SmartBite Synthetic Cafeteria Food Demand Dataset Documentation

Purpose:

This document provides a comprehensive overview of the synthetic dataset generated for the SmartBite AI/ML project, including its structure, generation logic, rules, and statistical properties. The dataset simulates realistic college cafeteria food demand patterns in an Indian educational institution, focusing on vegetarian meals, to support forecasting models, preference analysis, and dashboard development.

1. Project Context and Motivation

The SmartBite project aims to predict daily food demand in college cafeterias to minimize shortages and waste. Key objectives include forecasting quantities for each menu item across breakfast, lunch, snacks, and dinner; modeling traffic patterns influenced by time slots, weekdays, and events; analyzing student preferences for menu adjustments; and delivering insights via an interactive dashboard.

Given the unavailability of real POS logs or billing history (as noted in the project proposal), this synthetic dataset was created as a realistic proxy. It models historical cafeteria records over 4 years, incorporating variations in student attendance, dish popularity, seasonal trends, and special occasions. The dataset adheres to principles of realism for Indian college messes (e.g., vegetarian focus, high demand for staples like paneer dishes), while ensuring scalability and utility for machine learning tasks such as linear regression, XGBoost, LSTM, and Prophet.

The generation process prioritizes:

- **Realism**: Based on typical Indian college behaviors (e.g., higher lunch participation, Friday/Monday peaks, festival spikes).
- **Variability**: Dynamic elements like variable menu sizes and clustered events to mimic real-world unpredictability.
- **Privacy and Ethics**: Fully synthetic with no personal identifiable information (PII).
- **ML Readiness**: Large scale (~33,500–34,000 records) with rich temporal coverage (4 full years) and highly realistic distributions, providing more than sufficient data for robust training and testing of forecasting models without overfitting.

2. Dataset Overview

- **Dataset Name**: smartbite_cafeteria_demand_2022_2025.csv
- **Time Period Covered**: January 1, 2022, to December 31, 2025 (1,461 days, including leap years).
- **Total Records**: 33,567 rows (average ~23 rows per day, varying due to dynamic menu sizes across Breakfast 5–8, Lunch 6–10, Snacks 1–2, and Dinner 5–9 dishes).
- **Data Type**: Fully synthetic, generated via scripted simulation in Python.
- **Granularity**: One row represents one dish served in one meal type on one specific date.
- **Target Variable**: Servings (integer count of portions ordered/served for that dish).
- **Format**: CSV file, comma-delimited, with headers.
- **Purpose**:
  - Train ML models for demand forecasting (e.g., per dish, per meal).
  - Perform EDA on patterns (e.g., weekday vs. weekend, event impacts).
  - Test event-driven spikes and seasonal adjustments.
  - Augment or replace real data for early prototyping.

Column Structure

The dataset consists of the following 9 columns:

Column Name	Description	Data Type	Example Value
Date	Calendar date in YYYY-MM-DD format.	Date	2024-03-15
Day	Day of the week (full name).	String	Monday
Meal_Type	One of four meal sessions: Breakfast, Lunch, Snacks, Dinner.	String	Lunch
Dish_Name	Name of the dish served in that meal session (100% vegetarian).	String	Paneer Butter Masala
Servings	Total number of servings taken for the dish (target variable; students may take multiple).	Integer	2700
Student_Count	Total students available on campus that day (base for calculations).	Integer	1120
Event_Flag	1 if the day has a campus event/festival/sports day, else 0.	Integer	0
Vacation_Flag	1 if the day is a vacation/holiday/semester break, else 0.	Integer	0
Total_Eaters	Number of students who actually participated in the meal.	Integer	890

3. Generation Logic and Rules

The dataset was generated using a simulation script that models daily cafeteria operations. Each day starts with a base student count, adjusts for events/vacations, calculates meal participation, selects dishes with rotation constraints, and distributes servings based on popularity and per-eater averages. Key rules ensure realism, variability, and alignment with Indian college mess dynamics.

3.1 Base Population and Adjustments

- **Student Count Range (Normal Days)**: 2,800–4,200 students (realistic for a mid-sized Indian college, including hostellers and day scholars).
- **Yearly Growth**: Slight increase of 3% per year to simulate enrollment trends (e.g., ~2,800 in 2022 → ~3,600 by 2025).
- **Adjustments**:
  - **Events**: Increase by +10–25% (e.g., for visitors during fests).
  - **Vacations/Holidays**: Decrease by -40–70% (low attendance).
- **Overall Distribution**: Student counts vary daily but trend upward over time.

3.2 Meal Participation Logic

- **Meal Types**: Fixed to four: Breakfast, Lunch, Snacks (evening light items), Dinner.
- **Normal Day Participation Rates**:
  - Breakfast: 48–68%
  - Lunch: 78–96% (highest, as most students eat on campus midday).
  - Snacks: 65–80% (light evening session).
  - Dinner: 62–88%
- **Adjustments by Day Type**:
  - **Events**: +10–25% across meals (higher engagement).
  - **Vacations**: -20–35% (reduced turnout).
  - **Weekday Boosts**: Friday and Monday lunches/dinners get a +5% random boost to simulate peak days.
- **Total Eaters Calculation**: Student_Count × Adjusted Participation % (integer-rounded for realism).

3.3 Dish Selection and Rotation

- **Dish Pools**:
  - Breakfast: 33 items (e.g., Poha, Idli, Aloo Paratha – light, quick-prep options).
  - Lunch/Dinner: 67 shared items (e.g., Paneer Butter Masala, Veg Biryani, Dal Tadka – mains, sides, desserts).

    - Snacks: 46 light items (e.g., Maggi, Sandwich, Tea – 1–2 per day, focusing on evening quick bites).
- **Number of Dishes per Meal (Variable)**:
    - Breakfast: 5–8
    - Lunch: 6–10
    - Snacks: 1–2 (as specified for evening snacks like Maggi or Bread Pakoda with Tea/Coffee).
    - Dinner: 5–9
- **Rotation Rule**: No dish repeats more than 2–3 times per week per meal type (tracked via a recent history buffer of ~21 items, approximating a week).
- **Event Days**: Menus remain identical to normal days (no special additions), but quantities increase due to higher eaters.

3.4 Dish Serving Logic

- **Average Dishes per Eater** (Multi-Serving Allowed):
    - Breakfast: 1.8–2.5
    - Lunch: 2.5–3.5
    - Snacks: 1.2–1.8
    - Dinner: 2.3–3.0
- **Total Servings Needed**: Total_Eaters × Average Dishes per Eater (rounded).
- **Distribution**: Servings allocated proportionally based on dish popularity (power-law normalized probabilities), then clipped.
- **Per-Dish Constraints**:
    - **Minimum**: ≥ 0.6 × (Total_Eaters / Number of Dishes in Meal) – Ensures no unrealistic zeros.
    - **Maximum**: ≤ 1.8–2.5 × (Total_Eaters / Number of Dishes) – Caps over-serving; sides/snacks capped lower.
- **Power-Law Popularity**:
    - Top ~15 dishes (e.g., Paneer Butter Masala, Veg Biryani, Palak Paneer) account for ~50% of servings (boosted scores: 600–1,500).
    - Tail dishes (e.g., Khichdi, Karela Bharta, Lauki ke Kofte) low-demand (scores: 80–150).
    - Ensures realism: Popular items like paneer dominate, while khichdi cannot match paneer popularity.

3.5 Event and Vacation Distribution

- **Event Days**: ~ 9–11% of days (~130–160 total).
    - Clustered around real Indian festivals (e.g., Diwali, Holi, Janmashtami) and campus events (e.g., Feb–March fest season, October sports weeks).
    - Probability: Base 6% random + high (70–50%) in clusters.
    - Never overlaps with vacations.
- **Vacation Days**: ~ 7–8% of days (~100–120 total).

    - Clustered in summer breaks (May 20–June 20), winter breaks (Dec 20–Jan 10), and individual holidays (e.g., Republic Day, Independence Day).
    - Includes semester breaks and national festivals for low attendance.

3.6 Additional Realism Patterns

- **Weekday/Weekend Variations**: Higher participation on Fridays/Mondays; lower on weekends.
- **Seasonal Trends**: Lighter demand in summer (heat/vacations); spikes during festivals.
- **No Zeros/Negatives**: All values positive and realistic.
- **Vegetarian Focus**: 100% vegetarian dishes (~146 unique across pools), with ~70% North Indian, 20% South Indian, 10% Continental/Indo-Chinese veg.
- **Power-Law Imbalance**: Mimics real preferences (e.g., staples high, niche items low).

---

4. Statistical Properties

- **Servings Distribution**: Mean ~ 450–600 per dish; skewed high for popular items (e.g., Paneer: ~ 3,500 avg on busy days); low for rares (~100–300).
- **Class Imbalance**: Dish popularity follows power-law (top 20% dishes = 60–70% servings).
- **Event/Vacation Impacts**: Events boost servings by ~20–50%; vacations drop by ~50–70%.
- **Correlations**: Positive between Total_Eaters and Servings; negative with Vacation_Flag.
- **Outliers/Missing Values**: None – fully populated; outliers simulate real spikes (e.g., festival days).
- **Data Augmentation Potential**: Matches patterns from public datasets (e.g., Kaggle food demand sets) for baseline blending.

---

5. Constraints and Validation

- **Differential Privacy**: Not explicitly applied but inherently private as synthetic.
- **Scalability**: Script regenerates/scales easily; add years or adjust params for extensions.
- **Validation Approach**: Aggregates checked for realism (e.g., lunch > dinner > breakfast; no excessive repeats).
- **Limitations**: Simplified weather/semester effects (implicit in flags); assumes uniform student preferences.
- **Ethical Notes**: No real data used; promotes waste reduction in cafeterias.

This dataset is ready for immediate use in the SmartBite project. For questions or modifications, refer to the generation script `/notebooks/SmartBite_dataset.ipynb`.