# Term Project Proposal

## 1. Dataset

\* Motivation for Dataset Selection

: **Food waste** has become a **critical issue** in achieving the Sustainable Development Goals (SDGs) in today’s world. This project aims to systematically **analyze food waste** through data, and from this analysis, **propose efficient operational strategies and waste reduction plans**.

**Furthermore**, it explores the potential to implement an incentive system **based on the amount of food waste generated**, enabling restaurants or operational units to receive rewards according to their waste reduction performance, thus providing a practical and scalable management approach.  
  
The dataset has the following characteristics:  
- Includes categorical features (`staff\_experience`, `waste\_category`) and numerical features (`meals\_served`, `temperature\_C`, etc.)  
- Contains some missing values and wrong data

- Total number of samples: 911

* Main Columns:

-> Numerical Features:

- day\_of\_week : Numeric representation of the day

- special\_event : Whether a special event was held  
 - meals\_served: Number of meals served on that day.  
 - kitchen\_staff: Number of kitchen staff working that day  
 - temperature\_C: Kitchen temperature (Celsius)  
 - humidity\_percent: kitchen or cafeteria humidity  
 - past\_waste\_kg: Record amount of food waste generated(kg) in past

-> Categorical Features:

- date : The date of food waste measurement

- staff\_experience: Average years of experience among kitchen staff

- waste category: Type of waste( food, packaging)

## 2. Project Description

**:** To gain a deeper understanding of the causes of food waste, this project performs feature-grouped clusteringand per-cluster regression analysisinstead of applying a single model to the whole dataset.

1. Grouping Features by Their Roles

- Environmental Factors : temperature\_c, humidity\_percent

- Operational Factors : meals\_served, special\_event, day\_of\_week

- Human Resource Factors : kitchen\_staff, staff\_experience

Each group reflects a different type of influence on food waste (e.g., spoilage vs operational inefficiency).

2. Score Calculation and Normalization

Within each group, features are first standardized. Then, a regression model is used to calculate a score that represents the group's overall impact for each record (e.g., high temperature and humidity → high environmental score). After that, all scores are normalized again to maintain a consistent scale across groups.

3. Clustering Based on Group Scores

Using the three scores [environmental\_score, operational\_score, human\_score], clustering is applied to identify records that share similar patterns of influence.

For example:

- Cluster 0 : High environmental influence

- Cluster 1 : High operational influence

- Cluster 2 : High human resource influence

This step allows us to separate records by the likely cause of food waste.

4. Regression Analysis Within Each Cluster

For each cluster, a separate regression model is trained to predict food\_waste\_kgbased on the original features.  
This allows us to:

- Analyze which features are most important within each cause-based cluster

- Compare model performance across clusters

- Understand whether specific causes (e.g., high humidity vs understaffing) lead to predictable patterns of food waste

5. Evaluation

- Cross-validation for regression models:  
To ensure that the regression models built for each cluster generalize well and are not overfitted to the training data, we apply k-fold cross-validation.  
This helps evaluate the model's performance (e.g., using RMSE or R²) across multiple splits of the data, improving the robustness and reliability of the results.

## 3. Statistical Description of the Dataset

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Column Name | Mean | Standard Deviation | Minimum | Maximum |
| meals\_served | 375.41 | 502.81 | 100.00 | 4730.00 |
| temperature\_C | 22.19 | 8.92 | -10.37 | 60.00 |
| humidity\_percent | 60.76 | 17.33 | 30.12 | 89.98 |
| food\_waste\_kg | 44.84 | 27.93 | 10.82 | 274.33 |

- Missing Values : temperature\_C, staff\_experience, past\_waste\_kg (missing randomly (~8% of data))

-Wrong data : temperature\_C(too high (60°C) or too low (-10°C)), meals\_served(10x higher than the normal range)

\* We held a meeting on April 24th to set up the data set and direction of the project.

