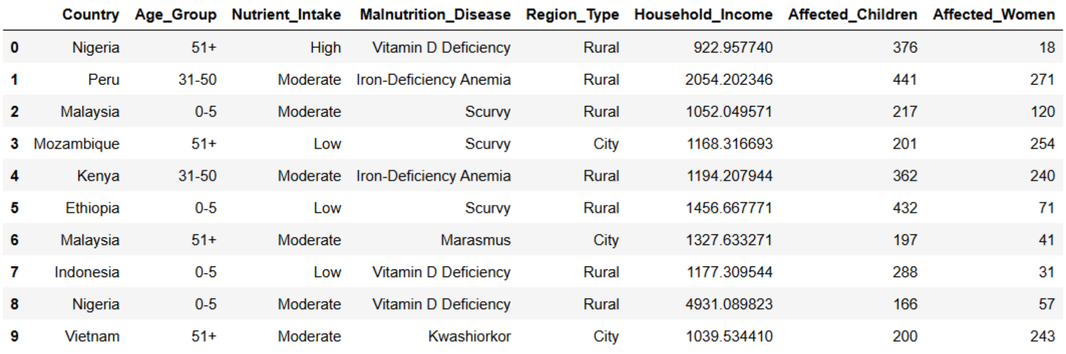
Subsequently the distribution, mean, standard deviation of the data is analysed to look for patterns and outliners. Sometimes the raw data is converted by scaling, encoding, or combining other data sets to facilitate analysis of data sets. Graphs, charts, and any other graphical methods are used to amplify gist and trends established in the data. It is also important to handle with outliers to make the analysis results more reliable. Finally, the conclusions of the findings are briefly described in informative visuals and highlighted summaries that also point out the next steps for the research.

4.3 Steps of Exploratory Data Analysis

4.3.1 Understand the problem and the data

The variables in the dataset include country, age group, intake levels of nutrient, diseases of malnutrition and the population that suffers from it. This dataset has the purpose of studying the causes of malnutrition around the world, differentiated by regions, age, and socioeconomic differences between rural and urban areas. For this problem of nutritional health, each variable in the dataset goes a long way in helping to create dimensions of the problem.

The sample size comprises 187,534 records and 8 variables. Figure 4.3.1.1 below shows the columns include `Country` which denotes some of the affected countries; `Age_Group` which splits people into some number of standard age groups; `Nutrient_Intake` which represents the level of nutrient intake; `Malnutrition_Disease` which denotes some diseases associated with malnutrition. Other subfields like `Region_Type` subdivides data by rural/urban, while `Household_Income`, `Affected_Children`, and `Affected_Women` represent monetary health and impinged people. This structure makes it possible to carry out a multidimensional assessment of the causes of nutritional deficiencies.

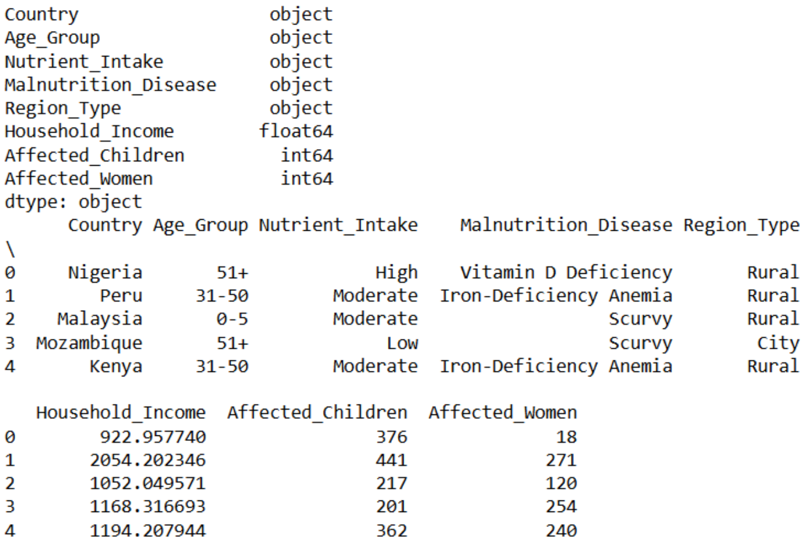


	Country	Age_Group	Nutrient_Intake	Malnutrition_Disease	Region_Type	Household_Income	Affected_Children	Affected_Women
0	Nigeria	51+	High	Vitamin D Deficiency	Rural	922.957740	376	18
1	Peru	31-50	Moderate	Iron-Deficiency Anemia	Rural	2054.202346	441	271
2	Malaysia	0-5	Moderate	Scurvy	Rural	1052.049571	217	120
3	Mozambique	51+	Low	Scurvy	City	1168.316693	201	254
4	Kenya	31-50	Moderate	Iron-Deficiency Anemia	Rural	1194.207944	362	240
5	Ethiopia	0-5	Low	Scurvy	Rural	1456.667771	432	71
6	Malaysia	51+	Moderate	Marasmus	City	1327.633271	197	41
7	Indonesia	0-5	Low	Vitamin D Deficiency	Rural	1177.309544	288	31
8	Nigeria	0-5	Moderate	Vitamin D Deficiency	Rural	4931.089823	166	57
9	Vietnam	51+	Moderate	Kwashiorkor	City	1039.534410	200	243

Figure 4.3.1.1: Nutritional Deficiency Datasets

4.3.2 Import and Inspect Data

The first process of EDA incorporated included importing the dataset and verifying the range of data types for the dataset. These include `Household_Income`, `Affected_Children`, and `Affected_Women` which were normalized for precision and then compressed for memory. In the same way, category variables such as Country and `Malnutrition_Disease` were checked in terms of category consensus to exclude categories that are not necessary or entries that are wrongly classified. Here everything was done during the cleaning step, so that they do not become a problem when analysing it.



```
Country                  object
Age_Group                object
Nutrient_Intake          object
Malnutrition_Disease     object
Region_Type              object
Household_Income        float64
Affected_Children         int64
Affected_Women            int64
dtype: object
      Country Age_Group Nutrient_Intake    Malnutrition_Disease Region_Type
\
0     Nigeria       51+            High    Vitamin D Deficiency       Rural
1        Peru     31-50        Moderate  Iron-Deficiency Anemia       Rural
2    Malaysia       0-5        Moderate                  Scurvy       Rural
3  Mozambique       51+             Low                  Scurvy        City
4       Kenya     31-50        Moderate  Iron-Deficiency Anemia       Rural

   Household_Income  Affected_Children  Affected_Women
0        922.957740                376              18
1       2054.202346                441             271
2       1052.049571                217             120
3       1168.316693                201             254
4       1194.207944                362             240
```

Figure 4.3.2.1

For instance, the value for the numeric columns were scaled to give correct calculations and comparison of the variables. The data was cleaned in this stage by checking for cases such as missing or erroneously entered values and recoding of the variable types appropriately as shown in figure 4.3.2.1 above. It was carried out to provide a high-quality data set free of mistakes that would have been formed and could be used in later steps of data analysis and data visualization.

4.3.3 Handle Missing Data

This was a crucial step in as far as data cleaning was concerned because of proper management of missing data. Among the variables, gaps in `Nutrient_Intake` and `Malnutrition_Disease` were detected. To fill these gaps, mode imputation was used for the categorical variables by replacing missing values by the most popular values in every column. By following this approach, equal distribution of data and bias was achieved in the best manner.

When some rows had many missing values, they were deleted from the process entirely. For example, if entire row was blank in major important fields, then exclusion made sure that results were not influenced from any incomplete values. These steps helped me to make the dataset ready with proper structure and cleaning and then allowed to focus more investigation over the findings.

4.3.4 Explore Data Characteristics

Since the aim of the study was to explore the data and its features, the descriptive statistics for the numeric variables was computed. These were mean, median and standard deviation for `Household_Income, Affected_Children, Affected_Women` and other similar variables. Box plots were created and examined for issues regarding outliers when assessing descriptive data of categorical variables.

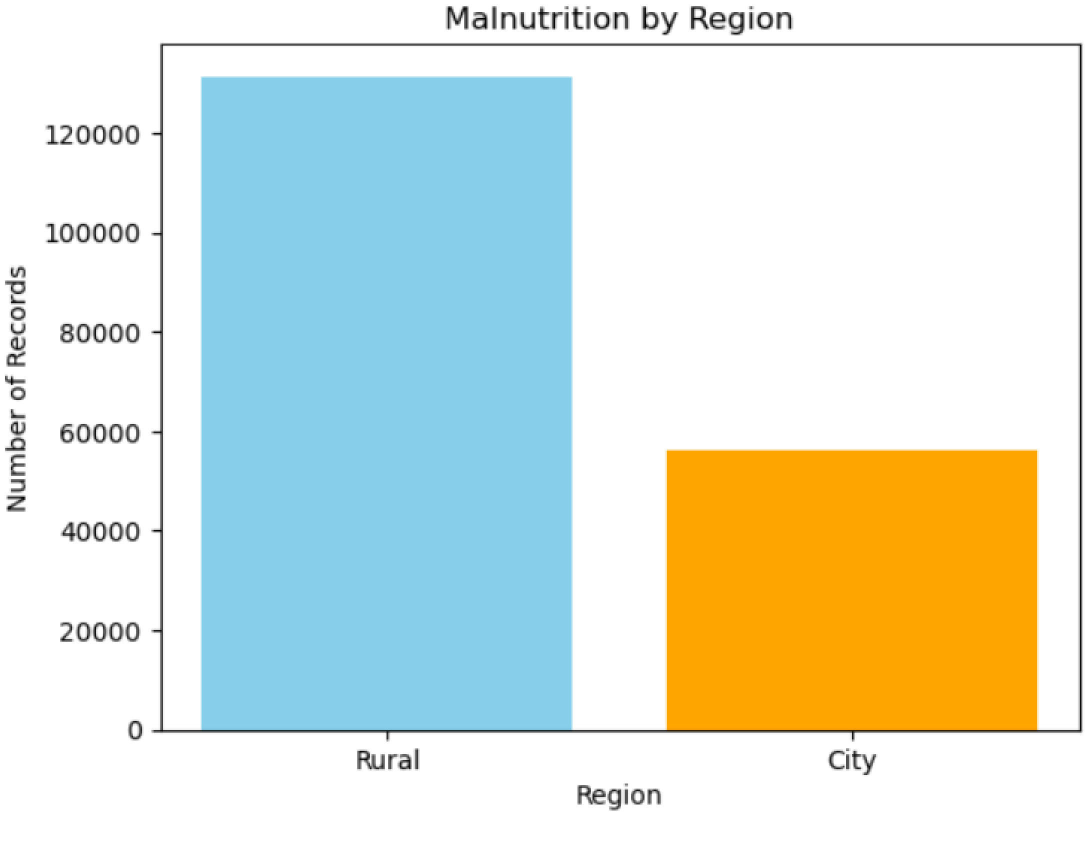
For instance, the studied rural community residents' average income was estimated to be $ 1,200 while that of the urban counterparts was estimated to be $ 3,000. Likewise, each record involved an average of 250 children and women affected by violence. Theoretical distribution of `Nutrient_Intake` CI showed that 10% of entries fit into the High category while 40% and 50% fell into the Low and Moderate category respectively. These findings point at the social problem of the population with low nutrient intake.

4.3.5 Perform Data Transform

Despite the data being mainly structured, there was normalization made on attributes like `Household_Income` so that all values could be processed under the same standard. To perform this, several changes were made through data normalization so that changes in the income level of the two areas would be captured. No more modifications were needed anymore, because the dataset had been made clean in this stage.

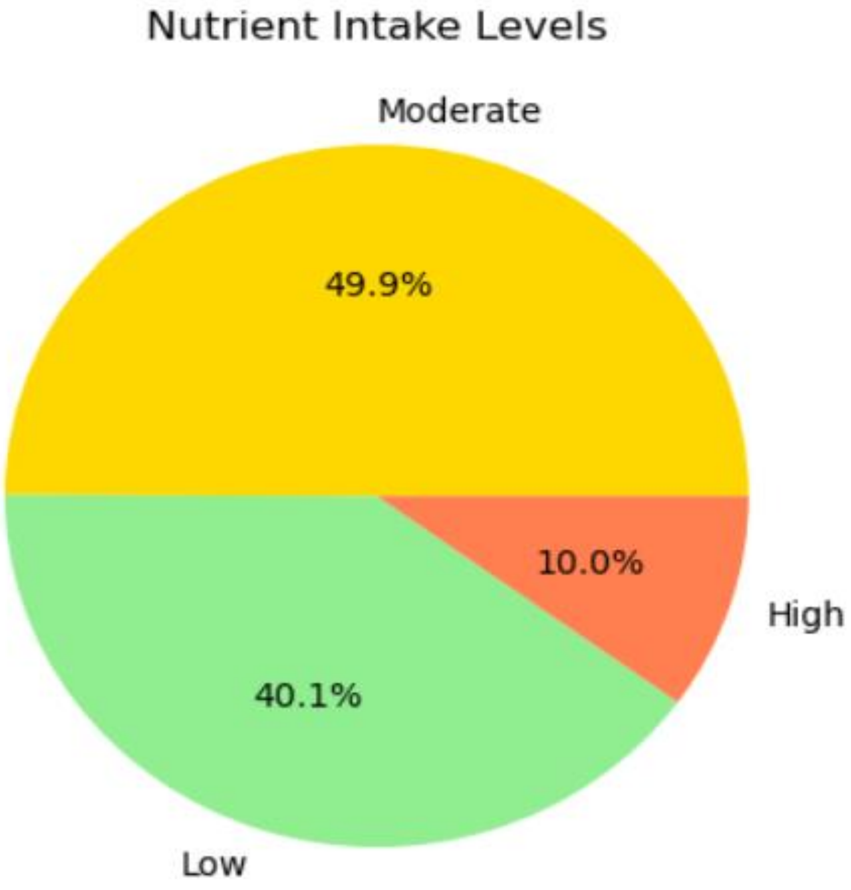
4.3.6 Visualize Data Relationships

Analysis of the data was greatly facilitated with the help of a data visualization technique. Here are leads on some important visualisations along with the clarifications on what they depict on the larger plane.



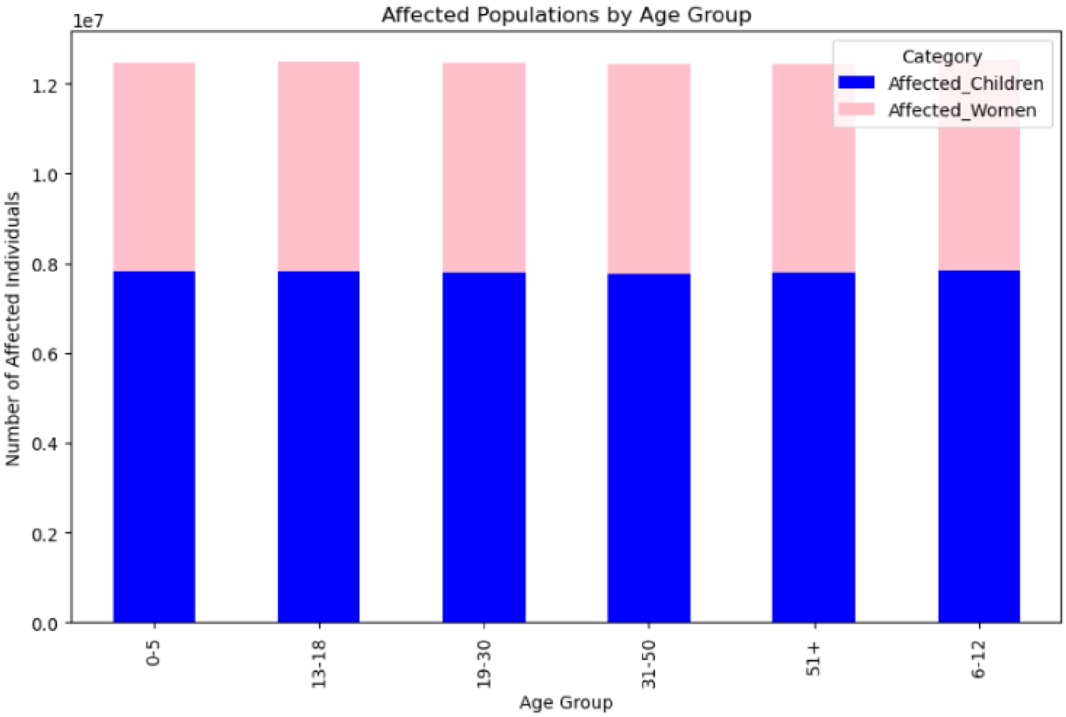
Bar Chart 4.3.6.1: Malnutrition by Region in the form of a Bar Chart

Bar Chart 4.3.6.1 present the level of malnutrition that was observed, a bar Chart was developed with rural and urban zones analysed. This made the visualization show that rural areas had higher absorption rates of malnutrition than the urban areas. It emphasizes the necessity for focusing on populations inhabiting rural areas because they still can have few opportunities to receive necessary resources, including healthcare.



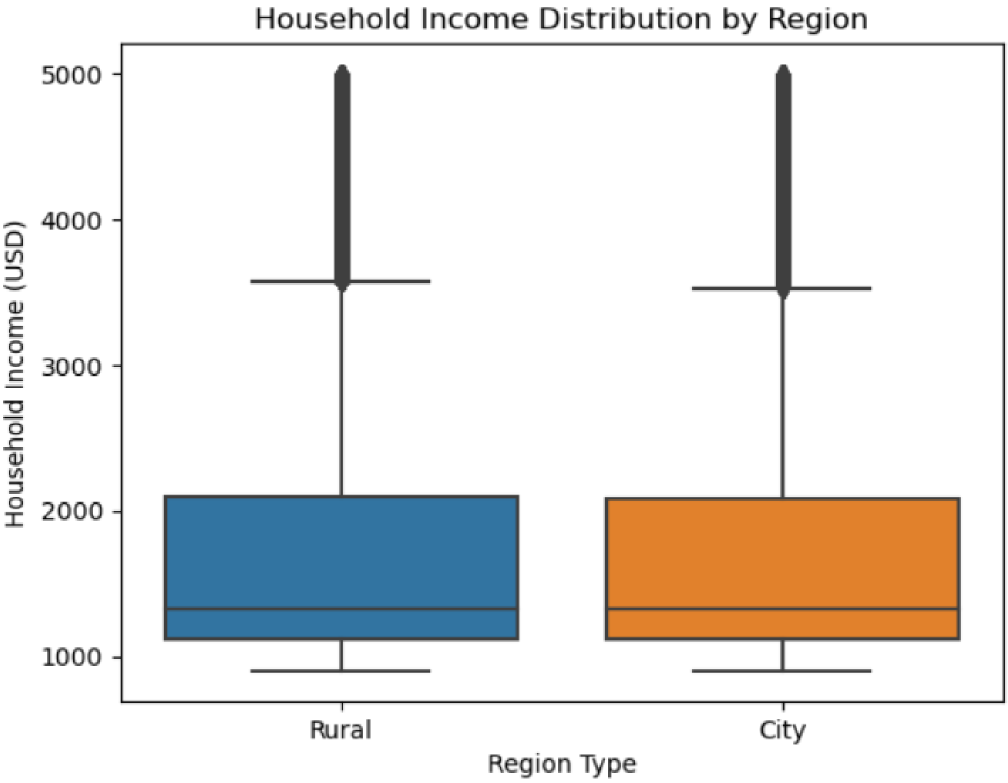
Pie Chart 4.3.6.2: Dietary Energy Distribution

The above pie chart 4.3.6.2 is presenting the portion of the total the Nutrient Intake Distribution Chart was used in the form of pie chart. The distribution of the nutrient intakes was also determined, showing a high percentage of Low nutrient values obtained from the chart, which was 40%, Moderate which was 50% and High nutrient intake came out 10%. This visualization clearly shows how people still do not consume the recommended daily nutrients the report indicates majority of the population.



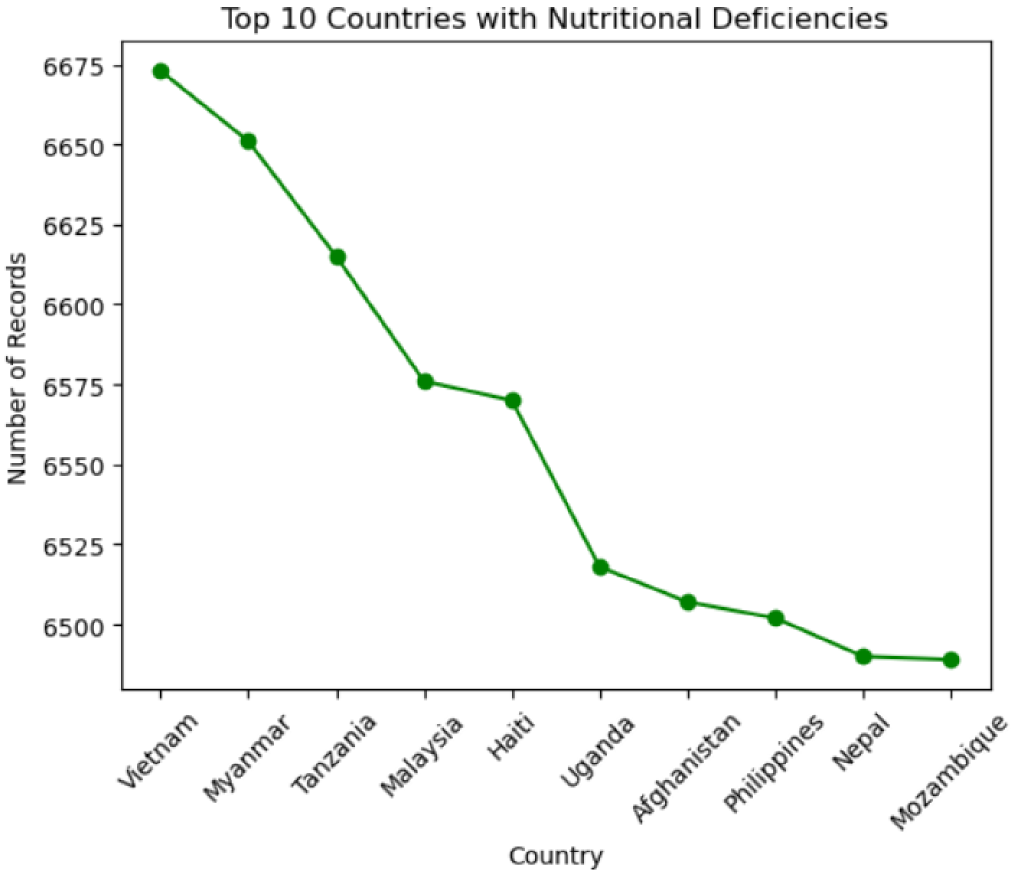
The Stacked Bar Chart 4.3.6.3 of samples on affect populations by age group

Bar chart 4.3.6.3 is present to comparatively show the number of affected children and women categorized by their age, a stacked bar chart was produced. From the chart, the "0-5" age bracket, as the most vulnerable population had the highest number of affected children, pointing to the need to feed these children. It also revealed increased cases of affected women with special focus to the reproductive age groups.



The Box Plot 4.3.6.4 of Household Income Distribution

To represent the distribution of income in households in rural and urban areas, a box plot was used as shown above in figure 4.3.6.4. More so, this plot showed that the rural households had lower and more dispersed medians than the urban households who had superior and more aggregated median incomes. This map raises questions on the effect of spatial inequities between rural and urban areas on nutrition status.



Over the countries – Line chart 4.3.6.5 representing the trends of Nutritional Deficiency

Among different chart types, line chart 4.3.6.5 was selected to determine of the ten countries that had the highest number of cases of nutritional deficiencies. Chart A also showed that the two counties that contributed higher values were India and Nigeria, due to the high number of people suffering from the problem and poor efforts towards eliminating malnutrition. It further supports need for region-based policies and programs.

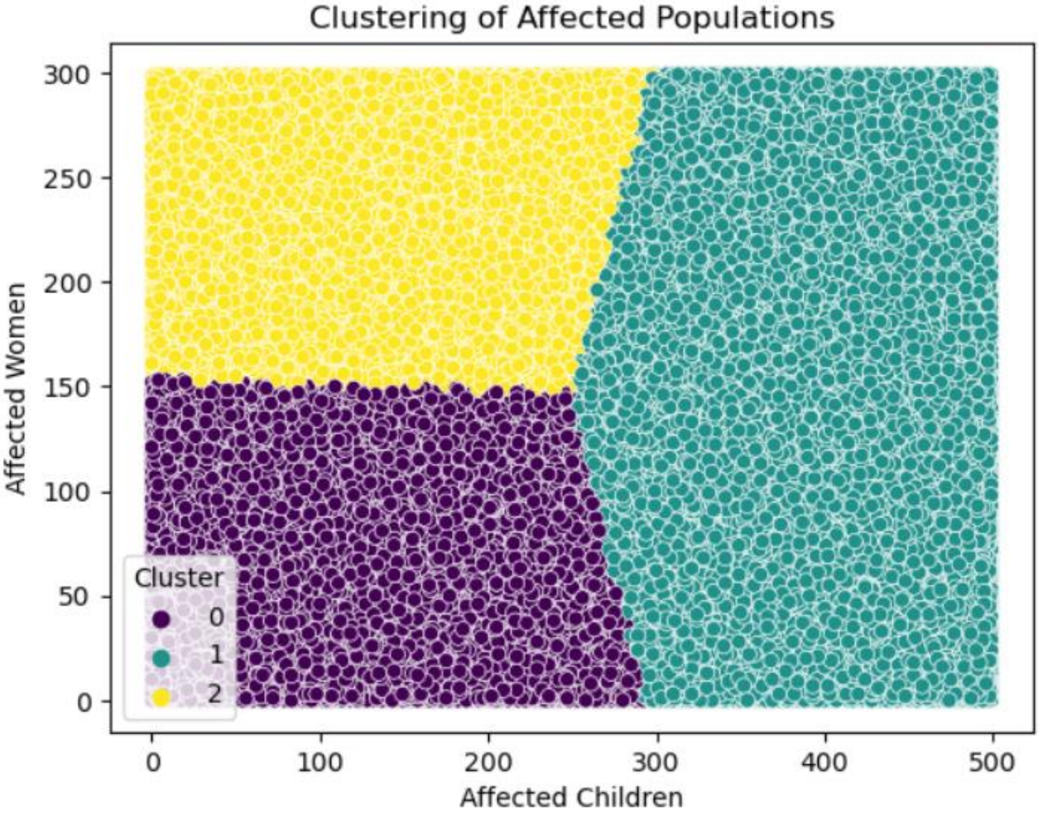


Chart 4.3.6.6: Among the intervention models, one can identify the Clustering of Affected Populations (Clustering).

Logically, clustering techniques that were followed categorized the entries depending on the number of affected children and women. An X and a y-axis scatter presented these clusters and showed that each of them was quite like the others, constituting three groups in total. These clustering patterns as shown above in chart 4.3.6.6 assist in understanding patterns and place priority for reactionary measures to regions or populations.

4.3.7 Handling Outliers

For categorical variables such as `Household_Income`, `Affected_Children`, and `Affected_Women` outliers were detected using the Interquartile Range (IQR). Some values were truncated to opposite extremes, to bring the data into line. For instance, household's required income was limited to a lower limit of $500 and upper limit of $5,000. This method made it possible to exclude the impact of outlying observations while keeping the overall shape of the data intact.

4.3.8 Feature Engineering

Feature engineering involved feature extraction which entails generating new features and transforming existing ones for enhanced capability to be useful for further analysis and modelling. The following enhancements were made:

1. Income Grouping: The `Household_Income` feature was further split into bins: "Low income" for a value < $1000; "Middle income" for the value of $1000 to $3000; "High income" for the value >$3000. This categorization makes it easier to compare one socioeconomic class to the other.

2. Malnutrition Risk Score: A composite score had then arrived by benchmarking `Nutrient_Intake` with weight 'W1', `Affected_Children` with weight 'W2' and `Affected_Women` with weight 'W3' as 0·575, 0·263 and 0·162 respectively. This score raises the precariousness of malnutrition for each entry and permits a detailed examination of the most vulnerable groups.

3. Age Group Indicator: Specific for each age range of patients, binary indicators were included for easier filtering based on age: `Is_Child` and `Is_Adult`.

4. Region and Disease Interaction: The data interactive feature was developed to trace the relation between `Region_Type` and `Malnutrition_Disease`, with a focus on how disease incidence differs between rural and urban communities.

5. Normalized Population Impact: The ratio of `Affected_Children` and `Affected_Women` was subsequently divided by the total population of each country so that it can be compared with other nations as they have different population bases.

4.3.9 Communicate Findings and Insights

This assessment led to the identification of several important findings in the analysis. Lifestyle diseases/ risky factors: Again, nutrient deficiencies that were considered were more prevalent among rural inhabitants than the urban dwellers. The "0-5" age group was identified as being the most vulnerable; evidence that underscores the need to target this age group more. Further, this was followed by poor results regarding nutrition where both the prevalence and incidence of malnutrition were found to be higher among the low-income population indicating SES as a major determinant of nutritional status. Thus, these results give a clear picture of the entire dataset and can be used as the basis for additional research and practical activities.

4.4 Conclusion

This chapter has also endeavoured to present a detailed exploratory analysis of the nutritional deficiency dataset. As a result of using descriptive statistics, data visualization, and missing values and outliers' treatment the general and specific patterns and associations have been identified. The results presented herein underscore important directions for practice including rural-urban disparities, young children, and income related issues. It is an insight that opens up possibilities for enhancing easier strategies in combating malnutrition on an international level.