Global Food Waste

Project 4 - Group 1 Write up

Arya Maredia
Emily Wimmer
Daniel Sterling
Caleb Meinke
Kerry Oostdyk

Table of Contents:

Introduction

As global citizens concerned about sustainability, we are focused on understanding the true scale of food waste across the world. Food waste is an unfortunate byproduct of consumption habits, affecting our environment, society, and economy. The story behind food waste is complex from vast amounts of waste from developed countries to hunger challenges in places where food is scarce. We supplemented our dataset with several sources from the World Bank to learn about how hunger is related to food waste. Our project aims to explore food waste from 2018 to 2024, focusing on how much food is wasted across different countries and how these numbers fluctuate over time.

We analyzed a comprehensive dataset from Kaggle, Titled *Global Food Wastage Dataset (2018-2024)*[1]. The dataset tracks food waste across multiple countries over several years. Several notebooks are linked to the dataset, giving us a jumpstart on the conclusions we could draw. The overall dataset gives us insights on the amount of food wasted globally and breaking waste patterns down by region and year. The web app produced as a part of this project can be found at: https://cymbalofjoy.pythonanywhere.com/

Objectives

Our primary objectives for analyzing this dataset include:

1. What countries were the most and least responsible for food waste?
2. Determine if the Covid virus increased or decreased how much food was wasted globally?
3. Were there any individual countries influenced by the virus to change their food wasting habits?

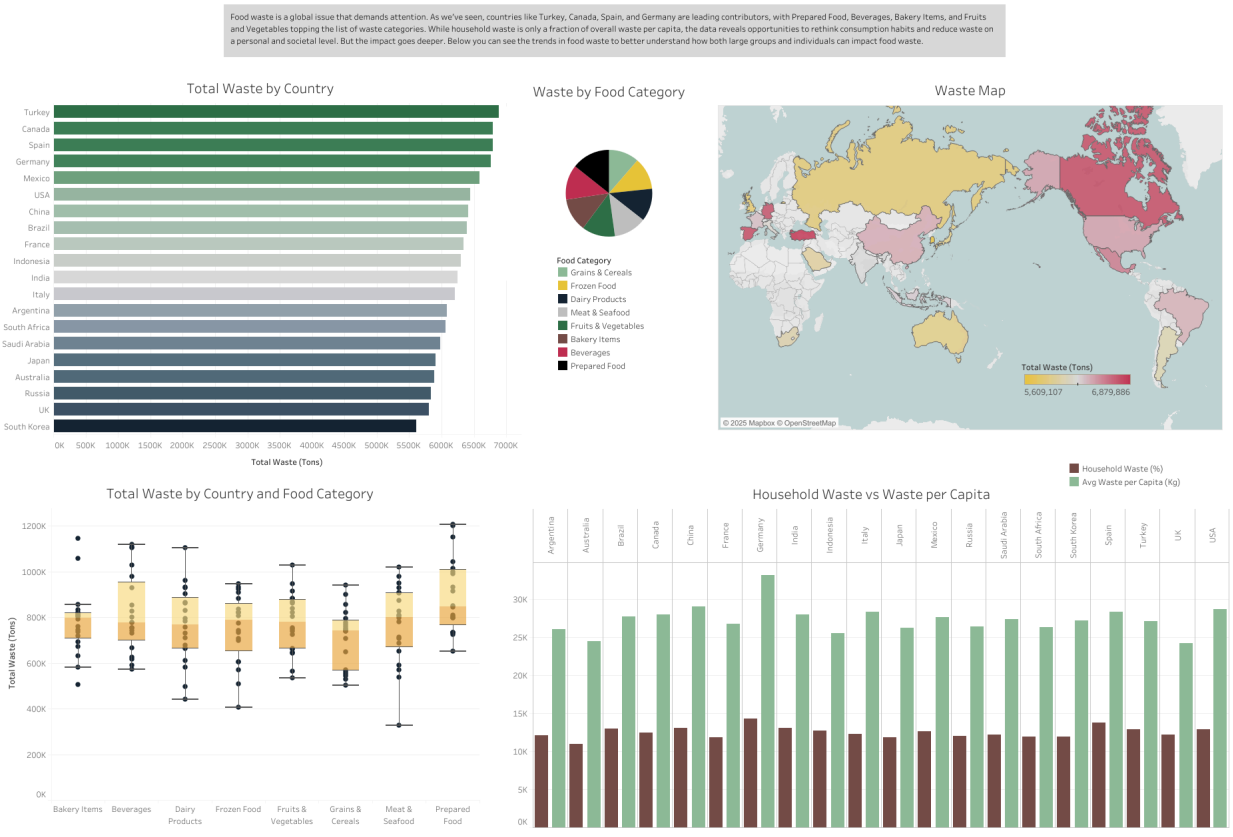
While we expected these questions to reveal different insights, the findings would offer valuable details to help us better understand the global impact of food waste.

[1] https://www.kaggle.com/datasets/atharvasoundankar/global-food-wastage-dataset-2018-2024/data

Data Cleaning

The dataset we selected was already quite clean from the start, requiring minimal preprocessing. All data cleaning was performed within our machine learning notebooks. There were no missing values, and the data types were already correctly assigned. The only adjustment made was standardizing the column names by converting them to lowercase and removing spaces. Overall, the dataset required very little cleaning.

Tableau Dashboard 1

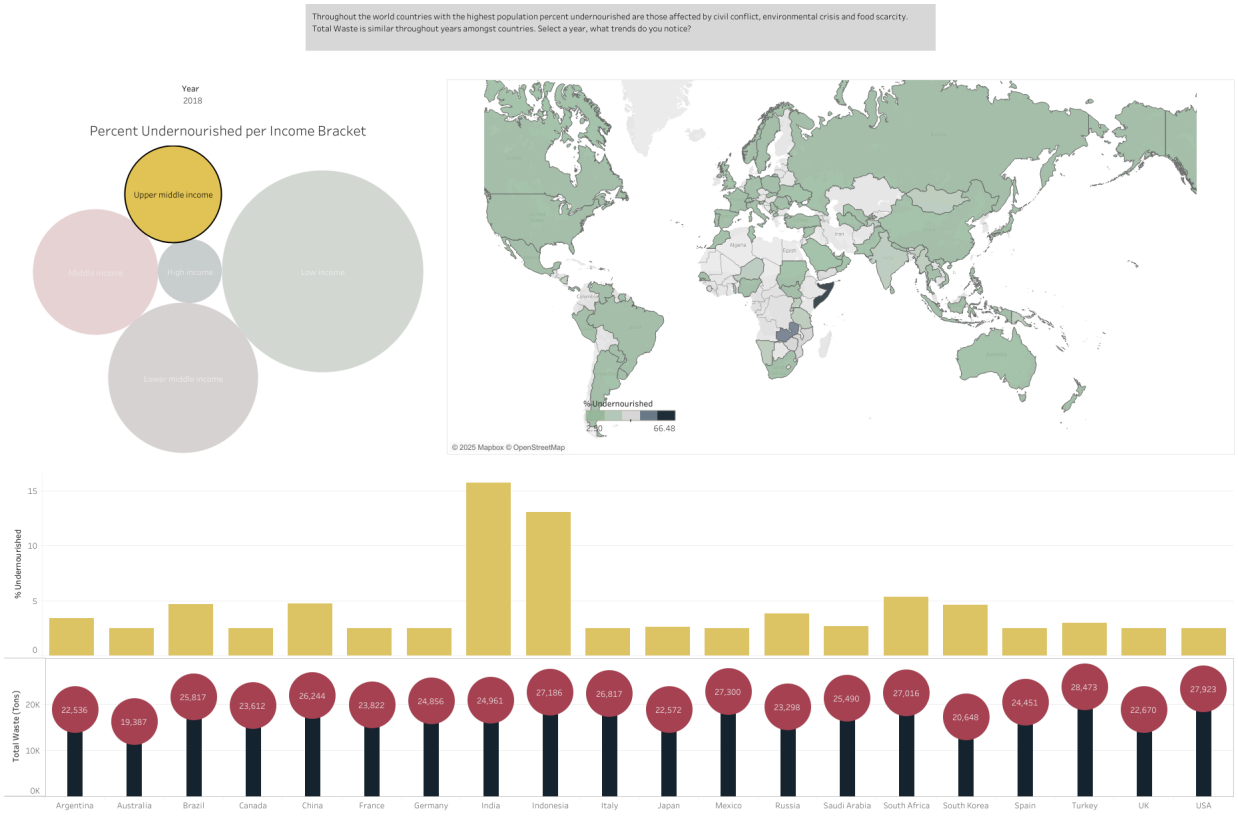


Our first dashboard is designed to display the Total Waste by Country and Food Category. We created a horizontal bar chart that lists all the countries in our data set and displays them in

descending order from most to least amount of Total Waste (Tons). After, we created a Pie chart that displays the Food Categories from our data. From this we see Turkey, Canada, Spain, and Germany represent the top 4 countries that contributed to Total Waste. We can also see that Prepared Food, Beverages, Bakery Items, and Fruits & Vegetables contribute the most to Total Waste.

To see if individuals were a leading factor in food waste, we also created another bar chart that compares Household Waste with Average Waste per Capita. In most cases Household Waste is almost equal to or even less than the Average Waste per Capita. We finished the Dashboard with a map that represents each Country, Average Waste per Capita, Population, and Total Waste.

Tableau Dashboard 2



Our second dashboard is shown above. The focus of this dashboard is to compare hunger to food waste. The data shows countries like India and Indonesia have some of the highest

percent of undernourished population, but comparatively similar food wastage. Total waste seems to be consistent throughout the years, including Covid years. Interestingly enough, food wastage is arguably consistent across the globe despite population disparity.

With this visualization, the user can dynamically change the years from 2018-2024 to compare world hunger and food waste. Through a map, you are able to visualize percent undernourished worldwide and contrast this to the total waste with our lollipop chart. In addition we can see the disparity in hunger based on income class. Unsurprisingly, the bubble chart shows that lower income classes have a higher hunger rate.

Machine Learning

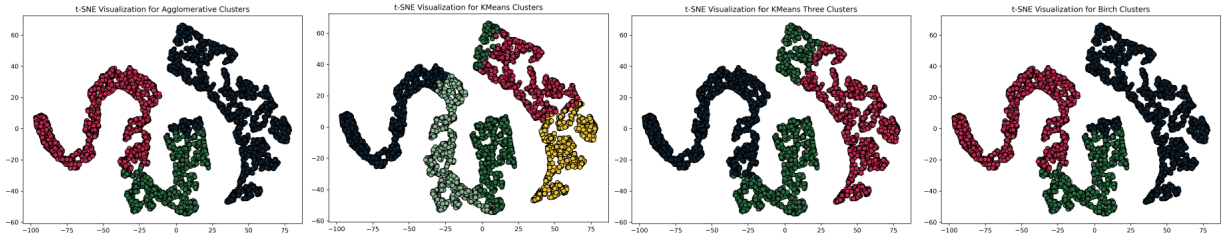
Overview

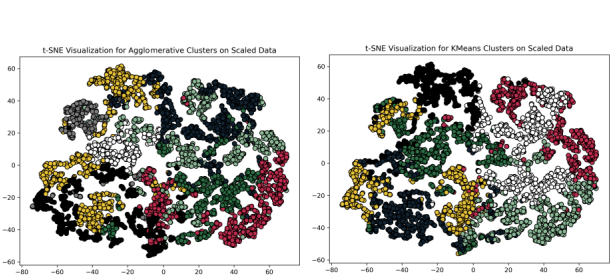
Our machine learning takes place in 4 parts and several different notebooks. To start our machine learning we analyzed our data as an unsupervised learning problem with the goal of exploring cluster analysis. We did this both on scaled data and unscaled data. We also analyzed the data set as a supervised learning problem. The focus was trying to predict the total food waste in tons, based on the other columns in our data set. The final notebook is a supervised learning experiment on the data dropping economic loss as we predicted that our target was directly related to the column.

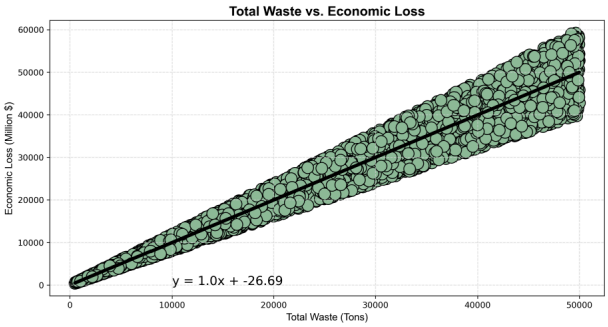
The unscaled data shows stronger groups than the unscaled data.

Economic Loss is derived from Total Waste and created an effective predictive model.

KMeans clustering effectively grouped our data. The Birch algorithm created 3 distinct groups.





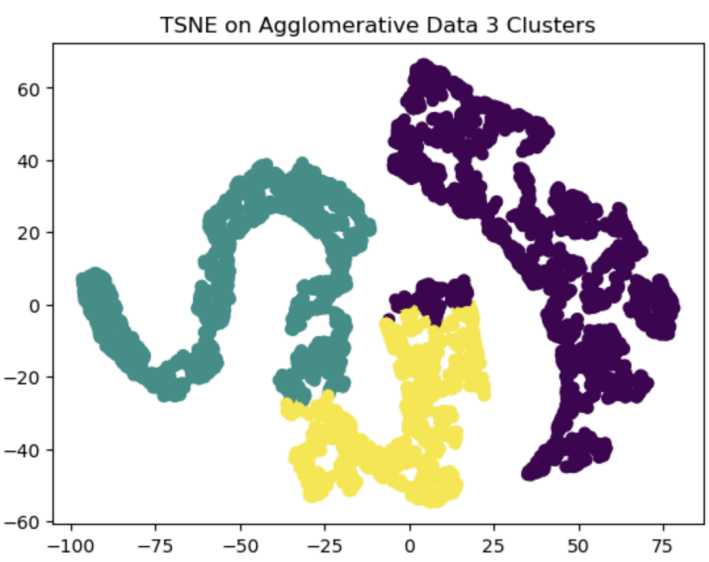


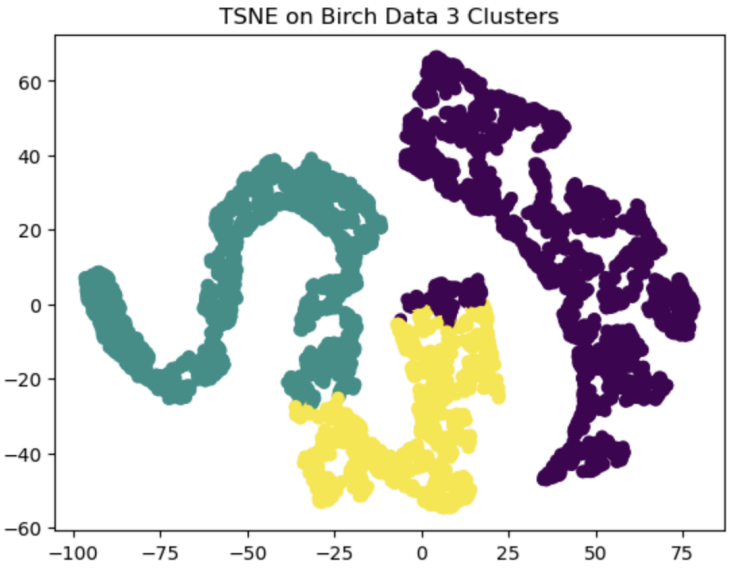
Unsupervised Unscaled

For the Unsupervised Unscaled step we explored the potential for clustering countries based on food waste related metric. First we loaded and examined the dataset, confirming there weren't any missing values and that most columns are numeric, except for "Food Category" which we dropped. Selecting key numerical features and applying "TSNE" for dimensionality reduction and visualizations. We applied TSNE on the unscaled data, producing scatter plots showing some structure but lacked clear cluster separation. After we standardized the data using "Standardscaler" ensuring equal weighting across features. Rerunning TSNE and then plotted the results. Scaled visualization showed more distinct grouping, offering better clustering potential. We generated a heatmap of correlations to understand feature relationships, showcasing a strong correlation between "Total Waste" and "Economic Loss".

In the unsupervised clustering analysis, we applied three different algorithms Kmeans, Agglomerative Clustering and BIRCH. We used these to identify grouping in the unscaled dataset, and was evaluated using a range of cluster numbers. Performance then was assessed through three key metrics, Inertia, Silhouette score, and Calinski Harabasz Index (VRC). For Kmeans, the inertia plot showed a sharp drop, suggesting an elbow point around those values, while silhouette and VRC scores peaked. This indicates decent but diminishing cluster quality beyond that. Agglomerative Clustering and BIRCH showed similar performance trends, with the highest silhouette scores then steadily decreasing afterwards, while VRC values continued to rise. This implied tighter intra cluster similarity but possibly over segmentation. K= 2 to 4 appears to be a reasonable range for meaningful clustering in this dataset, with minor variations between algorithms. Without scaling however, features with leather ranges like economic loss or population may dominate the clustering results, potentially screwing the structure of the data.

After identifying optimal k values the code visualizes the KMeans clustering results using TSNE. The fact that unscaled data is used and clusters, or scaling and feature engineering might be needed for better performance. Clustering is then applied using Agglomerative Clustering and BIRCH, both with an optimal number of clusters set to 3. Each method, a copy of the dataset (X) is made then fitted to the data. The resulting cluster labels are added as a new column called “clusters” to the dataset, allowing the countries to be grouped based on similarities in waste production, population, economic loss and other features. We used TSNE to determine if the data forms meaningful clusters, projecting the high dimensional data into two dimensions. Which enables a scatter plot where each point is colored by its assigned cluster. We assessed whether the clusters are well-separated or if the data lacks clear cluster structure. Our experiments are shown in the figures below.





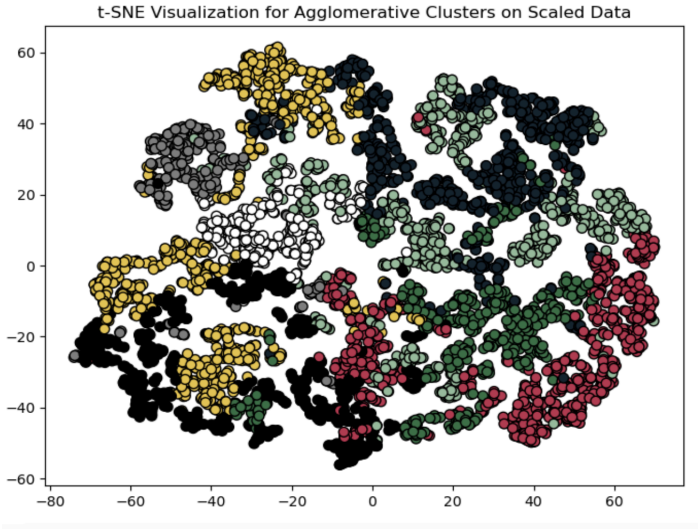
Unsupervised Scaled

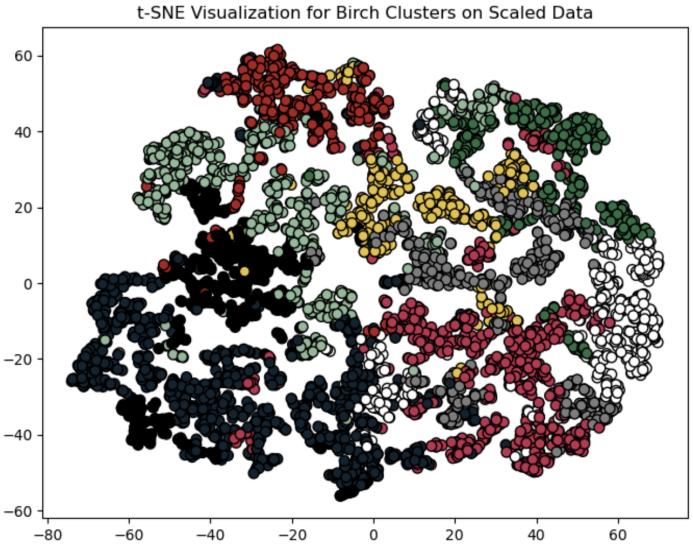
In our Unsupervised Scaled experiment we explored whether the dataset can be meaningfully clustered by appling TSNE. TSNE will first run on the unscaled numeric features to check for any natural clustering patterns in the raw data. After, the dataset is standardized using “StandardScaler”, transforming each feature to have a mean of 0 and a standard deviation of 1. It’s an important step for distance based methods like clustering, TSNE is then applied again on the scaled data and the resulting scatter plot reveals more structured patterns compared to the unscaled version. A heatmap of the correlation matrix is generated to examine the relationship between numeric values. Helping in identifying which features strongly correlated and can influence clustering results or suggest potential redundancy in the data.

For our unsupervised clustering, a range of clutter counts (K 2-19) is tested, and calculates the three metrics previously mentioned. KMeans shows a gradual decrease in inertia with increasing k, and its silhouette and VRC scores peak around 2 to 3 clusters, this suggests this might be an optimal range. Agglomerative and BIRCH clustering both show highest silhouette and VRC scores at K = 2, with steadily declining values as K increases. The indication the data likely contains is two primary clusters, regardless of the method used, clusters may lead

to overlapping or poorly separated groupings. Clustering is possible and meaningful, especially when limited to fewer clusters.

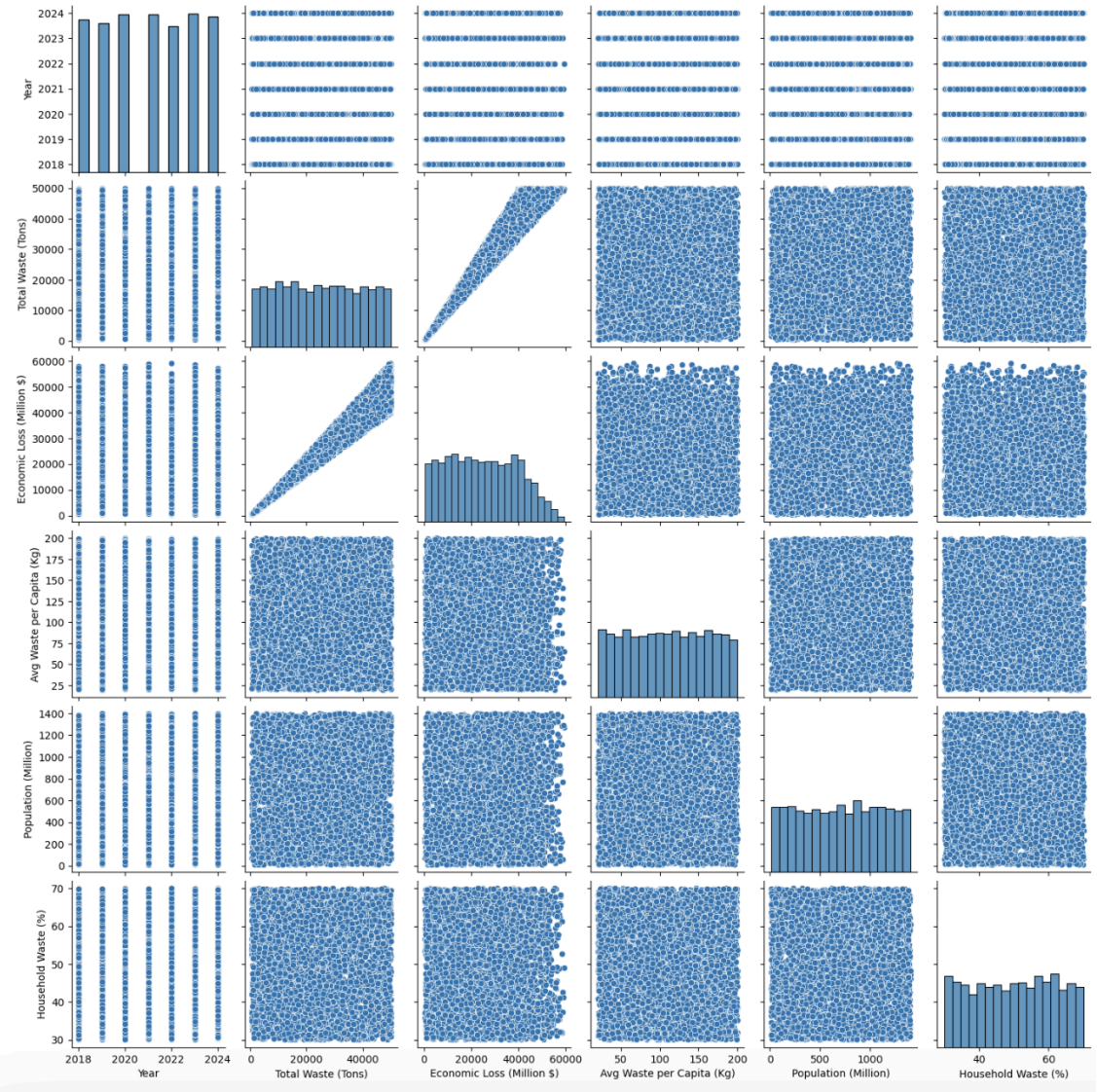
Finally, we performed a series of clustering and visualization tasks using scaled feature data (X) and then built a linear regression model to explore the relationship between two variables from a separate dataset (df). The results are visualized using TSNE. A linear regression analysis is conducted to determine how “Total Waste (Tons)” relates to Economic Loss (Million $). It calculates the regression line and visualizes it alongside a scatter plot overlaying the linear equation on the graph. The final plot is customized with color schemes for clarity and visual appeal.





Supervised (with Economic Loss)

We started setting up preprocessing pipelines to prepare a dataset on food waste and its economical impact. As the image below shows, we also did a pair plot to see correlations in the dataset. This is where we originally noticed that Total Food Waste correlates to Economic Loss.



The dataset included both numerical and categorical features. We identified and grouped the numerical features: “Total Waste (Tons)”, “Economic Loss (Millions $)”, “Avg Waste per Capita (Kg)”, “Population (Millions)” and “Household Waste (%)”. They are handled using a pipeline that imputes missing values with the median and then standardizes the data using “StandardScaler”, which ensures all features are on the same scale. No binary features are in this

dataset, however a placeholder pipeline is defined in case any appear later. The pipeline will fill in the missing value with the most frequent entry then encode the values numerically using “Original Encoder”. Categorical features “Year” and “Food Category’ are processed with another pipeline imputation missing values with the most frequent entry and applied “one hot encoder”. “One Hot Encoder” converts each unique category into its own binary column, helping machine learning models to interpret categorical variables effectively. All three pipelines then get combined into a single “ColumnTransformer”, ensuring that the appropriate transformations are applied to each type of feature during modeling. We applied the preprocessing to transform the dataset and construct a correlation analysis to explore relationships between features.

After transforming the data we created a dataframe with the processed values, then visualized the feature correlation using a heatmap and found that “Economic Loss” has a stronger positive linear correlation with “Total Waste (Tons)”, while others showed weaker or negligible correlations. This helps us understand the data before processing with model training. To prepare the data for machine learning by separating the features (X) from the target variable (Y), which is “Total Waste (Tons)”. We then split up the dataset into testing sets and training to evaluate model performance. We then built two classification models using pipelines, “Logical Regression” and “Random Forest Classifier”. Ensuring that the necessary columns were present in the dataset, we then ran a function “doClassification” to train and evaluate both models. This process allowed us to compare model performance and identify which classifier better predicted total food waste.

Supervised (without Economic Loss)

In addition to our experiment with including “Economic Loss” we decided to drop the feature and see how it affected our data set. We started off by loading the data in Pandas. The dataset contains 5,000 entries with seven columns, each entry representing a specific country. Our next step was to then drop a column named “Economic Loss (Million $)”, due to the strong correlation with "Total Waste (Tons)." See the Figure on the next page.

Country	Year	Food Category	Total Waste (Tons)	Avg Waste per Capita (Kg)	Population (Million)	Household Waste (%)
Australia	2019	Fruits & Vegetables	19268.63	72.69	87.59	53.64
Indonesia	2019	Prepared Food	3916.97	192.52	1153.99	30.61
Germany	2022	Dairy Products	9700.16	166.94	1006.11	48.08
France	2023	Fruits & Vegetables	46299.69	120.19	953.05	31.91
France	2023	Beverages	33096.57	104.74	1105.47	36.06
...	...	...	...	...	...	...
France	2021	Beverages	47524.74	77.41	1087.46	39.73
Australia	2021	Beverages	32337.72	194.35	1336.32	64.83
China	2018	Meat & Seafood	20640.96	21.04	16.13	31.23
Australia	2021	Beverages	26566.64	197.14	1086.17	69.95
France	2024	Bakery Items	8860.27	51.50	879.67	54.27

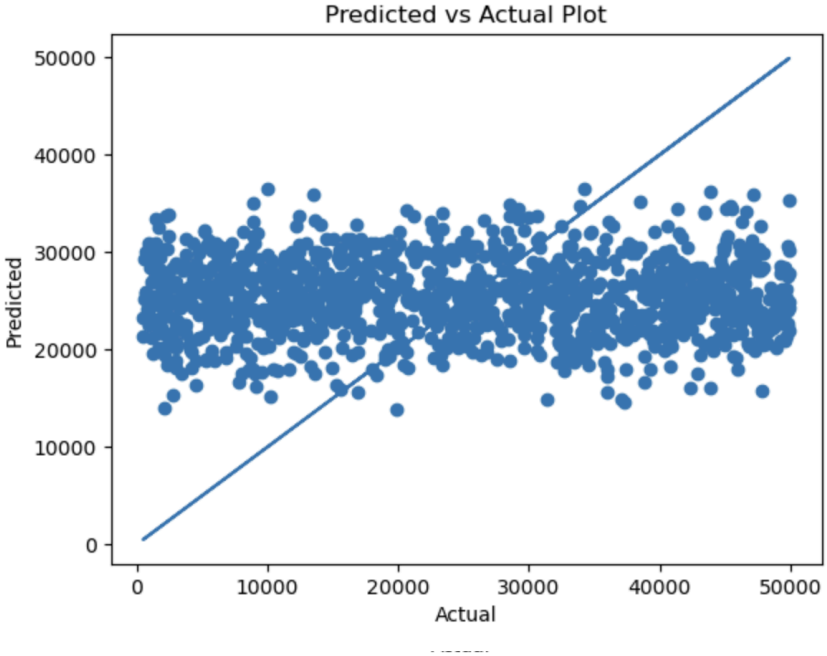
We then preprocess the data by defining and transforming different types of features for machine learning. Categorizing the features into numeric, binary and categorical groups. For numeric features we assign missing values with median and scale the data with Standard Scaler. For binary features we assign missing values with the most frequent value and apply “OrdinalEncoder”. For categorical features we assign missing values with the most frequent value and use “OneHotEncoder” to create binary columns for each unique category. Transformations are applied using Pipeline and Column Transformer. We then apply fit_transform to the transformations, retrieving the transformed data and getting the feature name using “get_features_names_out”.

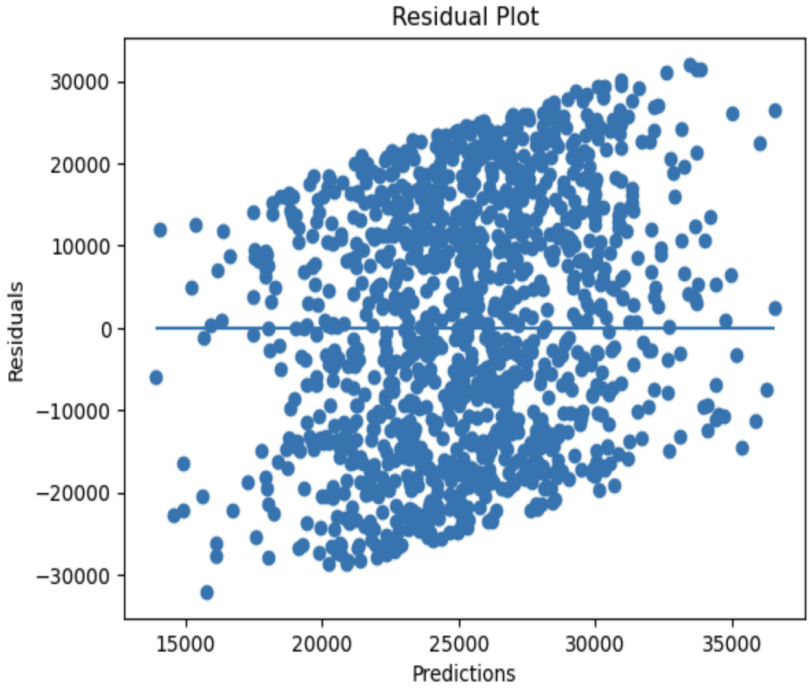
We started on preparing the data for training a machine learning model by splitting it to features and labels and partitioning the dataset in training and testing sets. We separated the features (x) from the target variable (y). Target Variable being “Total Waste (Tons)”, then dropping this column from the features Dataframe (x). Using “train_test_split” we divide the data into training and testing sets. We specified that 25% of the data will be reserved for testing and 75% will be used for training. “X_train_head () is used to preview the first few rows of the training feature, it’s important to note however that these features are still in their raw, unscaled, uncoded form and no preprocessing has been applied yet.

The function “doRegression” first fits the given model to the training data, then it makes predictions on both the training and test datasets. It then calculates and prints performance metrics for both training and test sets: R^2 (coefficient of determination), mean squared error (MSE), root mean squared error (RMSE), and mean absolute error (MAE). The metrics help assess the performance and accuracy of the model. It generated two plots, first being "Predicted vs Actual Plot" and second one being "Residual Plot". "Predicted vs Actual Plot" visualizes the data the actual and predicted values for the test set. "Residual Plot" shows residuals (errors) between predictions and actual values to help fix potential model issues.

A machine learning pipeline is created using “Pipeline” from “ScikitLearn”. The pipeline first applies a preprocessing step, then uses a “LinearRegression” model as a regressor. After the model initializes, the “doRegression” function is to train and evaluate the model on test datasets and training. Output showing performance metrics for test sets and training including R^2, MSE, RMSE, and MAE.

The dataset is then confirmed by checking necessary columns, especially “Total Waste (Tons)” designed as the target variable (Y). The features (X) are defined by dropping “Total Waste (Tons)” and “Economic loss (Million $)” from the dataset. A “Random Forest Regression” model is then initialized in a pipeline including previous defined preprocessing steps. The training metrics indicate that the model fits the training data very well with an R^2 of 0.85, performing poorly on test data with negative R^2, showing overfitting. Feature importance is then extracted on trained “Random Forest” model understanding which features have the most influence on predictions. The top three most important features are "Year", "Avg Waste per Capita (Kg)", and "Food Category", while having the most predictive power for estimating total food waste.


Predicted vs Actual Plot
Predicted
Actual
0
10000
20000
30000
40000
50000


Residual Plot
Residuals
Predictions
30000
20000
10000
0
−10000
−20000
−30000
15000
20000
25000
30000
35000

Machine Learning Predictive Model

To conclude our machine learning experiments, we selected a Random Forest model to predict Total Food Waste in Tons, using Economic Loss as a key feature. The inclusion of Economic Loss was essential in creating a functional and predictive model. Through our experimentation, we found that no models provided accurate predictions without incorporating Economic Loss.

Our evaluation using R-squared values indicated that models excluding Economic Loss tended to overfit or performed worse than random guessing. In contrast, the Random Forest model showed a slightly better fit compared to a standard linear regression model and achieved a near-perfect fit when Economic Loss was included.

Bias and Limitations

The Kaggle dataset used in this project was relatively small, containing only 5,000 rows and 8 columns. This posed several challenges. Notably, many countries were excluded, limiting the global scope of the analysis. To address these gaps, we incorporated supplemental datasets into our Tableau dashboards. Additionally, we observed that economic loss appeared to be closely correlated with total waste, suggesting it may have been derived from that column. Upon further examination, it became evident that the dataset had likely been pre-cleaned, with several original columns removed prior to publication.

Conclusions and Future Work

Economic loss emerged as a strong predictor of total food waste. This may be because it is derived from the total waste in tons. Further investigation into the original dataset would be needed. In addition, household waste represents only a portion of global food waste. Corporations create nearly half of the food waste created. Furthermore, food waste levels do not vary drastically between countries, even when accounting for population differences. Factors such as ease of access—especially in developed nations—may contribute to increased waste.

Looking ahead, seeing the expanded dataset would provide a more comprehensive understanding of global food waste patterns. In addition, collecting granular data at the city level could help identify high-impact areas, enabling more targeted interventions and meaningful change where it's needed most.

Works Cited

1. Soundankar, Atharva. “Global Food Wastage Dataset 2018–2024.” *Kaggle*, https://www.kaggle.com/datasets/atharvasoundankar/global-food-wastage-dataset-2018-2024/data

2. “GDP Growth (Annual %).” *World Bank*, https://data.worldbank.org/indicator/NY.GDP.MKTP.KD.ZG.

3. “Prevalence of Undernourishment (% of Population).” *World Bank*, https://data.worldbank.org/indicator/SN.ITK.DEFC.ZS.

4. “Color Palette Generator: Coolors.” *Coolors*, https://coolors.co/142532-c13050-2e7046-764f4a-eac33b-8eba99.

5. Baker, Marta, et al. “DS Project 4 Group 02.” *GitHub*, https://github.com/martabaker/ds_project_4_group_02.

6. Fadilah, R. (2023). *Global Food Waste Analysis and Prediction* [Kaggle Notebook]. Kaggle. https://www.kaggle.com/code/rizkyfadilah37/global-food-waste-analysis-and-prediction

7. “ChatGPT.” https://chatgpt.com/.

8. “Copilot.” *Microsoft*, https://copilot.microsoft.com/.