Week-1 Planning Report: Food Quality Data Exploration in Food Processing

Data Analytics Team

December 2025

Document Scope

This Week-1 report focuses only on planning for data exploration using publicly available datasets. It includes a short introduction, a shortlist of data sources, a basic EDA methodology (concept level), a simple 6–7 hour timeline for Week-1, and expected challenges with practical mitigations. It intentionally excludes team resourcing, costs, privacy deep-dives, multi-week roadmaps, and predictive modeling.

Contents

1 Introduction

Food processing value chains generate diverse data from raw materials, process parameters (e.g., temperature, humidity, time), quality control tests, and regulatory records. Structured exploration of such data supports food safety, quality consistency, and operational improvements [?, ?]. Public datasets can bootstrap an academic/industry-style exploration framework even without proprietary factory data [?, ?].

Week-1 objective: define a practical plan for exploratory analysis using open data, not to execute full analysis. Deliverables include a focused data-source shortlist, a basic EDA approach, and a minimal execution plan for the next stage.

2 Publicly Available Data Sources (Shortlist)

The following sources are appropriate for Week-1 planning (free or open-access):

2.1 Kaggle — Food Freshness Dataset

Image-based dataset with freshness annotations for fruits/vegetables; useful for studying visual quality/freshness patterns and shelf-life proxies [?].

2.2 Open Food Facts

Global, community-driven database with product-level nutrition, ingredients, allergens, and additives; provides CSV exports and GitHub snapshots for reproducible analysis [?].

2.3 Data.gov — Food Safety/Inspection Catalog

A catalog of food-safety related datasets (e.g., inspections, sampling, recalls) in CSV/JSON; enables inspection outcomes and violation trend exploration [?].

2.4 USDA ERS National Data Sets

Contextual supply/consumption datasets that help analyze category-level trends and seasonality relevant to quality outcomes [?].

Week-1 Deliverable. Select 3–5 datasets from the above based on relevance to the chosen quality dimension (e.g., labeling/nutrition; freshness/shelf-life; inspections/safety) and document links, formats, size, and key fields.

3 Basic Methodology (Concept Level)

No execution in Week-1; only the blueprint of exploratory techniques.

3.1 Data Structure Understanding

Record dataset shapes (rows, columns), data types, units (e.g., °C, g/100g), date formats, and identifier fields (e.g., product code, brand, category) [?, ?].

3.2 Descriptive Summaries

Compute conceptually planned summaries:

- Numeric: mean, median, min, max, standard deviation for key metrics (e.g., calories, sodium, protein).
- Categorical: frequency counts (e.g., additives, allergens, inspection outcomes).

These summaries guide pattern discovery in later weeks [?].

3.3 Basic Visual Explorations

Plan simple plots for later execution:

- Distributions: histograms/boxplots for nutrients or defect-related measures.
- Time trends: if dates are present (e.g., inspection counts by year/quarter).

Only plan in Week-1; execution follows in subsequent weeks [?].

3.4 Outlier Flagging

Define rules to flag unusual values (e.g., very high sodium/sugar; rare additive combinations). Outliers are flagged, not removed at this stage [?].

4 Week-1 Strategy and Timeline (6–7 Hours)

A concise plan suitable for assignment delivery:

Day 1 (2 hours)

- Define focus quality dimension:
  - Option A: Nutritional quality & labeling (Open Food Facts).
  - Option B: Freshness/shelf-life (Kaggle Food Freshness).
  - Option C: Safety/inspections (Data.gov).
- Draft a 1–2 page requirements note listing key metrics (e.g., sugar, sodium, additives; freshness labels; inspection score types) [?, ?, ?].

Day 2 (2–2.5 hours)

- Explore and compare 3–5 candidate datasets (size, columns, geography, years, access method).
- Create a comparison table: dataset name, link, type (image/tabular), food category, potential use [?, ?, ?, ?].

Day 3 (1.5–2 hours)

- Freeze final shortlist (3 primary datasets).
- Environment plan: Google Colab + Python or Excel for initial review; organize downloads (local/Drive) [?].
- Write a quick audit plan for each dataset: row count, missing% check, key columns list [?, ?].

Week-1 Output. Dataset shortlist with links and roles; written EDA plan; minimal environment and audit plan documented.

5 Expected Challenges (Week-1 Level)

Relevance. Public datasets may skew to consumer/label viewpoints rather than plant-floor QC; document the exploratory/academic scope and note how the same framework can later adapt to internal QC data [?].

Documentation/Cleanliness. Community/GitHub datasets vary in documentation quality; start a lightweight data dictionary draft (column names + short meaning) in Week-1 [?, ?].

References

1. USDA Economic Research Service (ERS). National data sets useful in food and nutrition assistance research. Retrieved 2025. http://www.ers.usda.gov/.[10]
2. Data.gov Catalog: Food-safety datasets. https://catalog.data.gov/dataset/?tags=food-safety.[11]
3. Kaggle: Food Freshness Dataset. https://www.kaggle.com/datasets/ulnnproject/food-freshness-dataset.[12]
4. Open Food Facts data and example EDA repositories (GitHub). https://github.com/kgeorgiev42/OpenFoodFacts-EDA.[13]
5. Overleaf LaTeX templates (report/technical). https://www.overleaf.com/latex/templates.[1]
6. LaTeX/Document Structure (Wikibooks). https://en.wikibooks.org/wiki/LaTeX/Document_Structure.[5]