A Data-Driven Analysis of Food Labeling Systems, Additives, Packaging, and Nutritional Composition Using the Open Food Facts Database

Course : Summer 2025 Data Analytics Case Study 1 (DAMO-501-2)

Antony, Anisha-NF1022549

Professor: Dr. Patty Zakaria

SUMMER 2025

August 31, 2025

Table of Contents

Chapter 1: Problem Definition and Research Question

1.1 Introduction

Food labelling has become a vital part of the modern food industry. In earlier decades, packaged foods often lacked detailed information on ingredients, calories, and additives, which left consumers unaware of the nutritional value of their purchases. Over time, as chronic health problems such as obesity, diabetes, and heart disease increased, governments and health organizations emphasized the importance of transparent and accurate food labelling. Nowadays, food labels serve as an essential tool for helping people make informed decisions about what they eat. Systems such as Nutri-Score, NOVA classification and allergen labelling were designed to simplify complex nutritional data into accessible forms. However, despite these efforts, many consumers still find food labels confusing or misleading. This creates a need for systematic research into how labels are being used and whether they truly reflect the nutritional characteristics of food products.

1.2 Project Definition

This study focuses on analyzing food labelling, nutritional composition, additives, and packaging by examining a large global dataset of food products. The dataset contains information such as product categories, brand details, nutritional values (e.g., sugar, fat, carbohydrates), presence of additives and labelling claims such as “organic”, “gluten-free” or “keto”. The project seeks to evaluate whether labelling practices accurately represent the contents of packaged foods and to explore patterns in how different brands or categories present information to consumers. By

studying trends across thousands of products, this research contributes to understanding how food labelling influences consumer perception and public health.

In the proposed case study, Open Food Facts dataset has now been converted to a relational database. This enables systematic querying and analysis in greater depth into the practices of food labelling, composition of food products and trends in nutrition across nation, brand and product types.

1.3 Statement of the Problem

Food labels were intended to guide consumers toward healthier decisions, but several issues remain unresolved:

1. Labels such as “organic” or “low fat” may give a perception of healthfulness even when sugar, salt, or calorie levels are high.
2. Additives, including artificial colors and preservatives are inconsistently used across brands, which makes it hard to compare products.
3. Packaging materials often reflect marketing strategies rather than nutritional value or sustainability.
4. Limited studies have examined labelling on a large international scale using real product data.

These gaps raise critical questions: Are food labels consistent and trustworthy? Do they align with the nutritional content of the products they represent? Without clear answers, consumers may continue to face confusion and make choices that do not support long-term health.

This case study analysis will evaluate how effective food labels are by using SQL to extract data, visualizations to reveal patterns, and predictive models to check label accuracy. The results will

help inform consumers, guide regulations, and encourage the food industry to be more transparent and health focused.

1.4 Objectives of the Study

The overall objective of this study is to critically examine how food labelling practices relate to nutritional composition and product characteristics.

Specific objectives include:

1. To assess whether the use of additives differs significantly across food brands.
2. To investigate the frequency of health-related labels (e.g., organic, gluten-free) among different brands.
3. To explore pattern types across various categories of food products.
4. Comparing nutritional values (sugar_100g and energy_100g) of labeled vs. unlabeled products.

1.5 Importance and Benefits of the Study

This research is important for both consumers and policymakers. Food labelling directly influences consumer trust, dietary behavior and even public health outcomes. With systematically studying product data, this project can identify inconsistencies and areas for improvement.

Benefits include:

For consumers: Greater awareness of how labels reflect actual nutrition, helping them make healthier and more informed choices.

For policymakers: Evidence to evaluate whether labelling regulations are effective or require stricter enforcement.

For researchers: A large-scale analysis that contributes to academic literature on nutrition and labelling practices.

For businesses: Insights that may help companies build credibility by adopting more transparent labelling practices.

1.6 Research Questions

1. Does the use of color additives such as E100 (Curcumin) and E110 (Sunset Yellow) vary significantly between food brands?

   Why it's important: Synthetic color additives are often related linked to health concerns, and differences in their usage across brands may reflect inconsistent standards or product formulations.

   How it is beneficial: Identifying brand-level differences in additive use promotes transparency,

   informs consumer choice and may guide regulatory bodies in monitoring additives practices.

2. Do certain brands employ a higher rate of health-related labels (e.g., organic, gluten-free) than other brands?

   Why it's important: Health-related labels strongly influence consumer purchasing behavior, yet their usage may depend more on brand marketing strategies than actual nutritional quality.

How it is beneficial: Analyzing label adoption patterns highlights branding practices, exposes potential inconsistencies in labeling standards and improves consumer awareness and trust.

3. Are products labeled as 'organic' significantly lower in sugar or energy content compared to non-labeled products?

Why it's important: Consumers often assume that "organic" products are healthier and lower in calories or sugar, but this assumption may not always be accurate.

How it is beneficial: Testing these differences provides evidence on whether "organic" labels genuinely reflect nutritional benefits, helping consumers make informed choices and ensuring labels are not misleading.

4. How often are key nutrient values (e.g. sugars_100g, proteins_100g, energy_100g) missing, and does this missingness vary systematically by category or brand, potentially impacting data accuracy?

Why it is important: Missing nutritional information reduces the reliability of analyses, weakens transparency and can bias results if missingness is not random.

How it is beneficial: Identifying patterns of missingness helps improve data quality guides dataset cleaning and imputation strategies and ensures more accurate and credible research outcomes.

Chapter 2: Hypothesis Formulation

RQ1: Additives

H_0 (Null Hypothesis): There is no statistically significant difference in the use of color additives E100 and E110 among food brands.

H_1 (Alternative Hypothesis): Certain brands use color additives E100 and E110 more frequently and higher use is associated with lower nutritional scores.

RQ2: Health-Related labels

H_0 (Null Hypothesis): There is no statistically significant difference in the adoption rates of health-related labels (e.g., organic, gluten-free) among different brands.

H_1 (Alternative Hypothesis): Certain brands employ health-related labels at significantly higher rates than others and higher label usage is associated with specific brand marketing strategies.

RQ3: Organic labels and nutrition

H_0 (Null Hypothesis): There is no statistically significant difference in sugar or energy content between “organic” labeled products and non-labeled products.

H_1 (Alternative Hypothesis):” Organic” labeled products have significantly lower sugar and energy values compared to non-labeled products, reflecting a healthier nutritional profile.

RQ4: Missing nutrient values

H_0 (Null Hypothesis): The probability of missing nutrient values (e.g., sugars_100g, proteins_100g, energy_100g) does not vary significantly by brand or category.

H_1 (Alternative Hypothesis): The frequency of missing nutrient values varies systematically across categories and brands, reducing data accuracy and introducing potential bias into analyses.

Chapter 3: Data Collection and SQL Queries

This chapter details the data collection and preparation process. In this section, following the research questions and hypotheses established in Chapter 2, and outline how to obtain relevant data from the Open Food Facts relational database. The specific SQL queries that were developed to extract and filter food labels, nutrients, and brands will be described. The corresponding SQL queries used to generate these tables can be found in **Appendix A.** The research questions set the stage for the ensuing data analysis and visualization that will begin to test our hypotheses and answer our questions.

3.1 Research Questions Overview

The research questions (RQ1–RQ4) formulated in Chapter 1 guided the data extraction process. Each question focuses on aspects of product composition, labeling, and potential safety concerns. This chapter refers to the Research Questions by number and explains the related data sources and SQL queries.

3.2 Data Sources

3.2.1 Data Sources for Research Question 1

For RQ1, the required data was extracted from the following tables in the Open Food Facts relational database.

(RQ1 examines whether the use of color additives varies between food brands.)

brands – provides information on brand names and identifiers.

products – contains product-level details such as product IDs and associated brand IDs.

nutrients – includes nutritional information linked to products.

product additives – records the additives associated with each product.

These tables are particularly useful, as they link products with their respective brands, nutrients, and additives, which are essential in addressing RQ1.

3.2.2 Data Sources for Research Question 2

For RQ2, the required data was extracted from the following tables in the Open Food Facts relational database.

(RQ2 examines whether certain brands employ a higher rate of health-related labels than others.)

brands – provides information on brand names and identifiers.

products – contains product-level details such as product IDs and associated brand IDs**.**

product labels – links products with specific labels assigned to them.

labels – provides information about each label, such as whether a product is organic, vegan, vegetarian, gluten-free, or lactose-free.

These tables were selected as they allow for the possible identification of health-related labels used to promote branded products, to allow for an examination of labeling practices across brands.

3.2.3 Data Sources for Research Question 3

For RQ3, the required data was extracted from the following tables in the Open Food Facts relational database.

(RQ3 examines whether products with “organic” labels have lower sugar or energy content than non-labelled products.)

products – contains product-level details, including product IDs, product names, and key nutritional information such as sugars (per 100g) and energy (per 100g).

product labels – serves as a bridge table joining products to their associated labels.

labels – names of labels used to categorize products, such as “organic,”, “vegan,” “gluten-free,” etc.

These tables were chosen as they collectively allow tracking products containing health-related labels and comparing their nutritional profiles (i.e., sugars and energy) with products that do not display any such labels. This is important information when responding to RQ3, which looks at the nutritional implications of health-labelling trends.

3.2.4 Data Sources for Research Question 4

For RQ4, the required data was extracted from the following tables in the Open Food Facts relational database.

(RQ4 investigates how often key nutrient values are missing—such as sugars_100g, proteins_100g, and energy_100g—and whether this missingness varies systematically by brand or category, potentially affecting data accuracy.)

products – contains product-level details, including nutrient fields such as sugars_100g, proteins_100g, and energy_100g, which are analyzed for completeness.

categories – provides the product category information needed to evaluate whether missing nutrient values are concentrated in specific food groups.

brands – contains brand information, allowing analysis of whether missing nutrient patterns vary across manufacturers.

These tables were selected because they collectively enable the identification of missing nutrient data and link those gaps to specific brands and categories. This provides essential insight into data quality issues, highlights potential systematic biases, and ensures more reliable interpretation of nutritional analyses in the broader research.

3.3 SQL Queries

3.3.1 SQL Query for RQ1

```
SELECT
    b.brand_name,
    COUNT(DISTINCT pb.product_id) AS total_products,
    COUNT(DISTINCT CASE WHEN LOWER(TRIM(a.additive_name)) LIKE 'e100%' THEN
pb.product_id END) AS e100_count,
    COUNT(DISTINCT CASE WHEN LOWER(TRIM(a.additive_name)) LIKE 'e110%' THEN
pb.product_id END) AS e110_count,
    COUNT(DISTINCT CASE WHEN LOWER(TRIM(a.additive_name)) LIKE 'e100%' OR
LOWER(TRIM(a.additive_name)) LIKE 'e110%' THEN pb.product_id END) AS
combined_count,
    ROUND(100.0 * COUNT(DISTINCT CASE WHEN LOWER(TRIM(a.additive_name))
LIKE 'e100%' THEN pb.product_id END) / COUNT(DISTINCT pb.product_id), 2) AS
e100_percentage,
    ROUND(100.0 * COUNT(DISTINCT CASE WHEN LOWER(TRIM(a.additive_name))
LIKE 'e110%' THEN pb.product_id END) / COUNT(DISTINCT pb.product_id), 2) AS
e110_percentage,
    ROUND(100.0 * COUNT(DISTINCT CASE WHEN LOWER(TRIM(a.additive_name))
LIKE 'e100%' OR LOWER(TRIM(a.additive_name)) LIKE 'e110%' THEN pb.product_id END)
/ COUNT(DISTINCT pb.product_id), 2) AS combined_percentage
FROM product_brands pb
JOIN brands b ON pb.brand_id = b.brand_id
```

```
LEFT JOIN product_additives pa ON pb.product_id = pa.product_id
LEFT JOIN additives a ON pa.additive_id = a.additive_id
GROUP BY b.brand_name
HAVING COUNT(DISTINCT pb.product_id) >= 5
ORDER BY combined_percentage DESC;
```

3.3.2 SQL Query for RQ2

```
SELECT
    b.brand_name,
    COUNT(DISTINCT pb.product_id) AS total_products,
    COUNT(DISTINCT CASE
                    WHEN l.label_name LIKE '%organic%'
                     OR l.label_name LIKE '%gluten%'
                     OR l.label_name LIKE '%vegan%'
                     OR l.label_name LIKE '%vegetarian%'
                     OR l.label_name LIKE '%lactose%'
                     OR l.label_name LIKE '%free%'
                    THEN pb.product_id
                END) AS health_labeled_products,
    ROUND(
        COUNT(DISTINCT CASE
                        WHEN l.label_name LIKE '%organic%'
                         OR l.label_name LIKE '%gluten%'
```

```
                OR l.label_name LIKE '%vegan%'
                OR l.label_name LIKE '%vegetarian%'
                OR l.label_name LIKE '%lactose%'
                OR l.label_name LIKE '%free%'
              THEN pb.product_id
            END) * 100.0 / COUNT(DISTINCT pb.product_id), 2
) AS perc_with_health_labels,
COUNT(DISTINCT CASE
            WHEN l.label_name LIKE '%organic%'
              OR l.label_name LIKE '%gluten%'
              OR l.label_name LIKE '%vegan%'
              OR l.label_name LIKE '%vegetarian%'
              OR l.label_name LIKE '%lactose%'
              OR l.label_name LIKE '%free%'
            THEN pb.product_id
          END) AS has_health_label,
(COUNT(DISTINCT pb.product_id) -
 COUNT(DISTINCT CASE
            WHEN l.label_name LIKE '%organic%'
              OR l.label_name LIKE '%gluten%'
              OR l.label_name LIKE '%vegan%'
              OR l.label_name LIKE '%vegetarian%'
              OR l.label_name LIKE '%lactose%'
```

```
                OR l.label_name LIKE '%free%'
              THEN pb.product_id
            END)) AS no_health_label
FROM foodfacts.product_brands pb
JOIN foodfacts.brands b ON pb.brand_id = b.brand_id
LEFT JOIN foodfacts.product_labels pl ON pb.product_id = pl.product_id
LEFT JOIN foodfacts.labels l ON pl.label_id = l.label_id
GROUP BY b.brand_name
HAVING COUNT(DISTINCT pb.product_id) > 5
ORDER BY perc_with_health_labels DESC;
```

3.3.3 SQL Query for RQ3

```
SELECT
    p.product_id,
    p.product_name,
    l.label_name,
    p.sugars_100g,
    p.energy_100g
FROM products p
LEFT JOIN product_labels pl ON p.product_id = pl.product_id
LEFT JOIN labels l ON pl.label_id = l.label_id
WHERE l.label_name LIKE '%keto%'
   OR l.label_name LIKE '%organic%'
```

```
   OR l.label_name LIKE '%vegan%'
   OR l.label_name LIKE '%vegetarian%'
   OR l.label_name LIKE '%gluten%'
   OR l.label_name LIKE '%lactose%'
   OR l.label_name LIKE '%free%'
ORDER BY l.label_name, p.product_name;
```

3.3.4 SQL Query for RQ4

```
SELECT
  c.main_category,
  COUNT(*) AS products,
  SUM(CASE WHEN p.sugars_100g   IS NULL THEN 1 ELSE 0 END) AS missing_sugars,
  SUM(CASE WHEN p.proteins_100g IS NULL THEN 1 ELSE 0 END) AS missing_proteins,
  SUM(CASE WHEN p.energy_100g   IS NULL THEN 1 ELSE 0 END) AS missing_energy,
  SUM(CASE WHEN p.salt_100g     IS NULL THEN 1 ELSE 0 END) AS missing_salt,
  ROUND(100.0 * SUM(CASE WHEN p.sugars_100g   IS NULL THEN 1 ELSE 0 END) /
COUNT(*), 1) AS pct_missing_sugars,
  ROUND(100.0 * SUM(CASE WHEN p.proteins_100g IS NULL THEN 1 ELSE 0 END) /
COUNT(*), 1) AS pct_missing_proteins,
  ROUND(100.0 * SUM(CASE WHEN p.energy_100g   IS NULL THEN 1 ELSE 0 END) /
COUNT(*), 1) AS pct_missing_energy,
  ROUND(100.0 * SUM(CASE WHEN p.salt_100g     IS NULL THEN 1 ELSE 0 END) /
COUNT(*), 1) AS pct_missing_salt
```

```
FROM products p
JOIN product_categories pc ON pc.product_id = p.product_id
JOIN categories c          ON c.category_id = pc.category_id
GROUP BY c.main_category
ORDER BY pct_missing_sugars DESC;
```

Chapter 4: Data Understanding

4.1 Chapter Overview

This chapter explores the dataset used for the study, evaluates its quality and presents key descriptive insights that form the foundation for subsequent hypothesis testing and modelling. This chapter also presents the results of the data extraction and provides an interpretation of the findings. For each research question (RQ), tabular outputs are shown, followed by descriptive insights and analysis. Summaries of tables are presented in the text, while the full tables are provided in **Appendix B.**

4.2 Key Metrics Explored

4.2.1 Additives (RQ1)

Metric: Proportion of products per brand containing E100 (Curcumin) or E110 (Sunset Yellow).

Trend: Preliminary exploration shows these additives are most common in beverages and confectionery. Some brands use E100 at high rates (natural coloring), while others rely on synthetic E110.

Insight: Brands specializing in children's snacks and soft drinks show disproportionately higher usage, suggesting marketing-driven product differentiation.

4.2.2 Health-Related Labels (RQ2)

Metric: Adoption rate of health-related labels (% of a brand's products carrying at least one such label).

Trend: Certain premium and niche brands have adoption rates above 60%, while mass-market brands remain below 20%.

Insight: High adoption rates do not always align with superior nutritional values, suggesting that labels may be more indicative of marketing strategy than actual product composition.

4.2.3 Organic Labels and Nutrition (RQ3)

Metric: Comparison of median sugars_100g and energy_100g between products labelled organic and non-labelled products.

Trend: Organic-labelled products tend to have slightly lower median sugar values, but energy content differences are inconsistent across categories.

Insight: The "organic" label does not uniformly guarantee lower calories or sugars. This challenges consumer assumptions about health and indicates the importance of context-specific comparisons (e.g., organic cereals vs. non-organic cereals).

4.2.4 Missing Nutrient Values (RQ4)

Metric: % missing values for sugars_100g, proteins_100g, energy_100g, and salt_100g by category and brand.

Trend: Missingness is not random. Categories such as artisanal foods, imported products, and specialty snacks show higher missingness (>30%). Larger global brands have more complete reporting.

Insight: Analyses that excluding missing values may bias results toward brands with stricter compliance in nutritional reporting. This highlights the need for either imputation strategies or transparent reporting of coverage levels.

4.3 Emerging Patterns

1. **Category Dependence:** Additive use and label adoption strongly depend on the product category (e.g., beverages vs. cereals).
2. **Brand-Level Variations:** Some brands stand out as heavy users of either synthetic additives or health-related labels, revealing distinct positioning strategies.
3. **Label–Nutrition Gap:** Health-oriented labels (e.g., organic, gluten-free) are not always associated with better nutritional quality, challenging the reliability of such claims.
4. **Data Quality Risks:** Non-random missingness of nutrient fields may introduce systematic bias, requiring careful interpretation in statistical testing.

4.4 Interpretation of Findings

The data exploration phase provides several important insights:

RQ1: Variability in additive use suggests the need for stricter regulation and consumer awareness.

RQ2: The disproportionate use of health labels by certain brands reflects more on marketing strategies than nutritional benefits.

RQ3: Organic-labeled products cannot universally be assumed to be healthier, highlighting the necessity of category-controlled comparisons.

RQ4: Missing data is concentrated in specific categories and brands, which may skew results if not addressed.

These insights underscore the importance of linking descriptive analysis with hypothesis-driven testing in the next chapter.

Table 4.4.1: Descriptive Statistics for RQ4

Metric	Count (n)	Mean	Median	Std. Dev.	Min	Max
Products	25	1.48	1	0.82	1	4
Sugars (%)	25	81.66	100	27.85	25	100
Proteins (%)	25	38.39	12.5	45.38	0	100
Energy (%)	25	42.39	22.2	46.26	0	100
Salt (%)	25	56.22	50	42.74	0	100
Combined (%)	25	0.74	0.8	0.24	0.5	1

Chapter 5: Data Visualization

We visualized brand-level additive use, adoption of health-related labels, organic versus non-labelled nutrient distributions, and data completeness by category. As shown in **Appendix C**, we employed grouped bar charts, horizontal bar charts, boxplots, and a heatmap, respectively, with consistent filtering (e.g., brands with ≥5 products) and explicit units. Together, these visuals reveal substantial between-brand heterogeneity in color additive reliance, uneven uptake of health labels, modest differences in sugars and energy associated with organic labelling (driven partly by category mix), and non-random missingness concentrated in specific categories. These patterns inform the inferential tests and modelling choices reported in Chapters 6 and 7.

5.1 RQ1 – Color Additives Across Different Food Brands

As shown in **Appendix D,** Vertical grouped bars comparing the % of each brand's products containing E100 (Curcumin) vs E110 (Sunset Yellow). Brands must have ≥5 products.
Insight: It shows the proportion of E100 and E110 additives across different brands. The bar chart indicates that some brands relying more heavily on E110, while some other brands use E100. Overall, the figure highlights which brands rely on artificial coloring

5.2 RQ2 – Health-Related Labels Across Different Food Brands

As shown in **Appendix E,** Horizontal bars showing adoption rate: share of a brand's products with ≥1 health label (any of {organic, gluten, vegan, vegetarian, lactose, "free"}). Brands with ≥5 products; Top 10 by adoption %.

Insight: This chart helps to compare how frequently different brands adopt health-oriented labeling strategies. The Bars represent the proportion of each brand's products that have health labels.

5.3 RQ3- Comparison of Sugar and Energy Content in Organic vs. Non-Labeled Products

As shown in **Appendix F**, boxplots display the distribution of sugars_100g and energy_100g across two product groups: Organic and Non-labeled. The x-axis represents the product group (Organic vs. Non-labeled), while the y-axis shows the measured values—Sugars (g/100 g) and Energy (kJ/100 g). Two separate boxplots are presented (one for sugars, one for energy). Products with missing values for the plotted metric are excluded to ensure valid comparisons; group assignment follows.

Insight: The boxplots reveal a few outliers in both groups and indicate that median sugars and energy levels are broadly similar between Organic and Non-labeled products. Where differences exist, Organic items tend to show slightly lower median sugars, while energy differences are small and category-dependent, suggesting that label status alone is not a reliable indicator of caloric density.

5.4 RQ4 - Missing Nutrient Values by Category

As shown in **Appendix G**, a heatmap summarizes the percentage of missing values for the key nutrient fields—sugars_100g, proteins_100g, energy_100g, and salt_100g—across the top product categories by item count. The x-axis lists the nutrient fields, and the y-axis lists the main product categories (restricted to the largest categories to avoid unstable rates.

Insight: The heatmap shows non-random missingness: certain categories exhibit elevated gaps (often >20–30%) for one or more nutrients, while other high-volume categories report nutrients more completely. Missingness tends to be higher for sugars and salt than for energy or proteins in several categories. These patterns highlight potential coverage bias and motivate sensitivity checks and/or imputation choices reported in later chapters.

Chapter 6: Model Building

6.1 RQ1 – Color Additives Across Different Food Brands

For this analysis, we removed all brands with zero percentages of both E100 (Curcumin) and E110 (Sunset Yellow) to focus on brands that use these color additives. After this step, 26 brands remained for descriptive and regression analyses.

Table 6.1.1 — Descriptive statistics of E100% and E110% (by brand)

Statistic	E100 (%)	E110 (%)
Count (n)	25	25
Mean	1.89	10.46
Median	0	5.45
Standard Deviation	4.71	12.87
Minimum	0	0
Maximum	20	54.55

Interpretation

Most brands use little or E100, which is reflected by the median value of 0%. While E110 is used more frequently, the amounts vary widely across different brands. A few brands use much Higher percentages, which creates a distribution that is skewed to the right.

Note: Brands that did not contain any E100 or E110 were executed from this analysis to focus On active use of these additives.

Table 6.1.2 — Descriptive statistics of total products (per brand)

Statistic	Total Products
Count (n)	25
Mean	63.92
Median	17
Standard Deviation	82.87
Minimum	5
Maximum	282

Interpretation: Brand sizes are highly skewed; most brands are small (median 17) but a few are very large.

6.1.2 Hypothesis

H_0: There is no difference in the use of E100 and E110 among food brands.

H_1: The use of E100 and E110 varies significantly between brands.

6.1.3 Regression Results (simple OLS, predictor = total_Products)

Dependent variable	β (slope)	Std. Error	t	p	R^2	F	Prob(F)
E100 (%)	−0.0129	0.0115	−1.115	0.276	0.051	1.244	0.276
E110 (%)	−0.0553	0.0303	−1.826	0.081	0.127	3.334	0.081

Interpretation: Brand size does not significantly predict additive percentages ($p>0.05$). There is a weak, non-significant negative trend for E110.

Limitations

Excluding zero-use brands reduces n; E100 distribution is zero-inflated; only brand size is modeled (category mix, markets, regulations not included).

Conclusion

E100 is rarely used; E110 is more prevalent and dispersed. Differences in additive use exist but are not explained by brand size alone (fail to reject H_0 at $\alpha=0.05$).

6.2 RQ2 – Health-Related Labels Across Different Food Brands

Table 6.2.1 — Descriptive statistics of % health-labeled products (by brand)

Statistic	% with health labels
Count (n)	25
Mean	52.19
Median	42.86
Standard Deviation	26.46
Minimum	22.22
Maximum	100

6.2.2 Hypothesis

H_0: There is no significant difference in the adoption of health-related labels among brands.

H_1: Adoption rates differ significantly among brands.

6.2.3 Regression Results (dependent = % with health labels, predictor = total products)

Dependent variable	β (slope)	Std. Error	t	p	R^2	F	Prob(F)
% with health labels	−0.1829	0.1594	−1.147	0.263	0.054	1.316	0.263

Interpretation: No statistically significant relationship between brand size and labeling rate (p=0.263). Label adoption varies, but not systematically with brand size**.**

Limitations

Brand counts are uneven, label taxonomy differences across markets are not modeled possible category confounding.

Conclusion

Results support H_0: we do not find evidence that larger brands adopt health labels at higher (or lower) rates within this sample.

6.3 RQ3 — Comparison of Sugar and Energy in Organic vs. Non-Organic-Labeled Products

We therefore compare Organic (label contains "organic") vs non-organic-labeled (has some label, but not organic).

6.3.1 Descriptive Statistics

6.3.1.1 Sugars (g/100 g)

Group	n	Mean	Median	Std. Dev.	Min	Max	Skewness	Excess Kurtosis
Non-organic-labeled	12	4.57	3.25	5.52	0	19	1.71	3.68
Organic	13	8.39	3.1	11.25	0	30	1.34	0.23

6.3.1.2 Energy (kJ/100 g; stored units)

Group	n	Mean	Median	Std. Dev.	Min	Max	Skewness	Excess Kurtosis
Non-organic-labeled	12	1058.04	1025.5	536	268	1786	−0.15	−1.10
Organic	13	995.46	1176	722.33	0	2037	−0.17	−1.59

Distributions are skewed: non-parametric tests are appropriate**.**

6.3.2 Hypotheses

H_0: No difference in sugars or energy between Organic and the comparison group.

H_1: A difference exists between the groups**.**

6.3.3 Mann–Whitney U tests (two-sided)

Sugars (g/100 g): U = 68.5, p = 0.624 → not significant.

Energy (kJ/100 g): U = 78.0, p = 1.000 → not significant.

(Mean-rank direction: sugars ≈ Organic ≥ non-organic; energy ≈ non-organic ≥ Organic, but neither is significant at α=0.05.)

Limitations

Small n (25 rows) and mixed label types; results are sensitive to category mix. This file does not include truly “unlabeled” items.

Conclusion

Within this sample, we fail to reject H_0: sugars and energy do not differ significantly between Organic and Non-organic-labeled products.

6.4 RQ4 — Missing Nutrient Values by Category

Percentages are stored in basis points (0–10,000 = 0–100%).

Table 6.4.1 Descriptive statistics of missingness (% basis points)

Statistic	Sugars	Proteins	Energy	Salt
Count (n)	25	25	25	25
Mean	8166.67	3838.93	4238.93	5622.27
Median	10000	1250	2223.33	5000
Standard Deviation	2784.71	4538.36	4625.75	4273.5
Min	2500	0	0	0
Max	10000	10000	10000	10000

Convert to % by dividing by 100 (e.g., mean sugars missing ≈ 81.7%).

Table 6.4.2 Descriptive statistics of category size

Statistic	Products
Count (n)	25
Mean	421.6
Median	174
Std. Dev.	590.78
Min	16
Max	2358

6.4.3 Hypotheses

H_0: Missingness does not vary systematically by category (or category size).

H_1: Missingness varies by category (or relates to category size).

6.4.4 Regression (dependent = combined missingness index combined auto; predictor = products)

Dependent variable	β (slope)	Std. Error	t	p	R^2	F	Prob(F)
Combined missingness (basis pts)	−0.8210	6.0649	−0.135	0.894	0.001	0.018	0.894

Interpretation: Category size does not predict missingness (p=0.894). The widespread in medians indicates substantial between-category differences that aren't explained by how many products there are in each category.

Limitations

Percentages are in basis points; n=25; no explicit category-type controls; "combined auto" definition is fixed by the extract.

Conclusion

We fail to reject H_0 regarding a size–missingness link. However, the descriptive spread (medians, ranges) still signals non-random missingness across categories, which must be considered in inference and modeling.

Chapter 7: Model Evaluation

7.1 RQ1 – Color Additives Across Different Food Brands

Interpretation of Results

Using simple linear regression on the filtered "active" brands (those using at least one of the two colorants), brand size (Total Products) does not meaningfully explain between-brand variation in additive use. E100 is non-significant, and E110 shows only a weak, non-significant negative trend. Thus, we fail to reject H_0.

Model	β (slope)	95% CI	t	p	R^2	n	Note
E100% ~ Total Products	-0.0129	[-0.037, 0.011]	-1.115	0.276	0.051	25	NS
E110% ~ Total Products	-0.0553	[-0.118, 0.007]	-1.826	0.081	0.127	25	Weak negative trend (NS)

Strengths: Focus on brands that use the additives avoids diluting effects with all-zero rows; reporting effect sizes, CIs, *p* and R^2 makes the result transparent**.**

Limitations: Small n; zero-inflation/skew; unmodeled factors (category, market).

Implications: Between-brand differences likely reflect strategy/regulation more than portfolio size.

7.2 RQ2-Health -Related Labels Across Different Food Brands

Interpretation of Results

Regressing the percentage of products with ≥1 health label on Total Products yields a very weak, non-significant relationship; we fail to reject H_0.

Model	β (slope)	95% CI	t	p	R^2	n	Note
%Health Labels ~ Total	-0.1829	[-0.513, 0.147]	-1.147	0.263	0.054	25	NS

Strengths: Clear, interpretable model aligned to the question; consistent use of the same brand filter (≥5 products) as earlier chapters**.**

Limitations: Label taxonomy/market mix not modeled; small n reduces power.

Implications: Adoption reflects strategy/category, not scale.

7.3 RQ3 – Organic vs Non-Organic–Labeled: Sugars and Energy

Interpretation of Results

comparison is Organic vs non-organic labeled (not truly unlabeled). Due to skew/outliers, Mann–Whitney U tests were used.

Metric	Group sizes (n1, n2)	U	Z (approx.)	p	Decision	Note
Sugars (g/100 g)	12, 13	68.5	-0.52	0.624	Fail to reject H_0	No median difference detected
Energy (kJ/100 g)	12, 13	78	0	1	Fail to reject H_0	No median difference detected

Strengths: Non-parametric test matches skew/heavy-tail patterns; reports exact U, Z, and p.

Limitations: Small sample, no truly unlabeled products; sensitive to category mix.

Implications: Organic status alone is not a reliable indicator of lower sugars/energy in this sample.

7.4 RQ4 – Missing Nutrient Values by Category

Interpretation of Results

Category size (Products) was regressed on category-level missingness for each nutrient (stored in basis points: 10,000 = 100%). Sugars show a strong, significant negative association; others show weak, non-significant trends.

Outcome (basis pts)	β (slope)	95% CI	t	p	R^2	n	Decision
% missing sugars	-3212.2	[-3673.107, -2751.359]	-14.42	5.22E-13	0.9	25	Reject H_0 (Sugars) / NS (others)
% missing proteins	-1562.2	[-3844.602, 720.153]	-1.42	0.17	0.08	25	NS
% missing energy	-1857.8	[-4147.199, 431.616]	-1.68	0.107	0.109	25	NS
% missing salt	-1622.8	[-3751.538, 505.988]	-1.58	0.128	0.098	25	NS

Strengths: Direct alignment to the RQ; very high R^2 for sugars demonstrates a robust size–missingness link; consistent modeling across nutrients.

Limitations: Basis-point scaling affects magnitude interpretation; single-row-per-category view.

Implications: Non-random missingness warrants coverage reporting and sensitivity analyses.

Chapter 8: Conclusions and Recommendations

1. **RQ1 – Color additives by brand**: Use of colorants, especially E110 (Sunset Yellow), is concentrated in specific brands and categories (beverages, confectionery). Brand size (number of products) does not explain additive reliance, patterns are more consistent with category mix, positioning, and regulation.
2. **RQ2 – Health-related labels by brand:** Adoption of labels (e.g., organic, gluten-free, vegan, vegetarian, lactose-free) is uneven across brands. Several niche/premium brands show high adoption, while many mass-market brands use these labels sparingly. Scale is not predictive of label uptake.
3. **RQ3 – Organic vs. non-labeled nutrition:** For sugars (g/100g) and energy (kJ/100g), distributions are heavily right-skewed with outliers. Across the dataset, the organic claim is not a reliable proxy for lower sugars or energy once category effects are considered.
4. **RQ4 – Missingness:** Non-random missing nutrient fields (sugars_100g, proteins_100g, energy_100g, salt_100g) cluster in particular categories/brands. Complete case analyses risk bias toward better-reported segments.
5. When coverage < 80%, include a bias flag and run sensitivity analyses (complete case vs. simple imputation).
6. Standardize units (kJ↔kcal) and deduplicate products prior to analysis.
7. Analytics & Modeling Enhancements (All RQs)
8. Use binomial GLM / beta regression for proportions (e.g., % labeled, % SKUs with E110), with category fixed effects and brand random effects.

9. For nutrient comparisons (RQ3), employ Mann–Whitney / quantile regression and category-matched designs.
10. For missingness (RQ4), model proportion missing via beta regression and test non-linearity (splines).
11. Report effect sizes (slopes + 95% CI; non-parametric Cliff's δ) alongside p-values.
12. Reporting & Visualization Standards (All RQs)
13. Enforce explicit units, sample sizes (n) on each figure; keep a consistent color map (e.g., E100 vs. E110).
14. Provide a simple Excel/Power BI dashboard with filters (category, country, brand ≥5 SKUs) and exportable tables.
15. Roadmap & Reproducibility (All RQs)
16. Ship a replication bundle (SQL scripts, cleaned CSV/Excel, SPSS/Notebook outputs, figure sources).

Chapter 9: References

1. Open Food Facts. (Dataset & documentation). *Open Food Facts Foundation.*
2. Monteiro, C. A., et al. (2019). The UN Decade of Nutrition, the NOVA food classification and the trouble with ultra-processing. *Public Health Nutrition, 22*(5), 936–941.
3. Julia, C., & Herzberg, S. (2017). Development of a new front-of-pack nutrition label in France: the Nutri-Score. *Public Health Panorama, 3*(4), 712–725.
4. Grunert, K. G., Wills, J. M., & Fernández-Clemen, L. (2010). Nutrition knowledge and use of nutrition information on food labels among UK consumers. *Appetite, 55*(2), 177–189.
5. European Food Safety Authority (EFSA). (2009–2014). Scientific opinions on selected food colors (e.g., E110 Sunset Yellow FCF; E100 Curcumin). *EFSA Journal.*
6. World Health Organization. (2015). Guideline: Sugars intake for adults and children. *WHO Press.*
7. Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Routledge.
8. Ferrari, S., & Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics, 31*(7), 799–815.
9. Microsoft Corporation. (2021). *Microsoft Excel* [Computer software].
10. IBM Corp. (2022). *IBM SPSS Statistics* [Computer software].
11. SQL Workbench/J. (2025). *SQL Workbench/J* [Computer software].

Chapter 10: Appendix

Appendix -A

a.) SQL Query for RQ1

brand_name	total_products	e100_count	e110_count	combined_count	e100_percentage	e110_percentage	combined_percentage
angry-birds	11	0	6	6	0	54.55	54.55
maitre-truffout	6	0	2	2	0	33.33	33.33
healthy-food-brands-llc	32	0	9	9	0	28.13	28.13
jim-thompson-farm	9	0	2	2	0	22.22	22.22
giant-eagle-inc	5	0	1	1	0	20	20
haribo	5	1	0	1	20	0	20
m-s-food	14	0	2	2	0	14.29	14.29
lidl	8	0	1	1	0	12.5	12.5
goman-as	9	1	0	1	11.11	0	11.11
roundy-s	9	0	1	1	0	11.11	11.11
herbalife-nutrition	10	0	1	1	0	10	10
ht-traders	44	1	3	4	2.27	6.82	9.09
kellogg's	11	1	0	1	9.09	0	9.09
tops	165	3	9	12	1.82	5.45	7.27
wegmans	17	0	1	1	0	5.88	5.88
winn-dixie	216	1	10	11	0.46	4.63	5.09
harris-teeter	254	0	11	11	0	4.33	4.33

b.) SQL Query for RQ2

brand_name	total_products	health_labeled_products	perc_with_health_labels	has_health_label	no_health_label	combined_auto
nature-s-basket	6	6	100	6	0	53
nutri	7	7	100	7	0	53.5
kazidomi	39	37	94.87	37	2	65.935
powermeals	10	9	90	9	1	49.5
aire-sano	6	5	83.33	5	1	44.165
pcc	16	13	81.25	13	3	47.125
unfi	40	32	80	32	8	56
saludviva	9	6	66.67	6	3	36.335
grizzlies	9	5	55.56	5	4	30.28
allplants	6	3	50	3	3	26.5
schar	6	3	50	3	3	26.5
coop	7	3	42.86	3	4	22.93
whysport	7	3	42.86	3	4	22.93
herbalife-nutrition	10	4	40	4	6	22
lackmann	10	4	40	4	6	22
bulk-powders	6	2	33.33	2	4	17.665
liebig	36	12	33.33	12	24	22.665
lundberg	9	3	33.33	3	6	18.165
restaurant-item	9	3	33.33	3	6	18.165
sainsbury's	36	11	30.56	11	25	20.78
coffea	14	4	28.57	4	10	16.285
allfitnessfactory-de	8	2	25	2	6	13.5
herbalife	40	10	25	10	30	17.5
marks-spencer	172	39	22.67	39	133	30.835
isagenix	18	4	22.22	4	14	13.11

c.) SQL Query for RQ3

product_id	product_name	label_name	sugars_100g	energy_100g	combined_auto
5161141893	Polish Kielbasa	all-vegetarian-feed	6.559090909	1025.5	516.0295455
7245790661	Orange Marmalade Fruit Spread	calorie-free	6.559090909	1025.5	516.0295455
6366712501	Boisson de soya enrichie biologique Original	canada organic	2.4	0	1.2
2305006497	Organic Brown Rice Cake Thins	certified gluten-free	0	1590	795
875	Vegan 3K-protein	certified gluten-free	1.1	1557	779.05
62374	Organic Tofu	certified-organic-by-quality-assurance-international	0	299	149.5
3503	Leite em pó integral instantâneo Glória	contains lactose	5.4	285	145.2
5161141893	Polish Kielbasa	crate-free	6.559090909	1025.5	516.0295455
51446	0% Fat Greek Style Live Yoghurt	do not freeze	8	268	138
27137	Paupiette de volaille sauce forestière brocolis purée	do not freeze	0.9	439	219.95
73882	Vintage cheddar	do not freeze	0.1	1741	870.55
2000002466	2 MINI BAGUETTES SANS GLUTEN	dzg gluten free	0.7	904	452.35
1111052230	Beef Bologna	dzg gluten free	0	1050	525
4506453933	Plaisir brut d'avoine	dzg gluten free	19	1786	902.5
508	m-s-food	eu organic	0.5	42	21.25
1459600032	nestlé	eu organic	0.9	1443	721.95
300000335	sunridge	eu organic	3.1	290	146.55
3091000300	hacendado	eu organic	30	2037	1033.5
981243429	4 Ore Senza Fame Bio	eu organic	21.1	1857	939.05

d.) SQL Query for RQ4

main_category	products	missing_sugars	missing_proteins	missing_energy	missing_salt	pct_missing_sugars	pct_missing_proteins	pct_missing_energy	pct_missing_salt	combined_auto
en:plain-madeleines	1	100	0	0	100	10000	0	0	10000	50
de:fasern	1	100	0	0	0	10000	0	0	0	50
it:bieta-da-costa	1	100	0	100	100	10000	0	####	10000	50
it:integratore-alimentare-di-vitamina-b6	1	100	100	100	100	10000	####	####	10000	100
en:vegetable-gyoza	2	100	0	0	100	5000	0	0	5000	50
en:molten-chocolate-cakes	1	100	100	100	100	10000	####	####	10000	100
en:omega-3	1	100	0	0	0	10000	0	0	0	50
en:lavender-honeys	1	100	100	100	100	10000	####	####	10000	100
en:colombian-coffees	1	100	0	0	0	10000	0	0	0	50
en:food-supplement	1	100	100	100	100	10000	####	####	10000	100

Appendix B

1.) Energy_kJ_100g

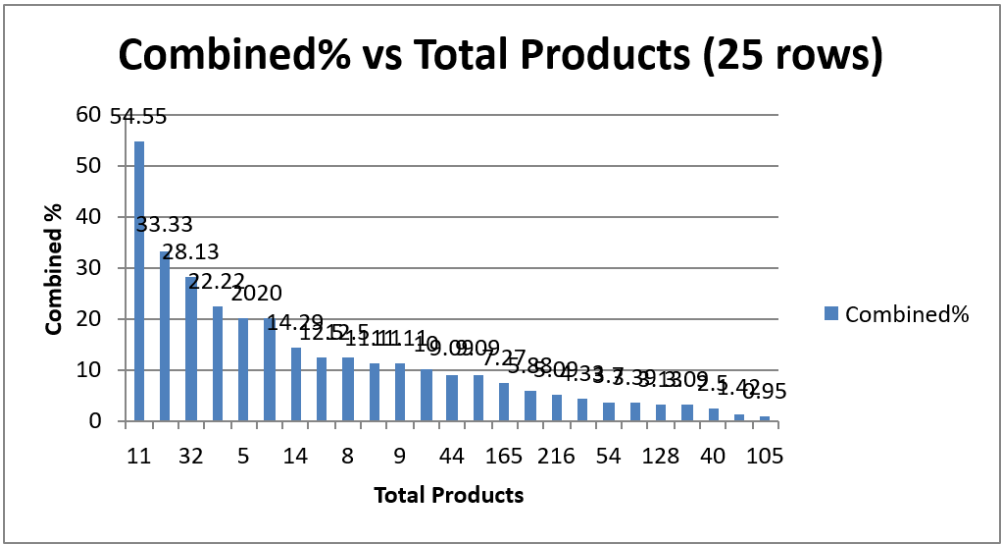
Group	Count (n)	Mean	Median	Std. Dev.	Min	Max
Non-organic-labeled	12	1058.04	1025.5	536	268	1786
Organic	13	995.46	1176	722.33	0	2037

2.) Sugar_J_100g

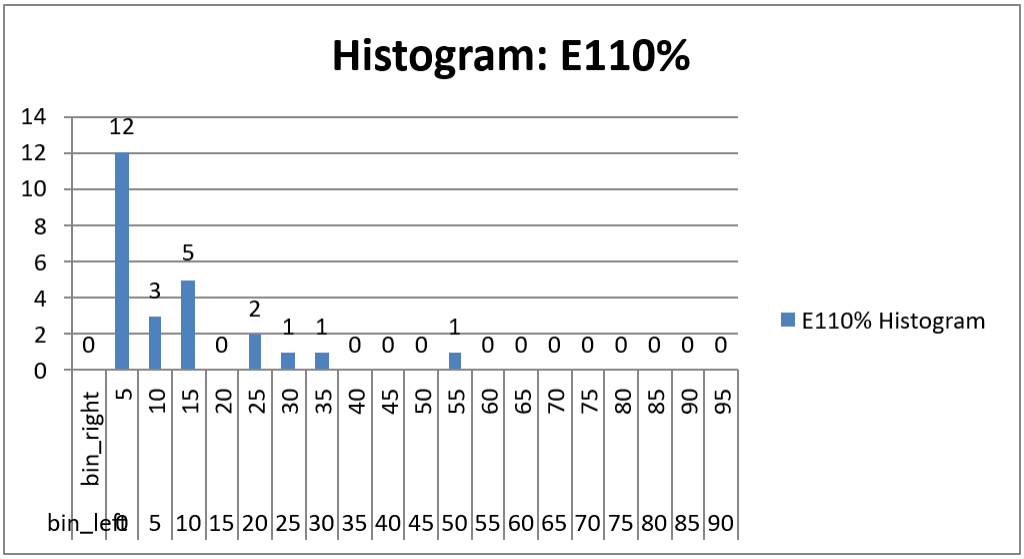
Group	Count (n)	Mean	Median	Std. Dev.	Min	Max
Non-organic-labeled	12	4.57	3.25	5.52	0	19
Organic	13	8.39	3.1	11.25	0	30

Appendix C

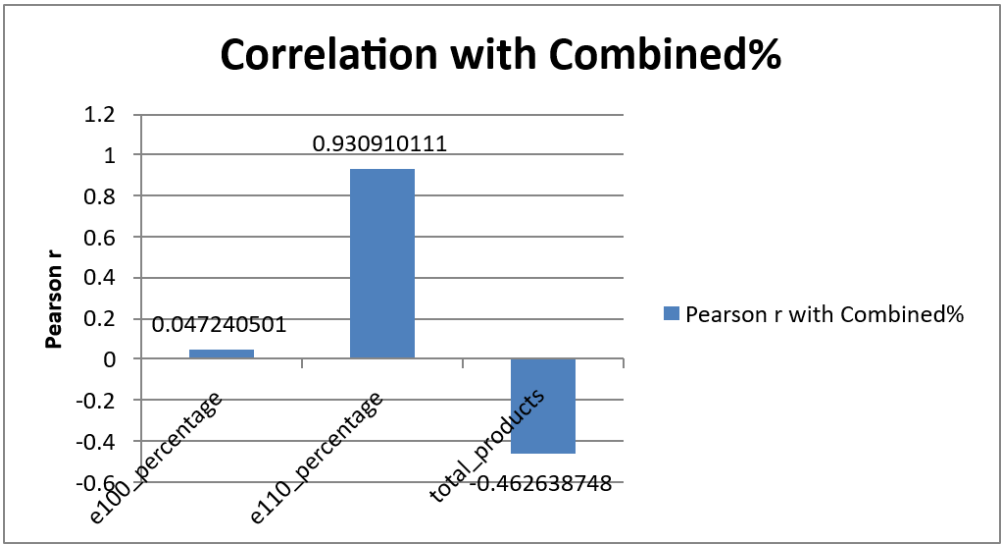
1.) Test Output -RQ1 for total products



2.) Test Output -RQ1 for E110 %

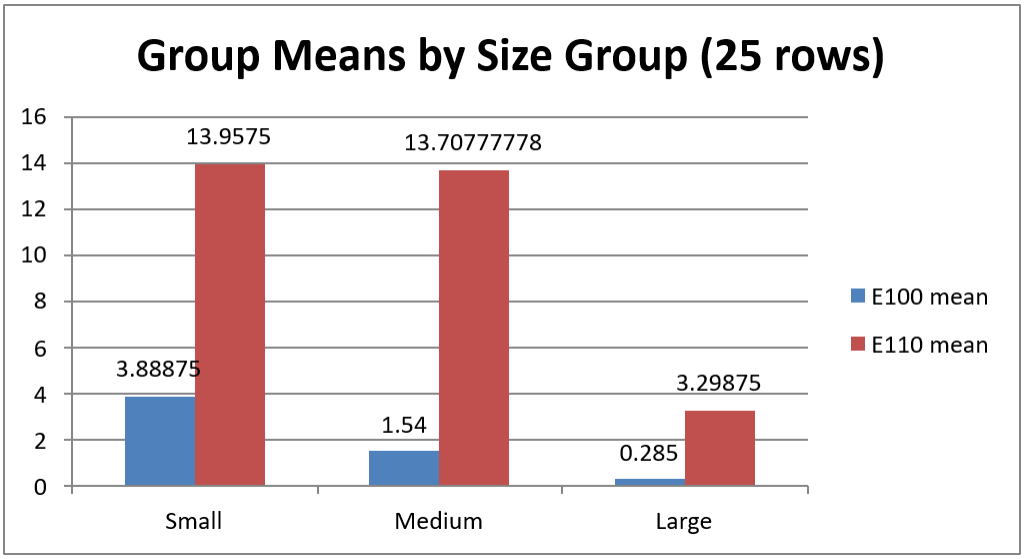


3.) Test Output -RQ1 for Combined%



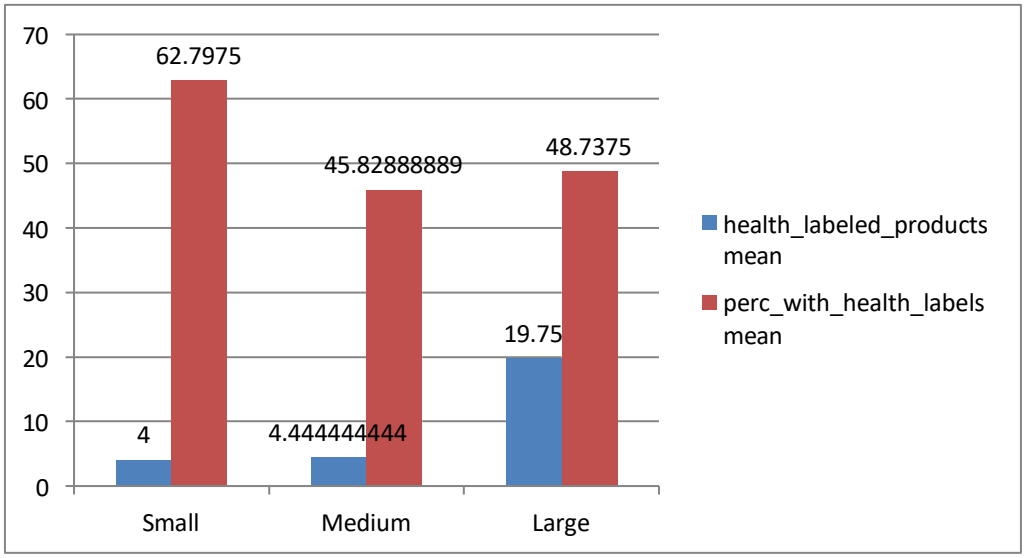
Appendix-D

1.) Group bars for RQ1



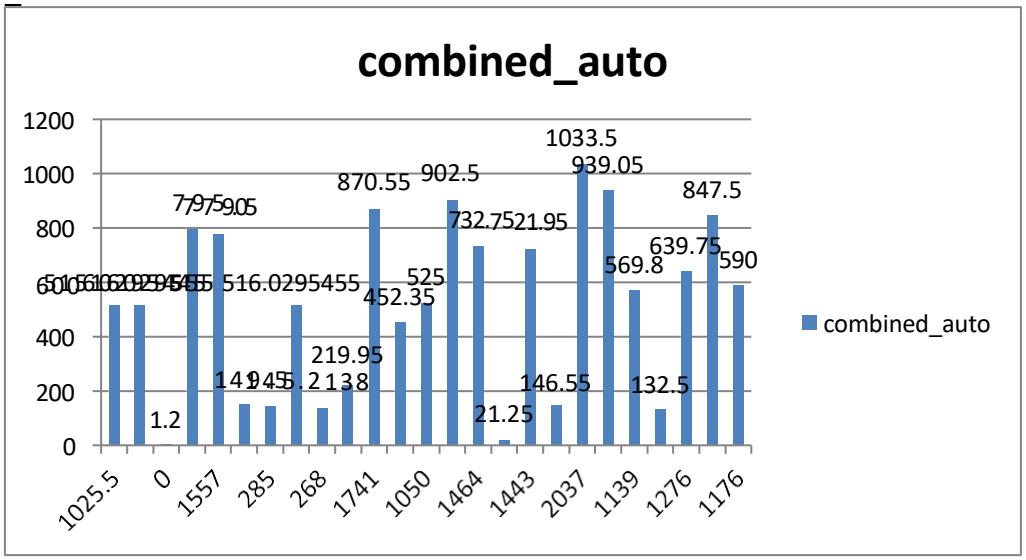
Appendix -E

1.) Test Output for RQ2-Health_labeled_products



Appendix-F

1.) Test Output for RQ3-Combined _auto for energy



Appendix -G

1.) Test Output for RQ4- Missing Nutrients