Project Report: Exploratory Data Analysis (EDA) on Food Service Data

1. Introduction

This project aims to analyze a food service dataset to uncover insights into operational efficiency and food waste management. The dataset includes variables such as the number of meals served, kitchen staff, environmental conditions (temperature and humidity), day of the week, special events, past food waste, staff experience, and waste categories. The goal is to clean the data, explore patterns, test hypotheses, and provide actionable recommendations to optimize operations and reduce food waste.

1.1. Data Overview and Inspection

In this section, we perform basic exploratory steps to understand the structure of the dataset, including data types, missing values, duplicate entries, and summary statistics for numerical columns. The following commands are used:

- **df.head()** – Displays the first five rows to get an initial view of the data.
- **df.columns** – Lists all column names to understand available features.
- **df.shape** – Returns the number of rows and columns, giving a sense of dataset size.
- **df.info()** – Provides data types, non-null counts, and memory usage to assess structure and missing values.
- **df.isnull().sum()** – Shows the total number of missing values in each column.

1.2. Data Type Correction and Conversion

After inspecting the dataset, we identified data type issues in three columns. Accurate data types are essential for correct analysis, so we proceed with necessary conversions.

- The **date** column needs to be converted to datetime format.
- The **kitchen_staff** and **special_event** columns should be integers, but they are currently of object type due to inconsistent entries (e.g., 'ten' instead of 10, 'eleven' instead of 11, and 'One ' instead of 1).

To address this, we first clean the inconsistent text values, then convert the columns to the appropriate integer data type.

2. Data Cleaning

The dataset consists of **1,822 entries** and **11 columns**. The cleaning process primarily focused on handling missing values and ensuring data consistency. Below is a summary of the approach used:

2.1 Missing Values:

Below table shows the percentage of missing values:

Column	Missing %
meal_served	1.76
kitchen_staff	0.99
humidity_percent	0.88
past_wastage_kg	0.88
staff_experience	18.50
wastage_category	1.15

2.1.1 Numerical Columns

- **meals_served**: Imputed using the **median** due to **skewed distribution**.
- **kitchen_staff**: Imputed using the **median** because there is no of staff member and it cannot be in point values.
- **humidity_percent** and **past_waste_kg**: Filled using the **mean** due to **normal distribution** and minimal missing values.

2.1.2 Categorical Columns

- **staff_experience**: Had ~18.5% missing values. These were filled with "Unknown Staff" to retain the row and introduce a new category.
- **waste_category**: Had only ~1.15% missing values. This was filled using the **mode** as the most frequent and representative category.

2.2 Duplicates: No duplicate rows were identified in the dataset.

3. Exploratory Data Analysis (EDA)

3.1. Summary Statistics

The dataset's numerical columns were analyzed:

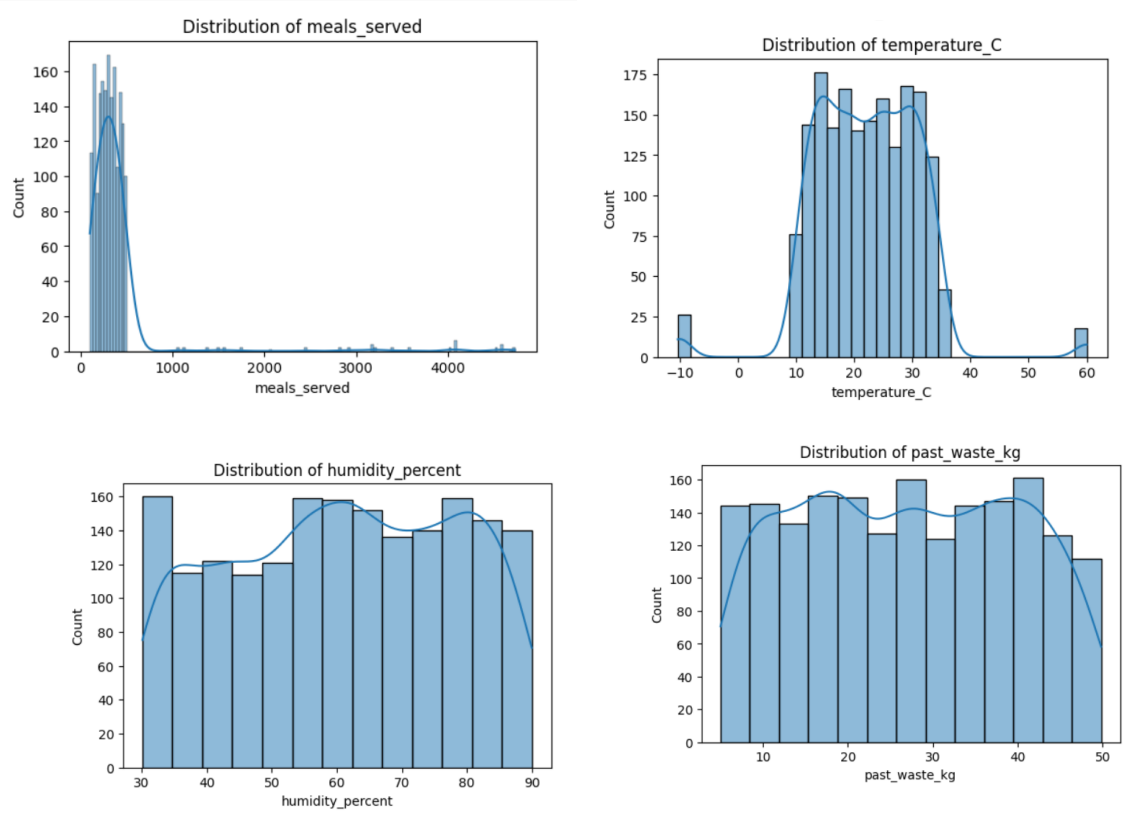
- **meals_served:** Mean = 373.51, Median = 306, Std = 494.79, Min = 100, Max = 4730.
- **kitchen_staff:** Mean = 22.18, Median = 22, Std = 8.91, Min = 5, Max = 60.
- **temperature_C:** Mean = 22.18, Median = 22.15, Std = 8.91, Min = -10.37, Max = 60.
- **humidity_percent:** Mean = 60.79, Median = 61.63, Std = 17.32, Min = 30.12, Max = 89.82.
- **past_waste_kg:** Mean = 26.99, Median = 26.82, Std = 12.79, Min = 5.00, Max = 49.80.

The wide range in meals_served and the presence of negative temperatures indicate potential outliers or data entry errors.

3.2. Visualizations

3.2.1 Histograms

Histograms are used to visualize the distribution of numerical variables. They help us understand the spread, skewness, and presence of outliers in the data. This is important for selecting appropriate statistical methods (mean vs. median), transformations, or identifying potential anomalies. The following are the histograms result:

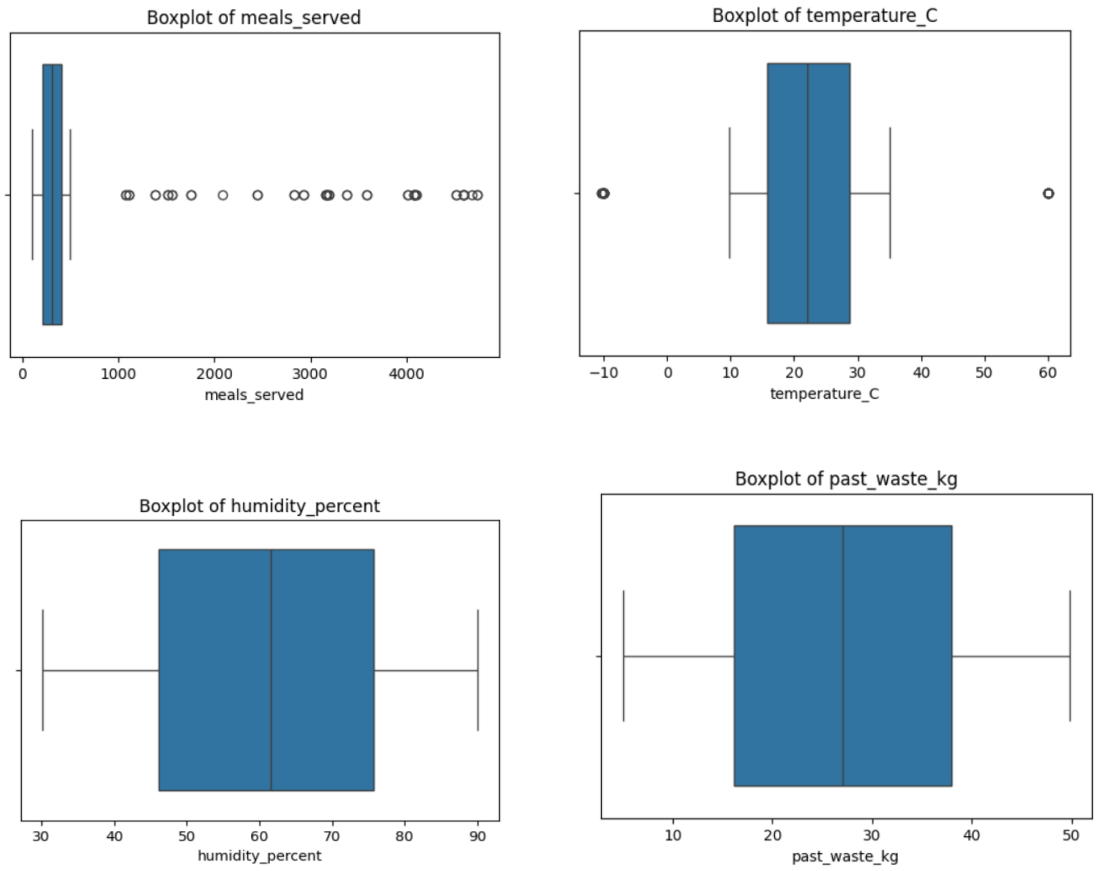


3.2.1.2 Histogram Analysis

The histogram plots indicate that **meals_served** is **right-skewed**, suggesting the presence of higher values or outliers. In contrast, the distributions of **temperature_C**, **humidity_percent**, and **past_waste_kg** appear to be **approximately normal**, indicating a balanced spread around the mean.

3.2.2 Box Plots

Box plots (also known as **box-and-whisker plots**) are useful for visualizing the spread and skewness of numerical data, as well as identifying outliers.
They display the median, interquartile range (IQR), and any data points that fall significantly outside the normal range (outliers shown as individual dots).



3.2.2.1 Box Plot Insights

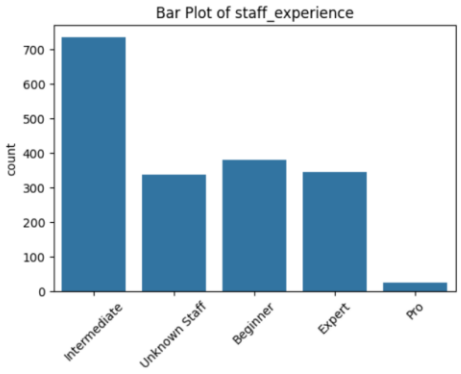
The box plot for **meals_served** reveals a significant number of outliers — **46 in total** — indicating unusually high values. To handle this, we **capped the values at the upper bound** to reduce the impact of extreme data points.

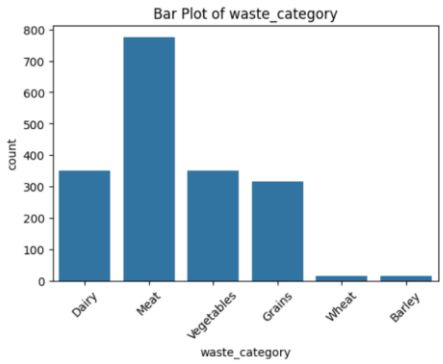
For **temperature_C**, only **2 mild outliers** were observed, which were retained as they appear to be **naturally occurring values**.

No outliers were detected in **humidity_percent** and **past_waste_kg**, suggesting a consistent distribution in those variables.

3.2.3 Bar charts

Bar charts are used to visualize the frequency or count of categories in categorical variables. They help identify dominant categories**,** imbalances**, or** patterns across groups.





3.2.3.1 Bar Chart Observations

- **Staff Experience**
  - **Intermediate** staff dominate the workforce, with over **700 entries**, making them the most common experience level.
  - **Beginner** and **Expert** categories are nearly equal, each contributing around **400 entries**, indicating a balanced mix.
  - **Unknown Staff** still represents a significant group, suggesting some data quality gaps or unrecorded entries.
  - **Pro** level staff are extremely rare, indicating either limited hiring of highly experienced professionals or data entry inconsistencies in that category.
- **Waste_Category**
  - **Dairy, Vegetables** and **Grains** have similar levels of waste, each contributing around **350 entries**, suggesting they are also regularly used and wasted..
  - **Meat** is the most commonly wasted category, with over **700 entries**, indicating it may be a key focus area for waste reduction.
  - **Wheat** and **Barley** appear to be **rarely used or wasted**, with very few entries, possibly due to limited menu use or better inventory control.

4. Correlation Analysis

To examine relationships between numerical variables, a correlation **heatmap** was created, followed by **scatter plots** and **statistical tests** (P-value) for deeper insight.

4.1 Key Observations from Heatmap:

- No strong correlations were found; all coefficients were below |0.5|.
- **meals_served** and **past_waste_kg** showed a weak positive correlation (0.042).
- **temperature_C** and **past_waste_kg** had a very weak negative correlation (-0.021).
- **day_of_week** and **special_event** also showed a very weak negative correlation (-0.041).

4.2 Exploring Research Questions:

1. Is there a correlation between the number of meals served and food waste?
    - Scatter plot indicated no strong visual trend.
    - However, a P-value test showed:
        - Correlation: 0.0417
        - P-value: 0.075
    - Although the correlation is weak, the P-value is slightly above 0.05, suggesting the relationship is not statistically significant at the 5% level, but may warrant further investigation.
2. Does temperature or humidity influence food waste?
    - Temperature vs. food waste:
        - Correlation: -0.0214
        - P-value: 0.361
        - The P-value is much higher than 0.05, confirming no statistically significant relationship.
    - Humidity was not significant in correlation nor tested further due to lack of visual trend.

In final word we can say that **Heatmap** and **scatter plots** suggest **no strong linear relationships**. **P-value** tests support the visual findings, indicating **meals served** may have a **weak association** with **food waste**, but **temperature does not**.

5. Hypothesis Testing

5.1. Impact of Kitchen Staff on Food Waste

- **Hypothesis**:
  - **Null (H0):** No relationship between kitchen staff and food waste.
  - **Alternative (H1):** Kitchen staff significantly affects food waste.
- **Test**: Pearson correlation.
- **Result**: Correlation = -0.082, p-value = 0.0004.
- **Conclusion**: Since the p-value < 0.05, we reject the null hypothesis (H_0). The result shows a statistically significant but weak negative correlation, suggesting that having more kitchen staff slightly reduces food waste.

5.2. Special Events and Food Waste

- **Hypothesis**:
  - **Null (H0):** No difference in food waste between special event and non-event days.
  - **Alternative (H1):** Food waste is higher on special event days.
- **Test**: Independent t-test.
- **Result**: T-statistic = 0.28, p-value = 0.77.
- **Conclusion**: Since the p-value > 0.05, we fail to reject the null hypothesis (H_0). This indicates that there is no statistically significant difference in food waste between event and non-event days.

6. Key Insights and Recommendations

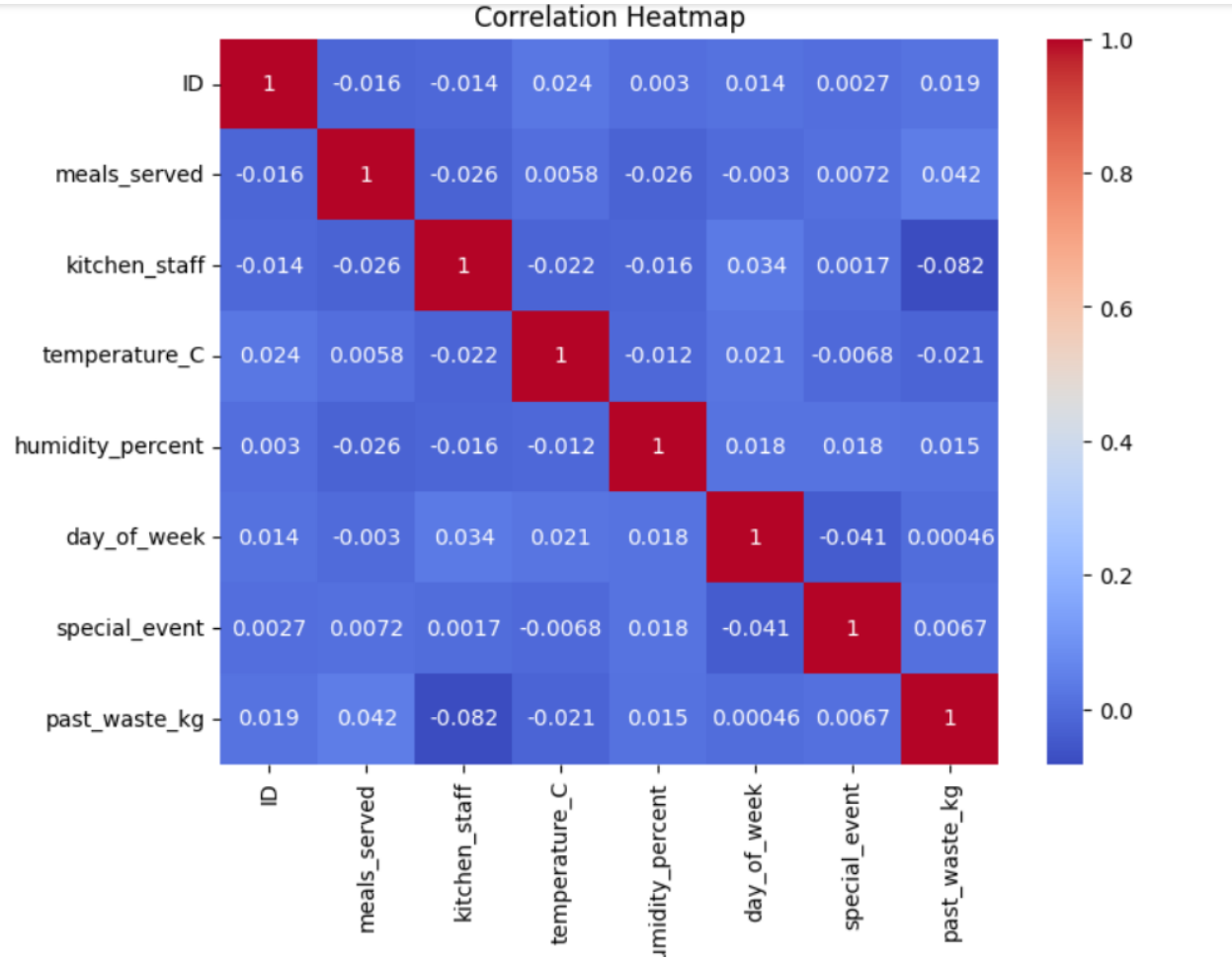
- **Staffing Optimization**: The weak negative correlation between kitchen_staff and past_waste_kg suggests that increasing staff slightly reduces waste. Optimize staffing levels to balance operational efficiency and waste reduction.
- **Environmental Factors**: Temperature and humidity have no significant impact on food waste, so no adjustments are needed based on weather conditions.
- **Event Management**: Special events do not significantly increase food waste, indicating effective event planning. However, monitoring portion sizes during events can further minimize waste.

7. Conclusion

This EDA revealed weak relationships between variables like kitchen staff and food waste, with no significant impact from special events or environmental factors. Recommendations include optimizing staffing and monitoring event planning. Limitations include the **dataset's small size** and potential **unrecorded variables** (e.g., menu types) affecting waste. Future analysis could explore seasonal trends or incorporate additional variables.

7. Appendix

7.1 Heat map



7.2 Scatter Plots