# Iteration #01: Understanding the Dataset

Aadarsh Gaikwad, Deepak Udayakumar, and Venkat Srinivasa Raghavan

## 1. Source

- **Source:** The diet dataset is sourced from Kaggle, created for health and fitness recommendation systems.
- **Purpose:** To support the development of fitness recommender systems by tailoring diet recommendations to user health conditions, dietary preferences, and nutritional requirements.
- **Availability:** The dataset is publicly available on Kaggle at **Kaggle - Fitness Recommender Dataset**.

## 2. Structure and Metadata

- **Key Features:**
  - **Diet Dataset:** `Meal_Id`, `Name`, `Category`, `Description`, `Veg_Non`, `Nutrient`, `Disease`, `Diet`, `Price`.
  - **User Profiles Dataset:** `User_Id`, `Veg_Non`, `Nutrient`, `Disease`, `Diet`.
  - **Recent Activity Dataset:** `User_Id`, `Meal_Id`, `Rated`, `Liked`, `Searched`, `Purchased`, `Timestamp`.
- **Dataset Sizes:**
  - Diet Dataset: 512 rows $\times$ 9 columns.
  - Exercise Dataset: 3,864 rows $\times$ 12 columns.
  - Recent Activity Dataset: 963 rows $\times$ 7 columns.
- **Key Metrics:**
  - **Exercise Dataset:** Average Calories Burn (301.86 Cal), Duration (40.19 mins), BMI (26.8), and Intensity (5.46).
  - **User Profiles Dataset:** Average Age (43.61 years), Preferred Intensity (4.79), and Duration (50.31 mins).

- **Recent Activity Dataset:** Average Rating (0.50), Likes (0.52), and Duration (63.82 mins).

- **API Availability:** No API is available for this dataset.

## 3. Missing Data

- **Are there any missing values? If so, how are they represented?**
  - **Exercise and Related Datasets:**
    - No missing values in any dataset.
  - **Diet and Related Datasets:**
    - **Diet Dataset**: 1 missing value in `Description`.
    - **Diet User Profile Dataset**: No missing values.
    - **Diet Recent Activity Dataset**: No missing values.

- **What strategies can be used to handle missing data?**
  - Upon analyzing the missing value in the `Description` column, we noticed that it pertains to only one row in the dataset. Since the `Description` column provides information about the dish,we believe its absence is unlikely to impact the overall analysis significantly. Therefore, we propose the following strategy:
    - We can safely drop the row with the missing `Description` value, given that it constitutes only a minimal portion of the dataset. This approach simplifies the handling of missing data without affecting the dataset's overall integrity or usability.

## 4. Anomalies

- **Are there outliers or anomalies? Do they indicate errors?**

- **How consistent is the data across features and observations?**

**Solution:**

- **Exercise Dataset:**

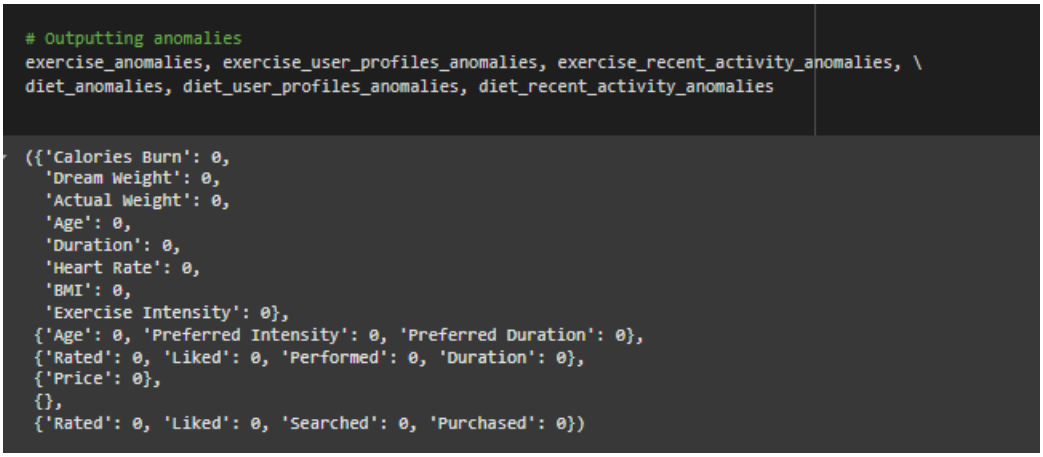


Figure 1: Distribution of anomalies.

  - No outliers detected in numerical features (`Calories Burn`, `BMI`, `Duration`, etc.).
  - Data is consistent across features and observations.

- **User Profiles Dataset (Exercise):**
  - No outliers detected in `Age`, `Preferred Intensity`, or `Preferred Duration`.
  - Data is consistent with no irregularities.

- **Recent Activity Dataset (Exercise):**
  - No outliers detected in `Rated`, `Liked`, `Performed`, or `Duration`.
  - Data is consistent and exhibits no anomalies.

- **Diet Dataset:**
  - No outliers detected in `Price`; data is consistent across features.

- **Diet User Profiles Dataset:**
  - No numerical features; outlier detection does not apply. Data is consistent for categorical features.

- **Diet Recent Activity Dataset:**

- Detected 6 anomalies in `Purchased`, indicating potential errors or extreme user behavior.
- Most data points align with typical user activity.

## 5. Bias

- **Representation:**
  - **Exercise Dataset:** Balanced gender distribution (Female: 50.8%, Male: 49.2%) and realistic metrics (e.g., calorie burn, duration). However, uniform distributions may oversimplify user behavior.
  - **Diet Dataset:** Skewed dietary preferences (Veg: 67.8%, Non-Veg: 32.2%) with limited regional or cultural diversity.
  - **Synthetic Datasets:** Idealized patterns, such as evenly distributed fitness goals and high purchase rates (80%), reduce real-world generalizability.
- **Potential Biases:**
  - Proprietary datasets lack variability in user demographics and regional diversity.
  - Synthetic datasets may overrepresent certain behaviors (e.g., purchases) and underrepresent others (e.g., searches).
- **Bias Mitigation:**
  - Introduce variability in synthetic datasets to better reflect real-world distributions.
  - Validate proprietary datasets using domain-specific studies to ensure relevance and accuracy.

## 6. Distributions

- **What are the distributions of numerical features? Are they skewed or normally distributed?**
- **Are there strong correlations between features? Could multicollinearity be a concern?**

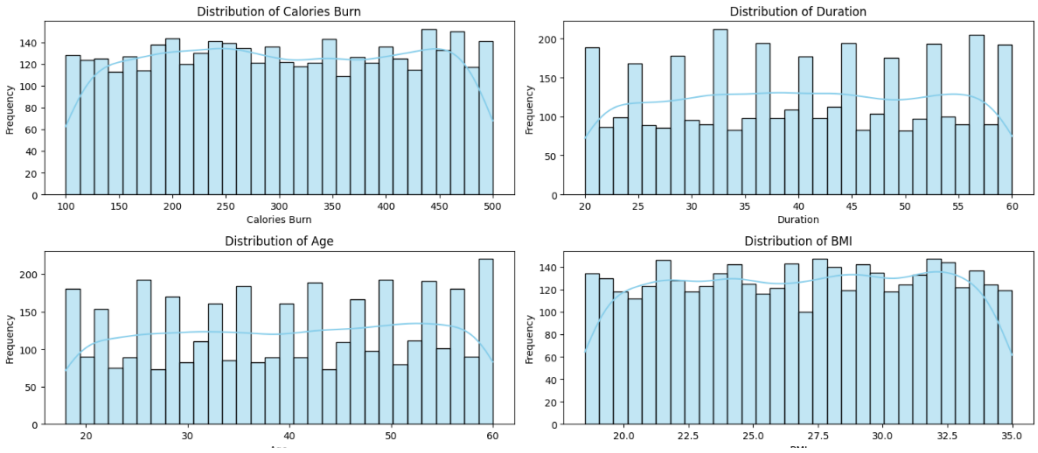


Figure 2: Distribution of numerical features.

**Exercise Dataset**

- **Distributions:**
  - **Calories Burn:** Nearly uniform, indicating exercises target a wide calorie range.
  - **Duration:** Slightly skewed, with most around 40 minutes.
  - **Age:** Uniformly distributed (18–60).
  - **BMI:** Slight left skew, centered near 26.8.
  - **Exercise Intensity:** Uniformly distributed (1–10).
- **Correlations:**
  - Weak overall correlations.
  - **Calories Burn**: Slight positive correlation with **Duration** (0.02), slight negative correlation with **BMI** (-0.03).
  - No multicollinearity concerns.

**Diet Dataset**

- **Distributions:**
  - **Price:** Slight right skew, most prices near the average ( 442.83).

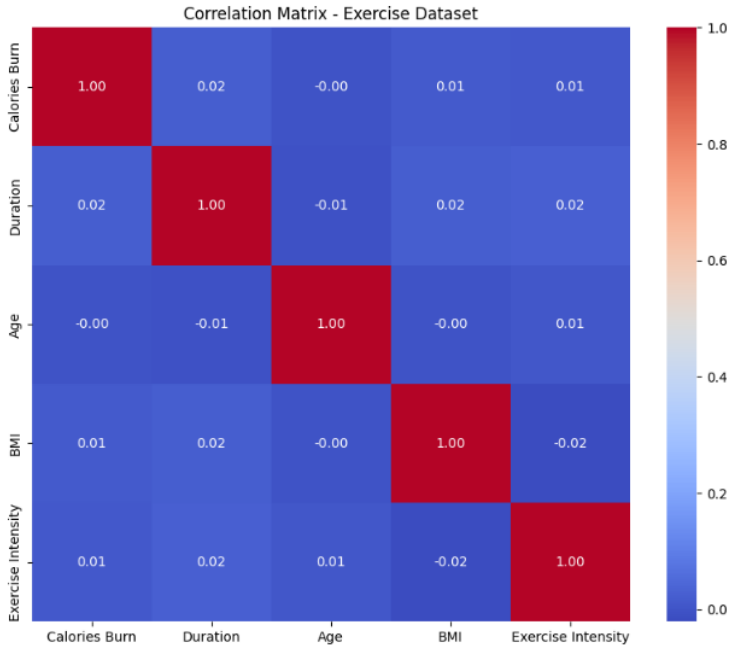


Figure 3: Correlation of numerical features.

- **Correlations:**
  - With only one numerical feature (`Price`), correlation analysis is not applicable.

## 7. Categorical Data

- **How many unique categories exist for each categorical variable?**
- **Are the categories well-balanced or imbalanced?**

**Solution:**

### Proprietary Datasets

The **Diet Dataset** contains several categorical variables, including `Veg_Non` (2 categories: Veg, Non-Veg), `Nutrient` (17 unique nutrients such as Iron and Calcium), `Disease` (124 unique disease associations), and `Diet` (141

unique diet types). The `Veg_Non` variable is skewed toward Veg (67.8%), while `Nutrient` is dominated by Iron (19.9%) and Calcium (12.9%). The `Disease` and `Diet` categories are highly imbalanced, with many combinations appearing infrequently.

The **Exercise Dataset** includes `Gender` (2 categories: Male, Female) and `Weather Conditions` (3 categories: Cloudy, Rainy, Sunny). Both `Gender` (Male: 49.2

### Synthetic Datasets

The **Diet User Profiles** dataset mirrors the `Veg_Non` and `Nutrient` categories of the proprietary diet dataset but adds fewer unique disease associations (47) and diet types (55). `Veg_Non` is more skewed toward Veg (69.4%), with Iron (21.4%) and Calcium (16.3%) dominating `Nutrient`. The **Exercise User Profiles** dataset includes `Gender` (Male: 53%, Female: 47%) and `Fitness Goal` (3 categories: Weight Loss, Muscle Gain, Endurance) with slight skew (Weight Loss: 38%, Muscle Gain: 32%, Endurance: 30%).

The **Recent Activity Datasets** for both diet and exercise primarily include binary interaction features (`Rated`, `Liked`, `Searched`, `Purchased`, `Performed`) with balanced or slightly skewed distributions. For example, in the diet activity data, 33.3% of interactions were rated, 30% were liked, and 80% involved purchases. Similarly, the exercise activity data shows positive ratings for 49.9% of interactions, while 51.6% of exercises were liked or performed.

**General Observations:** Proprietary datasets are realistic but show imbalances in dietary preferences (`Veg_Non`) and nutrient/disease distributions. Synthetic datasets display idealized patterns, with overrepresented categories like Veg and high purchase rates (80%) in the diet recent activity dataset.

## 8. Ethical Considerations

**Does the data contain sensitive or personally identifiable information (PII)?**

- **Proprietary Datasets:**
    - Both the **exercise dataset** and **diet dataset** do not contain any personally identifiable information (PII). The data focuses on general metrics like calorie burns, meal prices, and nutrients.

- **Synthetic Datasets:**
  - The **user profiles** and **recent activity datasets** include unique identifiers (`User_Id`) but no actual names, addresses, or other PII.
  - Since these are synthetic datasets, they do not involve real individuals, further eliminating risks related to PII.

**Are there any ethical concerns in using or publishing insights from this data?**

- **Proprietary Datasets:**
  - Insights derived from the data (e.g., meal recommendations or exercise plans) are unlikely to pose ethical concerns, as the data is anonymized and does not involve real individuals.
  - However, there may be a **responsibility to validate insights** against real-world applicability to ensure recommendations do not mislead or harm users.
- **Synthetic Datasets:**
  - While synthetic, the user behavior patterns (e.g., purchasing, liking, searching) might influence models, which could result in recommendations that do not generalize to real-world populations.
  - There is no direct ethical risk in publishing synthetic data, but **transparency** is required to clarify its limitations and scope.

## 10. Alignment with Goals

- **Does the dataset align with the project's objectives?**
- **Does it have the necessary features for the intended analysis or modeling?**

**Solution:**

The datasets align well with the project's objective of creating a personalized fitness and diet recommendation system. The proprietary **Exercise Dataset** includes key metrics like calorie burn, exercise duration, BMI, and intensity, essential for understanding user fitness needs. Similarly, the **Diet**

**Dataset** provides detailed information on meal nutrients, associated health conditions, dietary preferences, and prices, enabling tailored meal recommendations.

The synthetic datasets add value by capturing user-specific preferences (**User Profiles**) and behavioral data (**Recent Activity**) such as purchasing, rating, and liking. These datasets enhance the recommendation system by simulating user interactions and supporting both collaborative and content-based filtering approaches.

While the datasets are comprehensive and structured for predictive modeling, some limitations remain. Proprietary datasets lack representation of regional, cultural, and seasonal variations, while synthetic datasets idealize behaviors and may require refinement to better mimic real-world patterns. Addressing these gaps could further align the datasets with the project's goals.

## 12. Transformations

- **Does the data need preprocessing, such as normalization, standardization, or scaling?**
- **Are there opportunities to create new features that could improve the model or analysis?**

Data preprocessing was applied to ensure consistency across numerical features, and feature engineering was used to enhance the datasets for analysis and modeling.

- **Exercise Dataset:** StandardScaler was applied to numerical features such as `Calories Burn`, `Duration`, and `BMI`. A new feature, `Calories Per Minute` (`Calories Burn` / `Duration`), was added to represent exercise efficiency.
- **Diet Dataset:** `Price` was scaled using MinMaxScaler for normalization. A derived feature, `Nutrient Count`, was created to capture the number of nutrients listed per meal.
- **User Profiles (Exercise and Diet):** StandardScaler was used for features like `Age`, `Preferred Intensity`, and `Preferred Duration`. No additional feature engineering was required as these datasets capture user preferences.

- **Recent Activity (Exercise and Diet):** Binary features (`Rated`, `Liked`, `Performed`, `Purchased`) did not require scaling. A new feature, `Interaction Ratio`, was introduced to quantify the average interaction level per user.

**Observations:** Preprocessing ensures numerical features are on comparable scales, while engineered features like `Calories Per Minute`, `Nutrient Count`, and `Interaction Ratio` enhance the datasets' utility for modeling and analysis.

## 13. Data Encoding

- **How should categorical variables be encoded (e.g., one-hot, label encoding)?**
- **Are there any temporal or sequential features that require specific transformations?**

**Solution:**
Categorical variables are primarily encoded using one-hot encoding where needed for content-based filtering, while temporal features may require specialized transformations for time-aware models.

- **Exercise Dataset:** One-hot encoding is applied to features like `Gender` and `Weather Conditions` for content-based filtering. Collaborative filtering relies on interaction data, so encoding is unnecessary.
- **Diet Dataset:** Features like `Veg_Non`, `Nutrient`, `Disease`, and `Diet` are one-hot encoded for content-based recommendations. Encoding is not needed for collaborative filtering.
- **User Profiles (Synthetic):** One-hot encoding is used for categorical variables such as `Diet` and `Fitness Goal` to align with content-based filtering. These encodings do not benefit collaborative filtering.
- **Recent Activity (Synthetic):** Temporal features like `Hour` and `Weekday` can be transformed for time-aware recommendation systems but are not essential for general models.

## 14. Predictive Power

**1. Do the features contain sufficient predictive information for the target variable?**

The datasets provide sufficient predictive information for both content-based and collaborative filtering models:

- **Exercise Dataset:** Features such as `Calories Burn`, `Duration`, `Exercise Intensity`, and `BMI` support content-based filtering. Interaction data (`Rated`, `Performed`) from the **recent activity dataset** underpins collaborative filtering.
- **Diet Dataset:** Key features like `Nutrient`, `Disease`, `Diet`, and `Price` enable content-based recommendations. Interaction data enhances collaborative filtering.
- **Synthetic Datasets:** User preferences (`Fitness Goal`, `Preferred Duration`) complement hybrid recommenders by bridging content-based and collaborative approaches.

**2. Is feature selection or dimensionality reduction necessary?**

Feature selection and dimensionality reduction may improve model efficiency:

- **Content-Based Filtering:** High-dimensional categorical data (e.g., `Disease`, `Diet`) could benefit from feature selection or PCA to reduce sparsity and noise.
- **Collaborative Filtering:** Interaction matrices inherently focus on relevant data, negating the need for feature selection. For large datasets, matrix factorization (e.g., SVD) effectively reduces dimensionality.

## 15. Target Variable

- **If this is a supervised learning problem, what is the target variable, and is it well-defined?**
- **Is the target variable balanced, or does it require special handling (e.g., resampling)?**

**Solution:**

**1. Is there a target variable, and is it well-defined?**

The datasets provide well-defined binary target variables for classification tasks:

- **Exercise Dataset:** Target variables include:
  - `Liked`: Indicates user preference for an exercise (1: Liked, 0: Not Liked).
  - `Performed`: Indicates whether the exercise was performed (1: Performed, 0: Not Performed).
- **Diet Dataset:** Target variables include:
  - `Liked`, `Purchased`: Indicate user interaction with meals.

**2. Is the target variable balanced?**

- **Exercise Dataset:**
  - Targets like `Liked` ( 51.6% Liked) and `Performed` ( 48.4% Performed) are balanced, requiring no resampling.
- **Diet Dataset:**
  - `Liked` is imbalanced (30% Liked), and `Purchased` is heavily imbalanced (80% Purchased). Techniques like SMOTE or undersampling may address this imbalance.

## 16. Validation Strategy

- **How will you split the dataset for training, validation, and testing?**
- **Are there temporal or spatial dependencies that need to be preserved during splitting?**

**Solution:**

Datasets are split considering class balance and dependencies:

- **Exercise Dataset:**

- Recent Activity: Stratified split into 70% training, 15% validation, and 15% testing to maintain class distribution.
- Content-Based Filtering: Use k-fold cross-validation for features like `Calories Burn` and `Exercise Intensity`.
- Temporal Dependencies: For time-aware models, preserve sequential order in recent activity data.

- **Diet Dataset:**
  - Recent Activity: Use an 80%-20% stratified split, addressing imbalance with oversampling or undersampling.
  - Temporal Dependencies: Sequential splits to account for patterns in user behavior (`Timestamp`).

## 17. Data Leakage

- **Are there any risks of data leakage, where information from the test set inadvertently influences the model during training?**

**Solution:**

- **Risks:**
  - Data leakage may occur if interactions in the **recent activity datasets** (`Rated`, `Liked`, `Performed`, `Purchased`) are not split temporally, allowing future user behaviors to influence the training set.
  - Synthetic datasets may unintentionally encode patterns that align too closely with the test set.
- **Mitigation:**
  - Use **temporal splits** to ensure that future interactions do not influence predictions.
  - Maintain **user-level separation** between training and test sets in collaborative or hybrid filtering models.

## 18. Interpretability

- **Can the dataset and analysis provide interpretable insights for stakeholders?**
- **How will the results be communicated (e.g., visualizations, metrics)?**
- **Interpretability:**
  - The datasets are interpretable and provide actionable insights:
    - **Exercise Dataset:** Metrics like `Calories Burned` and `Duration` can be easily explained for personalized recommendations.
    - **Diet Dataset:** Insights on `Nutrient`, `Disease`, and `Diet Type` guide users in meal selection and planning.
- **Communication:**
  - Use visualizations such as bar charts, scatter plots, and heatmaps to present insights effectively.
  - Report metrics like **precision**, **recall**, and **F1-score** to evaluate model performance.
  - Provide outputs through user-friendly interfaces (e.g., dashboards or recommendation systems) to ensure clarity for stakeholders.

## 19. Limitations

- **What are the limitations of this dataset for the current project?**
- **What additional data would enhance the analysis?**

**Solution:**

- **Dataset Limitations:**
  - Proprietary datasets lack diversity, such as limited age ranges and cultural dietary variations.
  - Synthetic datasets idealize user behavior, reducing real-world generalizability.

- **Enhancements:**
  - Incorporate diverse, real-world user behavior data and regional dietary information.
  - Add motivational or psychological factors for richer, personalized recommendations.

## 20. Github Link

https://github.com/VenkatSR-14/nutribuddy