Food Waste Analysis Project Report

1. Introduction

This project focuses on analyzing food waste data collected from a kitchen operation environment. Our main goal is to understand the factors contributing to food waste and provide actionable insights to minimize it.

Dataset variables:

- **ID:** A unique identifier for each record.
- **date**: The date of the observation.
- **meals_served:** The number of meals served on that day.
- **kitchen_staff:** The number of kitchen staff working on that day.
- **temperature_C**: The temperature (in Celsius) on the recorded day.
- **humidity_percent:** The humidity percentage on the recorded day.
- **day_of_week**: The day of the week as a numeric value (0 = Sunday, 1 = Monday, etc.).
- **special_event:** A binary variable indicating whether a special event occurred (1 = event, 0 = no event).
- **past_waste_kg:** The amount of food waste in kilograms from previous days.
- **staff_experience:** The experience level of the kitchen staff (e.g., “Beginner”, “Intermediate”)
- **waste_category:** The category of food waste (e.g., "dairy", "meat").

2. Data Cleaning

- Dropped the 'ID' column since 'date' serves as a unique identifier.
- Duplicate rows were identified and removed.
- Corrected data types:
  - Columns like 'kitchen staff' and 'special events' had non-numeric entries which were converted into numeric.
  - 'Staff experience' and 'waste category' had inconsistent spellings (e.g., “Intermediate" and “intermediate" for the same category), which were standardized.
- Handled missing values:
  - Categorical columns were imputed with the mode.
  - Numerical columns were imputed with either mean or median based on their distribution (median for skewed data, mean for normal distributions)..

3. Exploratory Data Analysis (EDA)

Summary statistics was generated for all numerical columns.

Boxplots were used to detect outliers in the columns. Presence of outliers were also confirmed using the IQR method and capped within acceptable bounds.

Histograms and boxplots were used to visualize the distribution of numeric continuous columns and these visualizations show that the columns 'meals served' and 'temperature_C' are skewed and most of the data of 'meals served' lie between 200 and 400 with some outliers on the right while the 75% of data of 'temperature_C' lies between 17 ˚C and 30 ˚C. The other two columns ('humidity percent' and past_waste_kg) were normally distributed and spread across wide range of data.

Categorical data was visualized with countplots to observe the frequency of each category and from these visualization we can see that the 'special events' are rare and most of the staff had 'intermediate' level of experience. 'Meat' was the most occurring waste category.

4. Correlation Analysis

A heatmap used to show correlations between numerical variables. Amount of food waste showed moderate correlation with other variables and there wasn't any strong relationship between any two variables.

The bivariate analysis of food waste across 'staff experience' and 'waste category' showed that the beginner staff wasted the most food while the pro staff wasted the least and similarly the average amount of vegetable waste was most among the categories even though we saw in univariate analysis that meat waste had the highest count of entries but it doesn't indicate how much was wasted on average.

5. Key Insights and Recommendations

- **Staff Training:**
  - Beginner staff contribute to more waste which needs targeted training programs.
- **Special Events Planning:**
  - Higher waste during special events suggesting a need for better planning.
- **Environmental Factors:**
- Temperature & humidity had no significant impact on food waste.

6. Conclusion

This analysis reveals that staff experience and event planning are the primary drivers of food waste. Optimizing staffing and improving event forecasts can significantly reduce waste.

Full code, visualizations, and statistics are available in the in the Colab notebook.