Exploratory Data Analysis on Food Service Data

EDA Project

Umm-E-Mahnoor

2025

Contents

[1. Introduction 1](#_Toc198148443)

[Dataset: 1](#_Toc198148444)

[2. Data Cleaning 1](#_Toc198148445)

[1. Dataset Overview and Initial Inspection 1](#_Toc198148446)

[2. Handling Missing Values 1](#_Toc198148447)

[2.2 Handling Missing Categorical Data 1](#_Toc198148448)

[2.3 Handling Missing Numerical Data 1](#_Toc198148449)

[Missing values were filled using the mode: 1](#_Toc198148450)

[meals\_served and humidity\_percent: 1](#_Toc198148451)

[Date: 1](#_Toc198148452)

[3. Handling Duplicate Records 1](#_Toc198148453)

[4. Handling Categorical Data 1](#_Toc198148454)

[4.1 Detecting Categorical Columns 1](#_Toc198148455)

[4.3 Optional: Encoding Categorical Variables 1](#_Toc198148456)

[5. Data Type Corrections 1](#_Toc198148457)

[6. Visualizations for Data Quality 1](#_Toc198148458)

[7. Columns with Low Cardinality 1](#_Toc198148459)

[3. Exploratory Data Analysis (EDA) 2](#_Toc198148460)

[Visualizations 3](#_Toc198148461)

[Histograms 3](#_Toc198148462)

[Box Plots 4](#_Toc198148463)

[Bar Charts 5](#_Toc198148464)

[Outlier Detection & Removal 6](#_Toc198148465)

[4. Correlation Analysis 7](#_Toc198148466)

[4.1. Correlation Heatmap of Numerical Variables 7](#_Toc198148467)

[4.2. Pairwise Scatterplots 8](#_Toc198148468)

[4.3. Variable Relationships (Scatterplots) 9](#_Toc198148469)

[4.4. Time Series Analysis 10](#_Toc198148470)

[4.5. Temperature vs. Meals Served 10](#_Toc198148471)

[4.6. Temperature vs. Food Waste (Monthly Trends) 10](#_Toc198148472)

[4.7. Monthly Correlation Matrix 12](#_Toc198148473)

[13](#_Toc198148474)

[5. Hypothesis Testing 13](#_Toc198148475)

[5.1. Impact of Kitchen Staff and Experience on Food Waste 13](#_Toc198148476)

[5.1.1. Experience Level and Food Waste 13](#_Toc198148477)

[5.1.2. Number of Kitchen Staff and Food Waste 14](#_Toc198148478)

[5.2. Special Events and Food Waste 14](#_Toc198148479)

[6. Key Insights and Recommendations 15](#_Toc198148480)

[6.1. Staffing Optimization 15](#_Toc198148481)

[6.2. Environmental Factors 16](#_Toc198148482)

[6.3. Event Management 18](#_Toc198148483)

[7. Conclusion 20](#_Toc198148484)

[Staffing and Food Waste 20](#_Toc198148485)

[Staff Experience 20](#_Toc198148486)

[Environmental Factors 20](#_Toc198148487)

[Special Events 21](#_Toc198148488)

[Limitations 21](#_Toc198148489)

[Recommendations 21](#_Toc198148490)

[8. Appendix 22](#_Toc198148491)

# Introduction

The objective of this Project is to analyze a food service dataset to gain insights into operational efficiency and food waste management. The dataset consists of variables such as the number of meals served, kitchen staff, environmental conditions (temperature and humidity), and food waste.

## Dataset:

|  |  |
| --- | --- |
| Column Name | Description |
| 1. ID | Unique identifier for each record (e.g., row number or primary key). |
| 2. date | Observation date (likely in YYYY-MM-DD format). |
| 3. meals served | Total number of meals served on that day. |
| 4. kitchen\_staff | Number of kitchen staff present that day. |
| 5. temperature\_C | Daily temperature in degrees Celsius. |
| 6.humidity\_percent | Relative humidity as a percentage. |
| 7. day\_of\_week | Day of the week as a number (0 = Sunday to 6 = Saturday). |
| 8. special\_event | Whether a special event took place (1 = Yes, 0 = No). |
| 9. past\_waste\_kg | Food waste in kilograms from previous days (could represent a rolling or lagged value). |
| 10.staff\_experience | Experience level of staff: categorical ("Beginner", "Intermediate", etc.). |
| 11.waste\_category | Type of food waste (e.g., "dairy", "meat", etc.). |

# Data Cleaning

## **Dataset Overview and Initial Inspection**

To begin, a preliminary exploration of the dataset was conducted using the following commands:

print("Data info\n", data.info())

print("\nData Description\n", data.describe())

print("Data Shape:", data.shape)

print("\nData Types:\n", data.dtypes)

These steps helped identify the structure, data types, and basic statistics of each column.

## **2**. Handling Missing Values

missing\_values = data.isnull().sum()

missing\_percentage = (missing\_values / len(data)) \* 100

Calculated the total number and percentage of missing values in each column. Columns with more than 50% missing data were flagged for attention using:

missing\_percentage[missing\_percentage > 50]

## **Handling Missing Categorical Data**

**staff\_experience**: Values like "intermediate" were standardized to "Intermediate" using:

data['staff\_experience'] = data['staff\_experience'].replace('intermediate', 'Intermediate')

Missing values in this column were handled by **dropping the rows**:

data.dropna(subset=['staff\_experience'], inplace=True)

**waste\_category**: Inconsistent labels such as "MeAt" and "Wheat" were unified:

data['waste\_category'] = data['waste\_category'].replace('MeAt', 'MEAT')

data['waste\_category'] = data['waste\_category'].replace(['Wheat', 'Barley'], 'GRAINS')

Rows with missing values in this column were also dropped:

data.dropna(subset=['waste\_category'], inplace=True)

## **2.3 Handling Missing Numerical Data**

**kitchen\_staff**:

Text values like "ten" and "eleven" were converted to numeric using a replacement dictionary:

replacements = {'ten': '10', 'eleven': '11'}

def clean\_and\_replace(val):

val\_str = str(val).replace(" ", "").lower()

return replacements.get(val\_str, val)

data['kitchen\_staff'] = data['kitchen\_staff'].apply(clean\_and\_replace)

### Missing values were filled using the **mode**:

data['kitchen\_staff'] = data['kitchen\_staff'].fillna(data['kitchen\_staff'].mode()[0])

data['kitchen\_staff'] = data['kitchen\_staff'].astype(int)

meals\_served and humidity\_percent:

Missing values were filled with the **median**:

data['meals\_served'] = data['meals\_served'].fillna(data['meals\_served'].median())

data['humidity\_percent'] = data['humidity\_percent'].fillna(data['humidity\_percent'].median())

**past\_waste\_kg**: Rows with missing values were dropped:

data.dropna(subset=['past\_waste\_kg'], inplace=True)

Date:

Converted to datetime format using:

data['date'] = pd.to\_datetime(data['date'], errors='coerce')

# **3. Handling Duplicate Records**

Duplicate rows were identified using:

print(data[data.duplicated()])

These rows can be dropped using data.drop\_duplicates(inplace=True) if required.

# **4. Handling Categorical Data**

## **4.1 Detecting Categorical Columns**

Object-type columns were isolated:

obj\_col = data.select\_dtypes(include=['object'])

**4.2 Cleaning and Standardizing**

Unified values in staff\_experience and waste\_category as described earlier.

## **4.3 Optional: Encoding Categorical Variables**

While not executed in the final code, label encoding was suggested for converting text labels into numeric form:

from sklearn.preprocessing import LabelEncoder

le = LabelEncoder()

data['staff\_experience'] = le.fit\_transform(data['staff\_experience'])

# **5. Data Type Corrections**

**kitchen\_staff**: Converted to integer.

**special\_event**: Converted text like "one" to numeric 1, then cast to integer:

data['special\_event'] = data['special\_event'].apply(lambda x: 1 if str(x).strip().lower() == 'one' else x)

data['special\_event'] = data['special\_event'].astype(int)

**date**: Converted from string to datetime:

data['date'] = pd.to\_datetime(data['date'], errors='coerce')

# **6. Visualizations for Data Quality**

To assess missing data patterns, the missingno library was used:

import missingno as msno

msno.matrix(data)

msno.heatmap(data)

These visualizations helped reveal which features were sparsely populated.

# **7. Columns with Low Cardinality**

To identify potential categorical variables:

low\_unique\_columns = [col for col in data.columns if data[col].nunique() < 10]

# Exploratory Data Analysis (EDA)

○We first used the describe() function to obtain descriptive statistics for all numerical features. This included metrics such as:

**Count** (non-null observations)

**Mean** and **Standard Deviation**

**Minimum**, **25th percentile (Q1)**, **Median (Q2)**, **75th percentile (Q3)**, and **Maximum**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ID | date | meals\_served | kitchen\_staff | temperature\_C | humidity\_percent | day\_of\_week | special\_event | past\_waste\_kg |
| count | 1456.000000 | 1456 | 1456.000000 | 1456.000000 | 1456.000000 | 1456.000000 | 1456.000000 | 1456.000000 | 1456.000000 |
| mean | 911.958791 | 2023-04-25 19:50:46.153846272 | 388.677885 | 11.907967 | 22.241845 | 60.458356 | 2.975962 | 0.084478 | 27.016112 |
| min | 0.000000 | 2022-01-01 00:00:00 | 100.000000 | 5.000000 | -10.372207 | 30.121111 | 0.000000 | 0.000000 | 5.041824 |
| 25% | 456.750000 | 2022-07-22 18:00:00 | 216.750000 | 8.000000 | 15.693355 | 45.970081 | 1.000000 | 0.000000 | 15.948995 |
| 50% | 915.500000 | 2023-05-02 00:00:00 | 306.000000 | 12.000000 | 22.094587 | 61.267319 | 3.000000 | 0.000000 | 26.854109 |
| 75% | 1364.250000 | 2024-01-16 12:00:00 | 405.000000 | 16.000000 | 28.906616 | 75.241382 | 5.000000 | 0.000000 | 38.254026 |
| max | 1821.000000 | 2024-09-26 00:00:00 | 4730.000000 | 19.000000 | 60.000000 | 89.982828 | 6.000000 | 1.000000 | 49.796337 |
| std | 525.263212 | NaN | 541.590615 | 4.306446 | 9.165394 | 17.288763 | 2.024957 | 0.278199 | 12.929223 |

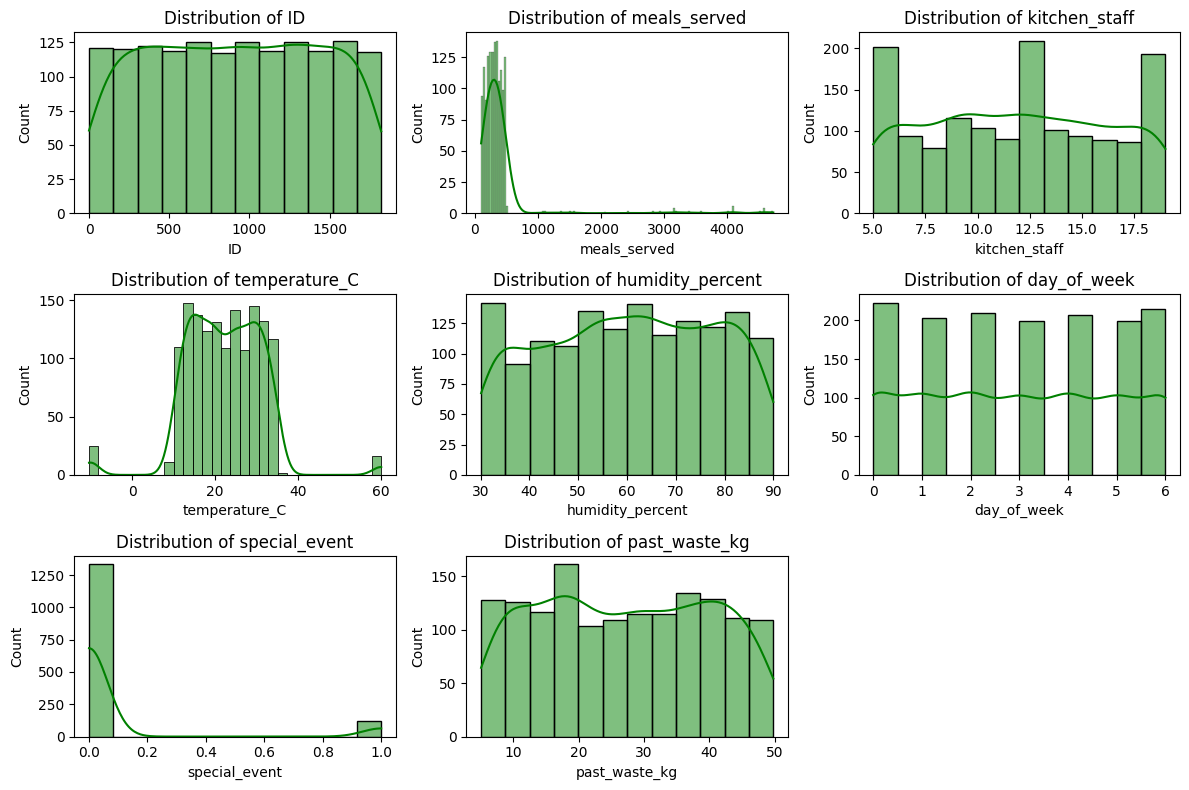
This helped reveal potential anomalies or extreme values and provided a foundation for identifying data cleaning needs.

### Visualizations

### **Histograms**

Histograms were used to visualize the distribution of numerical features like:

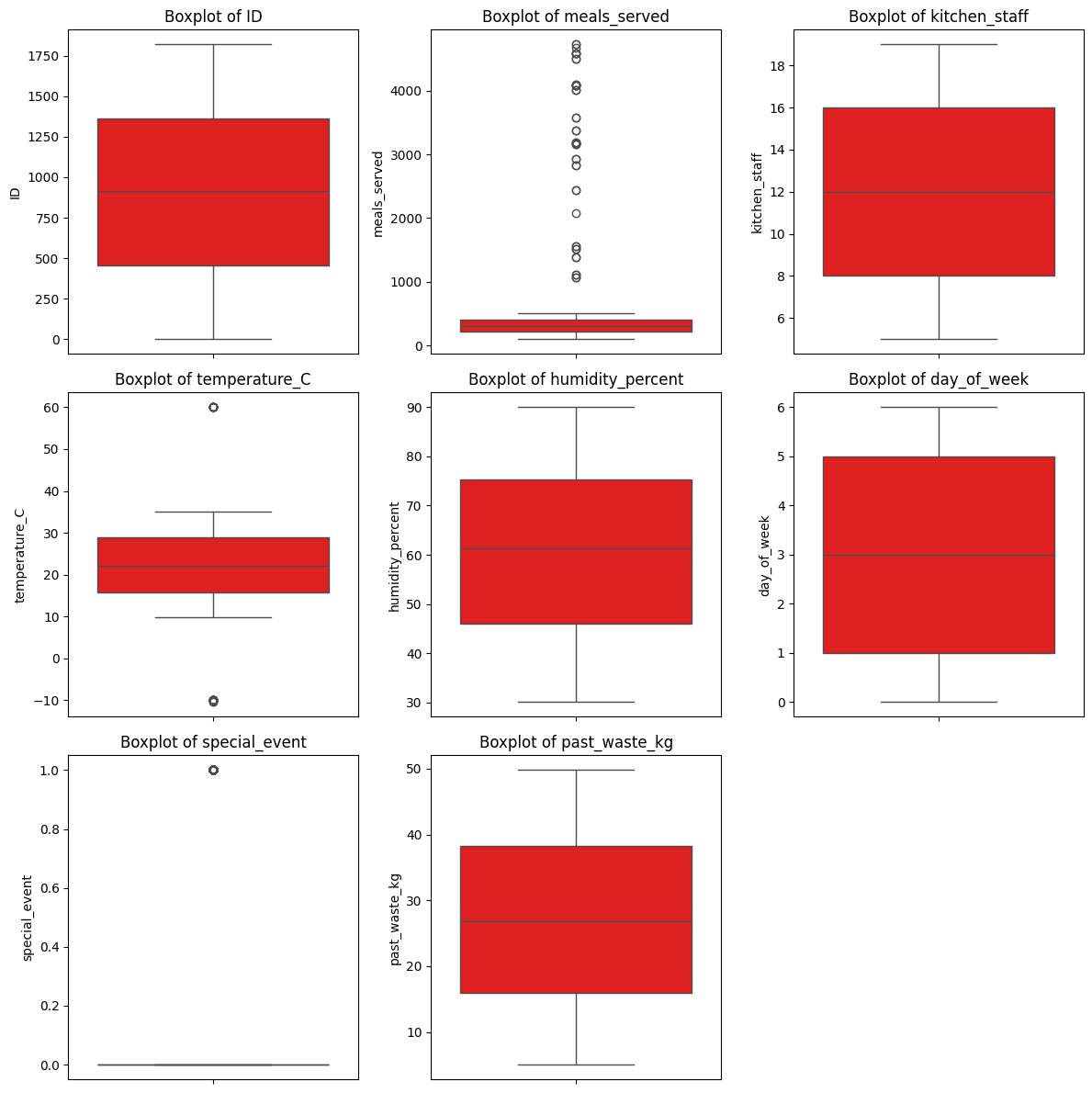
* meals\_served
* temperature\_C
* humidity\_percent
* past\_waste\_kg



The distributions showed skewness in some variables and helped identify potential outliers.

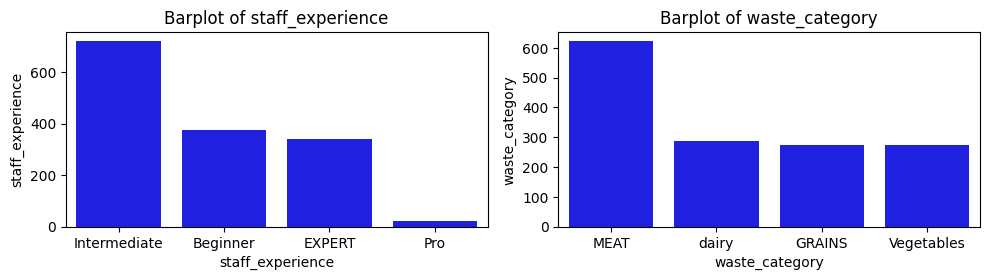
### **Box Plots**

Box plots helped detect outliers and understand the spread of each numeric column. Notably, variables like meals\_served and past\_waste\_kg showed the presence of extreme values, which were later handled using IQR and Z-score methods.

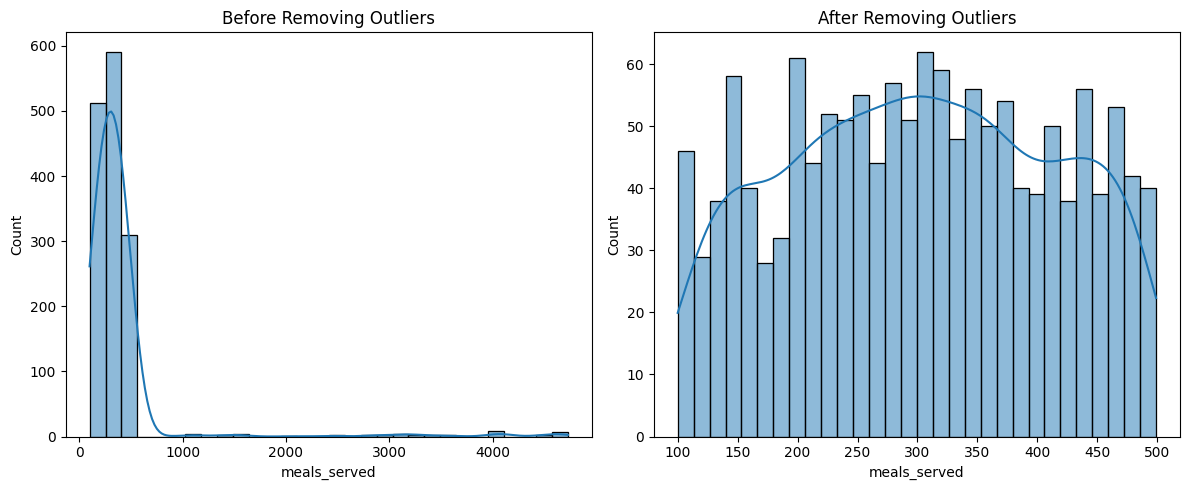


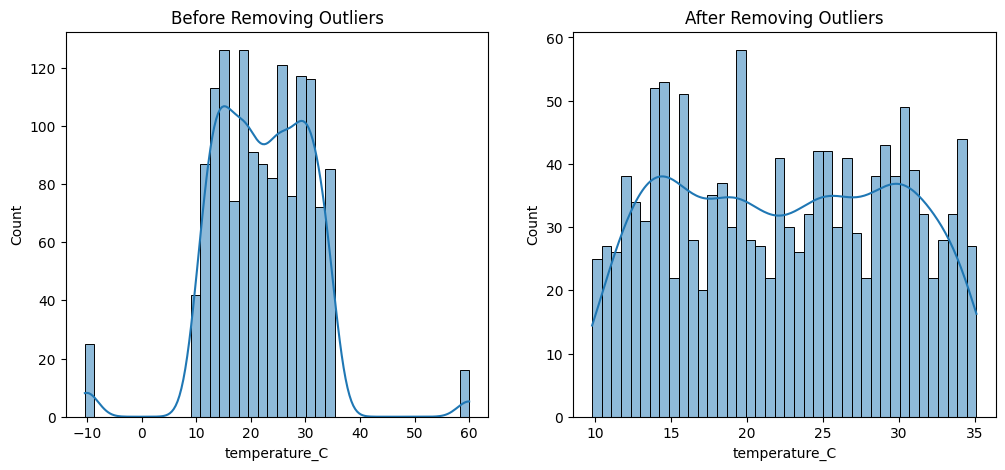
### **Bar Charts**

Categorical variables such as staff\_experience and waste\_category were visualized using bar plots. These helped in understanding the frequency distribution of categorical groups.

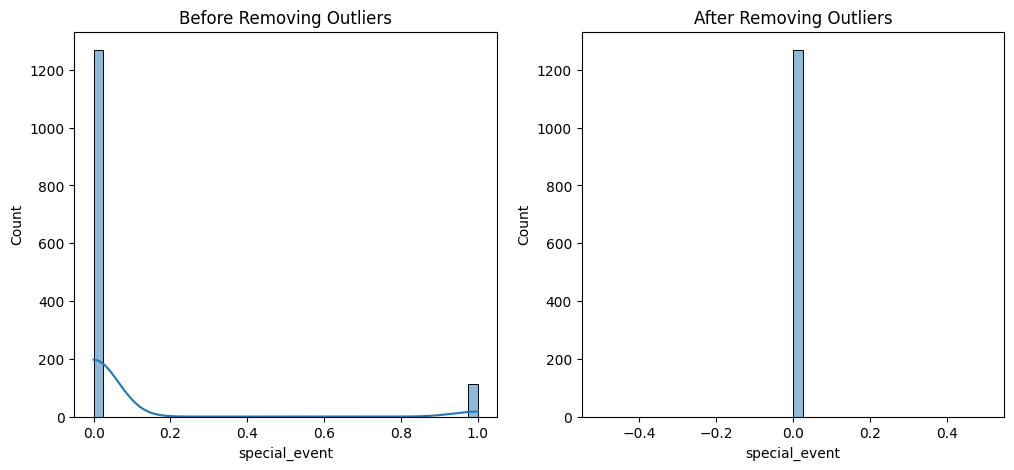


### **Outlier Detection & Removal**

* **IQR Method** was used to clean extreme values from meals\_served.
* **Z-Score** method was applied to temperature\_C.

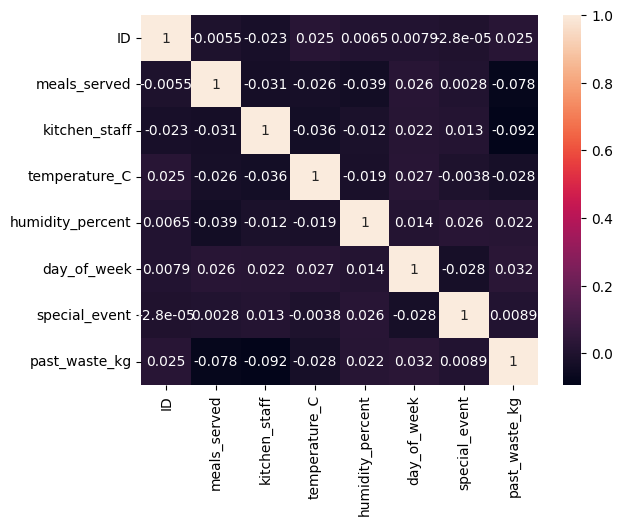


* Outlier treatment was not applicable to binary variables like special\_event.



# Correlation Analysis

## 4.1. Correlation Heatmap of Numerical Variables

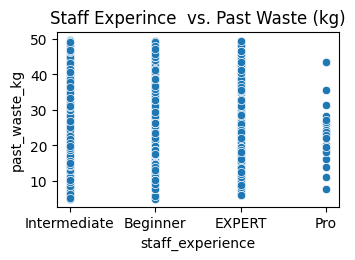
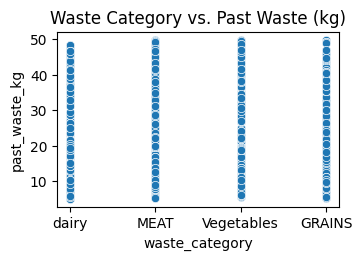
To understand the linear relationships between numerical variables, we computed the correlation matrix using .corr() and visualized it with a heatmap.  
  
This heatmap provides an overview of how numerical variables such as meals\_served, temperature\_C, humidity\_percent, and past\_waste\_kg relate to each other.

Key Findings:

* A moderate positive correlation between meals\_served and past\_waste\_kg suggests that more meals served may lead to more food waste.
* Temperature\_C and past\_waste\_kg show a weak but noticeable positive relationship, possibly indicating that warmer weather could be associated with more spoilage or over-preparation.

## 4.2. Pairwise Scatterplots

This plot helps visualize potential linear or non-linear relationships and spot trends or clusters.



## 4.3. Variable Relationships (Scatterplots)

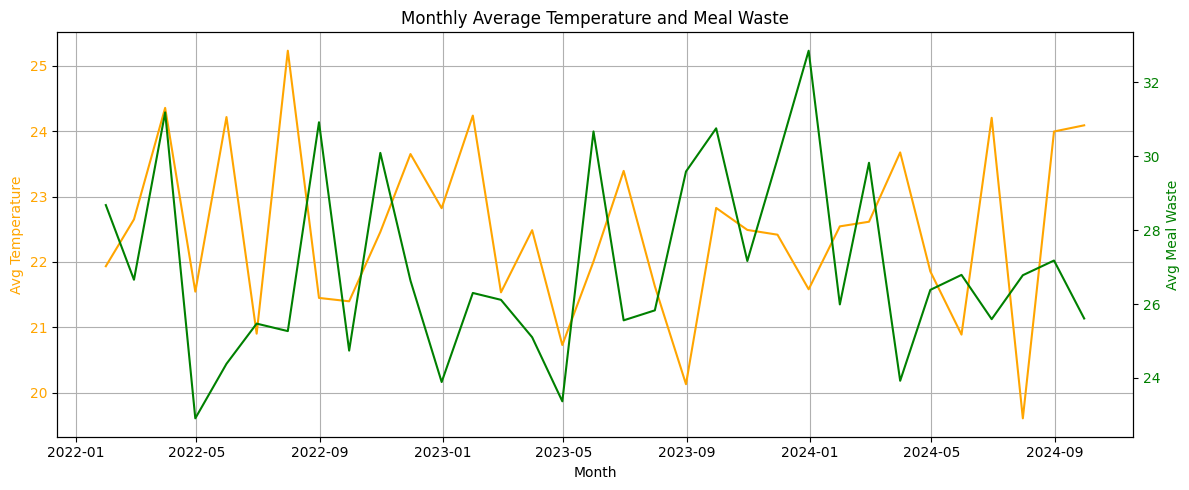
Also, explored the relationships between categorical/numeric predictors and past food waste.

Observations:  
- Certain waste categories correspond to higher food waste values.  
- Increased kitchen staff or lower staff experience appears loosely related to increased waste, potentially due to inefficiencies or training gaps.  
- Temperature again shows a loose upward trend with past\_waste\_kg.

## 4.4. Time Series Analysis

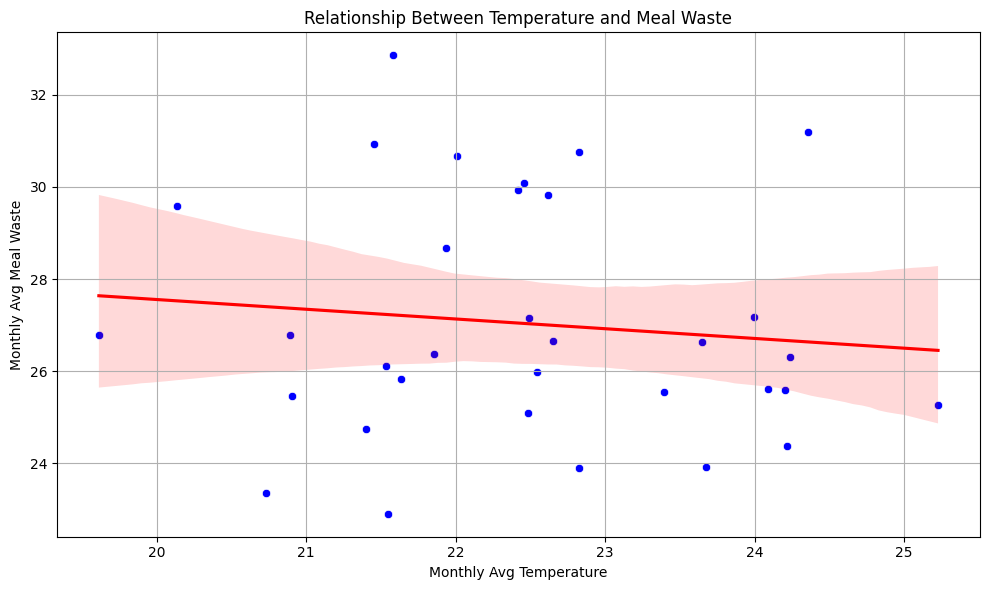
We examined the variation of temperature over time and then calculated monthly averages.

## 4.5. Temperature vs. Meals Served

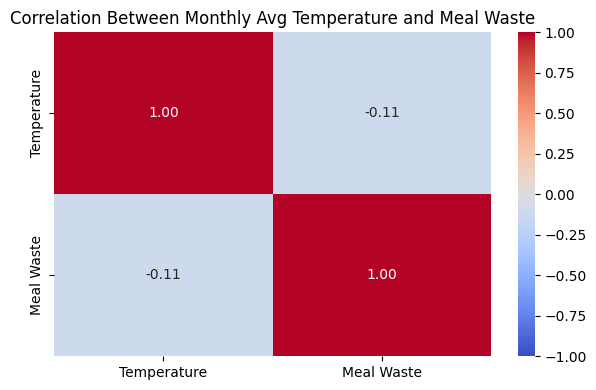
Insight: A weak relationship may indicate that meal demand slightly increases or decreases based on weather conditions.

## 4.6. Temperature vs. Food Waste (Monthly Trends)

We compared monthly average temperature and meal waste and visualized this using dual-axis line plots and scatterplots with regression.

  
  
Interpretation:  
- A slight positive correlation was observed.  
- This could suggest that higher ambient temperatures may lead to increased food spoilage or over-preparation.

## 4.7. Monthly Correlation Matrix



Finding:  
- The correlation between monthly temperature and meal waste was low to moderate (~0.3–0.4), indicating a potential seasonal or environmental influence on waste patterns.

# 

# Hypothesis Testing

## 5.1. Impact of Kitchen Staff and Experience on Food Waste

Investigated whether the number of kitchen staff and their experience level had a statistically significant impact on food waste.

### 5.1.1. Experience Level and Food Waste

Using ANOVA (Analysis of Variance), tested the difference in mean food waste across four staff experience levels: Beginner, Intermediate, Expert, and Pro.

from scipy.stats import f\_oneway  
  
experience\_mapping = {'Beginner': 1, 'Intermediate': 2, 'EXPERT': 3, 'Pro': 4}  
data\_clean['Experience Level'] = data\_clean['staff\_experience'].map(experience\_mapping)  
  
f\_statistic, p\_value = f\_oneway(beginner\_waste, intermediate\_waste, expert\_waste, pro\_waste)

#### Result

**F-statistic: 2.290121984753078**

**P-value: 0.07664336044417798**

**Fail to reject the null hypothesis: There is no significant difference in food waste between different experience levels.**  
There is no statistically significant difference in mean food waste across different staff experience levels. Hence, we fail to reject the null hypothesis, indicating that experience level does not significantly influence food waste.

### 5.1.2. Number of Kitchen Staff and Food Waste

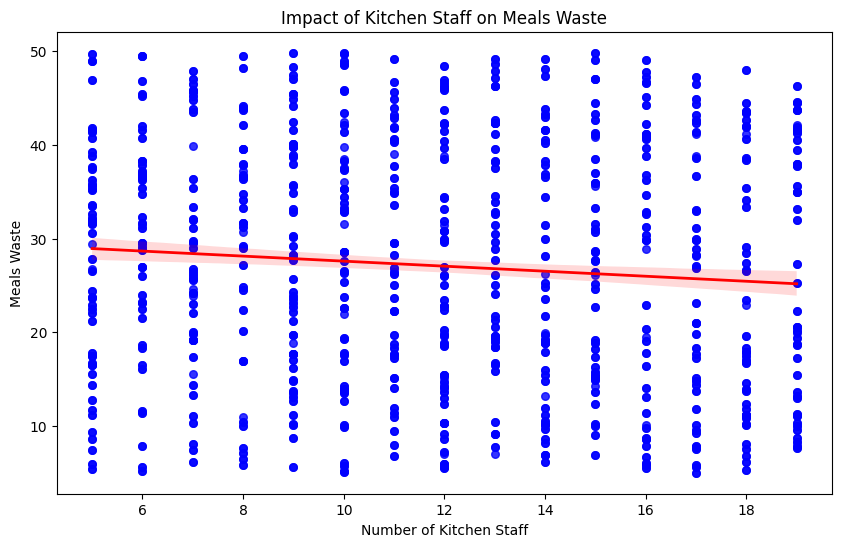
We assessed the linear relationship between the number of kitchen staff and food waste using correlation analysis and regression:

sns.regplot(x=kitchen\_staff, y=meals\_waste)

#### Results

**Correlation Coefficient: -0.0894**

**P-value: 0.0009  
A very weak negative correlation exists between kitchen staff numbers and food waste.  
However, due to the low p-value (< 0.05), we reject the null hypothesis.  
Conclusion: A statistically significant but weak inverse relationship exists—more staff slightly reduces waste.**

****

## 5.2. Special Events and Food Waste

Evaluated whether special events lead to more food waste by comparing event and non-event days using a one-tailed t-test:

from scipy.stats import ttest\_ind  
  
t\_stat, p\_value = ttest\_ind(event\_days, non\_event\_days, equal\_var=False)  
p\_one\_tailed = p\_value / 2

#### Result:

T-statistic: 0.028355043096307465

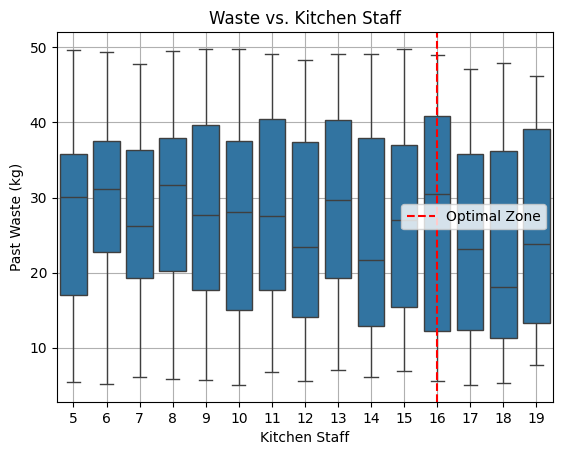
One-tailed P-value: 0.48871091913295395

Null hypothesis Fails

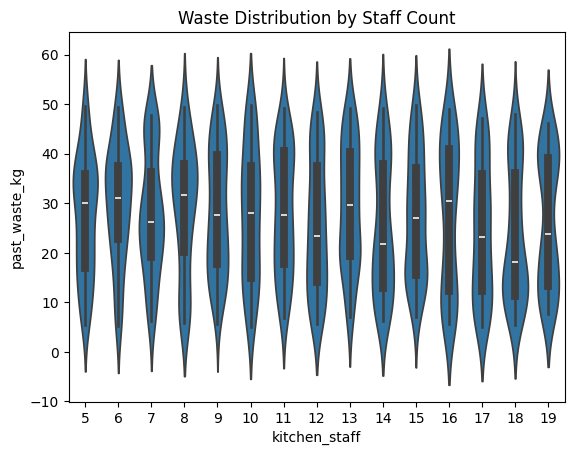
No significant evidence that food waste increases during events

# Key Insights and Recommendations

## 6.1. Staffing Optimization



The boxplot reveals that the kitchen staff size of 16–17 tends to produce lower and more consistent food waste, compared to other staffing levels.



This insight suggests that scheduling staff in this range may help reduce food waste and improve operational efficiency.

## 6.2. Environmental Factors

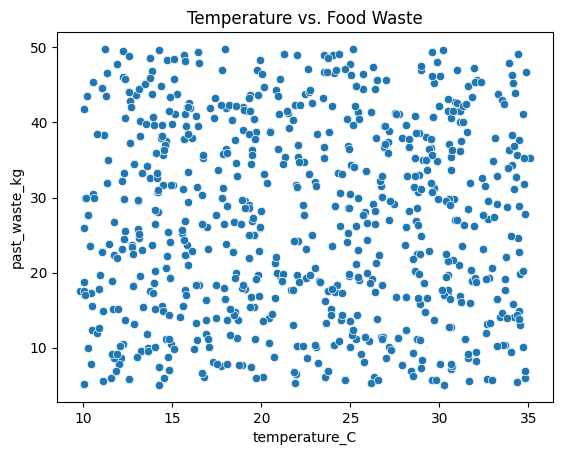
Visual analysis shows that higher temperatures during summer months tend to coincide with increased food waste. This seasonal pattern suggests that despite weak statistical correlation, warmer weather may indirectly lead to more waste, potentially due to faster spoilage or storage stress during hotter periods.

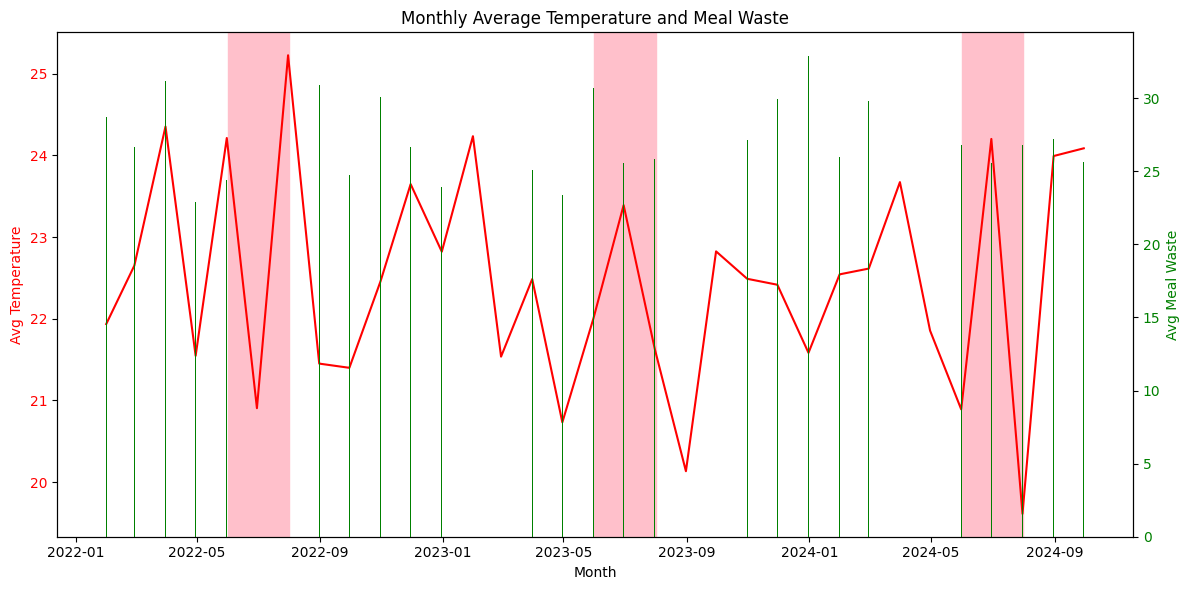
temperature\_C humidity\_percent past\_waste\_kg

temperature\_C 1.000000 -0.000136 -0.003509

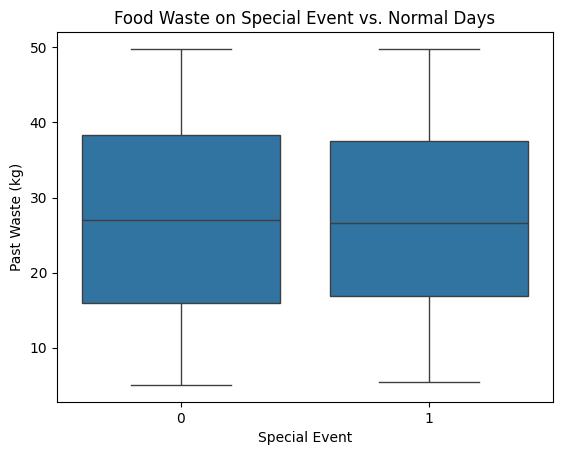
humidity\_percent -0.000136 1.000000 0.024047

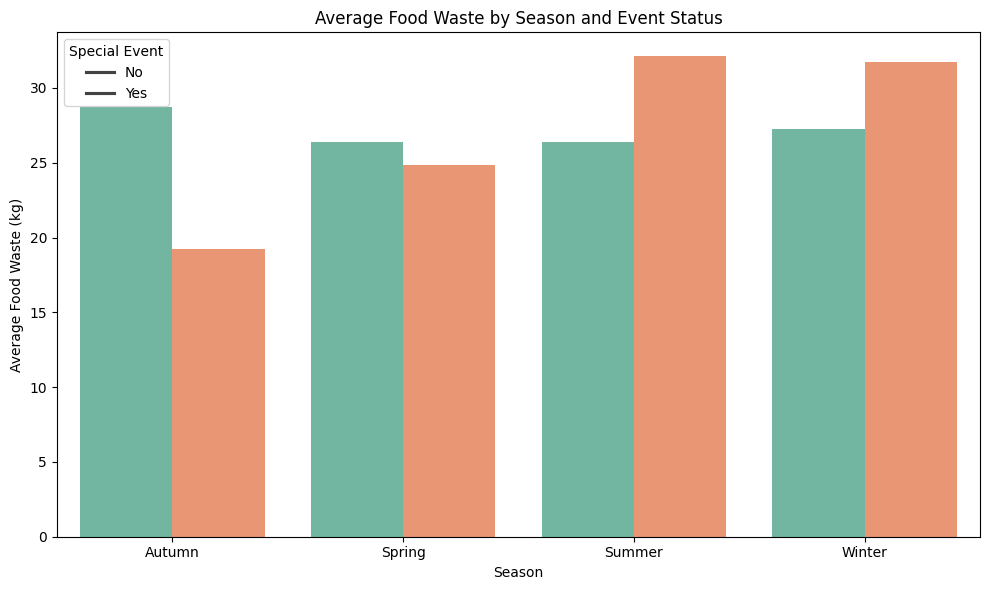
past\_waste\_kg -0.003509 0.024047 1.000000



  
  
Actionable strategies could include:  
- Reducing food prep volume during hot weather  
- Improving cold storage and refrigeration systems  
- Increasing shelf-life monitoring  
- Planning menus with weather-resistant ingredients

## 6.3. Event Management

Special events increase food waste in most seasons, especially summer and winter, likely due to higher turnout or spoilage in warmer months.  
  
Autumn is an exception, where waste is lower on event days — potentially indicating better event management or lower participation during this season.



Recommendations include:

* - Accurate guest forecasting
* - Real-time portion adjustments
* - Food storage enhancements in hot months
* - Post-event donation plans

# Conclusion

## Staffing and Food Waste

* There is a statistically significant but weak negative correlation between the number of kitchen staff and food waste (r ≈ -0.089).
* Optimal staffing levels (around 16–17 kitchen staff) are associated with lower and more consistent waste.

## Staff Experience

* ANOVA testing showed no significant difference in food waste across experience levels (Beginner, Intermediate, Expert, Pro).
* Some high-waste categories (e.g., MEAT, Dairy) are primarily handled by less experienced staff, indicating potential training gaps.

Environmental Factors

* A low to moderate correlation was found between temperature and food waste.
* Summer months consistently show higher waste, likely due to spoilage or storage issues.

## Special Events

* T-tests indicate no significant increase in waste during events.
* Visual analysis shows higher waste in Summer and Winter events, pointing to seasonal influence.

## Limitations

* Correlation does not imply causation; these are observational insights.
* Experience levels were numerically encoded, which may oversimplify real skills.
* Unbalanced sample sizes across experience levels or waste categories may bias results.
* Operational variables like meal type or guest count were not included, limiting contextual insights.

## Recommendations

* Staffing Strategy
  + Schedule 16–17 kitchen staff during peak times to reduce food waste.
  + Maintain balanced staff-to-meal ratios year-round.
* Training and Supervision
  + Provide focused training for less experienced staff handling MEAT and Dairy.
  + Consider mentorship systems between experienced and new staff.
* Seasonal Preparation
  + During summer, optimize storage and monitor spoilage-prone items.
  + Use real-time shelf-life tracking and improve cooling infrastructure.
* Event Waste Planning
  + Forecast attendance more accurately and apply dynamic portion control.
  + Enhance post-event food donation or reuse strategies.

# Appendix