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

Submitted by: Bilal Ahmed

Introduction

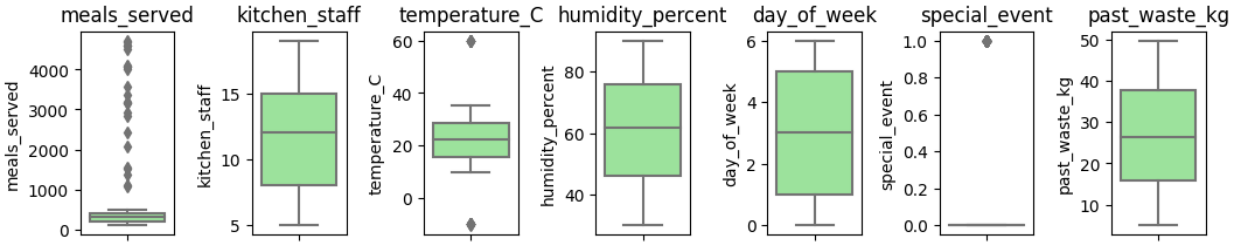
The goal of this project is to explore a food service dataset to understand how to improve efficiency and reduce food waste. The data includes things like the number of meals served, kitchen staff, temperature, humidity, and the amount of food wasted.

My project focuses on studying the dataset and explore different relationships while also making some new features using the given dataset. It is important to analyze this dataset so better strategies regarding staffing optimization, environmental factors and event management can be explored.

Data Cleaning

The given dataset originally had 1822 records with 11 variables. For data cleaning, the following steps were taken:

1. **Coherent dataset entries:** I transformed each string value of the dataset to lowercase.
2. **Duplicated rows:** I removed the unique identifier column so duplicated entries can be found out. I got 768 duplicated entries which I removed from my dataset.
3. **Missing values:** I coded a loop, that ran on each variable and gave me percentage of unique entries in each column. At this stage, I excluded, date variable from the loop. Also, it is when I noticed that some entries, such as 11, are also written as “eleven” in the column. I rectified the string entries and converted them into integers. If missing values were negligible, they were removed from the dataset. There was this one variable, staff experience, for which I imputed the missing values with the mode of the column as missing values percentage was relatively higher in this column.
4. **Consistent data types:** At this stage, I found out that some variables such as meals served and kitchen staff had entries in float. That is not possible as each staff member can never be in fractions as well as meal served are whole proper servings. Date variable was typecasted into dateframe and all other variables were typecasted into their respective datatypes.
5. **Outliers:** Only temperature was the variable which was clipped with its max and min value. Outliers for meals served were retained. Also, special event was a binary variable so it was retained too.



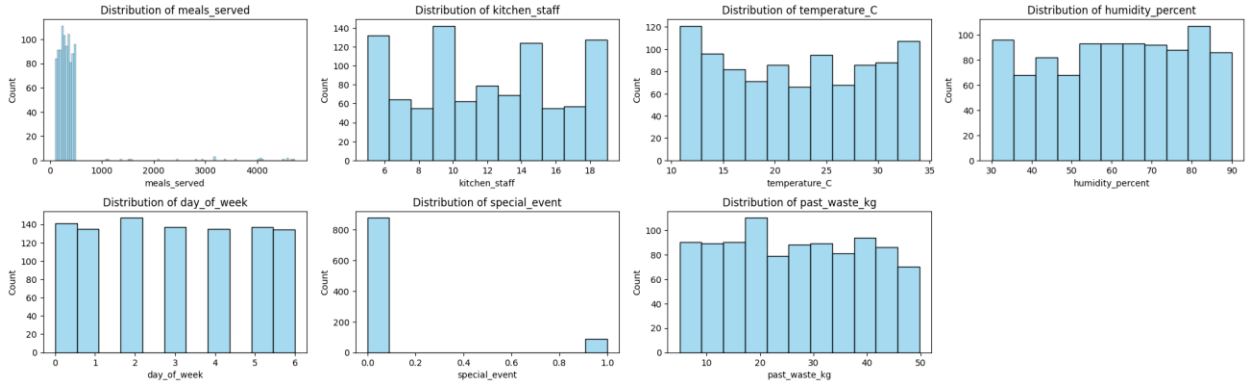
After data cleaning, I was left with only 966 records.

Exploratory Data Analysis

1. **Numeric Columns:** For exploratory data analysis, I developed a summary statistics table that gave me information in the form of numbers of the distribution of my dataset. The following is the table that only includes numeric columns of my dataset.

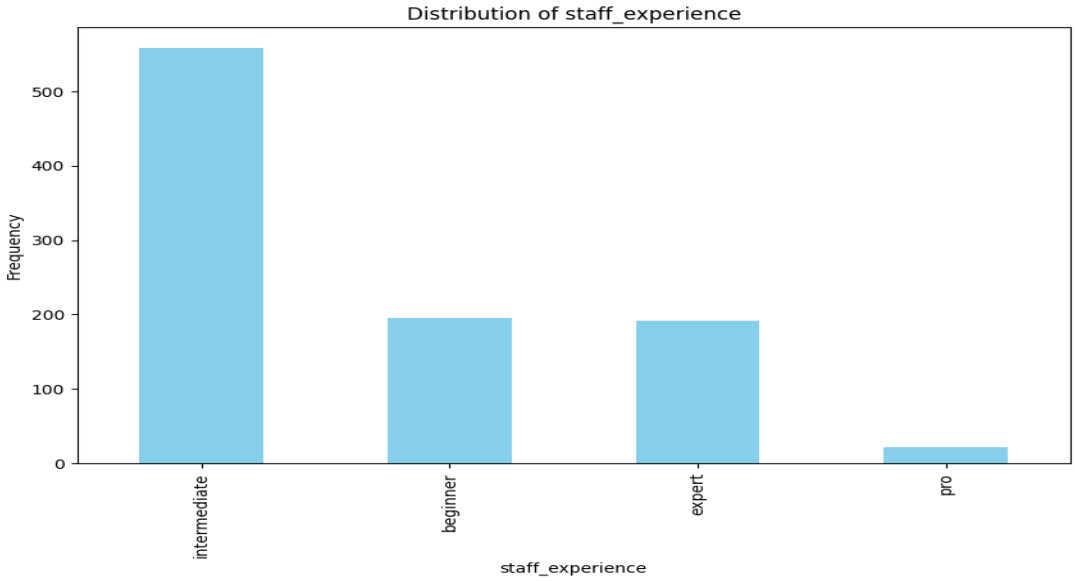
	meals_served	kitchen_staff	temperature_C	humidity_percent	day_of_week	special_event	past_waste_kg
count	966.0	966.0	966.000000	966.000000	966.0	966.0	966.000000
mean	369.3147	11.888199	22.222284	61.039105	2.969979	0.092133	26.737669
std	487.354653	4.284195	7.259670	17.343847	1.995884	0.289363	12.747442
min	100.0	5.0	10.795566	30.121111	0.0	0.0	5.008394
25%	207.0	8.0	15.700953	46.138856	1.0	0.0	15.891045
50%	303.5	12.0	22.094587	61.905365	3.0	0.0	26.458847
75%	406.0	15.0	28.760838	75.866047	5.0	0.0	37.885228
max	4730.0	19.0	33.997903	89.982828	6.0	1.0	49.803703

For the same, I developed visualizations too which can be found below:



If looked at the first distribution of meals served, we see that the data is skewed which is fairly understandable that higher count of servings happen relatively low as compared to low no. of meal servings. The other distributions are pretty much straightforward.

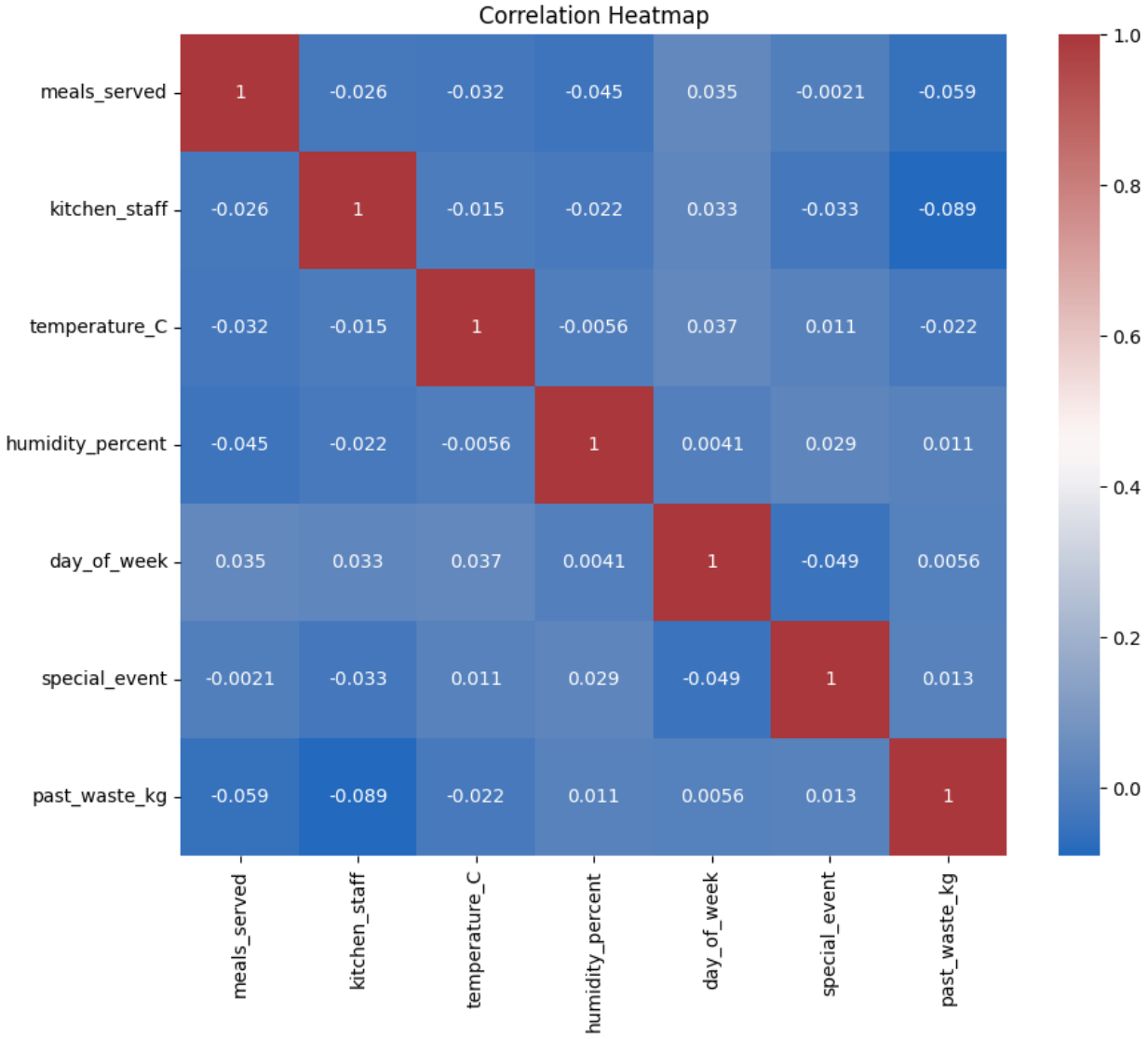
2. **Categorical columns:** The categorical columns, staff experience and waste category suggested that most of the staff members were of intermediate experience and meat has the largest contribution in waste respectively. Following are their visualizations:


Distribution of staff_experience
Frequency
0
100
200
300
400
500
intermediate
beginner
expert
pro
staff_experience


Distribution of waste_category
Frequency
0
50
100
150
200
250
300
350
meat
grains
dairy
vegetables
wheat
barley
waste_category

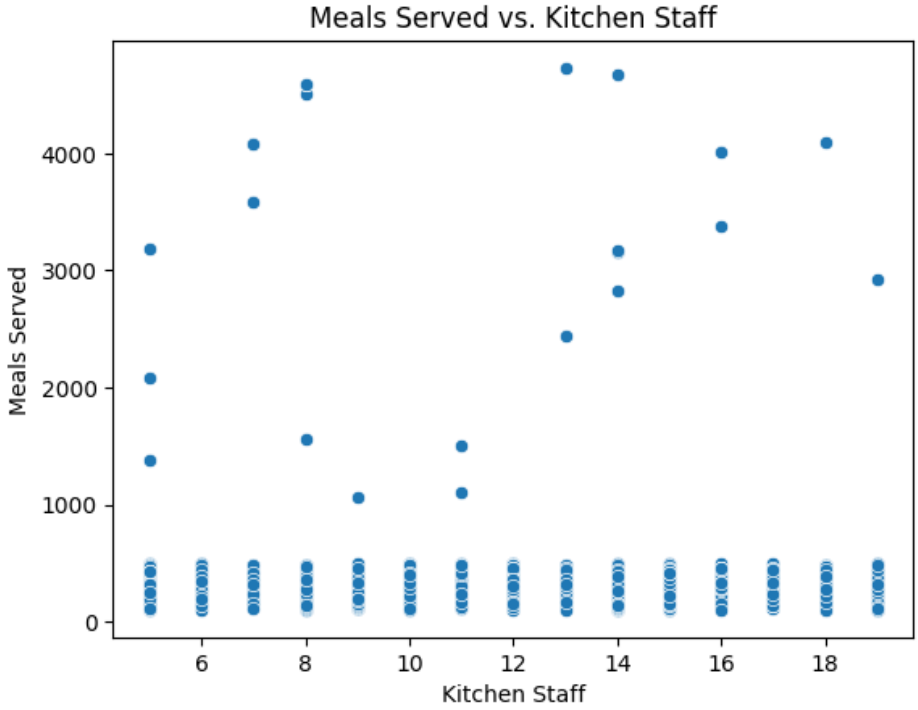
Correlation Analysis

There was no major correlation in between any variable. The following figure can be referred to know exact values:

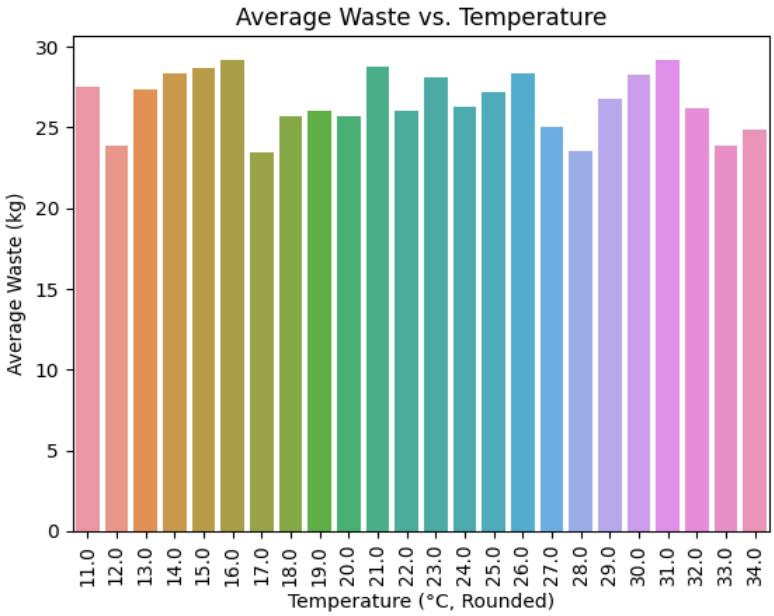


Key insights and recommendations

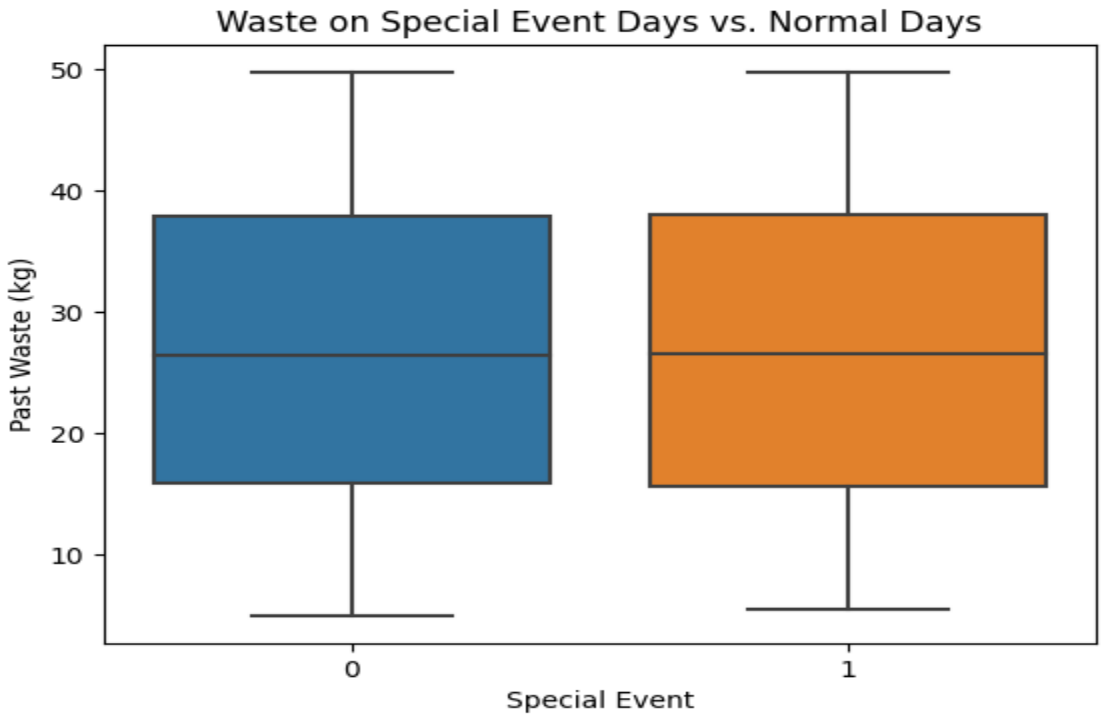
1. **Meals serving and kitchen staff relationship:** It was noted that the given data set suggests that there is not any relationship between meals served and kitchen staff. Following is the graph that illustrates this point as well:



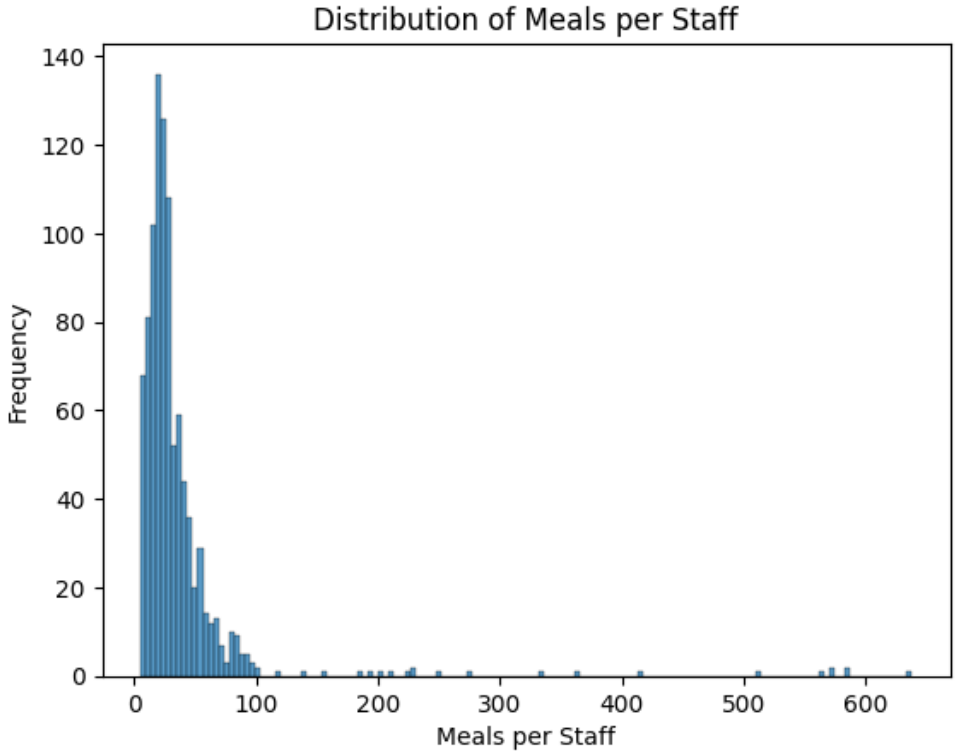
2. **Temperature and average waste relationship:** There was no relationship observed regarding varying temperatures with average waste.



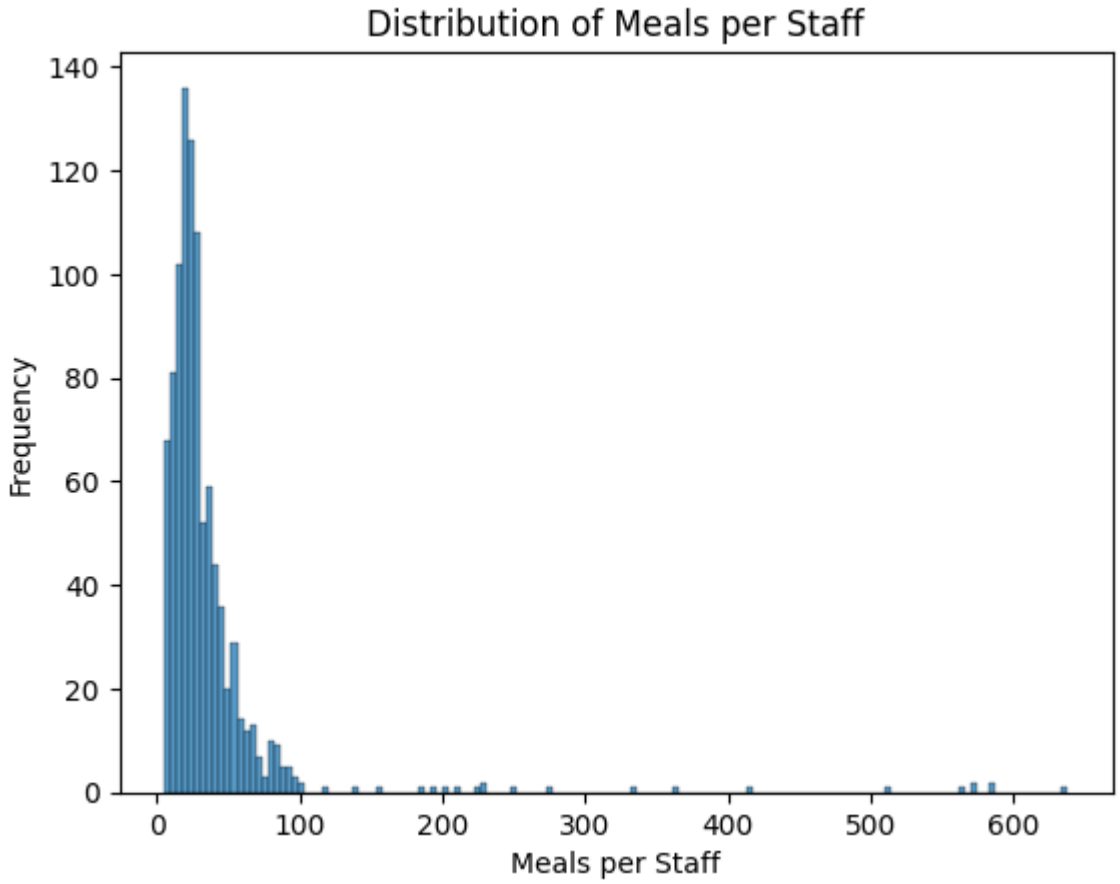
3. **Waste and special event relationship:** There was no difference in waste regardless of special event or not.



4. **Distribution meals per staff:** It was found that more the staff, the less meals per staff will be served.



5. **Average food waste and kitchen staff:** It was found that there is no relationship between kitchen staff and average food waste.



Conclusion

It was concluded that the given variables didn't have any major correlation among them, however; when some new features were found, it suggests some relationships. For further analysis, it is best to add in more features/variables and if the data volume can be increased. Both of these were sort of limiting the analysis.