Report #1

Classifier for Humanitarian Response

Exploration, Data Visualization, and Data Pre-processing

Authors: Ghiath Al Jebawi, Bercin Ersoz, Allaeldene Ilou, Caspar Stordeur

Index

1. Context

2. Objectives

3. Framework

4. Pre-processing and Feature Engineering

5. Visualizations and Statistics

6. Further Steps

1. Context

1.1 Project Context

In humanitarian crises, international non-governmental organizations (NGOs) face significant challenges in identifying needs and targeting communities due to the complexity and volume of data. Despite the availability of large datasets, data often remains underutilized, not transformed into actionable insights, or easily accessible for NGOs.

Syria, experiencing one of the largest humanitarian crises in recent decades, has a vast ecosystem of NGOs working to support communities. The availability of humanitarian data in Syria is substantial compared to other conflict zones, making it a suitable case for this project.

1.2 Technical Context

NGOs such as UNHCR and IMPACT REACH Initiatives have developed a consistent database over the past six to seven years, updated monthly, describing thousands of communities with internally displaced people (IDPs). The data includes thirteen key indicators (e.g., Health, Shelter, Education), each divided into numerous sub-indicators. The three most critical aspects of the data are:

- **Aid** provided to communities by NGOs.
- **Needs** defined by the communities themselves.
- **Vulnerability** criteria, such as the ratio of income to survival costs.

1.3 Scientific Context

The current process of aid provision is not systematic, especially during large-scale community displacements. Data is not easily summarized or utilized by local NGOs. The proposed model aims to base NGO interventions on three key aspects: aid, needs, and vulnerability.

1.4 Feasibility Context

The goal is to make the available data easily usable by creating a system that describes each community's needs and priorities for intervention. The model will predict future vulnerability trends and recommend interventions, optimizing the costly process of resource allocation. If successful, this model could significantly improve the efficiency of aid distribution, ensuring that resources reach the most vulnerable communities in the shortest possible time.

2. Objectives

2.1 Main Objectives

The primary objective is to transform available data (monthly time series over five years) into a predictive and classifier model. The model will learn from the data to predict future vulnerability trends and assist decision-makers in prioritizing aid and planning interventions.

2.2 Team Expertise

The team consists of four members with backgrounds in data analytics, data science, and statistics:

- **Ghiath Al Jebawi**: Data scientist with five years of experience in humanitarian work in Syria, particularly in IDP camps.
- **Bercin Ersoz**: Statistician with experience in data management, previously working at the Turkish Central Bank.
- **Allaeldene Ilou**: Freelancer with a background in Computer Science and Data Science, focusing on humanitarian and environmental projects.
- **Caspar Stordeur**: Expertise in data transformation and analysis, with a focus on demographics and community dynamics.

2.3 Communication and Project Ecosystem

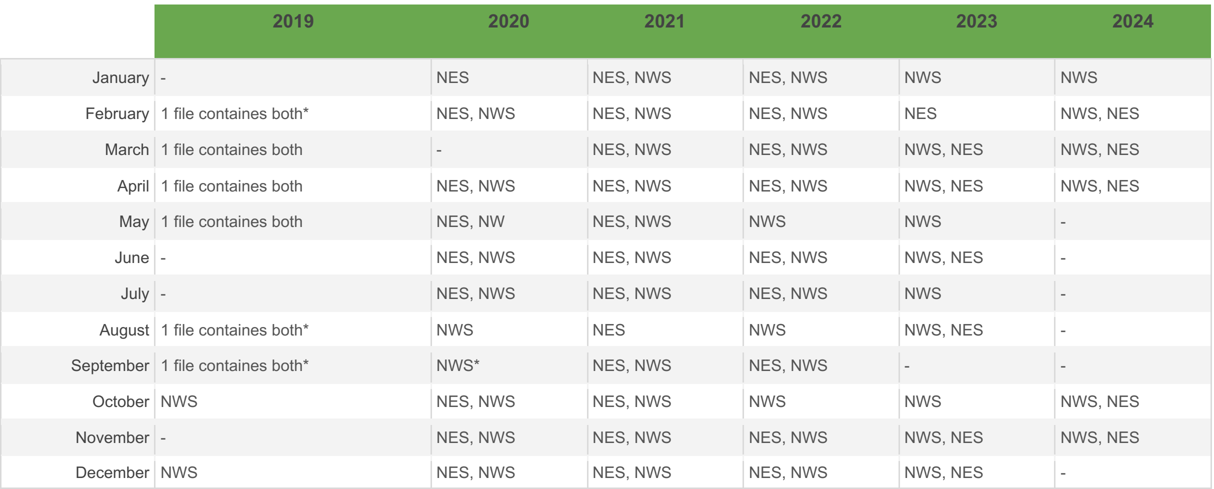
The team attempted to contact IMPACT REACH Initiatives and UNHCR for access to data and collaboration but faced challenges due to time constraints and logistical changes in Syria. The team is not aware of any similar projects and believes that this initiative could significantly improve the efficiency of humanitarian aid delivery.

3. Framework

3.1 Datasets and Volume

The primary dataset is the **Humanitarian Situation Overview in Syria (HSOS)**, which provides monthly data from 2019 to 2024 for **Northeast Syria (NES)** and **Northwest Syria (NWS)**.

This table shows the availability of data files:



	2019	2020	2021	2022	2023	2024
January	-	NES	NES, NWS	NES, NWS	NWS	NWS
February	1 file containes both*	NES, NWS	NES, NWS	NES, NWS	NES	NWS, NES
March	1 file containes both	-	NES, NWS	NES, NWS	NWS, NES	NWS, NES
April	1 file containes both	NES, NWS	NES, NWS	NES, NWS	NWS, NES	NWS, NES
May	1 file containes both	NES, NW	NES, NWS	NWS	NWS	-
June	-	NES, NWS	NES, NWS	NES, NWS	NWS, NES	-
July	-	NES, NWS	NES, NWS	NES, NWS	NWS	-
August	1 file containes both*	NWS	NES	NWS	NWS, NES	-
September	1 file containes both*	NWS*	NES, NWS	NES, NWS	-	-
October	NWS	NES, NWS	NES, NWS	NWS	NWS	NWS, NES
November	-	NES, NWS	NES, NWS	NES, NWS	NWS, NES	NWS, NES
December	NWS	NES, NWS	NES, NWS	NES, NWS	NWS, NES	-

An additional aspect is derived from the **Joint Market Monitoring Initiative (JMMI)** data, which focuses on the **Survival Minimum Expenditure Basket (SMEB)** across different regions of Syria. The analysis explores price trends, regional variations, and the relationships between various components of the SMEB. The JMMI was initiated by the Cash Working Group to better understand market dynamics and challenges in Syria. All time points for the regional subset are consolidated into a single file. For this analysis, we utilized the January 2024 files for **Northeast Syria (NES)** and **Northwest Syria (NWS)**. The final dataset comprises **21 indicators** and over **4,700 observations**.

Relevant indicator groups
We are interested in all variables that provide information about the 11 indicator groups that the HSOS surveyed.
The groups are:

1. Demographics
2. Shelter
3. Electricity & Non-food Items (NFIs)
4. Food Security
5. Livelihoods
6. Water, Sanitation, and Hygiene (WASH)
7. Health

8. Education
9. Protection
10. Accountability and Humanitarian Assistance
11. Priority Needs

Additionally, the **Joint Market Monitoring Initiative (JMMI)** dataset provides information on the Survival Minimum Expenditure Basket (SMEB) in different regions of Syria, focusing on price trends and regional variations.

3.2 Target Variable

Our main target variable candidate is **"Humanitarian assistance provided to IDP households in the community in the past 30 days."** This variable consists of 16 binary columns representing different types of assistance, such as Agricultural supplies, cash assistance vouchers or cash in hand, CP services, education, electricity assistance, explosive hazard risk awareness or explosive contamination removal, food and nutrition, gender-based violence (GBV) services, health, legal services, livelihood support, mental health, psychological support, non-food items (NFI), other, shelter, water sanitation and hygiene (WASH), and winterization. More details are shown in the visualization section.

3.3 Highlighted Features

The **HSOS dataset** captures key aspects of humanitarian assistance, community characteristics, and the challenges faced by IDP households. Notable features include detailed variables related to access to assistance (e.g., health, food, and shelter), priority needs, and barriers to receiving aid. Additionally, the dataset provides temporal context through month and year information, as well as spatial granularity with geographic identifiers such as **Governorate**, **District**, and **Community codes**. These features enable a comprehensive analysis across time, location, and types of assistance.

3.4 Data Limitations

The primary limitations of this project include the high proportion of **binary variables**, which can complicate data interpretation and increase dimensionality. Additionally, the dataset contains a significant amount of **missing values**, necessitating imputation or exclusion strategies that may introduce bias or reduce the overall dataset size. Another challenge is the presence of a large number of **categorical variables with low cardinality**, which require careful encoding to maintain meaningful relationships without inflating the feature space. Collectively, these factors pose challenges in achieving robust and generalizable insights from the data.

4. Pre-processing and Feature Engineering

4.1 Data Cleaning and Treatment

In total, we had 110 **HSOS** files. To ensure data consistency, we limited the dataset to the years **2021 to 2023**, resulting in **61 files**. Notably, there is no consistent catalog of questions across the files. These 61 files alone contained over **5,000 indicators**. To maintain consistency, we retained only those indicators that were present in 57 of the 61 files. Unfortunately, this process excluded indicators related to education, which we manually added back into our sample. The final analysis sample consists of **1,668 indicators** and **52,095 observations**.

For the **Market Data**, all time points for each region are consolidated into a single file. Similarly, to ensure consistency, we restricted the data to the years **2021 to 2023**. The files used for this analysis are the January 2024 files for Northeast Syria (NES) and Northwest Syria (NWS). The original dataset included 123 columns, featuring detailed prices for individual items. However, since the dataset also contained subtotals for these items, we decided to exclude the individual item prices and focus on the subtotals**,** totals, and categorical data. The final dataset contains **21 indicators** and over **4,700 observations**.

Additionally, approximately **four SMEB component columns** are planned to be added from the market dataset to the **HSOS dataset**.

For further analysis, we made an important decision regarding the HSOS data: to focus on the **IDP (Internally Displaced Persons) data** rather than the **resident population data**. This decision was made primarily to reduce data processing and preparation time. However, a fully functional classifier model should ideally consider both datasets. In communities where NGO aid is provided only to IDPs, tensions can arise with the resident population. As a result, NGOs have increasingly begun to treat IDPs and residents as a single community, expanding their needs assessments and aid delivery to include both groups proportionately. To streamline the dataset and focus on actionable insights, redundant or irrelevant columns, such as those ending with **"-Other"**, were removed.

Other pre-processing steps are summarized below:

Mapping and Encoding:

- **Binary responses** (e.g., Yes/No) were mapped to **1/0**.
- Unnecessary categories such as **"No Data"** or **"NA"** were replaced with **np.nan** to ensure usability.

- **Proportion and percentage ranges** were mapped using thresholds (e.g., 10%, 20%, 25%).

Categorical Variables:

- **Governorate** and **District** were preprocessed as follows:
    - **One-Hot Encoding** was applied to variables related to **Governorate** (6 unique values).
    - **One-Hot Encoding** was also applied to the **Month** variable to capture potential seasonal fluctuations in the data. This approach ensures the model can account for temporal variations without assuming a linear relationship between months.
    - **One-Hot Encoding** was implemented for the **Top Priority (Need)** variables to preserve their categorical nature. This allows the model to consider the unique importance of each need without imposing ordinal or linear relationships between them.
    - **Frequency Encoding** was used for variables related to **District** (23 unique values). This approach was chosen to avoid increasing the number of columns, which were already numerous.

4.2 Transformation, Normalization, and Standardization

Both the **HSOS** and **Market datasets** contain several price-related columns. These columns are planned to be normalized using **StandardScaler** prior to modeling. Numerical features with different scales, such as **expenditures** and **daily wages**, are normalized to ensure a fair comparison and to prevent differences in scale from disproportionately affecting the analysis. These transformations are critical to ensure that all features contribute appropriately to the analysis and that no single feature dominates due to its scale or coding. The second dataset includes **Survival Minimum Expenditure Basket (SMEB)** components across different regions of Syria.

Data were collected from the Northwest Syria (NWS) and Northeast Syria (NES) datasets, including median prices at the community level. Since each region uses different currencies, exchange rates (**TRY/USD** and **SYP/USD**) were included to standardize costs into a single currency (**USD**).

4.3 Dimension Reduction Techniques

Since working with too many columns reduces both the **speed** and **efficiency** of the model, we aim to reduce the number of columns in the next step. Although we have already removed unnecessary or redundant categorical columns, the dataset still contains nearly **1,000 numerical columns** after mapping.
For this reason, **dimension reduction techniques** are being considered, particularly given the large number of **binary**, **ordinal**, and **numeric features** in the dataset.

Nonlinear techniques such as **t-SNE** (t-Distributed Stochastic Neighbor Embedding) or **UMAP** (Uniform Manifold Approximation and Projection) are potential options.
These methods are effective at capturing complex, non-linear patterns in high-dimensional data, making them well-suited for feature visualization or preprocessing in our case.

5. Visualizations and Statistics

5.1 HSOS Dataset

5.1.1 Target Variable

The pie chart **(Figure 1)** visualizes the percentage distribution of general responses (as opposed to binary responses, where responses are text separated by commas) related to "Humanitarian assistance provided to IDP households in the community over the last 30 days."

The largest category, **Food and Nutrition (28.3%)**, highlights its prominence in humanitarian assistance. The category **NA - No humanitarian assistance reported (26.9%)** represents cases where no aid was reported, underscoring gaps in support. Other categories, such as **Health (15.8%)**, **WASH (10%)**, and **Cash assistance vouchers or cash in hand (8.4%)**, reflect varying priorities in assistance types. The **Other (10.5%)** category represents the sum of all categories below 5%.

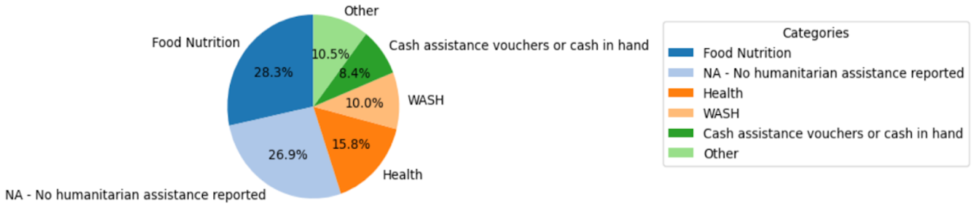


Figure 1: Percentage distribution for “Humanitarian assistance provided to IDP households in the community over the last 30 days”

Binary targets (Humanitarian assistance categories) showed no **multicollinearity**, confirming their independence **(Figure 2)**.

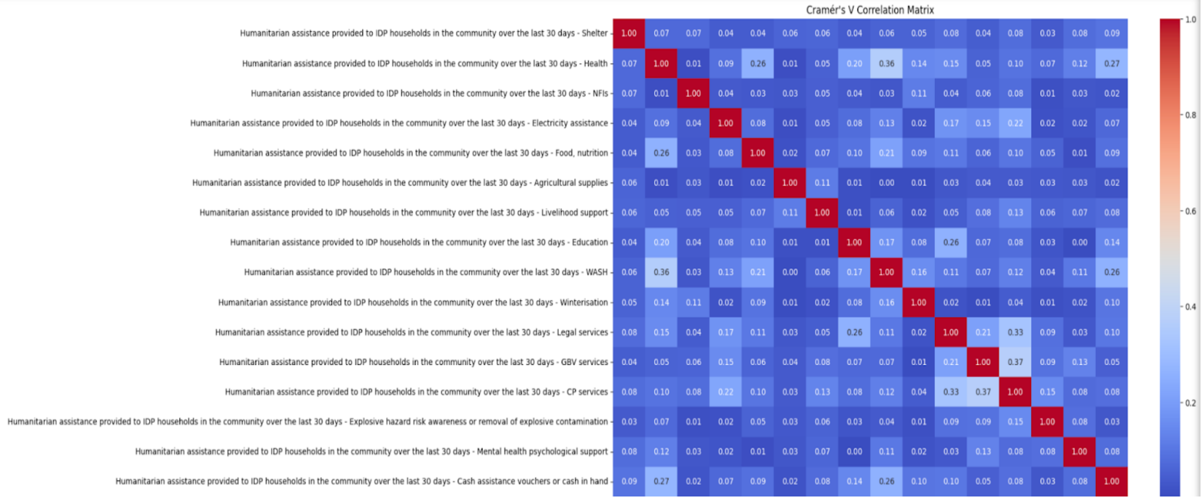


Figure 2: Camérs V Correlation Matrix - Target Variable

The stacked bar chart **(Figure 3)** illustrates the **monthly percentage distribution** of the types of humanitarian assistance provided to IDP households over the past 30 days. Binary target variables were grouped by month, and percentages were calculated to normalize the distribution for each month. The graph shows a consistent dominance of **food and nutrition assistance** across most months, while other categories, such as **health** and **WASH**, exhibit notable fluctuations. For example, WASH assistance spikes in certain months, likely due to seasonal needs or external factors such as infrastructure challenges. These trends highlight the **dynamic allocation of resources** and the potential impact of contextual factors on aid prioritization.

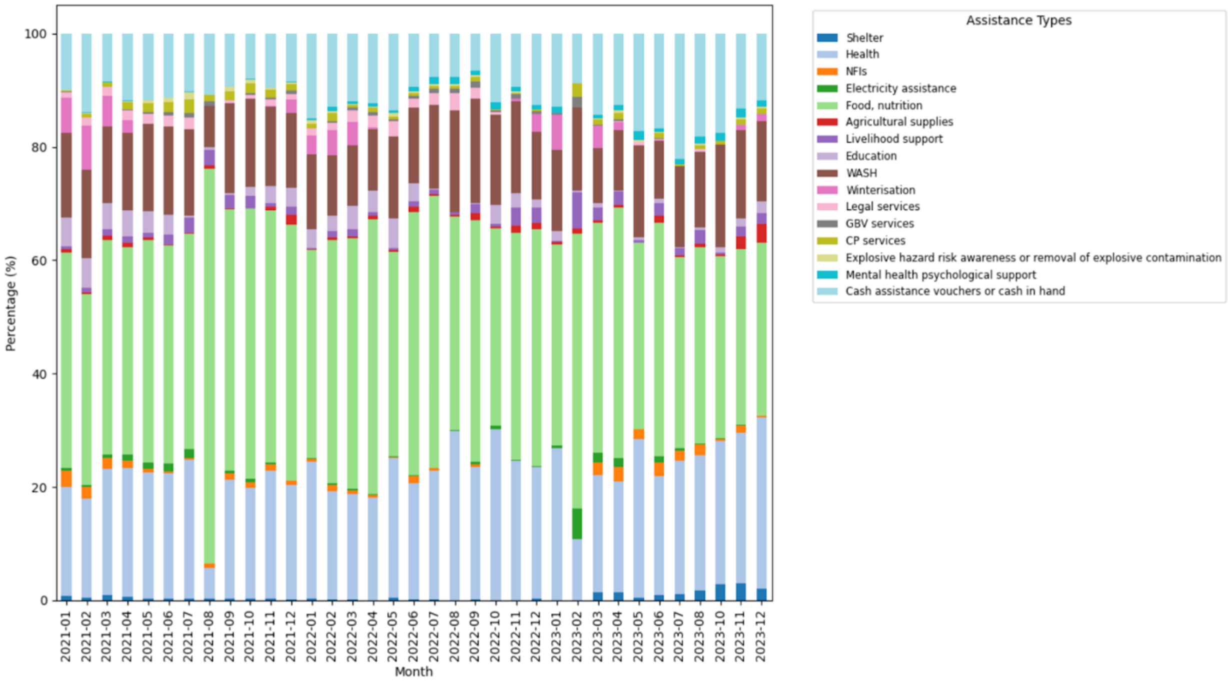


Figure 3: Monthly percentage distribution of types of humanitarian assistance provided to IDP households over the past 30 days

Time-series visualizations **(Figure 4)** revealed trends in support across months, providing insights into **seasonality** and **variation**. As a result, it was decided to encode the months using **One-Hot Encoding**and include them as features in the model.

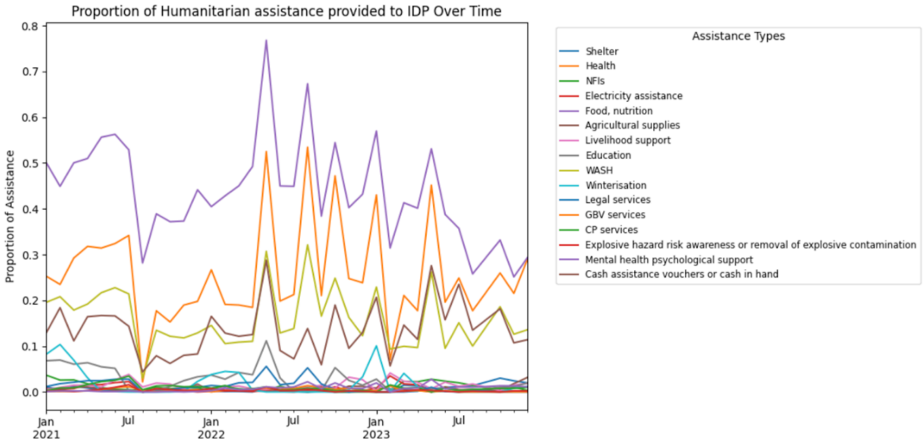
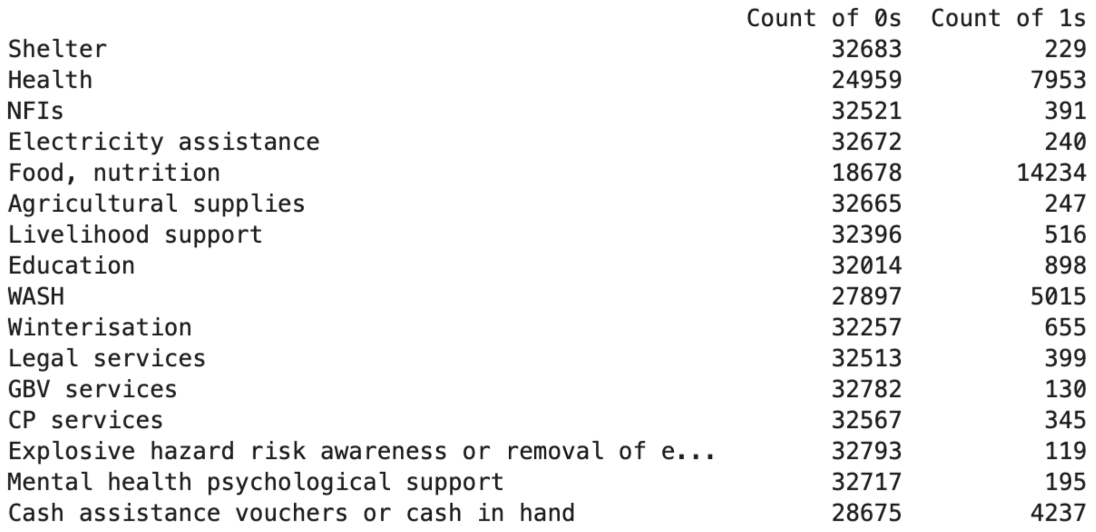


Figure 4: Proportion of Humanitarian assistance provided to IDP over time

The **Augmented Dickey-Fuller test** results, indicate **mixed stationarity** across variables. While some variables, such as health and cash assistance vouchers or cash on hand, show stationarity with p-values well below the significance level, others, such as shelter and awareness of explosive hazards, fail to reject the null hypothesis, suggesting **non-stationarity**.

Given the nature of the HSOS data, conducting time-series analyses over this short period is complex and was not pursued. Since our focus is on **classification** rather than **prediction**, the inclusion of **month** as a feature is considered sufficient to capture potential temporal patterns.

The **value counts table** clearly demonstrates that the dataset is **unbalanced**, with the majority of assistance categories dominated by **"0s"** (indicating no assistance provided). This imbalance highlights the potential challenge of **biased learning** in machine learning models. Models may become overly sensitive to the **majority class (0s)** and fail to accurately predict the **minority class (1s)**, which is often critical in humanitarian analysis. Algorithms that handle class imbalance well, such as **gradient boosting** or **random forests**, can be tuned to account for class weights. Additionally, applying **oversampling techniques** (e.g., **SMOTE**) to increase the representation of the minority class may be a viable solution. Finally, adjusting the **classification threshold** to optimize **recall** for the minority class can ensure that the model prioritizes the detection of critical instances where assistance is provided.

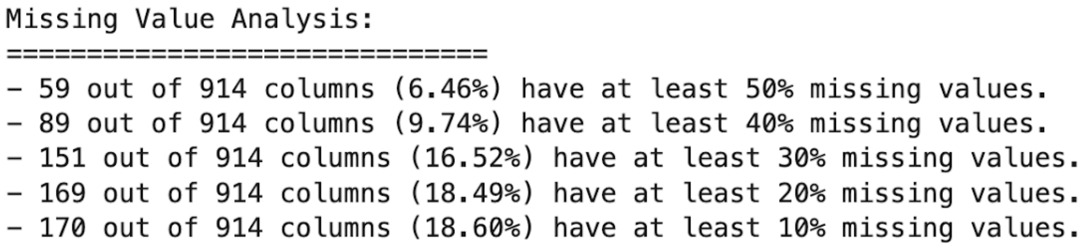


```
                                                    Count of 0s  Count of 1s
Shelter                                                   32683          229
Health                                                    24959         7953
NFIs                                                      32521          391
Electricity assistance                                    32672          240
Food, nutrition                                           18678        14234
Agricultural supplies                                     32665          247
Livelihood support                                        32396          516
Education                                                 32014          898
WASH                                                      27897         5015
Winterisation                                             32257          655
Legal services                                            32513          399
GBV services                                              32782          130
CP services                                               32567          345
Explosive hazard risk awareness or removal of e...        32793          119
Mental health psychological support                       32717          195
Cash assistance vouchers or cash in hand                  28675         4237
```

The **target matrix** contained **53 missing values** across each category. Given the small number of missing entries, these rows were **excluded** from the dataset rather than being imputed, to avoid introducing bias.

5.1.2 Feature Variables

Missing value analysis reveals a significant proportion of missing data in the dataset, necessitating careful consideration of feature inclusion during modeling. For **feature selection**, an appropriate threshold for missing values should be determined to strike a balance between retaining valuable information and avoiding the noise introduced by sparse data. A preliminary threshold of **20% missing values** appears reasonable, as it retains most of the dataset while excluding the most incomplete columns. However, this threshold can be revisited based on the specific requirements of the model and its sensitivity to missing data. It is recommended to apply the feature selection process **after train-test splitting** to avoid data loss. This approach ensures that the missing value threshold is applied independently to both the training and test sets, enhancing the model's robustness to unseen data.



```
Missing Value Analysis:
==============================
- 59 out of 914 columns (6.46%) have at least 50% missing values.
- 89 out of 914 columns (9.74%) have at least 40% missing values.
- 151 out of 914 columns (16.52%) have at least 30% missing values.
- 169 out of 914 columns (18.49%) have at least 20% missing values.
- 170 out of 914 columns (18.60%) have at least 10% missing values.
```

5.1.3 Demographic Features

The top pie chart **(Figure 5)** illustrates the percentage distribution of the assessed population across governorates. **Idleb** and **Aleppo** dominate, with similar proportions (**31.3%** and **30.8%**, respectively), while **Deir-ez-Zor** and other smaller regions contribute minimally to the dataset.
The second pie chart **(Figure 6)** shows the distribution across districts, with a significant proportion (**36.2%**) categorized as **"Other"** (the sum of categories below 5%), followed by **Ar-Raqqa (10.9%)**and **Harim (9.7%)**. This highlights the **geographic spread** and variability in representation at the district level, which includes **23 unique values**.

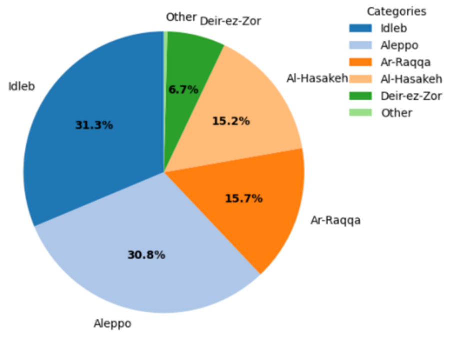


Figure 5: Percentage distribution across governorates

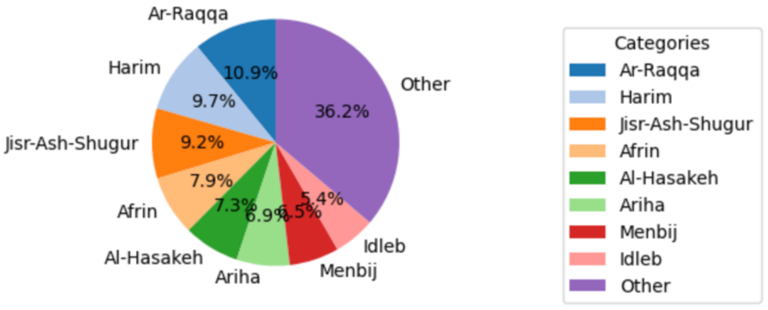


Figure 6: Percentage distribution across districts

5.1.4 Important Features

To analyze **priority needs**, three questions regarding the most critical needs (**Top 1, Top 2, Top 3**) were aggregated with weighted values of **3, 2, and 1**, respectively. The data was then reshaped using a **melt operation** and aggregated on a monthly basis to observe trends over time. From the resulting dataset, the **top five priority needs—Food, Health, Livelihoods, Shelter, and WASH**—were identified and plotted as a time series.

The **graph (Figure 7)** illustrates the **weighted monthly percentage distribution** of the top five needs over time. **Food** consistently appears as the dominant priority, followed by **Livelihoods** and **Shelter**, indicating that these are the most critical areas of need. **Health** and **WASH** display significant but smaller shares, with varying levels of importance depending on the month. The identified priority needs are expected to have a strong relationship with the **target variable**, as addressing these needs directly correlates with improving humanitarian outcomes for the IDP population.

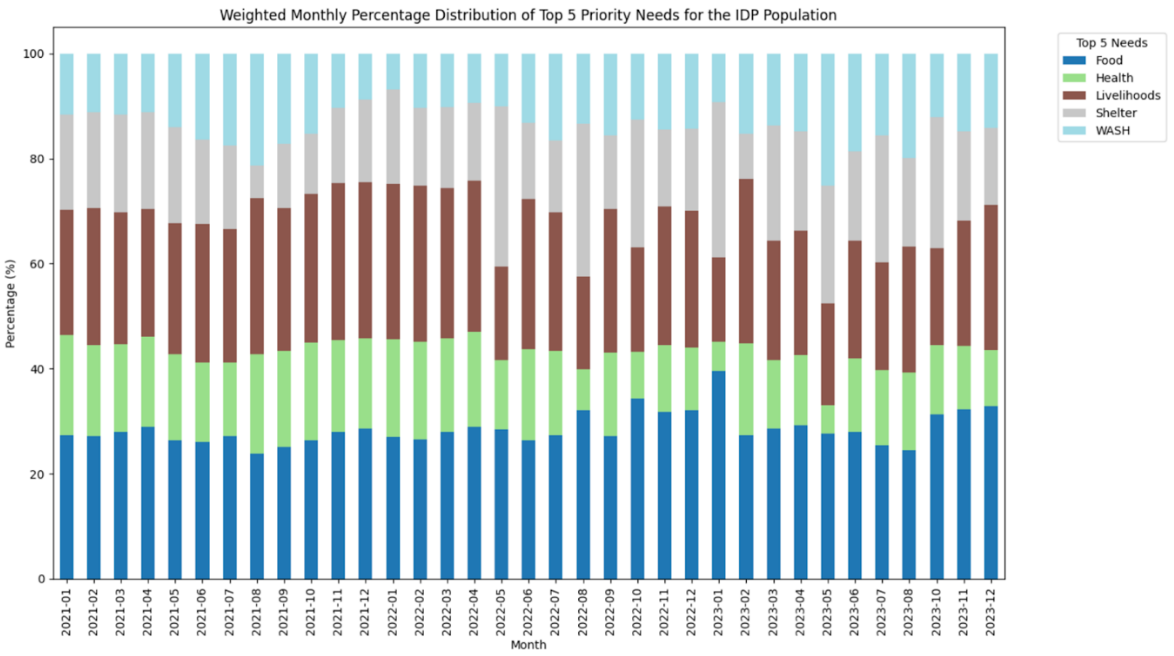


Figure 7: Weighted monthly percentage distribution of the top 5 Priority Needs for the IDP Population?

The pie chart **(Figure 8)** illustrates the percentage distribution of **IDP households' access to humanitarian aid** within the last 30 days. It shows that **59% of households** reported receiving aid. This distribution suggests that **access to humanitarian aid** may serve as a strong explanatory variable for understanding the provision of various types of assistance to IDP households in the community.

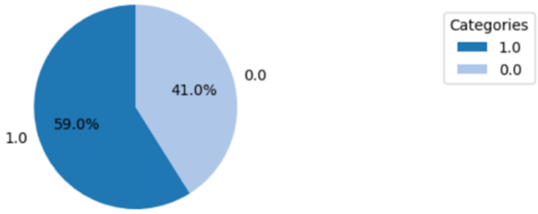


Figure 8: Percentage Distribution - Were IDP households in the community able to access humanitarian aid in the last 30 days?

The following graph **(Figure 9)** highlights the main challenges faced by IDPs in accessing humanitarian assistance.

The **"NA - No humanitarian assistance reported"** category (**22.6%**) indicates cases where no assistance was provided, and thus no challenges were reported.

Other challenges include **insufficient quantity (20.6%)**, **types of assistance not relevant (15.8%)**, and **not meeting eligibility criteria (8.1%)**. These factors could significantly impact the **distribution** and **accessibility of aid**, and may be related to how well aid is tailored to the specific needs of the community.

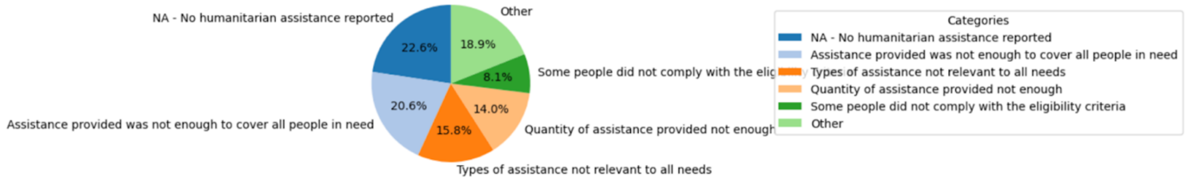


Figure 9: Percentage Distribution - Most important missing information identified by households?

The chart below **(Figure 10)** shows the percentage distribution of whether people in the location receive information about **humanitarian aid** and the **humanitarian situation**. The responses are almost evenly split, with **50.5%** indicating **"yes" (1.0)** and **49.5%** indicating **"no" (0.0)**. This balance suggests that a significant portion of the population lacks access to critical information about available assistance, which could hinder their ability to effectively seek and receive aid. There may be a potential relationship with the **target variable**, as individuals who are better informed may be more likely to access assistance.

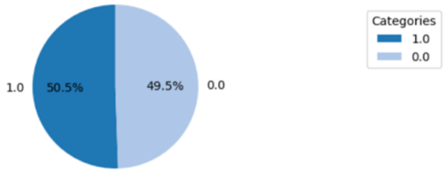


Figure 10: Percentage Distribution - Do people in the location receive information about humanitarian assistance and the humanitarian situation?

The graph below **(Figure 11)** shows the percentage distribution of the **most important missing information** identified by households. The largest category, **"Other"** (the sum of categories below 5%), accounts for **35.4%**, followed by **"How to find work" (18.7%)** and **"How to get more money/financial support" (18.6%)**. The relationship with the **target variable** may stem from the fact that households lacking critical information—particularly on how to register for assistance or obtain financial support—may face barriers in accessing needed humanitarian assistance.

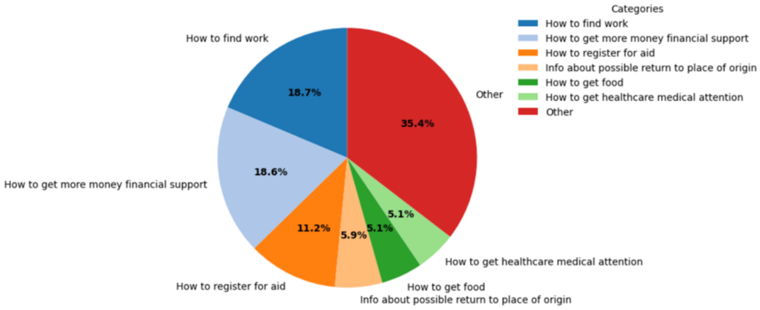


Figure 11: Percentage Distribution - Most important missing information identified by households?

The pie chart **(Figure 12)** illustrates the preferred methods of providing information to households in the location. **"Personal face-to-face"** is the most preferred method at **32.4%**, followed by "WhatsApp or other mobile phone-based platforms**"** at **28.5%**. This distribution highlights the importance of **direct and personal communication channels** in humanitarian efforts. The **target variable** may be influenced by the effectiveness of these communication methods in disseminating critical information, potentially impacting access to aid.

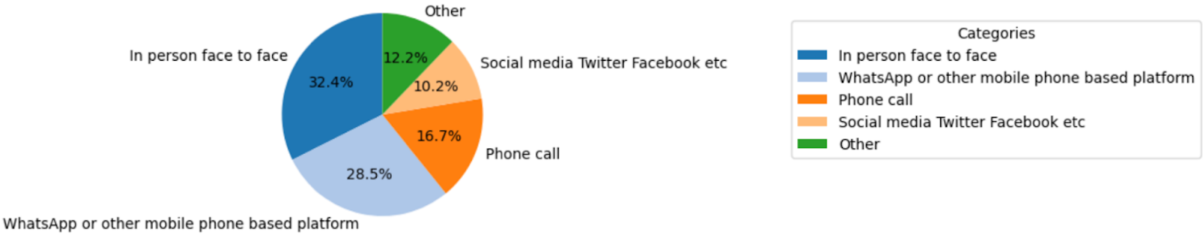


Figure 12: Percentage Distribution - Are households in the location aware of a humanitarian assistance feedback of complaints mechanism?

(Figure 13) shows that **62.8% of households** are not aware of a **feedback or complaint mechanism** for humanitarian assistance, while **37.2%** are aware. This lack of awareness can hinder **effective communication** and **problem-solving**, potentially affecting access to assistance.

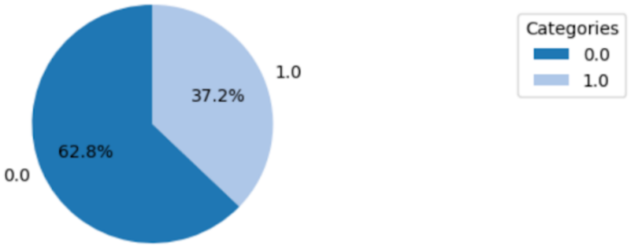


Figure 13: Percentage Distribution - Are households in the location aware of a humanitarian assistance feedback of complaints mechanism?

(Figure 14) shows the **monthly average daily wages** of unskilled workers in **Syrian pounds**, one of the few continuous variables in the HSOS dataset. Unlike other categorical variables, wages provide **quantitative insights** into economic conditions and their fluctuations over time. The significant increases and fluctuations observed may reflect **inflation**, **economic instability**, or changes in the demand for unskilled labor. Given its nature, this variable is hypothesized to have a **significant**

relationship with the **target variable**, as economic stability often influences both access to and the need for humanitarian assistance.

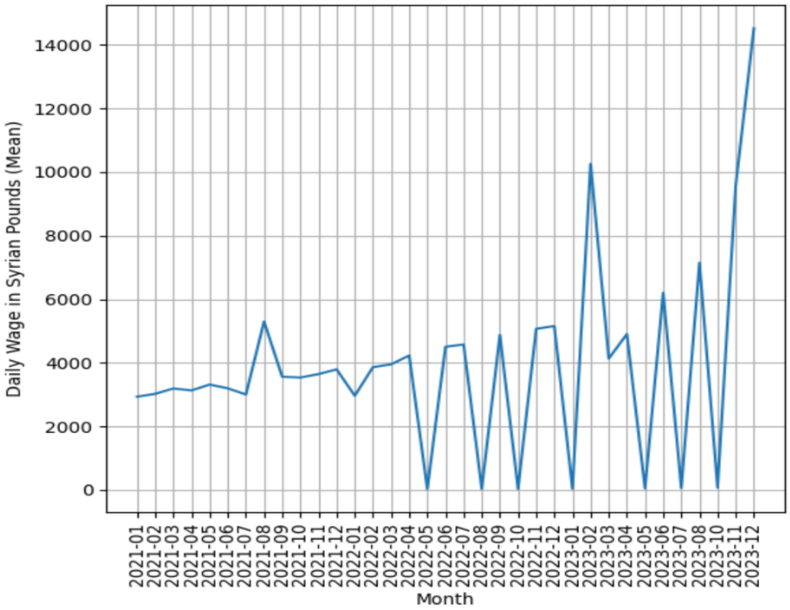
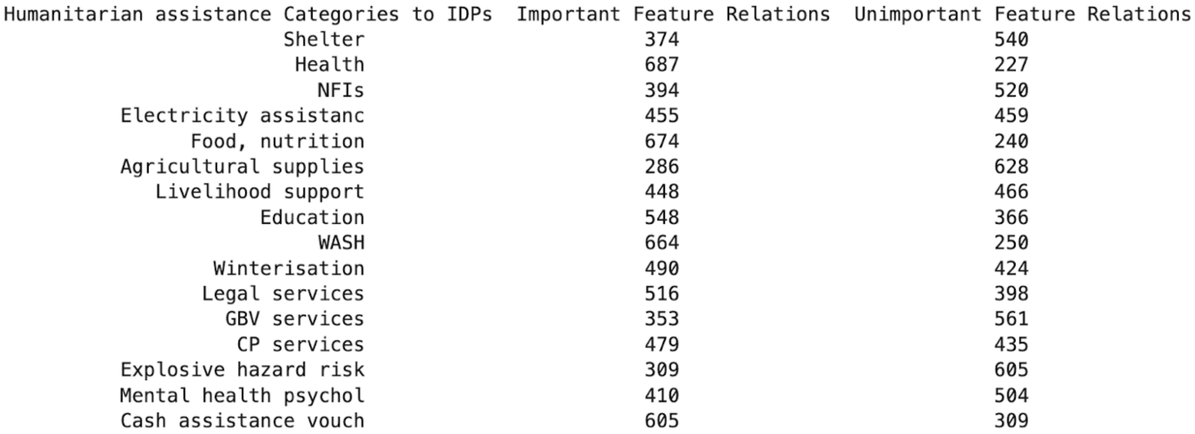


Figure 14: monthly average daily wages of unskilled Labour (IDPs)

5.1.5 Relationship between target variable and features

Cramer's chi-square test was performed to account for **non-linear relationships**. The number of **significant features** (p-value < 0.05) and **insignificant features**, based on the test statistics, are shown in the following table:


```
 Humanitarian assistance Categories to IDPs  Important Feature Relations  Unimportant Feature Relations
                         Shelter                        374                           540
                          Health                        687                           227
                            NFIs                        394                           520
           Electricity assistanc                        455                           459
                 Food, nutrition                        674                           240
           Agricultural supplies                        286                           628
              Livelihood support                        448                           466
                       Education                        548                           366
                            WASH                        664                           250
                    Winterisation                       490                           424
                  Legal services                        516                           398
                    GBV services                        353                           561
                     CP services                        479                           435
           Explosive hazard risk                        309                           605
           Mental health psychol                        410                           504
           Cash assistance vouch                        605                           309
```

Analysis of the characteristics of the **humanitarian assistance categories** provided to IDPs reveals varying degrees of **importance** and **irrelevance** in terms of their **predictive power**.

For example, categories such as **Shelter** and **Food/Nutrition** show a higher number of important feature relationships, suggesting that these areas have **strong associations** with other variables in the dataset. In contrast, categories such as **Agricultural Supplies** and **Explosive Hazard Risk Awareness** have a relatively higher proportion of unimportant feature relationships, indicating **limited or indirect connections** with other explanatory variables.

The conclusion drawn from the **five most important** and **five least important** relationships for each target variable can be summarized as follows

The pattern underscores the **complexity of humanitarian data**, where some categories are more robustly related to **contextual variables** (e.g., economic conditions or shelter dynamics), while others are likely more **situational** or **region-specific**.

Chi-square analysis further supports these findings, as high chi-square values in some categories indicate **statistically significant relationships** between variables, while others show negligible influence. This duality highlights the importance of **targeted modeling approaches**. For instance, reducing noise from irrelevant features for less connected categories and emphasizing key variables for categories with stronger associations is likely to improve model performance.

5.2 Market Dataset

The Joint Market Monitoring Initiative **(JMMI)** data focuses on the Survival Minimum Expenditure Basket **(SMEB)** across different regions in Syria. The analysis examines price trends**,** regional variations, and relationships between SMEB components.

The JMMI was developed by the Cash Working Group **(CWG)** to understand **market functionality** and **challenges** in Syria. It monitors prices and stock levels of basic commodities monthly to inform cash-based responses and **food assistance programs**.

The analysis focuses on **SMEB components** selected by the CWG, which represent the **minimum culturally-adjusted items** needed by an average six-person household in Syria per month. Data collection spans multiple **governorates**, with regular monitoring of selected retailers in major markets.

In this section, the **SMEB cost (total)** and its **components—food**, **non-food items**, **cooking fuels**, **water**, and **mobile data**—are analyzed.
To avoid **multicollinearity**, only the **components** are planned to be added to the model.

(Figure 15) already shows that **food** dominates the **SMEB calculations**. Over **80%** of the component analysis is food, followed by non-food items**,** cooking fuels**,** and water.

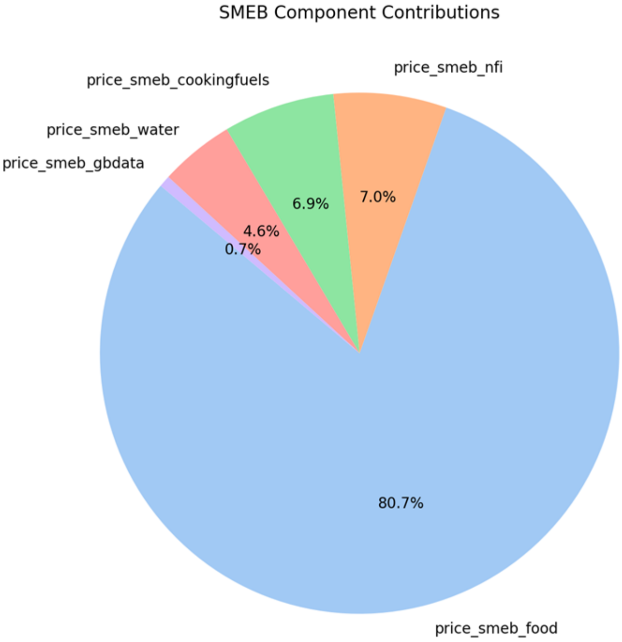


Figure 15: Overall Distribution of SMEB Components

The graph (Figure 16) illustrates **SMEB cost** and **component trends** over time for the **Northeast** and **Northwest** regions.

The **time series analysis** shows a similar pattern. Food has the largest share and a rising trend until mid-2023, when food prices drop abruptly. The other categories show a smoother trend with less fluctuation. The exchange rate used here for Syrian Pound/USD is not a floating rate but a fixed one.

The sharp drop in **July 2023**, with an **80% devaluation**, is the primary reason for this significant change. While the **Northeast** shows significant **volatility** with a sharp decline in mid-2023 due to **currency devaluation**, the **Northwest** exhibits more **stability**, with SMEB costs remaining relatively consistent over the observed period.

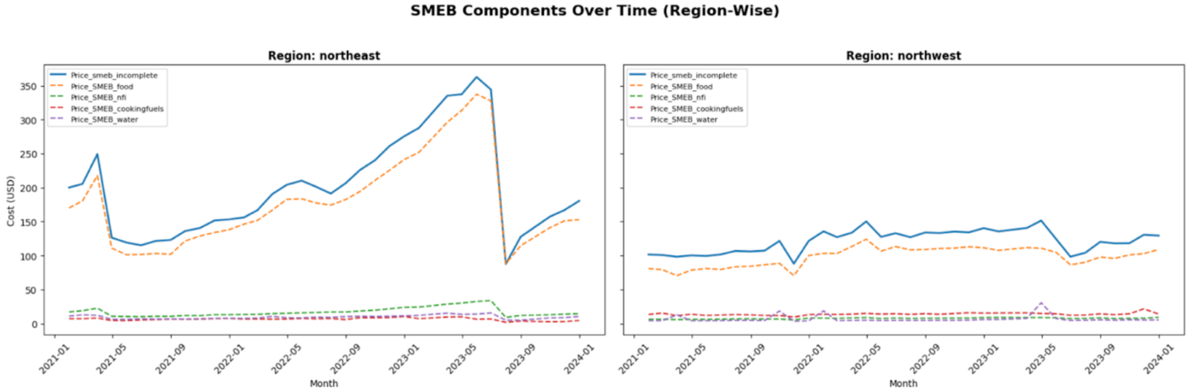


Figure 16: Overall Distribution of SMEB Components

Exact figures can be found in the table below:

Component	Mean	Standard Deviation
Food	128.46	57.53
NFI	11.19	6.54
Cookingfuels	11.04	9.15
Water	7.36	28.34
Mobile Data	1.14	1.83

The **correlation matrix (Figure 17)** shows that only **food** and **non-food items** have a **high positive correlation**. This could indicate **multicollinearity**, which is a problem for our future **model estimation**. **VIF results** also indicated **multicollinearity**.

A possible solution during the **modeling step** could be to exclude the **NFI values**. In **high-dimensional datasets**, multicollinearity can lead to **overfitting**, where the model captures **redundant information** from correlated features.

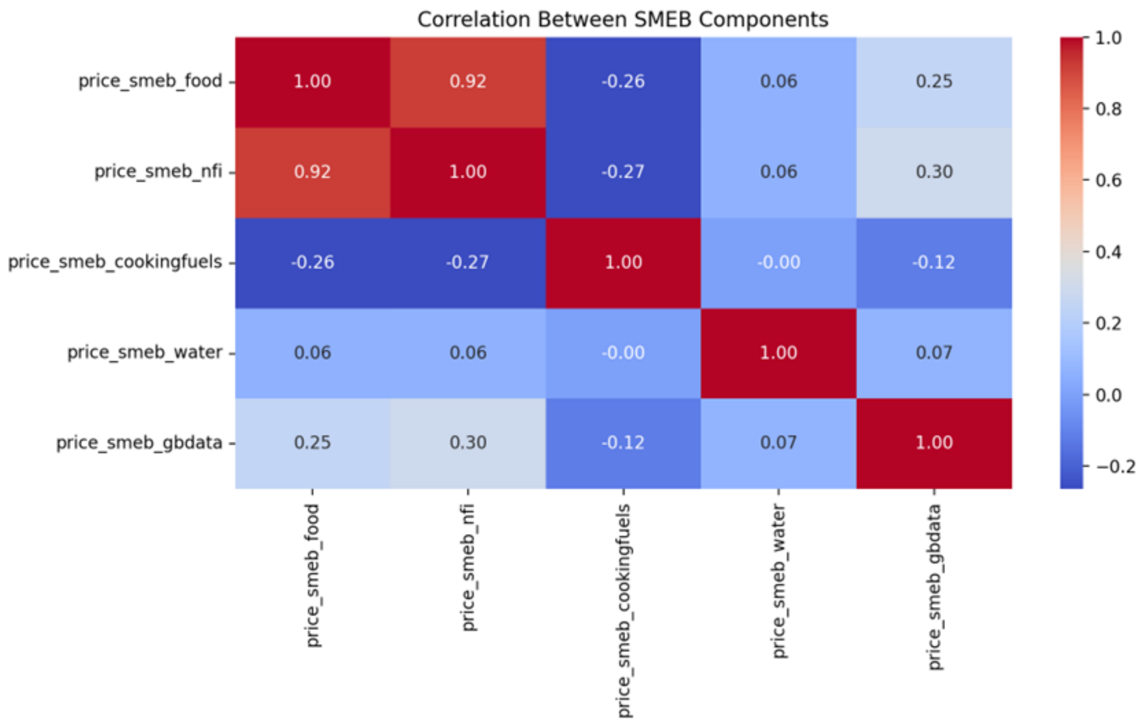


Figure 17: Correlation Matrix of SMEB Components

Key Insights:

- **High Multicollinearity:**
  - Variables such as **price_smeb_food**, **price_smeb_nfi** exhibit **high VIF values**, indicating significant **multicollinearity**.
  - This suggests potential **redundancy** in these variables for **regression models**.
- **Strong Correlations:**
  - **price_smeb_food** and **price_smeb_nfi** have a **correlation of 0.92**.

7. Further Steps:

1. **Merge Datasets:**
   - Combine the **market dataset** and the **HSOS dataset**, most likely at the **sub-county level**.
2. **Target Value Layout:**
   - **Merge coded variables** into a **0-1 coding** (e.g., aid provided vs. aid not provided).
   - Create a **Matrix Multi-Classification Problem**.
   - Build **16 separate models** with their corresponding features.
3. **Additional Steps:**
   - **Normalize numerical features**.
   - Find a method to **handle missing data**.
   - Address **class imbalance** in the target variable using techniques such as **Synthetic Minority Over-sampling Technique (SMOTE)**.
   - **Recursively eliminate features** to reduce dimensionality.