Modeling and Factor Analysis for Food Wastes

Group 1

Contents

01 PROJECT OVERVIEW

02 DATASET INSPECTION

03 DATA PREPARATION

04 ANALYSIS

05 EVALUATION

06 TEAM ROLES & LESSONS LEARNED

Project Overview

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.

Expected benefits by Data Analysis

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.

Image Source: https://www.hobsonsbay.vic.gov.au/Services/Waste-and-recycling/How-to-use-your-four-bins/Food-Garden-Light-Green-Bin/How-Much-Food-Do-You-Think-You-Throw-Away-Every-Year

Project Overview

End to End process 7

Object setting

- To solve the problem of food waste occurring in restaurants...etc.
- Predict the amount of food waste using actual operation log data and contribute to establishing a reduction strategy through this.

Data Inspection

- Data exploration using the same plots as Boxplot, Histogram, and Scatter.
- Suitability check - In predicting waste volume, various factors are logically related.

Data Preparation

- Data Value Changes cleaning dirty data, text preprocessing, discretization, data normalization, encoding
- Feature engineering - selection, creation

Data Analysis

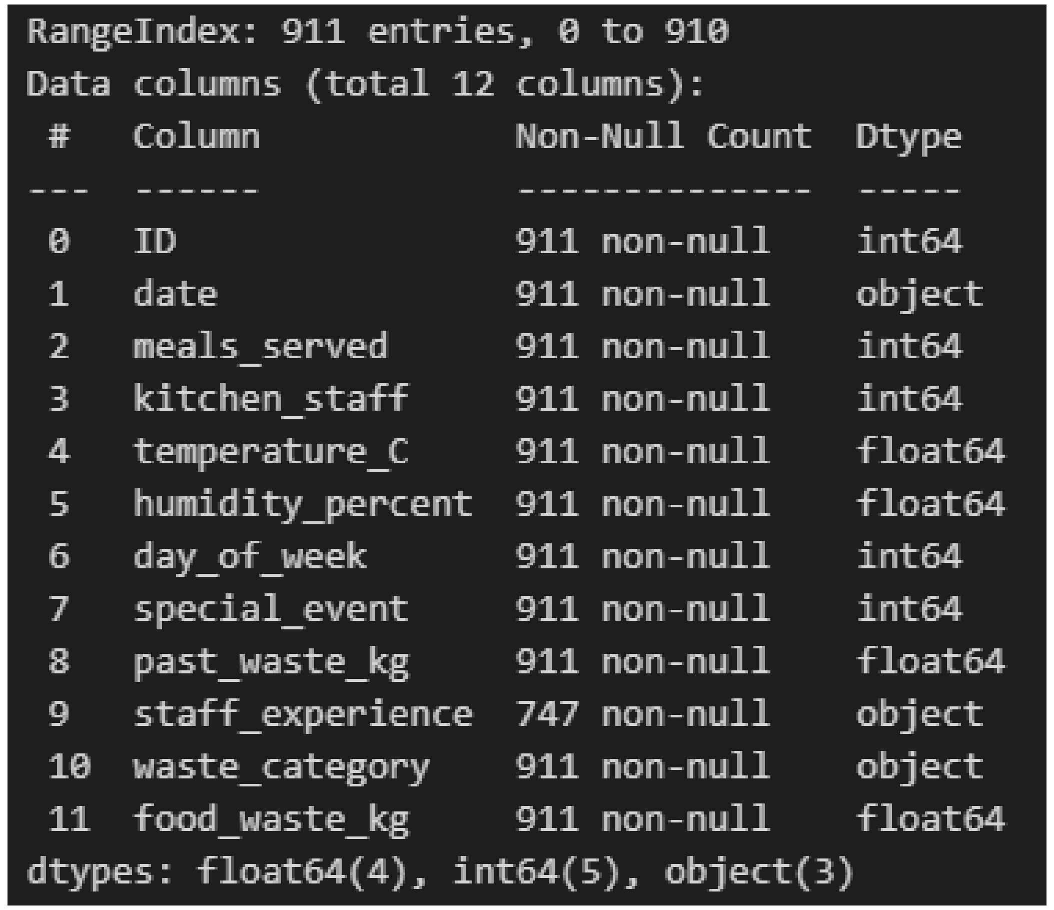
- Clustering - Kmeans : Deriving factor-based clusters
- Regression - Interpretation by applying Linear Regression and LGBM Regression respectively

Evaluation

- K-Fold Cross Validation : Evaluate model performance more precisely for each cluster

Dataset Inspection

- Contains some missing values and wrong data
- Total number of samples: 911



```
RangeIndex: 911 entries, 0 to 910
Data columns (total 12 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   ID                911 non-null    int64
 1   date              911 non-null    object
 2   meals_served      911 non-null    int64
 3   kitchen_staff     911 non-null    int64
 4   temperature_C     911 non-null    float64
 5   humidity_percent  911 non-null    float64
 6   day_of_week       911 non-null    int64
 7   special_event     911 non-null    int64
 8   past_waste_kg     911 non-null    float64
 9   staff_experience  747 non-null    object
 10  waste_category    911 non-null    object
 11  food_waste_kg     911 non-null    float64
dtypes: float64(4), int64(5), object(3)
```

Main Columns Descriptions

- **Numerical Features**

1. day_of_week : Numeric representation of the day
2. special_event : Whether a special event was held or not
3. meals_served: Number of meals served on the day
4. kitchen_staff: Number of kitchen staff working that day
5. temperature_C: Kitchen temperature (Celsius)
6. humidity_percent: Humidity of kitchen or cafeteria (%)
7. past_waste_kg: Average of wasted food generated in the past (Kg)
8. food_waste_kg: Amount of wasted food measured for a day (Kg)

- **Categorical Features**

1. date: The date of food waste measurement
2. staff_experience: Experience levels among kitchen staff (Beginner, Intermediate, Expert)
3. waste_category: Type of waste (Dairy, Meat, Vegetables, Grains)

Data Preprocessing

- Categorical data (‘staff_experience’, ‘waste_category)
  ：Converted to guilt and removed the replacement.

- Outlier - ‘temperature_C’, ‘meals_served’
- Missing Value - ‘temperature_C’, ‘paste_waste_kg’, ‘staff_experience’
- Encoding - ‘staff_experience’, ‘waste_category’

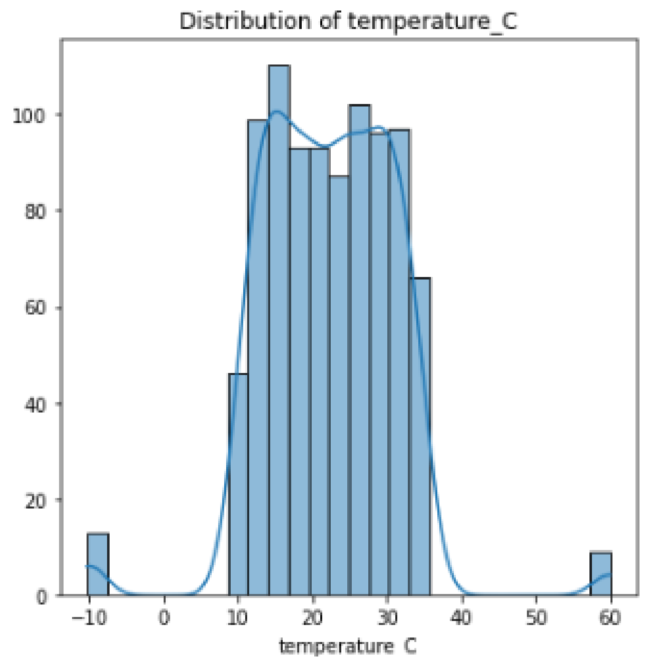
Data Preprocessing

Outliers

1. temperature_C

: Some temperature_C values are way too high (60°C) or too low (-10°C)

: In the "Temperature_C" column, calculate the most frequent value
from the values greater than 0 and less than 36
(To obtain robust central tendency measurements for temperatures within a normal range)
→ Identify values less than or equal to 0 or greater than or equal to 36 as outliers.
→ These outliers are replaced with the previously computed mode_temp.

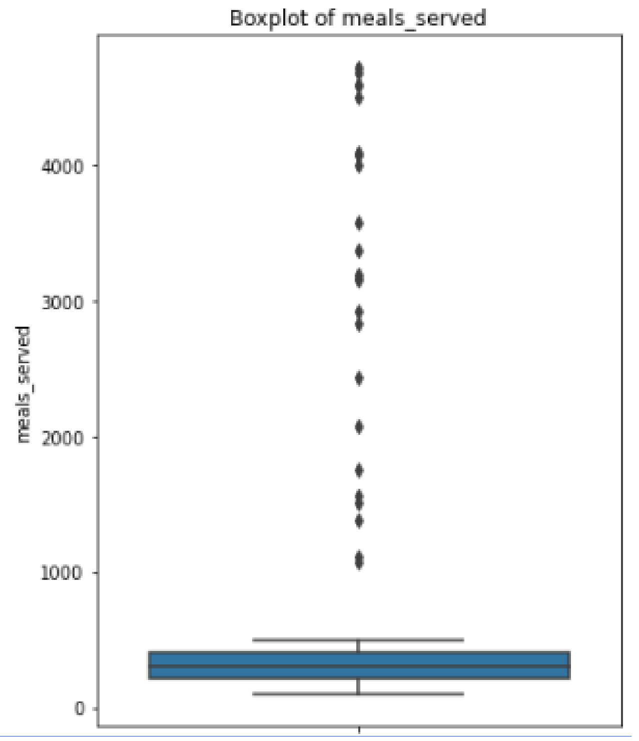


2. meals_served

: Remove all rows more than 500,

which are extremely high (about 10 times higher than the normal range)

: Rows exceeding 500 are a minority.
The specific event data was 0
but the values were extremely high and out of the normal range,
those rows were removed.



Data Preprocessing

Missing Value

1. temperature_C

: Any missing values (NaN) remaining in the column are filled with mode_temp.

: The mode was selected because it is sensitive to outliers and is less affected by outliers.

2. paste_waste_kg

: Any missing values (NaN) remaining in the column are filled with mean.

: The average was selected because it was normally distributed and reflected the overall trend well.

3. staff_experienced

: Considering the correlation between kitchen_staff and past_waste_kg.

: If there are few kitchen staff(<12) and a lot of waste(>20), it is processed as beginner,
if there are many kitchen staff(>17) and little waste(<15), it is processed as expert,
and the rests are processed as intermediate.

Data Preprocessing

Encoding

1. staff_experience

  ∶ begginer 0, intermediate 1, expert 2

  ∶ Since staff_experience represents an ordinal variable with a meaningful order of experience levels, and the subsequent regression model interprets numeric values as having quantitative significance, we manually encoded the categories as beginner 0, intermediate 1, and expert 2.

2. waste_category

  ∶ Label encoding was done in alphabetical order.

Feature Engineering

Feature creation

- By standardizing and summarizing multiple features of each group (environment, operation, human), we created a meaning-based representative value (score).
- By performing clustering and regression modeling based on this score, we improved the predictive power and interpretability.
- Normalization within factors by MinMaxScaler and calculation(mean) of factor scores
  → Renormalize factor scores

env_features

temperature_C
humidity_percent

ops_features

meals_served
special_event
day_of_week

human_features

kitchen_staff
staff_experience

Modeling Strategy：Clustering + Regression Approach

Goal

- Food waste cause prediction
- heterogeneous patterns depending on various environmental/operational/human conditions.

Approaches

- Clustering
  Customized analysis and prediction
  ：Each cluster has different characteristics,
  customized solutions can be provided for each cluster.
- Linear Regression
- LGBM Regression

Data Analysis - Clustering

Clustering

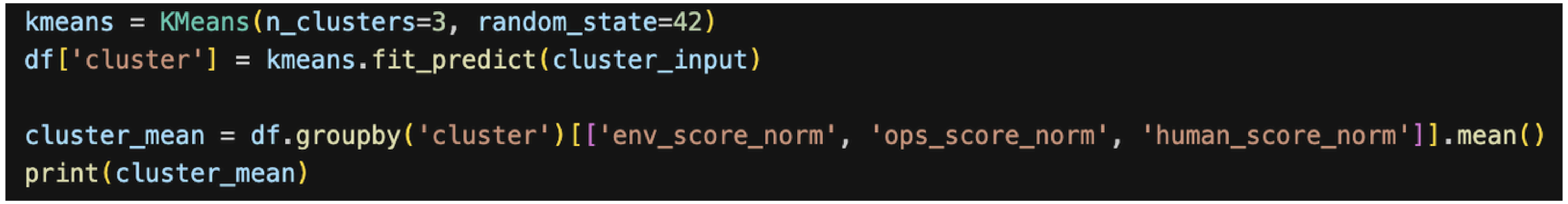
- KMeans

  : KMeans clustering based on normalized factor scores (n=3)

  : Store which cluster each data belongs to in df['cluster']

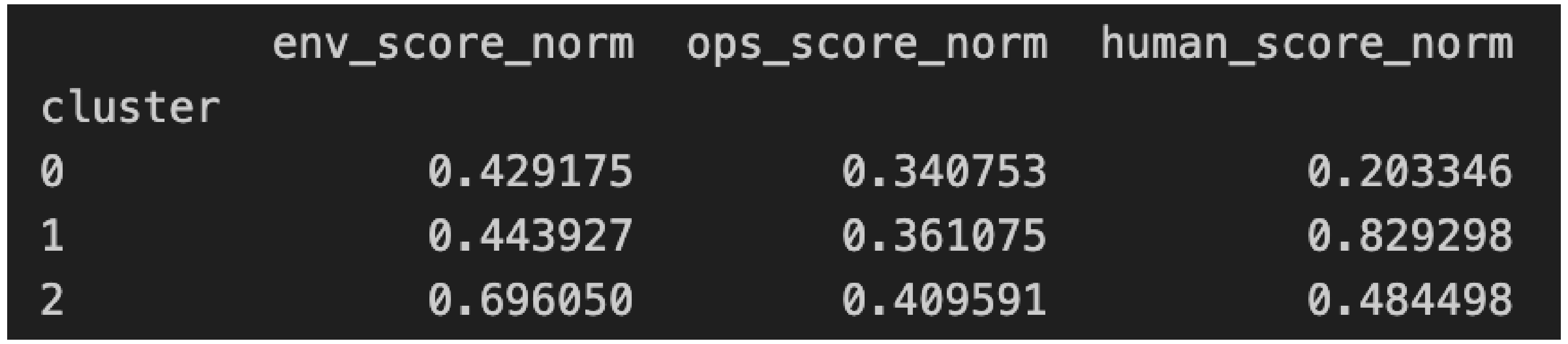
  : Calculate average score for each cluster

  → Use to identify cluster characteristics



```
kmeans = KMeans(n_clusters=3, random_state=42)
df['cluster'] = kmeans.fit_predict(cluster_input)

cluster_mean = df.groupby('cluster')[['env_score_norm', 'ops_score_norm', 'human_score_norm']].mean()
print(cluster_mean)
```

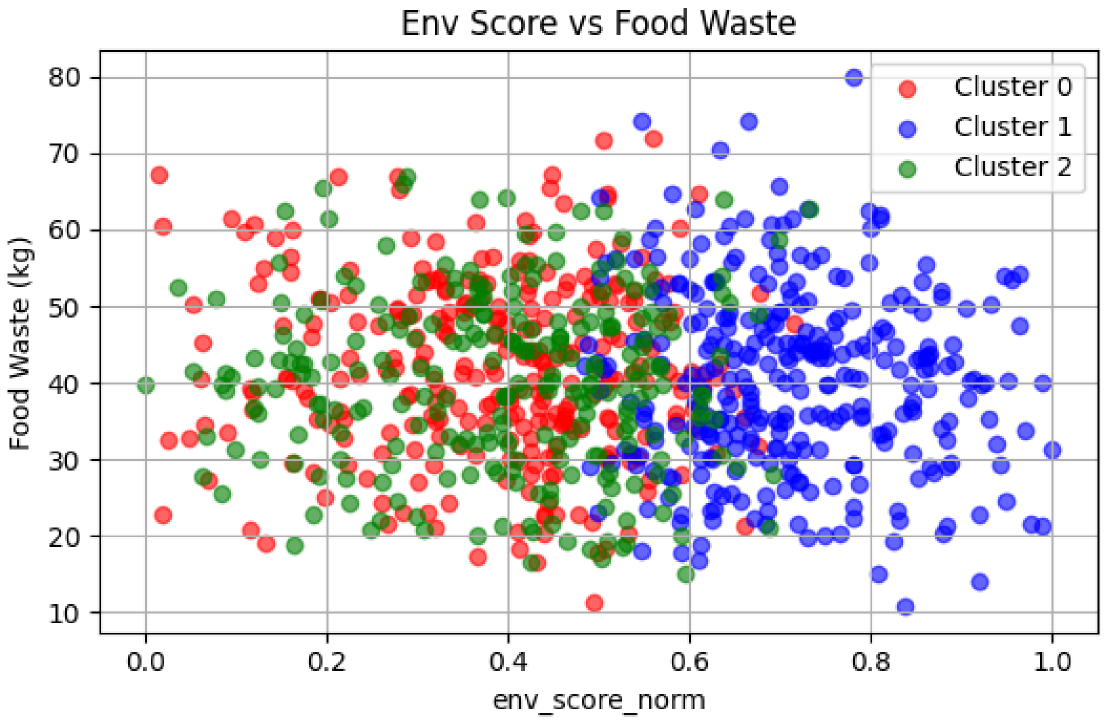


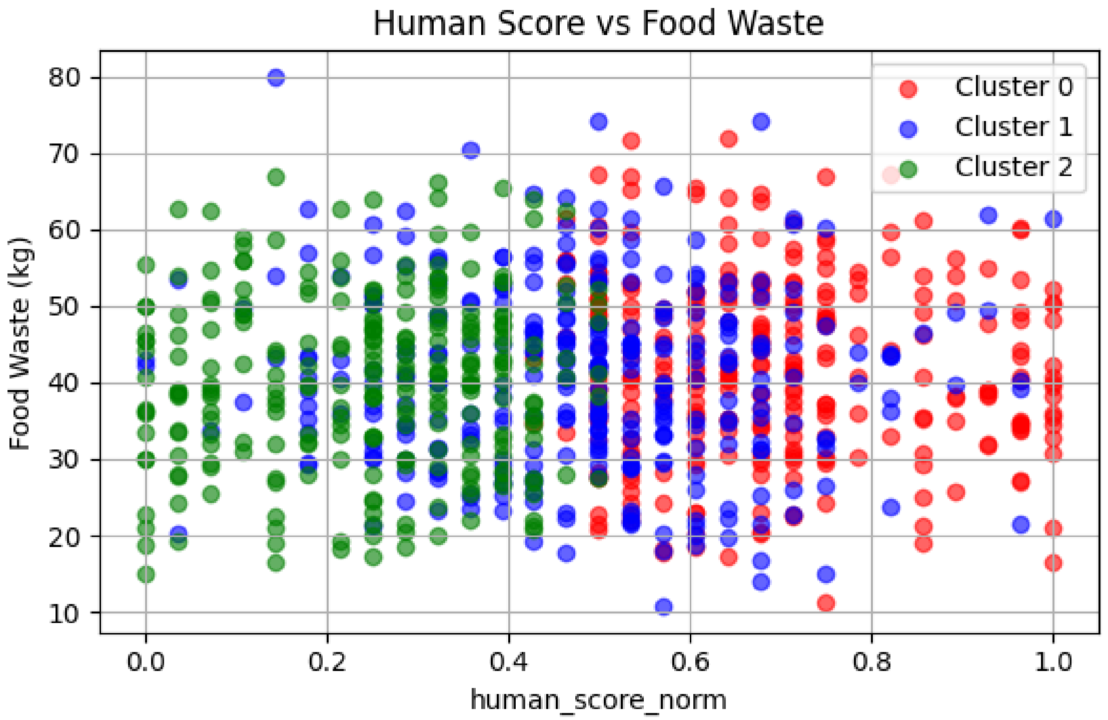
```
         env_score_norm  ops_score_norm  human_score_norm
cluster
0              0.429175        0.340753          0.203346
1              0.443927        0.361075          0.829298
2              0.696050        0.409591          0.484498
```

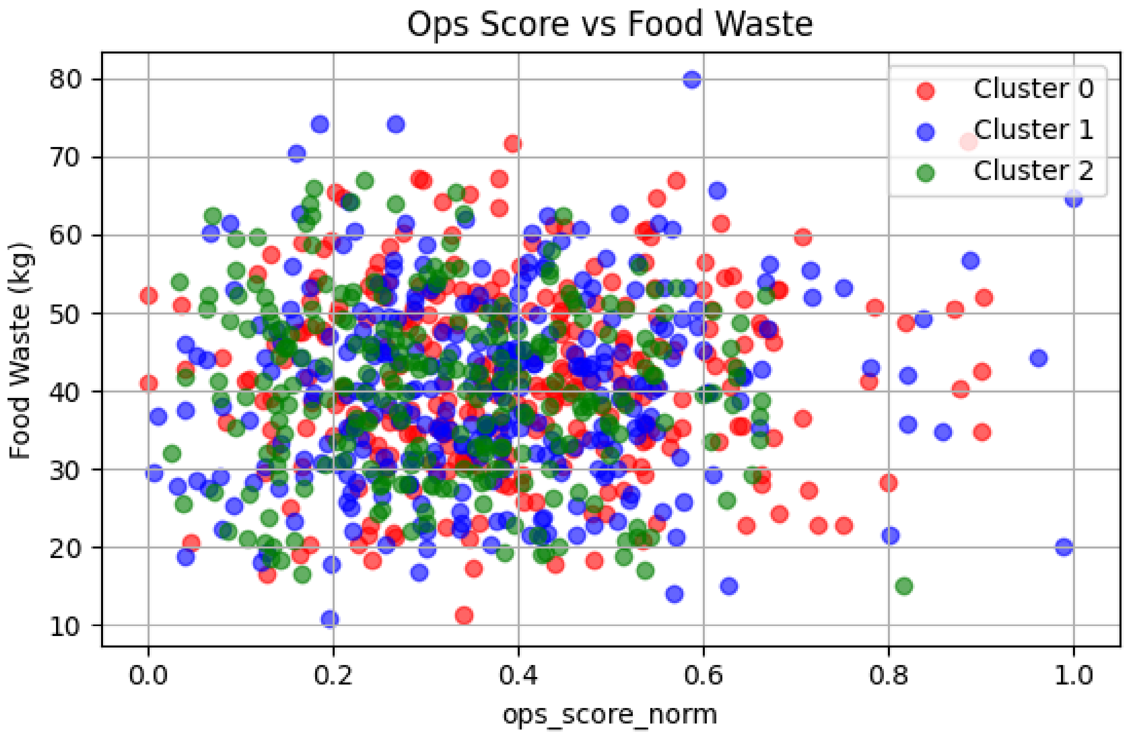
Data Analysis - Clustering

Cluster Visualization

- Cluster
  1. Cluster 0：Consists of clusters with high human_score
  2. Cluster 1：Configure clusters with high env_score
  3. Cluster 2：All factors are evenly distributed and all three scores are low


Env Score vs Food Waste


Human Score vs Food Waste


Ops Score vs Food Waste

Data Analysis - K-Means Clustering + Linear Regression

Why? used **Linear Regression** - Effectiveness

1. Interpretability of the Model

- **Coefficients** allow us to quantify how much and in what **direction each variable affects the target (food waste).**

2. Cluster-Specific Variable Influence

- Enables clear **comparison of impact factors across clusters**, helping to identify key drivers like human, environmental, or operational factors in each group.

3.Low Complexity, Fast Execution

- Linear regression assumes a simple relationship, making it quick to train and evaluate without complex tuning.

4. Effective with Normalized Data

- Works well with data scaled using MinMax or StandardScaler, making **coefficient magnitudes more comparable across features.**

Data Analysis - Linear Regression Result by Coefficient

Feature	Cluster 0	Cluster1	Cluster2
temperature	0.04	1.25	0.56
humidity_percent	0.09	1.73	-0.05
meals_served	6.08	6.14	7.13
special_event	2.84	3.95	2.61
day_of_week	-0.35	0.42	0.41
kitcen_staff	0.84	1.08	-0.38
staff_experience	-1.88	-0.99	-0.40

*Since **special_event** is a binary feature can result in a relatively large regression coefficient.

Cluster 0: Human-driven group

staff_experience has a strong negative coefficient(-1.8), **meaning more experience significantly reduce waste.**

Kitchen_staff is positively weighted(+0.08), implying **more staff may cause inefficiencies.**

Key Factor: Staff experience.

Cluster 1:Envrionmentally Sensitive Group

Temperature and **humidity** both have **positve** influence.

Special_event is most influential.

Key factor: Climate and events.

Cluster 2: All overall Lowest Group

meal_served has the strongest influence(+7.13), meaning food waste is highly correlated.

Negative coefficient for **kitchen_staff**(-0.38) indicates more staff slightly reduces waste.

Key factor: meal quantity

Data analysis - Linear Regression Conclusion

Cluster 0

- When the experience level of kitchen staff is low, food waste increases significantly, especially during special events.

Cluster 1

- This cluster is highly sensitive to environmental factors such as temperature and humidity, and is still significantly affected by special events.

Cluster 2

- The amount of meals served is the most critical driver of food waste, while the impact of staff size or experience is relatively minor.

Use Clustering & Regression

Why?

Q. Why did you use clustering and regression together?

- Data heterogeneity handling, local prediction model creation, accuracy improvement, ease of interpretation, etc...

How?

Q. How did you use K-Means and LGBM Regressor?

K-means: use MinMaxScaler to normalize all features to [0, 1].

LGBMRegressor : Divide the data into clusters, then split the training/validation data for each cluster using train_test_split().

Comparison with another answers-1

Compare **clustering** aspects

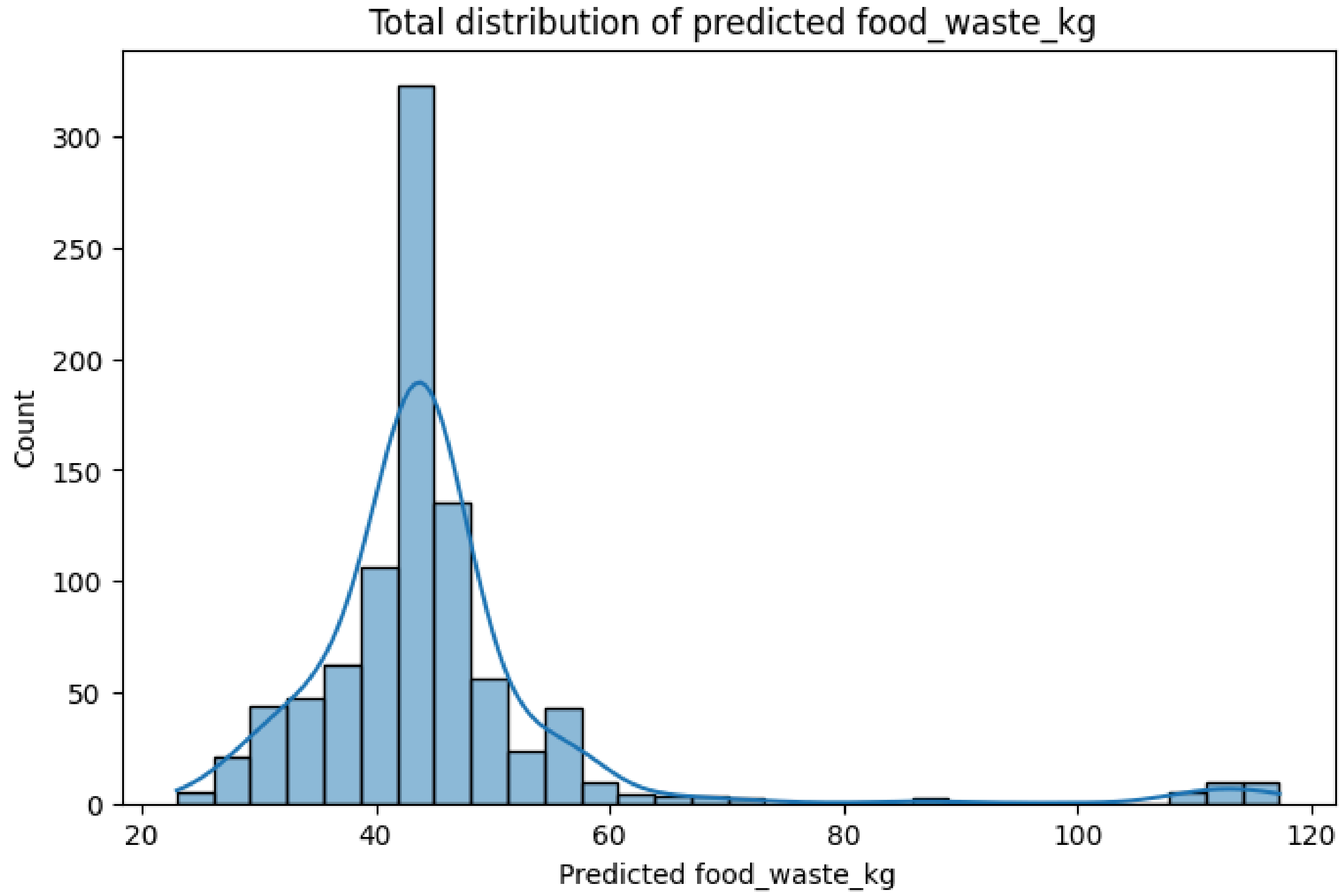
item	**Ours**	Answer
Apply clustering	Create clusters with KMeans (n_clusters=3)	No clustering
Leveraging clusters	Split data by cluster ID and apply a separate regression model to each cluster	Treat your entire data as one model
Benefits	Learn complex patterns with cluster-specific models	Simple structure with full integration processing without splitting data
Cons	The number of clusters (n=3) needs to be specified in advance, and improper division can actually degrade performance	Possible performance degradation when data is heterogeneous

Comparison with another answers-2

Compare **Regression** aspects

item	**Ours**	Answer
Model Type	LGBMRegressor (based on LightGBM)	GradientBoostingRegressor (based on sklearn)
Number of models	As many individual models as there are clusters (3)	Single model
How you learn	Train_test_split per cluster and then train separately	Pipeline training on full data
Hyperparameters	n_estimators=1000, learning_rate=0.01, etc. (conservative learning)	n_estimators=100, learning_rate=0.15 (aggressive learning)

Graph Analytics



What the graph tells us about each cluster

- Cluster 0 or 1 (more likely to be the center cluster)：Predicted values concentrated in the 40-50 kg range → representative of the central tendency of the overall data

- Cluster 2 (↑ likelihood of being a high-prediction group)：Rare but present large predictions in the 80-120 kg range

Evaluation

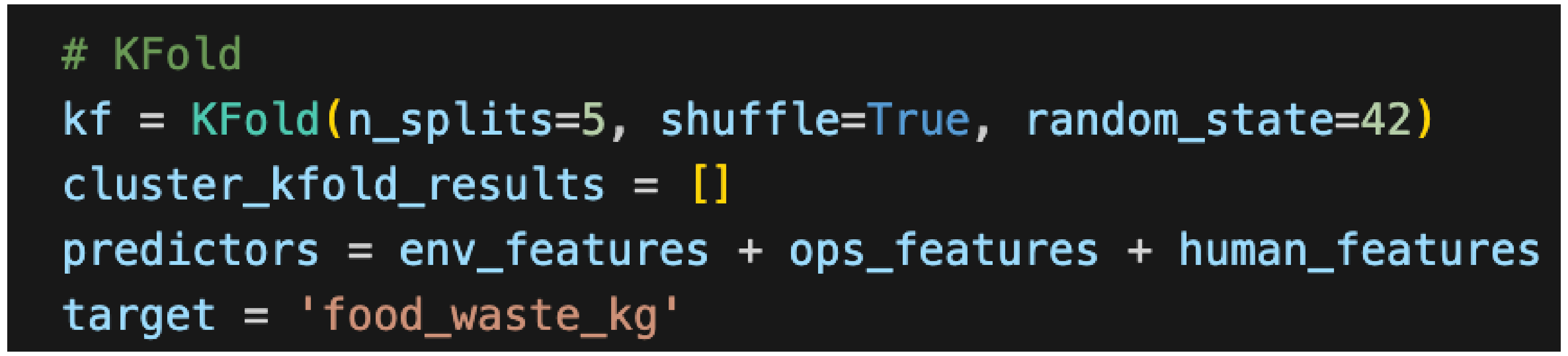
LGBMRegressor vs LinearRegression

Separation	LGBMRegressor	Linear Regression
Pros	Can learn non-linear relationships - Automatically reflects interactions between features based on trees - Supports automatic handling of missing values - Easy to analyze feature importance - Easy to control overfitting (early_stopping, max_depth, etc.) - Good for customized learning by clusters	Simple and fast to implement - Intuitive to interpret (utilizing regression coefficients) - Low risk of overfitting (especially on small datasets) - Can be utilized as a classic baseline model
Cons	Structure is complex and requires tuning - Learning can be slow - Interpretation is not intuitive (non-linear decision boundaries) - Per-cluster models increase administrative complexity	Cannot learn non-linear relationships - Cannot reflect interactions between traits - Very sensitive to handling missing values, outliers, and categorical variables - Has performance limitations with complex data

Evaluation - K-Fold (Linear Regression)

Pipeline

1. Unsupervised Clustering
2. Supervised Learning per Cluster
3. **Evaluation: K-Fold Cross -Validation**



```
# KFold
kf = KFold(n_splits=5, shuffle=True, random_state=42)
cluster_kfold_results = []
predictors = env_features + ops_features + human_features
target = 'food_waste_kg'
```

1.Data Splitting After Sturcture

- KMeans(n_cluster=3) was used to split the data into 3 Cluster(0,1,2)

2.Applying K-Fold within Each Cluster

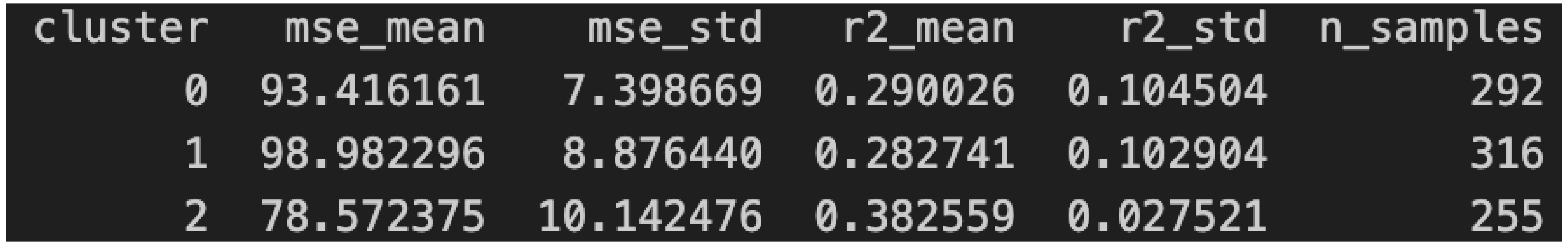
- For each Cluster:
  - split the data into features X and target y
  - Apply K-Fold(n_split = 5) to divide the data into 5 Folds
    -> Each cluster undergoes 5 rounds of training and validation

3.Iterative Training & Evaluation per Fold

- In each iteration:
  - Use X_train, y_train to train a LinearRegression model
  - Predict on X_test, evaluate using y_test

- Evaluation metrics:
  - MSE (Mean Squared Error)
  - R^2 (R-squared)

Evaluation - K-Fold (Linear Regression)



```
cluster   mse_mean    mse_std  r2_mean    r2_std  n_samples
      0  93.416161   7.398669  0.290026  0.104504        292
      1  98.982296   8.876440  0.282741  0.102904        316
      2  78.572375  10.142476  0.382559  0.027521        255
```

Cluster 0 - Human-Focused Cluster

- MSE: 93.4 / R^2: 0.2900
- Influenced by **human operational factors.**
- Although staff_experience is an ordinal variable (0, 1, 2), which can **limit expressiveness in linear models, its categorical nature supports clear, rule-based policy design.**

Cluster 1 - Envionment-Sensitive Cluster

- MSE: 98.9 / R^2: 0.2827 (Lowest performance)
- Environmental factors often behave **non-linearly** and with **high variability**, making them **harder to model accurately using linear regression.**
- While **special_event** appears influential, its **binary** nature may **exaggerate importance due to normalization**, and interactions with environmental variables **might be overlooked.**

Evaluation - K-Fold

Cluster 2 - All overall Lowest Group

- MSE: 78.6 / R^2: 0.3826 (best performance)
- Waste generation in this group is highly correlated with meal volume, forming a near-linear and predictable relationship.
- The strong signal-to-noise ratio and structural simplicity of this cluster enable **more accurate predictions** even with basic models.

Why the R^2 Scores & MSE were unstable..?

- The relatively unstable R^2 & MSE values are not due to flaws in the model itself, but rather stem from limitations in the data characteristics and feature design.
- Several key input features are discrete or binary, which makes it difficult to accurately capture the variance in the continuous target variable food_waste_kg.
- The model also does not account for feature interactions. In real-world restaurant operations, factors like temperature + event occurrence or staff count + meal volume interact and jointly impact food waste.
- However, **the linear model treats these variables independently, which limits explanatory power.**

Future Directions & Model Enhancement Strategy...

1. *Introduce <u>Non-Linear </u>Models (Random Forest, XGBoost, Neural Net...)

- Implementing models like Random Forest, XGBoost, or Neural Networks can capture complex relationships better and improve R^2 performance

2. Feature Engineering with Interaction Terms

(ex. kitchen_staff × special_event OR temperature × humidity)

- Currently we only considered each variable individually but we can create combinatorial variables to reflect potential relationships.

Team Roles & Lessons Learned

박유원 202135768 Data Research Visualization PPT	이동호 202239878 Evaluation Regression PPT	채민석 202334166 Data Analysis Regression PPT	이나영 202434800 Data Preparation Clustering PPT
이번 과제를 통해 데이터분석이 어떤 과정으로 이루어지는지뿐만 아니라 데이터분석의 필요성과 유용성, 그리고 소요되는 자원과 시간을 몸소 경험할 수 있었고 조원들과 함께 협력하여 문제를 해결하는 기회를 가져 전반적으로 지식과 학습법을 익힐 수 있는 유익한 시간을 보낼 수 있었습니다. 감사합니다.	데이터 전처리부터 해석까지 전체 end to end process를 조원들과 함께 경험하며, 데이터가 제공하는 구조적 통찰과 다양한 아이디어의 중요성을 절실히 느꼈습니다. 수업 시간에 배운 여러 내용들을 본 프로젝트에 최대한 녹여내려 노력하며 역량 강화에 도움이 되었습니다. 또한 주어진 결과를 다양한 관점으로 해석하는 법을 새롭게 배운 것 같아 많이 깨우쳤습니다. 감사합니다.	전처리된 dataset을 기반으로 전체적인 모델을 설계하는 역할을 맡았으며, 주어진 dataset에 수업시간에 언급되었던 K-means, LGBMRegressor를 직접 사용해보다 실무적인 경험을 할 수 있었습니다. 데이터 전처리와 군집별 모델 설계가 예측 정확도에 얼마나 큰 영향을 미치는지도 배울 수 있던 활동이었습니다. 감사합니다.	데이터 전처리 과정이 분석 단계와 함께 반복적으로 수정되며 진행되는 과정임을 체감했습니다. 또한 조원들과 다양한 관점에서 데이터를 논의하며 데이터를 획일적으로 바라보는 대신 다양한 패턴과 특성을 발견하고 이에 맞춰 분석하는 접근 방식이 깊이 있는 통찰을 도출하는 데 기여함을 새롭게 배울 수 있는 의미 있는 경험이었습니다. 감사합니다.

Thank you

The entire process and outcomes of our project are openly shared at
https://github.com/dongramiho/Datascience_termproject

Appendix

Plaenned(In proposal) vs Actual Approach

Initial Proposal

- Perform clustering by individual factors (Environment, Operations, Human)
- Conduct regression analysis within each cluster
  → Goal: Understand how each factor independently influences food waste

Challenges Encountered❗

1.Limited Dataset Size

- Total samples: 911
- After preprocessing: ~870 usable entries
- 3 clusters per factor × 3 factors = 9 clusters
- → Fewer than 100 samples per cluster → **Unstable training & reduced reliability**
-

2.Lack of Factor Interaction Control

- Clustering based on a single factor makes it difficult to control for variation in other factors
- Example: Clusters based on environmental conditions still include mixed operational and human characteristics

Final Implemented Approach🌟

- Performed clustering using combined features: Environment + Operations + Human
- Recognized that in real-world restaurant settings, these factors are interdependent, not isolated
- Grouped data based on similar overall operational conditions
- Then, conducted regression analysis within each cluster to identify key influencing variable.