Problem Statement - Food Demand Forecasting

Project Title:

Food Demand Forecasting using Time Series Analysis

1. Introduction

In today's fast-paced world, accurately predicting food demand is essential for minimizing food waste, optimizing supply chains, and ensuring customer satisfaction. Food demand forecasting helps businesses, restaurants, and retailers maintain optimal inventory levels by understanding consumption trends. This project leverages historical product demand data to predict future requirements using data preprocessing and analysis techniques.

2. Objective & Business Context

Efficient food demand forecasting plays a crucial role in minimizing waste, ensuring availability, and optimizing supply chain operations in the food industry. This project aims to preprocess real-world food transaction data collected from multiple CSV sources to make it suitable for accurate predictive modeling.

The goals of the preprocessing pipeline are to:

1. Merge raw CSV files into one unified dataset.
2. Handle missing and inconsistent values.
3. Detecting and removing duplicates.
4. Identify and remove statistical outliers.
5. Classify and analyze variable types and distributions.
6. Visualize the dataset using boxplots, histograms, bar graphs, and scatter plots.
7. Export a clean, ready-to-use dataset for time series forecasting and ML models.

3. Problem Description

The core issue addressed in this project is the challenge of maintaining the right inventory levels to meet dynamic customer demand. Incorrect forecasts can lead to overstocking, which increases storage costs and food spoilage, or understocking, which results in lost sales and unsatisfied customers. By cleaning and analyzing past demand data, we aim to create a reliable basis for predictive modeling in future stages.

4. Dataset Overview

Source: Kaggle (Food Demand Forecasting Dataset Link)

Files: 5 CSVs merged

Rows: ~521,822

Columns: 15

Key Features:

- center_id, city_code, region_code, center_type, op_area
- meal_id, category, cuisine, id, num_orders
- week, checkout_price, base_price
- emailer_for_promotion, homepage_featured

5. Initial Data Quality Assessment

Performed with df.info(), df.describe(), and df.isna().sum().

Summary of issues and actions:

Issue	Affected Rows	Action Taken
Missing city_code	521,745	Imputed with median
Missing center_type, etc.	~521,000+	Imputed (mode for categorical)
Duplicate Rows	Detected	Removed
Full NaN Rows	Detected	Removed
Statistical Outliers	Multiple columns	Removed via IQR

6. Variable Classification & Distribution

Each column was **categorized** and **analyzed** as follows:

Column	Type	Level	Notes
checkout_price	Numeric	Continuous	Skewed
num_orders	Numeric	Discrete	Skewed
emailer_for_promotion	Binary	Discrete	0 or 1 only
center_type	Categorical	Nominal	String categories
op_area	Numeric	Continuous	Mostly uniform

Distributions were examined using histograms, and skewness **improved** after **cleaning**.

7. Preprocessing Pipeline

Steps:

- Merge CSVs → Used pandas to merge 5 datasets.
- Explore Attributes → Data types identified.
- Handle Missing Values → Mode for object types, median for numeric.
- Remove Duplicates → Used df.drop_duplicates().
- Outlier Removal (IQR Method) → Applied to all numeric columns:

```
Q1 = df[col].quantile(0.25)
Q3 = df[col].quantile(0.75)
IQR = Q3 - Q1
df = df[(df[col] >= Q1 - 1.5 * IQR) & (df[col] <= Q3 + 1.5 * IQR)]
```

- Exported Clean Dataset → df.to_csv("CleanedDataset.csv", index=False)

8. Visualizations

- **Boxplots** (Before & After Outlier Removal) for all numeric columns
- **Histograms** for numeric columns
- **Bar graph** for top 20 most frequent products
- **Scatter plot** between Encoded Product Category and Order Demand

Note: All plots are included in the submission ZIP file.

9. Before vs After Comparison

Metric	Before	After	Change
Total Rows	521,822	~310,000	Reduced (cleaned)
Missing Values	Many	0	Fully resolved
Duplicates	Present	Removed	Cleaned
Outliers	Detected	Removed	Cleaned

10. Outcome & Future Work

The final dataset is clean, consistent, and structured for downstream machine learning and forecasting tasks. This preprocessing pipeline serves as a foundational step for building demand prediction models, inventory optimization tools, and advanced time series forecasting systems.

11. Contributors

Prepared By ***Group 5:***

- *M.Talha Qureshi (BSCS-23122)*
- *Assadullah Farrukh (BSCS-23213)*
- *Abdullah Hussain Yasim (BSCS-23008)*