Introduction:

This project focuses on analyzing a food service dataset to gain insights into operational efficiency and food waste management. The dataset includes information such as the number of meals served, kitchen staff count, environmental factors (temperature and humidity), special events, and food waste details categorized by type and past trends. The objective is to clean the data, identify patterns through visualization, handle missing or inconsistent values, detect outliers, and extract actionable insights. These insights aim to help optimize resource allocation, improve kitchen operations, and reduce food waste. The analysis also considers contextual variables like staff experience and special events that may influence food waste levels.

Summary of the dataset and key variables:

The dataset contains the following columns with data type mentioned below:

1. ID(int): A identifier or a primary key that's indicates each record uniquely.

2. date(date-format): The date of the observation.

3. meals_served(Numeric & Continuous) : The number of meals served on that day.

4. kitchen_staff(Numeric & Discrete): The number of kitchen staff working on that day.

5. temperature_C(Numeric & Continuous): The temperature (in Celsius) on the recorded day.

6. humidity_percent(Numeric & Continuous): The humidity percentage on the recorded day.

7. day_of_week (Categorical & Ordinal): The day of the week as a numeric value (0 = Sunday, 1 = Monday, etc.).

8. special_event (Categorical &Nominal): A binary variable indicating whether a special event occurred (1 = event, 0 = no event).

9. past_waste_kg(Numeric & Continuous): The amount of food waste in kilograms from previous days.

10. staff_experience(Categorical & Ordinal): The experience level of the kitchen staff (e.g., "Beginner", "Intermediate").

11. waste_category (Categorical &Nominal): The category of food waste (e.g., "dairy", "meat")

Data Cleaning:

Check Data types inconsistencies:

```
#check data types
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1822 entries, 0 to 1821
Data columns (total 11 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   ID                1822 non-null   int64
 1   date              1822 non-null   object
 2   meals_served      1790 non-null   float64
 3   kitchen_staff     1804 non-null   object
 4   temperature_C     1822 non-null   float64
 5   humidity_percent  1806 non-null   float64
 6   day_of_week       1822 non-null   int64
 7   special_event     1822 non-null   object
 8   past_waste_kg     1806 non-null   float64
 9   staff_experience  1485 non-null   object
 10  waste_category    1801 non-null   object
dtypes: float64(4), int64(2), object(5)
memory usage: 156.7+ KB
```

in which, kitchen staff data types should be numeric so converted it to numeric.

Data types inconsistencies

```
#convert full column in numeric
df['kitchen_staff'] = pd.to_numeric(df['kitchen_staff'].replace({'ten': 10, 'eleven': 11}), errors='coerce')
```

```
#check data types
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1822 entries, 0 to 1821
Data columns (total 11 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   ID                1822 non-null   int64
 1   date              1822 non-null   object
 2   meals_served      1822 non-null   float64
 3   kitchen_staff     1822 non-null   float64
 4   temperature_C     1822 non-null   float64
 5   humidity_percent  1822 non-null   float64
 6   day_of_week       1822 non-null   int64
 7   special_event     1822 non-null   object
 8   past_waste_kg     1822 non-null   float64
 9   staff_experience  1822 non-null   object
 10  waste_category    1822 non-null   object
dtypes: float64(5), int64(2), object(4)
memory usage: 156.7+ KB
```

Now data types issue is resolved.

Handle Missing Values:

```
missing_values=df.isnull().sum()
print("Missing Values", missing_values)

missing_value_percentage= (missing_values/len(df))  * 100
print("Missing values in percentage",missing_value_percentage )


#benchmark is if missingdata> 50% the  drop colum
```

```
Missing Values ID                    0
date                   0
meals_served          32
kitchen_staff         18
temperature_C          0
humidity_percent      16
day_of_week            0
special_event          0
past_waste_kg         16
staff_experience     337
waste_category        21
dtype: int64
Missing values in percentage ID                 0.000000
date                 0.000000
meals_served         1.756312
kitchen_staff        0.987925
temperature_C        0.000000
humidity_percent     0.878156
day_of_week          0.000000
special_event        0.000000
past_waste_kg        0.878156
staff_experience    18.496158
waste_category       1.152580
dtype: float64
```

Now in this picture , we have a lot of missing values in numeric as well as categorical columns, now resolving missing values in each column

Meals Served	Histogram of this numeric continuous column is left skewed then used median in place of missing values.
Kitchen Staff	Histogram of this numeric continuous column is slightly skewed then used mode in place of missing values.
Humidity Percentage	Histogram of this numeric discrete column is slightly skewed then used mode in place of missing values.
Past Waste Kg	Histogram of this numeric continuous column is slightly skewed then used mode in place of missing values.
Staff experience	For this Categorical ordinal column I firstly handle unclear category which pro and converted to Expert then I used “Unknown” in replace of missing values because 337 is a large number so if I added then create large difference between others values
Waste Category	For this Categorical nominal column I used mode in replace of missing values

Duplicates:

```
#finds duplicates
duplicates = df[df.duplicated()]
print(duplicates)
```

```
Empty DataFrame
Columns: [ID, date, meals_served, kitchen_staff, temperature_C, humidity_percent, day_of_week, special_event, past_waste_kg, staff_experience, waste_category]
Index: []
```

Exploratory Data Analysis (EDA)

Summary statistics

```
df.describe()
```

	ID	meals_served	kitchen_staff	temperature_C	humidity_percent	day_of_week	past_waste_kg
count	1822.000000	1822.000000	1822.000000	1822.000000	1822.000000	1822.00000	1822.000000
mean	910.500000	372.327113	11.909989	22.189280	60.522873	3.01427	26.830368
std	526.110413	490.505492	4.269558	8.919939	17.484150	2.00899	12.858876
min	0.000000	100.000000	5.000000	-10.372207	30.121111	0.00000	5.008394
25%	455.250000	212.250000	8.000000	15.684259	45.362854	1.00000	15.565114
50%	910.500000	306.000000	12.000000	22.115040	61.514385	3.00000	26.577480
75%	1365.750000	405.750000	15.000000	28.807494	75.755784	5.00000	37.978663
max	1821.000000	4730.000000	19.000000	60.000000	89.982828	6.00000	49.803703

The average number of meals served is approximately **372**, with a standard deviation of **490**, indicating a high variation across records. The number of kitchen staff range from **5 to 19**, with a mean of around **11.91**, suggesting moderately sized teams. The recorded temperatures range from **-10.37°C to 60°C**, with an average of **22.19°C**, which aligns with typical ambient conditions. Humidity levels have a mean of **60.52%**, reflecting moderately humid environments, while the day of the week (encoded 0–6) shows a fairly even spread. Past waste (in kg) have an average of **26.83 kg**, with a minimum of **5.01 kg** and a maximum of **49.80 kg**, revealing substantial variability in waste production. These insights provide a foundation for identifying trends and patterns in food service operations.

2.2 Visualization:

Histograms:

Meals Served

The histogram for meals served shows a right-skewed distribution, indicating that while most days had a moderate number of meals served, there are some days with exceptionally high values, suggesting occasional high-demand events.

Kitchen Staff

The distribution of kitchen staff appears roughly normal, with most values centered around 12 staff members. This supports the idea that kitchens are generally staffed with medium-sized teams.

Temperature (°C)

The temperature histogram is slightly left-skewed, with most values concentrated between 15°C and 30°C. A few extreme cold and hot days are present but rare.

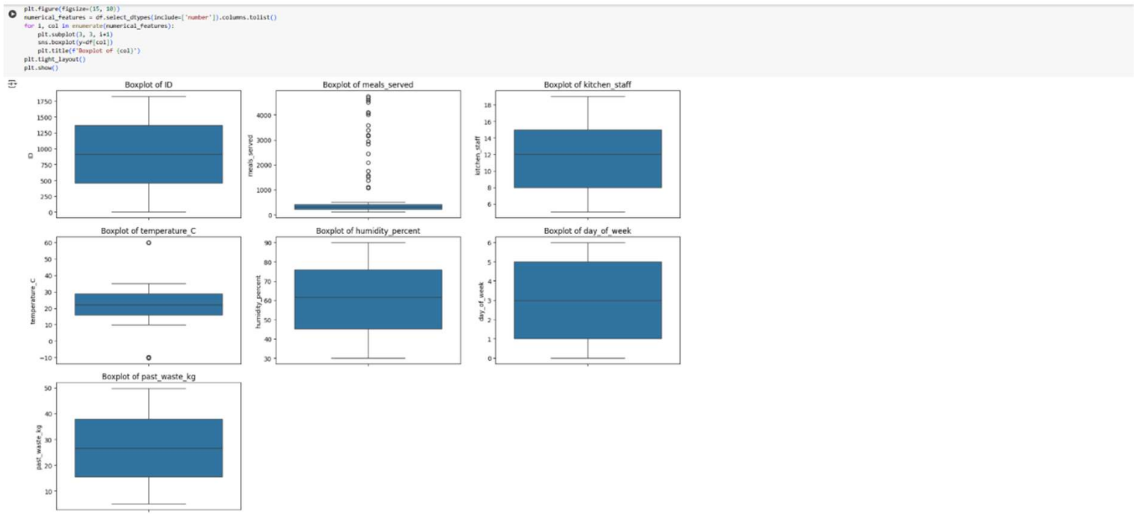
Humidity (%)

Humidity values are fairly normally distributed, clustering around the mean of approximately 60%, indicating stable humidity conditions most of the time.

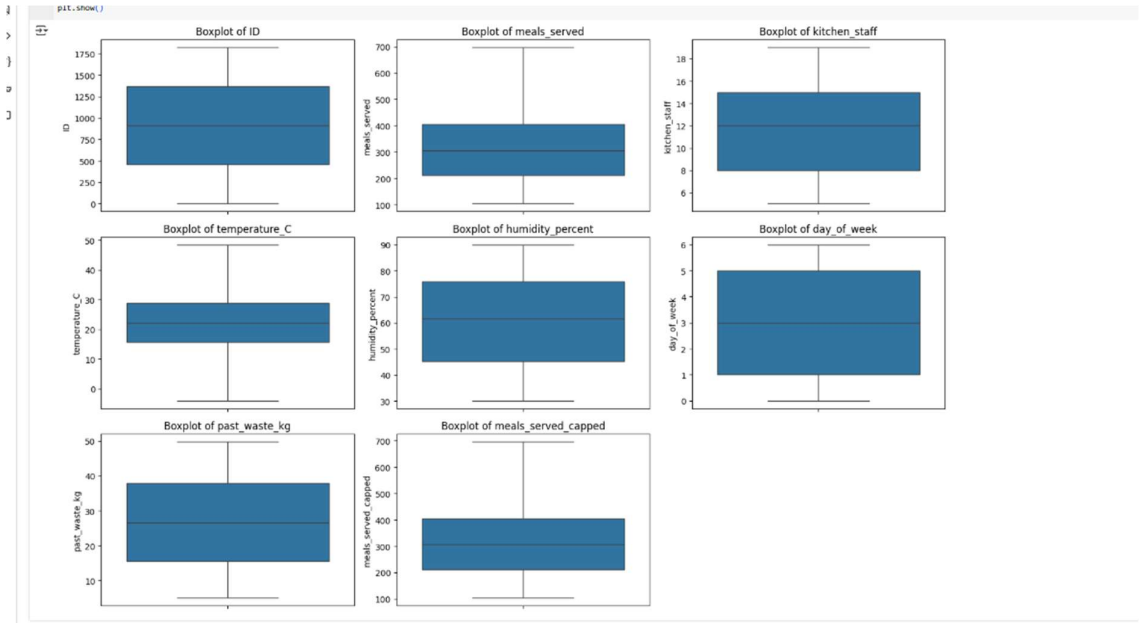
Past Waste (kg)

The past waste histogram shows a fairly wide spread, with some days having significantly higher waste. This may indicate inefficiencies or variations in meal planning and demand forecasting.

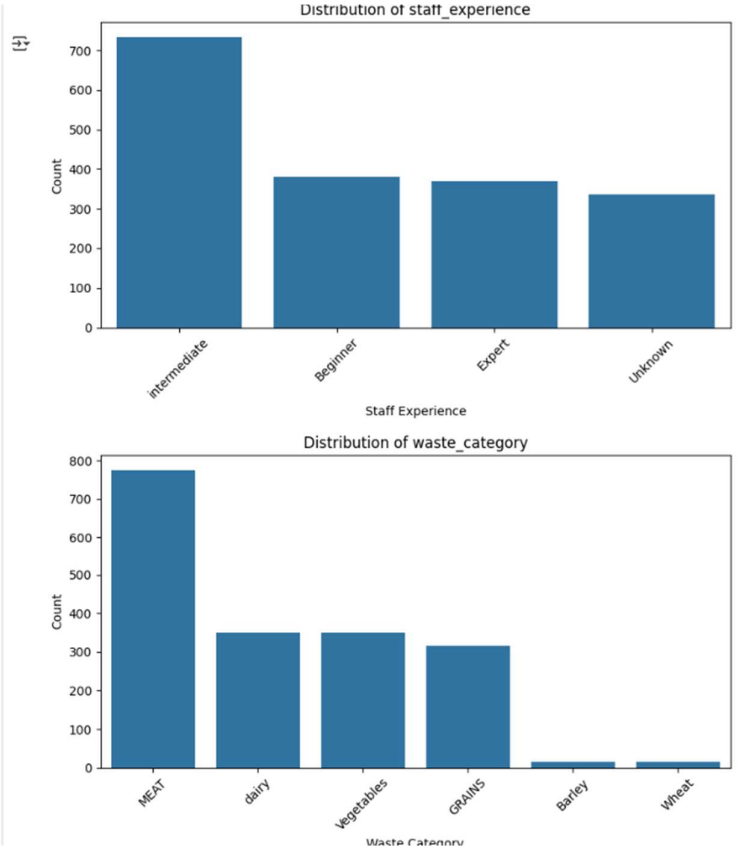
Box plots:



Here I am using cap techniques to remove outliers, huge amount of data will be loss so that's why using cap technique is the best option.


plt.show()
Boxplot of ID
Boxplot of meals_served
Boxplot of kitchen_staff
Boxplot of temperature_C
Boxplot of humidity_percent
Boxplot of day_of_week
Boxplot of past_waste_kg
Boxplot of meals_served_capped

Bar Charts:


Distribution of staff_experience
Count
intermediate
Beginner
Expert
Unknown
Staff Experience
Distribution of waste_category
Count
MEAT
dairy
Vegetables
GRAINS
Barley
Wheat
Waste Category

Explanation:

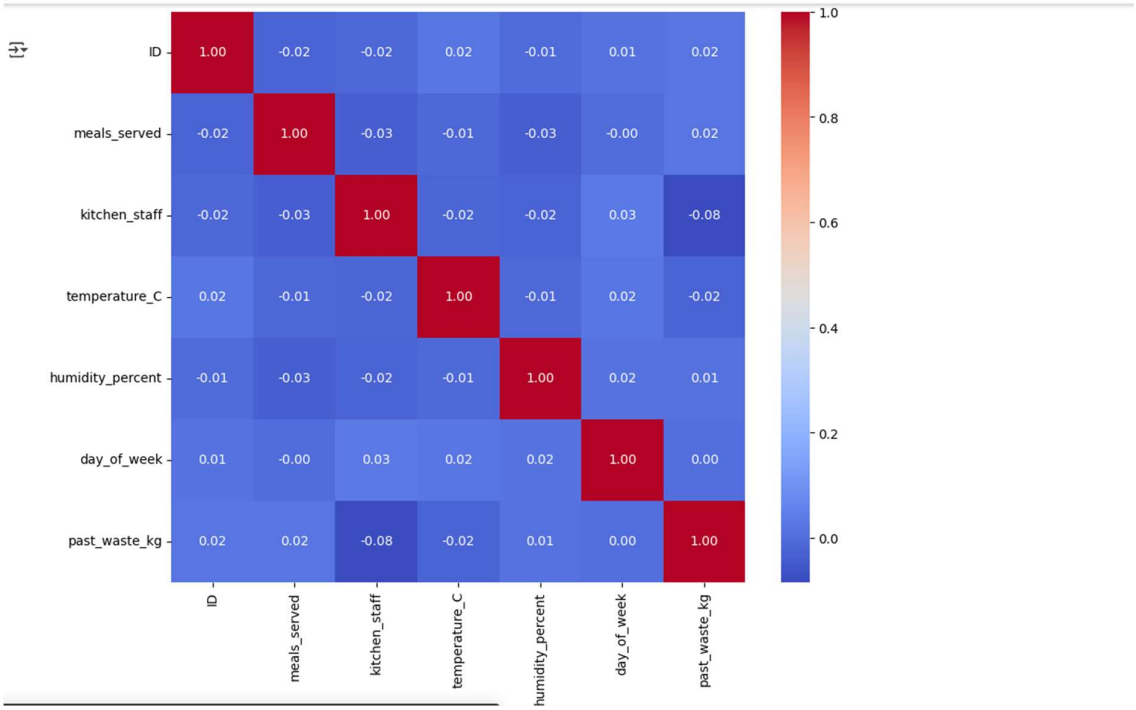
Staff Experience

- The most common experience level is "Intermediate", with over 700 entries.
- "Beginner", "Expert", and "Unknown" categories are more evenly distributed, each with around 350 entries.

Waste Category

- "MEAT" is by far the most frequently recorded waste category, approaching 800 counts.
- Other categories like "dairy", "vegetables", and "GRAINS" are moderately represented.
- Categories such as "Barley" and "Wheat" appear very rarely.

Correlation Analysis:



1. Is there a correlation between the number of meals served and the amount of food waste (past_waste_kg**)?**

- Correlation coefficient: 0.02
- Interpretation: This is a very weak positive correlation**,** essentially negligible. It suggests that there is no meaningful linear relationship between the number of meals served and food waste in this dataset.

2. Does temperature or humidity influence food waste?

- Temperature vs**.** past_waste_kg**:** Correlation = -0.02
- Humidity vs. past_waste_kg: Correlation = 0.01
- Interpretation: Both correlations are very close to zero, meaning temperature and humidity have no significant influence on food waste in this data.

There are no strong or even moderate correlations between food waste (past_waste_kg) and any of the variables listed, including meals served, temperature, or humidity.

Summary of Findings

The dataset displayed significant variability in key metrics such as meals served and past food waste. Data cleaning involved correcting data types, imputing missing values using median, mode, and placeholder categories, and addressing outliers through capping. Exploratory data analysis revealed:

- **Right-skewed** distribution in meals served, suggesting occasional high-demand events.
- **Moderate staffing levels**, with most days staffed by around 12 kitchen personnel.
- **Stable environmental conditions**, with temperature and humidity generally falling within normal ranges.
- **No significant correlations** between food waste and variables such as meals served, temperature, or humidity.

- **Most common waste category** was "MEAT", and "Intermediate" was the most frequent staff experience level.