**Exploratory Data Analysis (EDA) on Food Service Data**

# Introduction:

The objective of this Project was to analyse a food service dataset to gain insights into operational efficiency , food waste management, and environmental or staffing conditions in a kitchen or cafeteria setting.

The dataset captures a wide range of variables, including daily meal counts, staffing levels, weather conditions, staff experience, and special events, as well as detailed food waste data such as quantity and category.

**Dataset Overview**

The dataset contains the following columns:

1. **ID**: A unique identifier for each record.

2. **date**: The date of the observation.

3. **meals\_served**: The number of meals served on that day.

4. **kitchen\_staff**: The number of kitchen staff working on that day.

5. **temperature\_C**: The temperature (in Celsius) on the recorded day.

6. **humidity\_percent**: The humidity percentage on the recorded day.

7. **day\_of\_week**: The day of the week as a numeric value (0 = Sunday, 1 = Monday etc.).

8. **special\_event**: A binary variable indicating whether a special event occurred (1=event,

0= No event)

9. **past\_waste\_kg**: The amount of food waste in kilograms from previous days.

10. **staff\_experience**: The experience level of the kitchen staff(e.g.,"Beginner",

"Intermediate").

11. **waste\_category**: The category of food waste (e.g., "dairy", "meat").

# Data Cleaning

**Fixing Label Inconsistencies in Categorical Data:**

Both the Categorical columns ,had labels/values inconsistencies, with similar terms identified separately despite referring to the same category.

1. **Fixing inconsistent labels in 'Waste\_category'**

* ‘Meat’ had two labels i.e 'MeAt' and 'MEAT.
* Lower and Upper case inconsistencies were present .

To deal with the textual error, we first **standardized the text** to lowercase and stripped

any whitespace. Later **mapped** these cleaned strings to properly capitalized labels

using a self-created dictionary of correct labels.

1. **Fixing inconsistent labels in 'Staff\_Experience'**

* Same concerns were present as in ‘Waste\_category’, repeated labels due to capitalization errors. Used the same approach of standardizing the labels first and replacing them with correct labels using a dictionary.
* Snippet of the code used :

df["staff\_experience"]=df["staff\_experience"].str.strip().str.lower()

correct\_labels1={'beginner':'Beginner','intermediate':'Intermediate','expert':'Expert','pro':'Pro'}

df["staff\_experience"]=df["staff\_experience"].map(correct\_labels1)

**Fixing Label and Data Type Inconsistencies in Numerical Data:**

1. **Fixing inconsistent labels and data type of 'Special\_Event**

* Textual representation of number 1 was present in a numerical column. Replaced ‘One’ with 1.

1. **Data-type correction of 'Kitchen\_staff' column**

* The column consisted of numerical values i.e whole numbers, however the respective data type is "object". This misinterpretation of type arises because of presence of two string values, 'ten' and 'eleven'. Replaced them with its integer representation to fix the error.

1. **Data-type correction of 'Date' to 'datetime' format**

* ‘Date’ data type was set to object by default, converted it to correct ‘date\_time’ format.
* df['date'] = pd.to\_datetime(df['date'])

**Dealing with Missing Values:**

**Check for Missing Values**

Following code was used to check for missing values in all categories

* df.isnull().sum()

**Dealing with Missing Values**

1. *“Meals\_served”* column contained 1.75 % missing values. Plotted a histogram ,that exhibited skewed distribution. Imputed the missing values with column’s median value.

* median\_value = df['meals\_served'].median()
* df['meals\_served'] = df['meals\_served'].fillna(median\_value)

1. All three numerical columns *,"Kitchen\_staff", "Humidity\_percent"* and *"Past\_waste\_kg"* were normally distributed and contained 0.98 %,0.87% and 0.87% missing values respectively. The missing values were replaced with their respective mean values.

* df['past\_waste\_kg'] = df['past\_waste\_kg'].fillna(mean\_p)

1. Both ,*”Staff\_Experience”* & *“Waste\_ Category”* columns have skewed data distribution.

* 18% of missing values in 'Staff\_Experience' is a significant portion. Simply filling it with the mode would have exaggerated existing skew and misled models. Used placeholder ’Unkown’ instead ,to maintain integrity of data.
* 1.1% missing values in 'Waste\_category' could be dealt by replacing it with mode,however since the labels refer to specific food categories, mode could have misled the data. Replaced missing values with 'unknown' instead.
* df['staff\_experience']=df['staff\_experience'].fillna('Unknown')

**Checking for Duplicates:**

There were no duplicate rows in the data frame.

* df.duplicated().sum()

# Exploratory Data Analysis (EDA)

**Summary Statistics**

In built Panda function, Describe, was used to attain statistical summary of all numerical categories.

* numerical\_cols = df.select\_dtypes(include=['number'])
* print("\nStatistical Summary for Numerical Columns:")
* display(numerical\_cols.describe())
* **ID**: A unique identifier for each record. The dataset contains 1,822 entries, ranging from 0 to 1821.
* **date**: Timestamps of data entries ranging from January 1, 2022, to September 26, 2024. The mean timestamp is approximately April 22, 2023.
* **meals\_served**: Number of meals served per record. Values range from 100 to 4,730 with an average of ~372 meals.
* **kitchen\_staff**: Number of kitchen staff, ranging from 5 to 19 with a mean of ~11.9.
* **temperature\_C**: Recorded temperature in Celsius, ranging from -10.37°C to 60.0°C, with a mean of ~22.19°C.
* **humidity\_percent**: Humidity levels in percentage, ranging from ~30.12% to ~89.98%, with an average of ~60.79%.
* **day\_of\_week**: Numerical representation of the day of the week (0 = Sunday, 6 = Saturday), with a mean of ~3.01 indicating the middle of the week.
* **past\_waste\_kg**: Weight of past food waste in kilograms, ranging from ~5.01 kg to ~49.80 kg, with an average of ~26.99 kg.

**Visualizations**

**Histograms:**

A [histogram](https://en.wikipedia.org/wiki/Histogram) is a graphical representation commonly used to visualize the distribution of numerical data.

We plotted histograms for 4 numerical categories, ‘Meals\_Served’, ‘Temperature\_C’, ‘Humidity\_Percent’, and ‘Past\_Waste\_Kg’ .

Here's a breakdown of the distributions shown in the histograms:

1. **Humidity**:

* The data appears to have a fairly uniform distribution.
* Most values fall between **30% and 50% humidity**.
* There’s no extreme peak, meaning humidity is relatively stable .

1. **Past Waste**:

The distribution is **fairly uniform**, meaning waste levels do not concentrate in any

particular range.

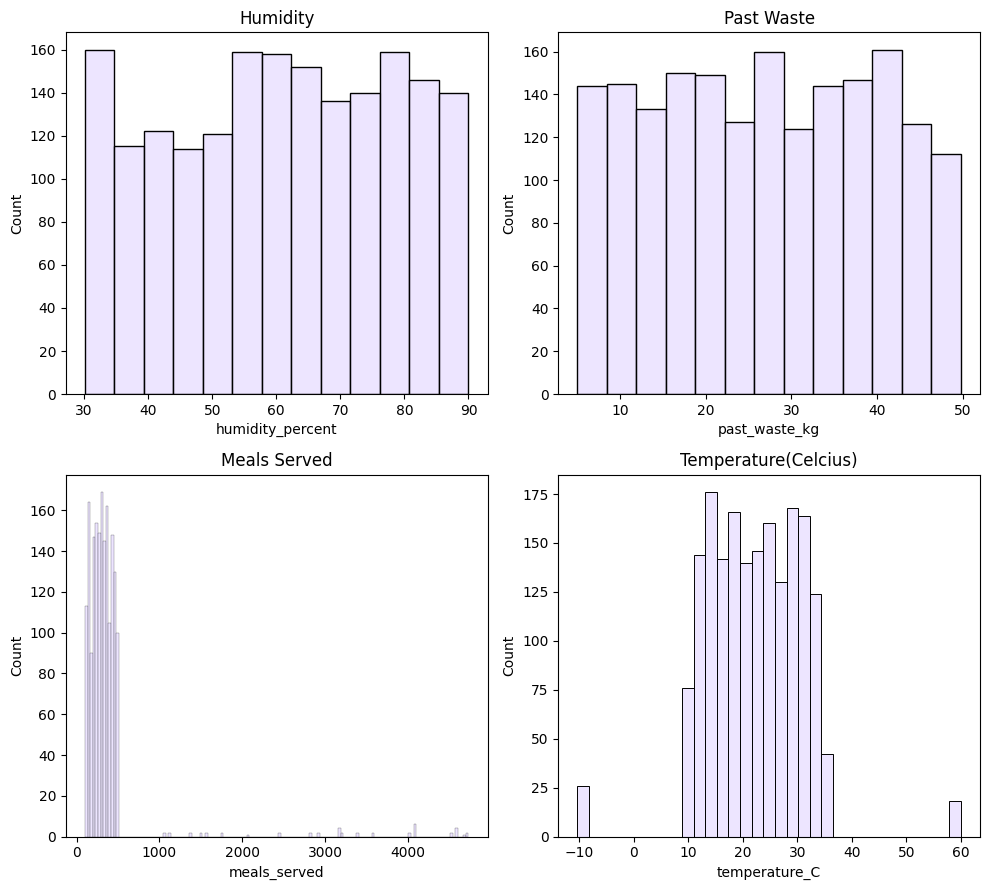
* Most waste values range between 30 kg and 50 kg .
* No strong skew or clustering is observed.

1. **Meals Served**:

* The distribution is highly skewed to the left.
* Most data points fall between 0 and 1000 meals served, indicating fewer instances of exceptionally high meal counts.
* There are very few cases where more than 1000 meals were served.

1. **Temperature (Celsius)**:

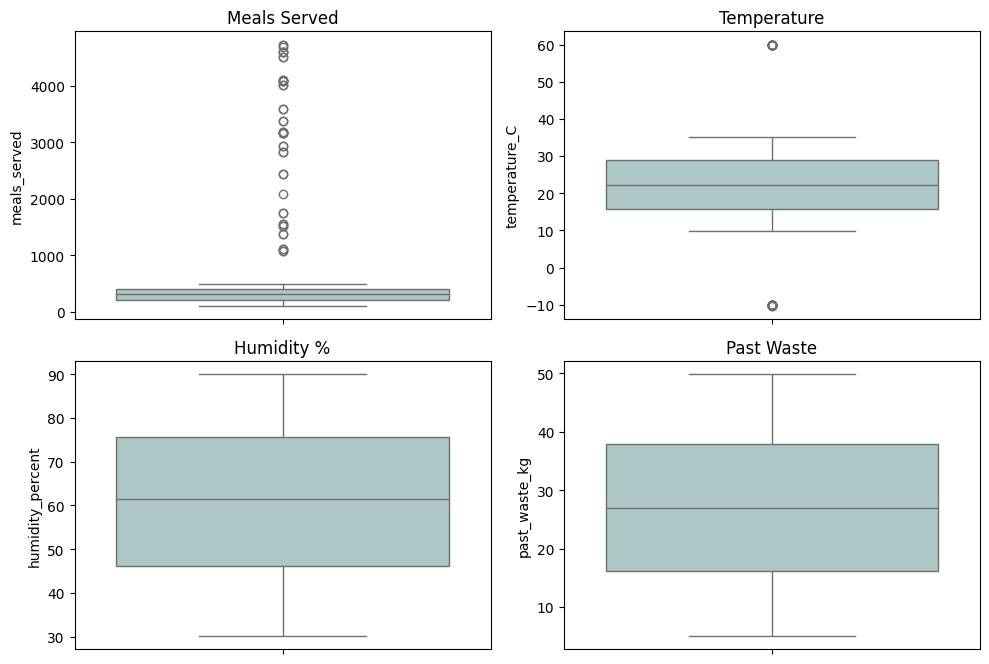
* The distribution resembles a bell shape, indicating a normal distribution.
* Most temperatures range between 10°C and 30°C, suggesting a common temperature range.
* Extreme values (below -10°C and above 50°C) are rare.



**Boxplots:**

Plotted box plots to visualize outliers in numerical categories.

**Meals Served** plot shows a large spread, with quite a few extreme values—suggesting occasional high demand spikes. Meanwhile, **Temperature** ranges significantly, but most readings seem clustered around comfortable levels. **Humidity** stays relatively stable, with fewer outliers, while **Past Waste** trends between 20 kg and 40 kg, possibly indicating a consistent pattern.



**Dealing with outliers:**

1. **Meals Served**

* A significant number of data points for meals\_served were above 1000, it's likely that those weren't errors but actual high-volume days. Removing or capping them might have thrown away important data input.
* Used log transformation to compress the high-end values while keeping relative differences.
* df['meals\_served'] = np.log1p(df['meals\_served'])

1. **Temprature(Celcius)**

* Extreme values (like the one around –10°C or 60°C ) are likely errors or rare cases not common in most real-world ambient conditions.
* Clipped the outliers with lower and upper bound values to maintain data integrity.
* Q1 = df['temperature\_C'].quantile(0.25)

    Q3 = df['temperature\_C'].quantile(0.75)

  IQR = Q3 - Q1

    upper\_bound=Q3+1.5\*IQR

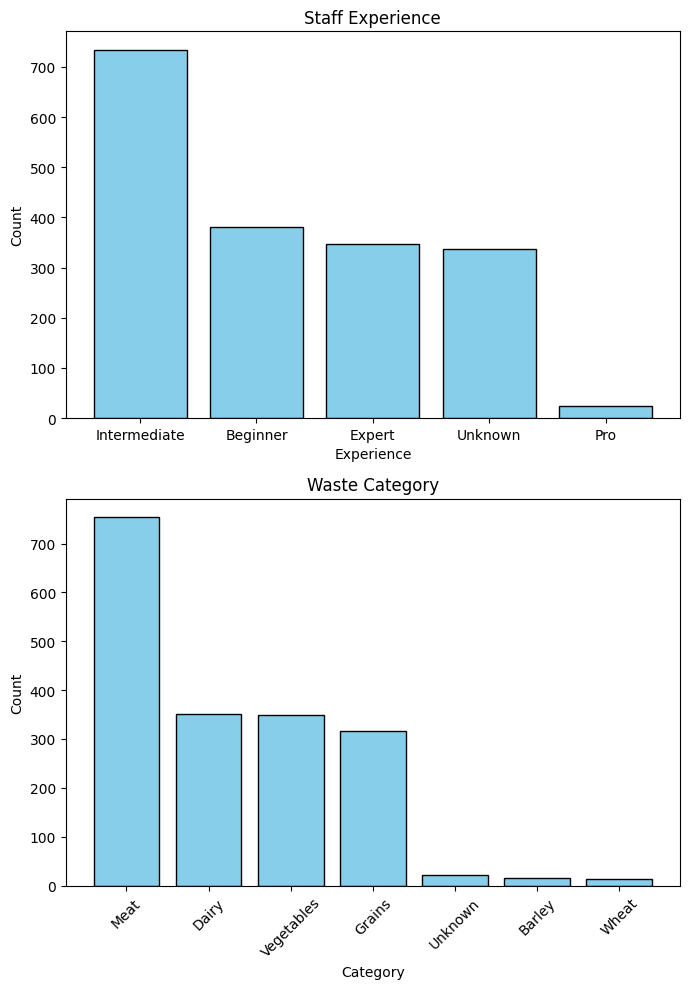
    lower\_bound=Q1-1.5\*IQR

    df['temperature\_C'] = df['temperature\_C'].clip(lower\_bound, upper\_bound)

**Bar Plots**

Bar plot is a recommended visualisation for categorical features. Constructed them for the following two categorical columns:

* Staff\_Experience
* Waste\_Category



**Staff Experience**

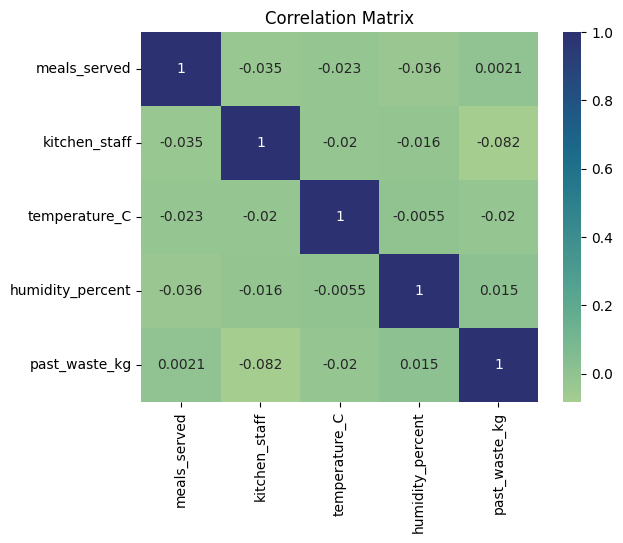
The data indicates that **Intermediate** experience is the most common, followed by **Beginner**, then **Expert**. Interestingly, a significant portion of staff members have **Unknown** experience levels, which might indicate incomplete records .

**Waste Category**

The **Meat** category has the highest recorded waste, which could indicate over-purchasing, wastage, or specific consumption patterns. **Dairy** and **Vegetables** also contribute a notable amount. Meanwhile, **Grains** make up a smaller proportion, and **Barley** and **Wheat** waste are minimal.

## Correlation Analysis

A heatmap is an ideal method to visualize correlation across your variables. By analysing the relationships between **meals served, temperature, humidity, staff experience, and food waste**, we c identified key patterns that influenced food consumption and waste levels.



Most of the correlations seemed weak, with values close to zero. That suggests that none of these variables strongly influence each other. However, there is a slightly negative correlation (-0.082) between kitchen staff and past waste, implying that a higher number of kitchen staff might be associated with a small reduction in food waste.

## Hypothesis Testing

**Impact of Kitchen Staff on Food Waste**

We tested whether the number of kitchen staff affects the amount of food waste for the following two hypothesis:

● **Null hypothesis (H0):** There is no relationship between the number of kitchen staff and food waste.

● **Alternative hypothesis (H1):** The number of kitchen staff significantly affects food waste.

**Approach Used**

* Grouped data, count of kitchen Staff, into three categories (Low/Medium/High staff)
* Compared their means using ANOVA to see if differences were real or random.

def categorize\_staff(n):

    if n <= 5:

        return 'Low'

    elif n <= 10:

        return 'Medium'

    else:

        return 'High'

df['staff\_level'] = df['kitchen\_staff'].apply(categorize\_staff)

import scipy.stats as stats

low = df[df['staff\_level'] == 'Low']['past\_waste\_kg']

medium = df[df['staff\_level'] == 'Medium']['past\_waste\_kg']

high = df[df['staff\_level'] == 'High']['past\_waste\_kg']

# Perform one-way ANOVA

f\_stat, p\_value = stats.f\_oneway(low, medium, high)

print(f"F-statistic: {f\_stat:.2f}")

print(f"P-value: {p\_value:.4f}")

**Conclusion**

* F-statistic = 4.16 means there's some variation in food waste between the groups.
* P-value is less than 0.05, which means the result is statistically significant

Rejecting the null hypothesis (H₀).

* There is a difference in average food waste between at least one of the groups (Low, Medium, or High).

**Special Events and Food Waste**

We tested whether food waste increased during special events:

* **Null hypothesis (H0):** There is no difference in food waste during special event days and no special event days.
* **Alternative hypothesis (H1):** Food waste is more on special event days.

We performed a t-test comparing the average food waste on days with and without special event.

from scipy.stats import ttest\_ind

event\_days = df[df['special\_event'] == '1']['past\_waste\_kg']

non\_event\_days = df[df['special\_event'] == '0']['past\_waste\_kg']

t\_stat, p\_value = ttest\_ind(event\_days, non\_event\_days, alternative='greater')

print(f"T-statistic: {t\_stat:.2f}")

print(f"P-value: {p\_value:.4f}")

**Conclusion:**

* T-statistic = 0.28 states that the difference between group means is small.
* P-value = 0.3881 is much greater than 0.05, which means the result is not significant.
* Null Hypothesis seemed valid here, thus proving that special event days have no effect on food waste.

# Key Insights and Recommendations:

**Key Insights**

1. ***Environmental Factors:***

* No Clear Trend was observed between *Humidity* and *Past Waste*,thus indicating no obvious correlation between humidity and food wastage.
* The *Temperature* gradient suggests that warmer temperatures and cooler temperatures are evenly distributed across all *Humidity levels*. This implies that temperature does not significantly impact food waste in a significant manner.

1. ***Event Management:***

* There is no significant difference between *Average meals* served on event and non-event days.
* The clustering of outliers on the right suggests that a few days had unusually high meals, probably big events or peak days.
* Significant surge in *Meal count* was observed during weekends.

1. ***Staffing Optimization***

* The correlation between kitchen staff and meals served is towards the lower value, indiacting that the number of meals prepared does not scale directly with staff numbers
* The median waste remains around 30 kg across all staff levels, suggesting that kitchen staffing does not strongly influence food wastage.

**Recommendations**

1. ***Staffing Adjustments Might Not Reduce Waste***: Since staff level doesn't strongly impact waste, optimizing staff based on efficiency and experience rather than waste reduction may be more beneficial.
2. ***Investigate Factors contributing in Food wastage***: Factors such as meal demand , inventory management, and special events might contribute more to waste variations than kitchen staff count.
3. ***Demand Forecasting & Meal Planning***: Use historical meal service trends (day of the week, special events, seasonal patterns) to predict demand more accurately.
4. ***Explore Waste Reduction Strategies***: Since waste levels remain similar across staff levels, efforts should be focused on improving menu planning, portion control, and food repurposing methods to minimize wastage. Professional with expertise in professional kitchen are required to be hired to better deal with food flow and waste output.
5. ***Promotions***
   * Offer discounts on frequently ordered meals to reduce waste and attract customers.
   * Create loyalty programs rewarding returning customers with discounts or free meals.
6. ***Implementation of automated tracking systems***

Implement automated tracking systems for meal production and waste to refine

operational strategies.

1. ***Optimising food storage methods*** Ensure proper food storage to minimize spoilage, especially in extreme temperature conditions. A market research might provide productive methods of storage and inventory management.

# Conclusion:

The correlation analysis reveals that environmental factors such as humidity and temperature do not significantly influence food waste levels, while event management and staffing have limited impact on meal production efficiency. The surge in meal counts during weekends highlights demand fluctuations that could be leveraged for productive forecasting and meal planning.

To optimize operations, emphasize on strategic staffing decisions, forecasting, automated tracking systems, and waste reduction initiatives. By refining menu planning, storage methods, food waste can be minimized while improving profitability. Implementing these solutions will enable a more sustainable and data-driven approach to food service management, ensuring that resources are utilized effectively and waste is significantly reduced.