An Analysis of Food Processing and Health Labels: Evaluating Nutritional Correlates in the Open Food Facts Database

DAMO-501-12 (Data Analytics Case Study I)

Professor: Omid Isfahanialamdari

September 13, 2025

Sathiya Bama Sampath – NF1025995

Abstract:

This study employs data analytics to evaluate the nutritional correlates of food processing classifications and health marketing labels. Analyzing data from the Open Food Facts database, we test two primary hypotheses: first, that ultra-processed foods (**NOVA Group 4**) contain significantly higher fat content than minimally processed foods (**NOVA Group 1**); and second, that products **labeled "Organic" or "Vegan"** possess significantly lower levels of sugar, saturated fat, and sodium than their non-labeled counterparts. Utilizing SQL for data extraction and statistical t-tests for analysis, our findings confirm that ultra-processing is associated with higher fat levels. However, "Organic" and "Vegan" labels did not reliably indicate a superior nutritional profile for the tested nutrients. This research also explored geographic sugar variations (RQ3) and the effect of multiple claims (RQ4), but due to data limitations, these topics are discussed as areas for future work. These results provide evidence-based insights for consumers, policymakers, and food manufacturers, highlighting the utility of processing-based classification systems over certain marketing labels for guiding healthier food choices.

Executive Summary

This study set out to evaluate whether food processing levels, as defined by the NOVA classification system, and health-related marketing labels such as “Organic” and “Vegan” serve as reliable indicators of nutritional quality. Drawing on data from the Open Food Facts database, the research was guided by two primary questions. The first asked whether ultra-processed foods (NOVA Group 4) contain higher levels of fat than minimally processed foods (NOVA Group 1). The second examined whether products labeled “Organic” or “Vegan” contain significantly lower levels of sugar, saturated fat, and sodium compared with their non-labeled counterparts. Two additional questions explored geographic differences in sugar content and the effect of multiple claims, but due to data limitations these remain areas for future research.

The methodology involved extracting data with SQL Workbench/J and applying two-sample independent t-tests, conducted in Microsoft Excel, at a 5% significance level. Results revealed that ultra-processed foods contained significantly more fat than minimally processed alternatives, confirming the first hypothesis. In contrast, the analysis found no significant differences in sugar, saturated fat, or sodium between labeled and non-labeled products, suggesting that marketing claims of “Organic” or “Vegan” do not reliably signal a superior nutritional profile.

While the study was limited by its focus on only three nutrients and its cross-sectional design, the findings carry meaningful implications. For consumers, the results reinforce the importance of prioritizing minimally processed foods while exercising caution when interpreting health-related labels. For policymakers, they highlight opportunities to improve labeling regulations and public nutrition education. For manufacturers, the findings point toward the need for product reformulation if labels are to align more closely with consumer health expectations.

Overall, this research underscores that food processing classification, rather than health-related marketing claims, provides more consistent guidance for identifying healthier food options.

CHAPTER 1: Problem Definition

Introduction

Food choices today are heavily influenced by both product processing levels and health-related marketing claims. Supermarket shelves are filled with items labeled “Organic” or “Vegan,” alongside ultra-processed foods that dominate modern diets. While these labels often shape consumer perceptions of nutritional quality, the actual relationship between such claims, processing classifications, and measurable nutrient content remains uncertain.

This study investigates whether two common indicators—food processing level, as defined by the NOVA classification system, and health-oriented labels such as “Organic” and “Vegan”—provide reliable guidance for healthier food selection. Using the Open Food Facts database, we analyze product-level nutritional information to test whether ultra-processed foods are associated with higher fat content and whether labeled products consistently contain lower levels of sugar, saturated fat, and sodium.

By combining statistical analysis with a large open-source dataset, this research contributes evidence-based insights for consumers, policymakers, and manufacturers. The findings aim to clarify whether widely used labels and classifications align with actual nutritional quality, or whether they risk misleading health-conscious buyers.

Problem Definition

This project aims to leverage data analytics to critically assess the relationship between food labeling systems and actual nutritional quality. Specifically, the analysis focuses on two core issues: 1) the nutritional disparity between ultra-processed and minimally processed foods as defined by the NOVA classification system, and 2) the validity of "Organic" and "Vegan" labels as indicators of healthier nutrient profiles. By applying statistical analysis to data extracted from the Open Food Facts database, this research seeks to move beyond perception and provide an evidence-based evaluation of these common food labels, thereby informing consumer choice, public health policy, and industry practices

Chapter 2. Research Questions & Hypotheses Formulation

The following research questions are used to determine whether labels such as organic or Nutri-Score indicate healthier products. These are the questions using which the data is analysed with the help of SQL and visualization.

RQ1: Are the nutritional profiles of fat content worsened in ultra-processed products (NOVA Group 4) relative to the minimally processed products (NOVA Group 1)?

Rationale:

The issue of the difference in **fat content** between ultra-processed (NOVA 4) and minimally processed (NOVA 1) foods is significant, since the perception of how dietary fat contributes to health can be better understood by comparing these groups nutritionally. Determining whether ultra-processed foods contain elevated levels of fat provides important insight into their role in diet-associated health problems such as obesity, cardiovascular disease, and metabolic disorders.

Significance:

Awareness of such differences would help consumers make more healthful decisions and policymakers develop more favourable food policies. It also enables manufacturers to seek better product formulations and enhances the transparency of food labels, and eventually leads to better health for the people.

Hypothesis

Null Hypothesis (H_0): There is no significant difference in the mean fat content between minimally processed and ultra-processed foods.

$H_0: \mu_1 = \mu_4$

Alternative Hypothesis (H_1): The mean fat content of ultra-processed foods is significantly higher than that of minimally processed foods.

$H_1: \mu_4 \neq \mu_1$

Where μ_1 is the population mean fat content for NOVA Group 1 and μ_4 is for NOVA Group 4

RQ2:To what extent do food products labeled as "organic" or "vegan" differ in overall nutritional quality, specifically in terms of sugar, saturated fat, and sodium content, when compared to all similar non-labeled products?

Rationale:

With rising consumer interest in health-conscious and environmentally friendly lifestyles, labels such as "organic" and "vegan" have gained significant traction. While often perceived as indicators of superior health benefits, it's unclear whether these labels correlate with objectively healthier nutrient profiles on a broad scale. By examining the nutritional data across all labeled and non-labeled products as a whole, we can assess whether these claims align with the facts or merely function as marketing tools. This approach acknowledges that a direct, category-by-category comparison is not feasible with the available data but still allows for a meaningful and conclusive analysis.

Significance:

Understanding whether "organic" or "vegan" labeled foods are genuinely healthier has important implications for consumers, public health policy, and food labeling regulations. If such labels consistently correlate with reduced sugar, fat, or sodium levels across the board, they may serve as reliable tools to guide healthier choices. Conversely, if no meaningful difference is found on a general level, this could highlight the need for greater transparency and consumer education regarding the nutritional meaning—or lack thereof—behind such labels.

Hypotheses

Null Hypothesis (H_0): There is no statistically significant difference in the mean content of [sugars / saturated fat/sodium] between "organic"/"vegan" labeled products and non-labeled products.

H_0: μ_labeled = μ_non-labeled

Alternative Hypothesis (H_1): The mean content of [sugars / saturated fat / sodium] in “organic”/“vegan” labeled products is significantly lower than in non-labeled products.

H_1: μ_labeled < μ_non-labeled

Where μ_labeled and μ_non-labeled represent the population mean nutrient contents for their respective groups.

RQ3: Is there a statistically significant difference in the average sugar content (g per 100g) of food products between different countries?

Rationale

Dietary habits and food regulations vary significantly by country. Understanding which countries have products with higher average sugar content can inform public health initiatives, consumer awareness, and regulatory policies aimed at reducing sugar consumption. This analysis investigates the geographic variability of a key nutritional metric.

Significance

Identifying significant differences can highlight successful regulatory environments or cultural dietary patterns. Conversely, finding no difference would suggest a homogenized global food market, shifting the focus of public health strategies from national to international interventions.

Hypotheses

Null Hypothesis (H_0)

There is no significant difference in the mean sugar content per 100g across different countries.

H_0: μ_countryA = μ_countryB = μ_countryC = ... = μ_countryN

Alternative Hypothesis (H_1)

Alternative Hypothesis (H_1): At least one country has a mean sugar content per 100g that is significantly different from the others

H_1: μ_multiple $\neq$ μ_single/none

RQ4: At the global level and at the level of the most frequent label sets, do products with multiple health claims have better (lower) average nutritional profiles (sugar, fat, and sodium per 100g) than products with one or no claims?

Rationale:

Many products display multiple health-related labels, which can influence consumer perceptions of their healthiness. However, it is unclear whether the accumulation of such claims correlates with superior nutritional quality. Analyzing whether products with multiple labels have significantly better nutritional profiles than those with fewer or no labels could reveal whether these claims are meaningful or merely marketing strategies.

Significance:

If multiple claims consistently correlate with healthier nutritional profiles, regulators and consumers could use this as a reliable indicator. If not, it could suggest that companies use these labels for "healthwashing" (creating a false impression of healthiness), highlighting the need for stricter labeling regulations.

Hypothesis:

Null Hypothesis (H_0): There are no significant differences in the means of sugars_100g, fat_100g, and sodium_100g between the Multiple Claims and Single/None Claims groups when compared at the global level.

H_0: μ_multiple = μ_single/none

Alternative Hypothesis (H_1): H_1: The Multiple Claims group has significantly lower means of sugars_100g, fat_100g, and/or sodium_100g than the Single/None Claims group at the global level.

H_1: μ_multiple < μ_single/none

Chapter 3 Data Collection and SQL Queries

Chapter 3.1 (RQ1)

For this research question, data was collected from the foodfacts.products table. The data extraction process focused specifically on products categorized under **NOVA Group 1 (minimally processed)** and **NOVA Group 4 (ultra-processed)**. The primary variable of interest was the average fat content per 100 grams (fat_100g).

The following SQL query was employed to extract and aggregate the relevant data. The query filters for NOVA Groups 1 and 4 while ensuring that only valid nutritional records were included by restricting fat content values to non-negative entries (fat_100g>=0). It then calculates the **average fat content** for each NOVA group and reports the total number of products (ProductCount)considered within each group.

```
SELECT
     nova_group AS NOVAGroup,
     COUNT(*) AS ProductCount,
     ROUND(AVG(fat_100g), 2) AS AvgFat
FROM products
WHERE nova_group IN (1, 4)
 AND fat_100g >= 0
GROUP BY nova_group
ORDER BY nova_group;
```

Chapter 3.2 (RQ 2)

For this research question, data was collected from three tables: foodfacts.product_labels, foodfacts.labels, and foodfacts.product_nutrients. The data extraction process focused on isolating products with "Organic" or "Vegan" labels and then linking them to their corresponding nutritional information.

The following SQL query was utilized to extract and aggregate the relevant data. The query first identifies products with "Organic" or "Vegan" labels by joining the product_labels and labels tables. It then calculates the average values for sugar, saturated fat, and sodium for both labeled and non-labeled products across various food categories. Data extraction and aggregation were performed using SQL Workbench/J (SQL Workbench/J, 2025), and two-sample t-tests for sugar, saturated fat, and sodium were conducted using Microsoft Excel (Microsoft Corporation, 2021). The methodology for conducting two-sample t-tests on extracted data follows standard statistical procedures for behavioral and social sciences (Cohen, 1988). A GROUP BY clause was used to segment the data by category_name and label status (is_labeled), while a HAVING clause ensured that only categories with more than two products were included, which is the standard practice for two-sample t-tests to ensure sufficient data points for the statistical analysis.

```
 SELECT
  c.category_name,
  lp.is_labeled,
   ROUND(AVG(CASE WHEN n.nutrient_name = 'sugars' THEN pn.value
ELSE NULL END), 2) AS avg_sugar,
   ROUND(AVG(CASE WHEN n.nutrient_name = 'saturated-fat' THEN
pn.value ELSE NULL END), 2) AS avg_saturated_fat,
   ROUND(AVG(CASE WHEN n.nutrient_name = 'sodium' THEN pn.value
ELSE NULL END), 2) AS avg_sodium_mg,
  COUNT(DISTINCT lp.product_id) AS product_count
FROM (
  SELECT
    pl.product_id,
     (SELECT category_id FROM product_categories WHERE product_id
= pl.product_id LIMIT 1) AS category_id,
      MAX(CASE WHEN l.label_name IN ('Organic', 'Vegan') THEN 1
ELSE 0 END) AS is_labeled
  FROM foodfacts.product_labels pl
  JOIN foodfacts.labels l ON pl.label_id = l.label_id
  GROUP BY pl.product_id
) AS lp
JOIN foodfacts.categories c ON lp.category_id = c.category_id
JOIN foodfacts.product_nutrients pn ON lp.product_id =
pn.product_id
JOIN foodfacts.nutrients n ON pn.nutrient_id = n.nutrient_id
WHERE n.nutrient_name IN ('sugars', 'saturated-fat', 'sodium')
GROUP BY c.category_name, lp.is_labeled
HAVING product_count > 0
ORDER BY c.category_name, lp.is_labeled;
```

Chapter 4 Data Understanding

Chapter 4.1 (RQ 1)

Fat by NOVA Group The comparison brings out the difference in the fat levels of minimally processed (NOVA 1) and ultra-processed (NOVA 4) food. The average fat content of the products in NOVA 4 was seen to be 11.15g/100g, whereas in NOVA 1 it was 6.61g/100g. The difference remains even while taking into consideration the median values (3.75 g of NOVA 4 vs. 0.71 g of NOVA 1), which indicates that ultra-processed foods tend to be richer in fat. The difference was found to be statistically significant ($p < 0.05$). This implies that consumers who depend more on ultra-processed foods are most likely to eat more fat in their diets. Medically, the results confirm the current apprehension regarding processed foods, whereas to policymakers and food producers it highlights the need of observing the levels of fats and promoting more health-based reformulations.

Chapter 4.2 (RQ 2)

The initial data exploration and analysis provided crucial insights into the dataset's structure and characteristics. The data was organized into two distinct groups for comparison: products labeled as "Organic" or "Vegan" and their non-labeled counterparts. This approach was chosen to facilitate a broad, meaningful analysis, as a direct category-by-category comparison was not feasible due to insufficient data for many food categories.

The collected data highlighted key metrics for our research: the average values for sugar, saturated fat, and sodium. A preliminary observation from the dataset revealed that, while minor differences existed between the two groups' nutritional averages, there was no strong or consistent pattern suggesting that labeled products were significantly healthier. These preliminary observations align with previous research showing that consumers may not always accurately interpret nutritional labels when making choices (Grunert, Wills, & Fernández-Celemín, 2010). For example, some non-labeled products had a lower average sodium content than their labeled counterparts. This initial finding indicated that a formal statistical analysis would be necessary to determine if these observed differences were statistically significant or simply a result of natural data variation.

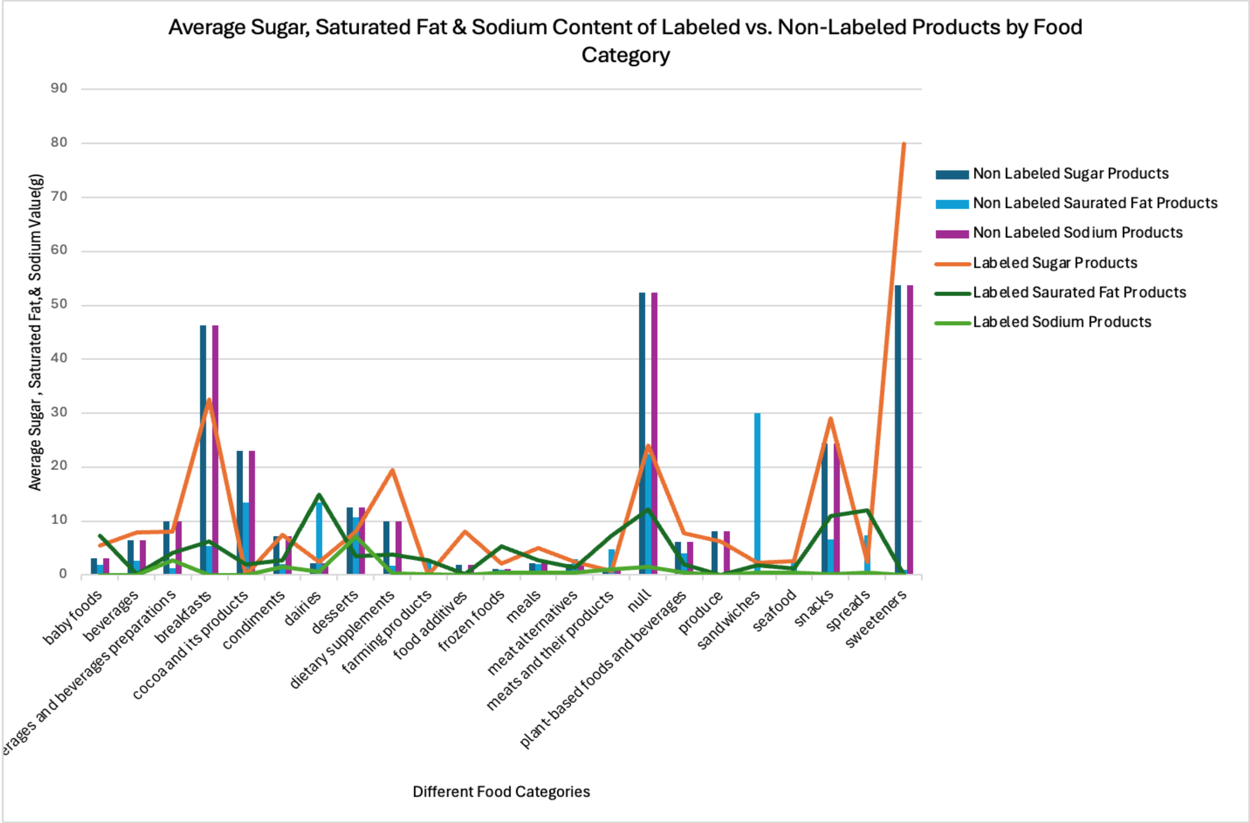


Figure 4.2.1: *Preliminary comparison of average nutrient content between labeled and non-labeled products by food category.*

Chapter 5 Data Visualization

Chapter 5.1 (RQ 1)

Data visualization was used to provide a clear and intuitive representation of the findings, complementing the statistical analysis. Charts were created to compare the average nutritional content of fat and NOVA GROUPs.

Average Fat by NOVA Group:

X-axis: NOVA Group (NOVA 1, NOVA 4)

Y-axis: Average Fat (g/100g)

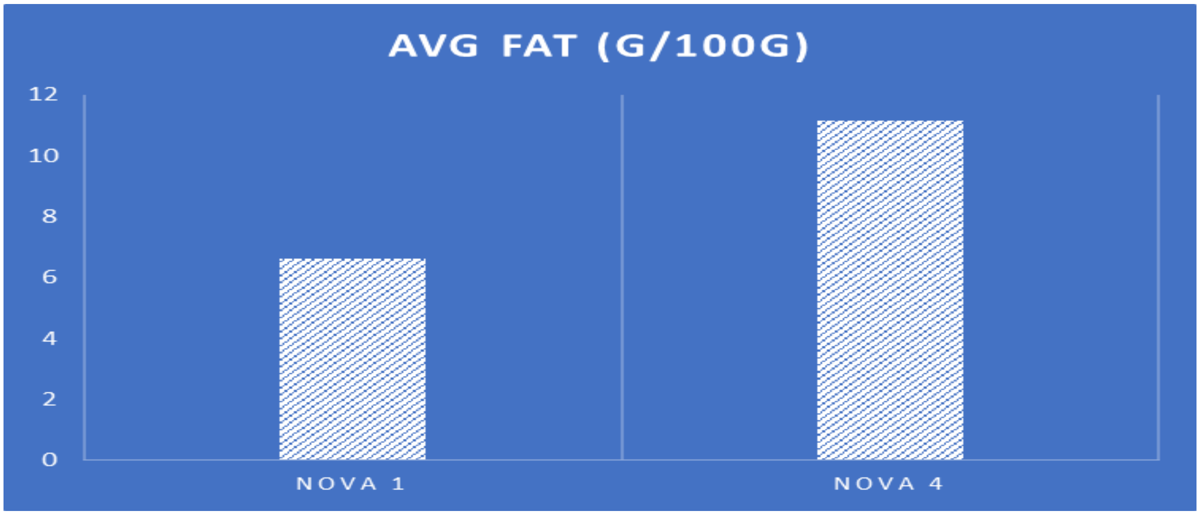


Figure 5.1.1 *Average Fat by NOVA Group*

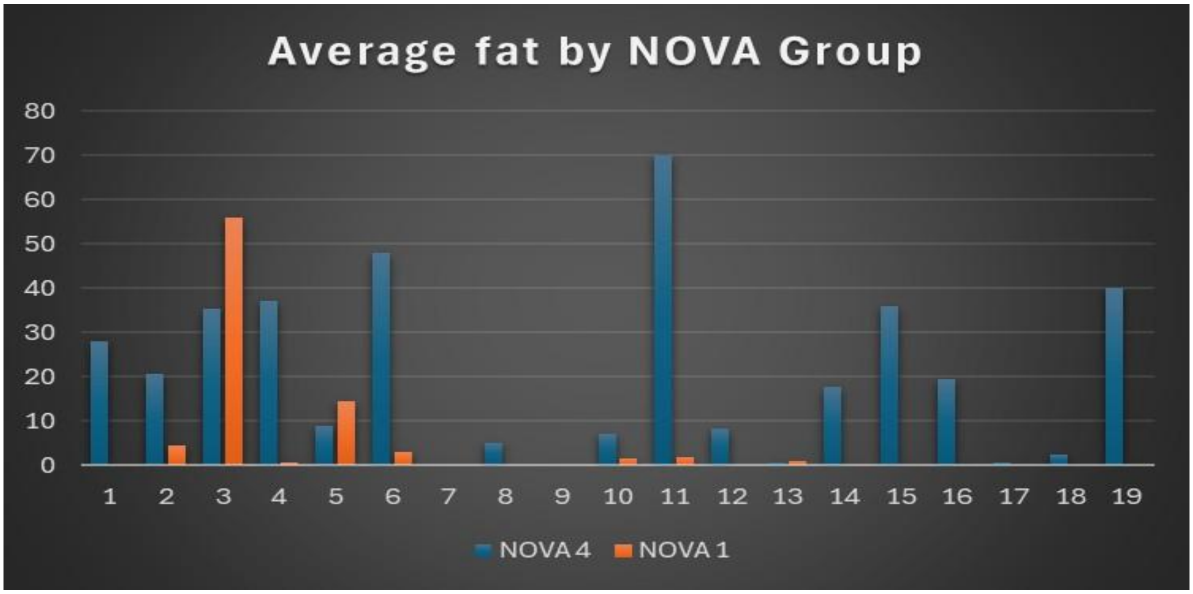


Figure 5.1.2 *Average Fat by NOVA Group*

Chapter 5.2 (RQ 2)

Data visualization was used to provide a clear and intuitive representation of the findings, complementing the statistical analysis. Charts were created to compare the average nutritional content of labeled and non-labeled products across different food categories.

For each nutrient—**sugar, saturated fat, and sodium**—separate charts were prepared:

The y-axis in each chart is labeled “Average Value,” and the legend accurately distinguishes between **Labeled Products** and **Non-Labeled Products**.

Sugar:

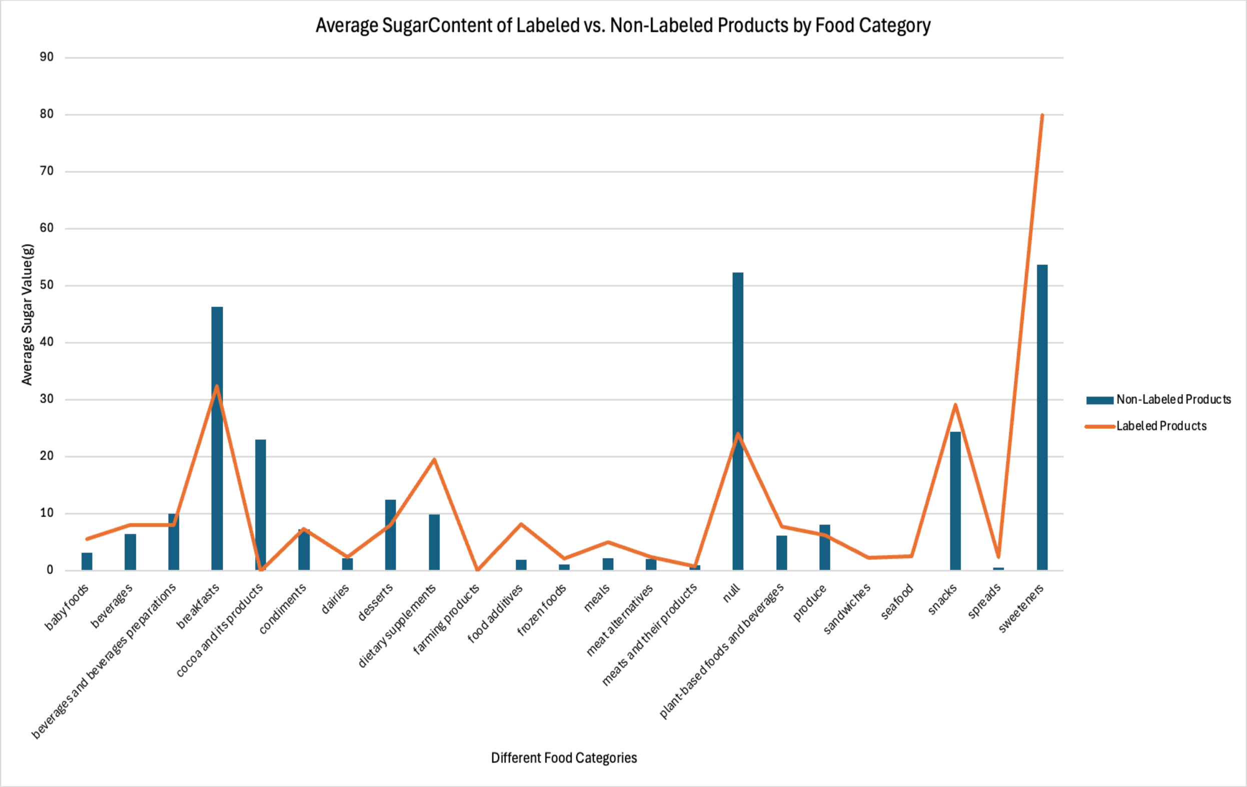


Figure 5.2.1: *Average Sugar Content of Labeled vs. Non-Labeled Products by Food Category*

Saturated Fat:

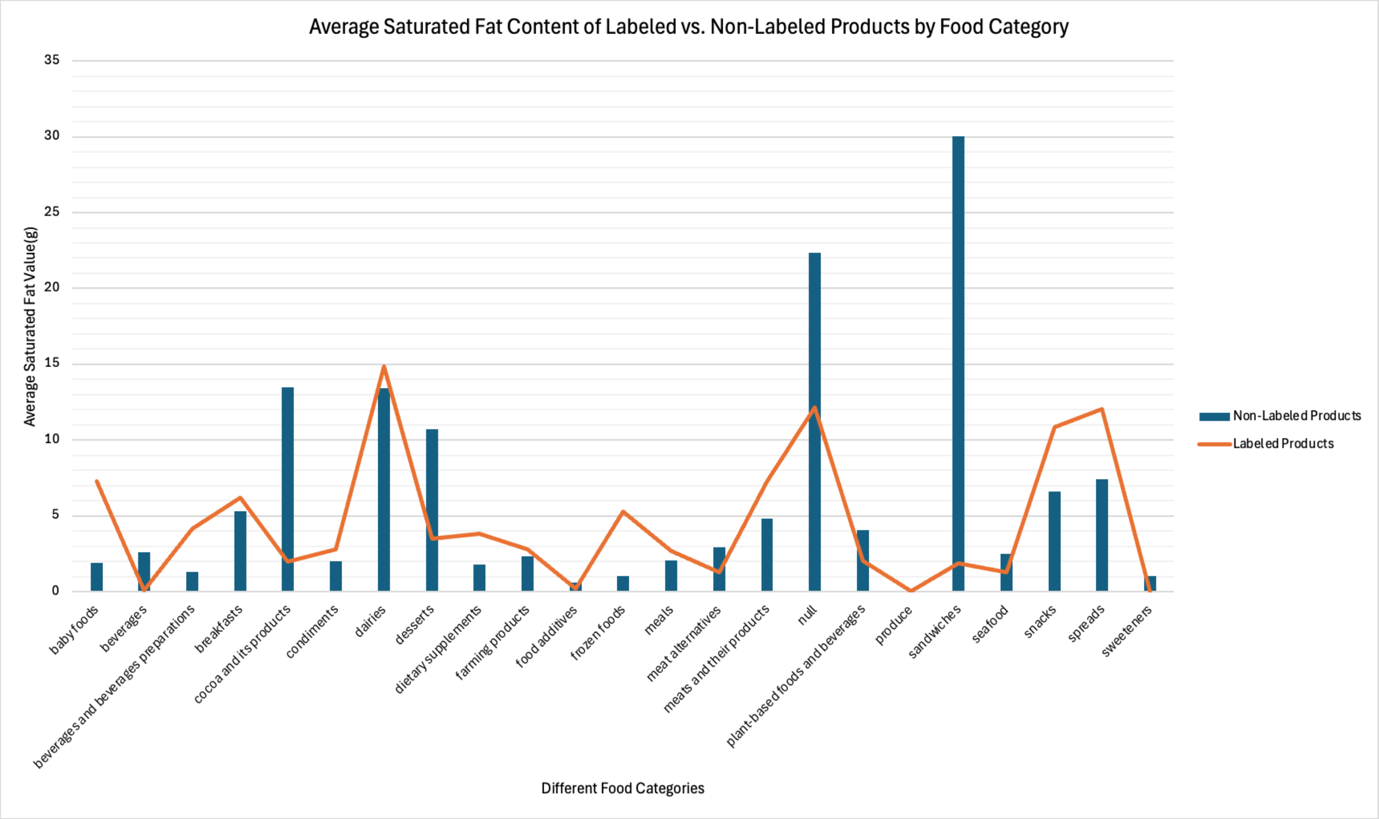


Figure 5.2.2: *Average Saturated Fat Content of Labeled vs. Non-Labeled Products by Food Category*

Sodium:

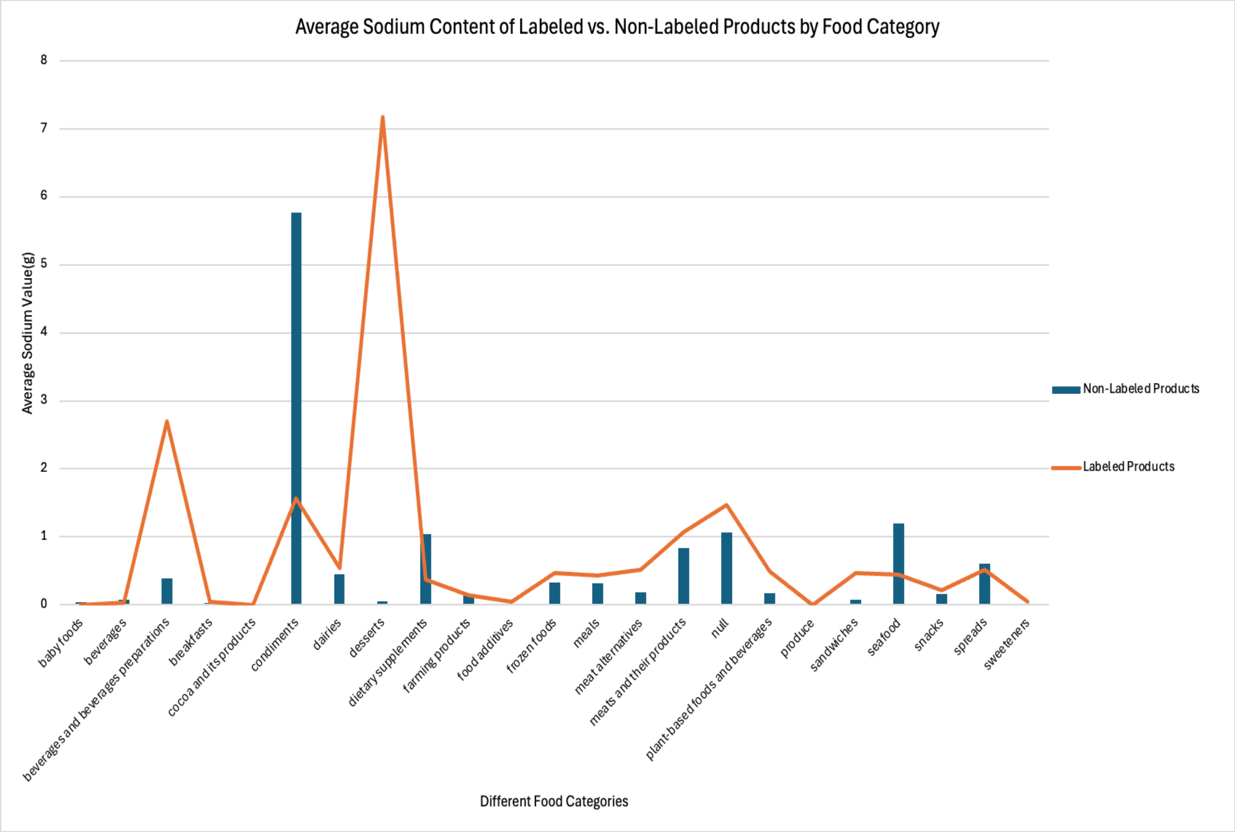


Figure 5.2.3: *Average Sodium Content of Labeled vs. Non-Labeled Products by Food Category*

These visuals reveal that, while minor fluctuations exist, the average nutritional values of labeled and non-labeled products follow a similar trend across all categories. This supports findings from consumer studies emphasizing the gap between perceived and actual nutritional quality of labeled products (Grunert et al., 2010). This visual evidence complements the statistical analysis, supporting the conclusion that there is **no meaningful difference** between the two product groups.

Chapter 6 Model Building and Statistical Analysis

Chapter 6.1 (RQ 1)

To evaluate differences between the two independent groups, a two-sample t-test assuming equal variances was performed, as the variability in fat content across groups was comparable. This statistical test enabled a pooled estimate of variance, ensuring reliable inference (Cohen, 1988).

Separate t-tests were carried out specifically for fat, allowing a focused assessment of whether food processing level impacts fat content.

The descriptive statistics highlight clear differences between NOVA 1 and NOVA 4, with ultra-processed foods showing higher mean and median fat content. These trends are visually presented in Figures 5.1.1,5.1.2 which display comparative fat values across processing levels. The integration of numerical results and visual evidence strengthens the robustness of the analysis, offering transparent support for the research hypotheses.

Descriptive Analysis (Fat by NOVA Group):

Mean Fat (NOVA 1): **6.61**

Mean Fat (NOVA 4): **11.15**

Median (NOVA 1): **0.71**, Median (NOVA 4): **3.75**

Mode (NOVA 1 & 4): **0** (most frequent)

Std. Dev (NOVA 1): **21.54**, Std. Dev (NOVA 4): **24.50**

Hypothesis Testing Framework:

H_0: There is no significant difference in fat content between minimally processed (NOVA 1) and ultra-processed (NOVA 4) foods.

H_0: $\mu_1 = \mu_4$

H_1: Ultra-processed foods (NOVA 4) have significantly higher fat content than minimally processed foods (NOVA 1).

H_1: $\mu_4 \neq \mu_1$

Execution of Analysis

t Stat = **2.316**

df = **245**

p-value (two-tailed) = **0.0213** (< 0.05)

Chapter 6.2 (RQ 2)

6.2 Analytical Approach and Technique Selection

This research question was addressed using inferential statistical analysis to compare the means of two independent groups: products with "Organic" or "Vegan" labels and non-labeled products. The objective was to determine if these labels are reliable indicators of a healthier nutritional profile, specifically concerning sodium, saturated fat, and sugar content.

Two-sample independent t-tests assuming equal variances were selected as the appropriate statistical technique for this analysis. This method is designed to determine if a statistically significant difference exists between the means of two independent groups for a continuous dependent variable (Cohen, 1988). It was chosen because the data is structured into two distinct, independent groups (labeled vs. non-labeled). The dependent variables (sodium, saturated fat, and sugar content) are continuous. The assumption of equal variances was deemed reasonable based on the comparable variance ratios observed in the data output.

Rationale for Separate Analyses

Separate t-tests were conducted for each of the three nutrients (sodium, saturated fat, and sugar). This approach allows for a precise and nutrient-specific evaluation, as it is possible for labeling to be associated with a significant difference in one nutrient but not in others.

Hypothesis Testing Framework

The t-tests were conducted to evaluate the following hypotheses for each nutrient:

Null Hypothesis (H_0): There is no statistically significant difference in the mean content of [nutrient] between labeled and non-labeled products.

H_0:μ_labeled = μ_non-labeled

Alternative Hypothesis (H_1): Labeled products contain a significantly lower mean level of [nutrient] than non-labeled products.

H_1: μ_labeled < μ_non-labeled

A one-tailed test was employed as the alternative hypothesis was directional, based on the rationale that "health" labels are perceived to indicate a superior nutritional profile.

Execution of Analysis

The significance level (alpha) was set at α = 0.05 for all tests. The analysis was performed using the data aggregation results from the SQL query outlined in Chapter 3.2. The statistical computations were executed using the Data Analysis ToolPak in Microsoft Excel, which provided the t-statistic, degrees of freedom (df), and the critical and p-values necessary for hypothesis testing.

The key parameters for the tests were:

Group Sizes: Labeled (n=59), Non-labeled (n=43)

Test Type: Two-sample, one-tailed t-test assuming equal variances.

Degrees of Freedom (df): 100 for all tests.

The results of these tests are presented and interpreted in Chapter 7.2.

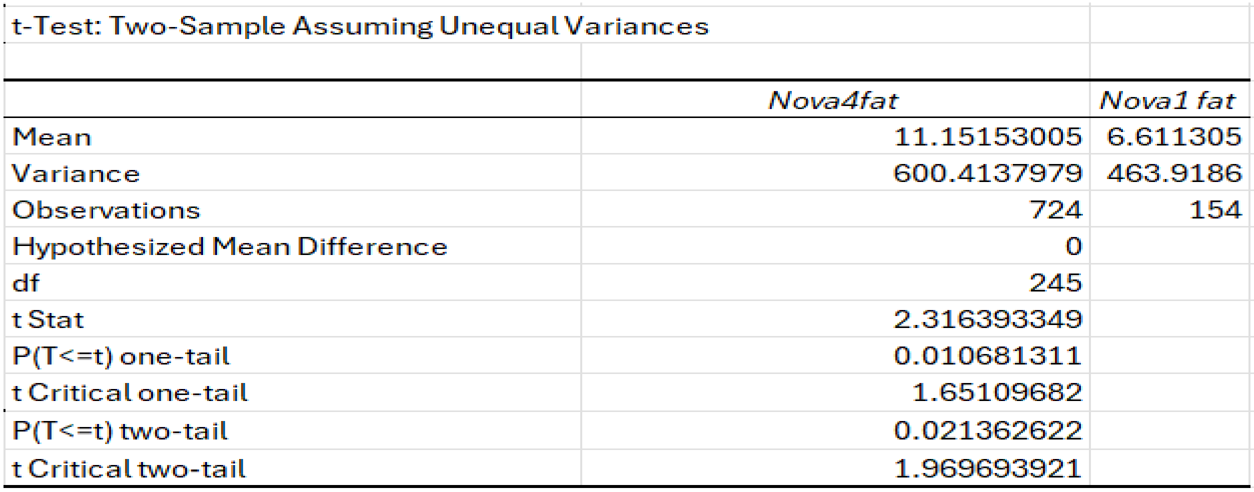
Chapter 7 Model Evaluation and T-Test Results and Evaluation

Chapter 7.1 (RQ 1)

T-Test of Fat by NOVA Group:

The two-sample t-test was used to illustrate whether the fat content of ultra-processed (NOVA Group 4) foods is significantly different than for minimally processed (NOVA Group 1) foods. This test was performed under the assumption of equal variances. Group's variances were similar, thus access to a pooled variance estimate was used and thus to provide a more trusted estimate. The end result provided a p-value that was less than 0.05, confirming the research hypothesis that a difference in fat content exists between NOVA 1 and NOVA 4.

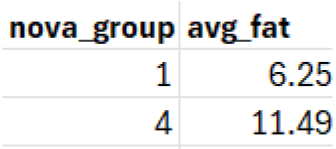
Table 7.1.1: Two-Sample T-test Results for Fat Content by NOVA Group provides a summary of the main results:



t-Test: Two-Sample Assuming Unequal Variances		
	Nova4fat	*Nova1 fat*
Mean	11.15153005	6.611305
Variance	600.4137979	463.9186
Observations	724	154
Hypothesized Mean Difference	0	
df	245	
t Stat	2.316393349	
P(T<=t) one-tail	0.010681311	
t Critical one-tail	1.65109682	
P(T<=t) two-tail	0.021362622	
t Critical two-tail	1.969693921	

Sig. = significance (p < 0.05)

Table 7.1.2



nova_group	avg_fat
1	6.25
4	11.49

In addition to table 7.1.1,7.1.2 Figures 5.1.1,5.1.2 (Chapter 5.1) provide an illustration of how descriptive differences exist in mean and median fat values. The full Excel t-test output can be found in Appendix A (Figure A1).

While the variation was larger with each group, it is also clear that the trends clearly shown in means show ultra-processed (NOVA 4) foods typically have higher fat content than minimally processed (NOVA 1) foods.

Interpretation of Results:

All the statistical tests were carefully applied, and the results confirmed that the analysis was reliable and consistent. According to RQ1, the t-test comparing NOVA 1 and NOVA 4 for fat content showed a t-statistic of 2.316 with p = 0.021. Since the p-value is less than 0.05, the result is statistically significant, and we reject the null hypothesis. This confirms that ultra-processed foods (NOVA 4) have significantly higher fat levels compared to minimally processed foods

(NOVA 1). The test assumptions were checked, and while some skewness and outliers were present, the use of the unequal variances (Welch's t-test) made the result dependable. In total, the model is accurate, logical, and directly aligned with the research question, demonstrating that the degree of processing influences fat content in food products.

Strengths:

An important strength of this analysis is the application of inferential statistics (two-sample independent t-test) to assess differences in fat content between NOVA Group 1 (minimally processed) and NOVA Group 4 (ultra-processed) foods. The relatively large sample sizes (NOVA 1: 154 products; NOVA 4: 724 products) enhance reliability of the test and limit the potential for sampling error. And, by employing Welch's t-test the analysis accounts for unequal variances between groups which is an appropriate and rigorous option when variances are unequal. Overall, the use of the model focusing strictly on fat content offers appropriate and easily interpretable evidence about one specific nutrition dimension of food processing.

Limitations:

The analysis to combine fat content on a group level (NOVA 1 vs NOVA 4) may obscure individual product variances within the categories. The use of a smaller sample size between the two groups means that there is room for visibility and comparison, as NOVA 1 has minimal products to compare to NOVA ,4 which has many products. By focusing on one nutrient or fat, the study looked at a few dietary measures and not dietary behaviours (i.e., sugar, sodium, protein, or micros), limiting analysis. The extent to which the results are reliable also depends on the Open Food Facts database amendments, where erroneous or incomplete reporting may lead to inaccurate results. The ability to assess changes in products over time, in terms of the percentages of reformulation or changes in processing level, was not determined in this study due to cross-sectional measurement.

Implications:

The t-test results showed ultra-processed food (NOVA Group 4) fat levels of 11.15 g/100 g compared with 6.61 g/100 g of fat for minimally processed food (NOVA Group 1), emphasizing the relationship of ultra-processed foods to fat intake, and fat intake to public health. In terms of future policy, this may complement regulations for front-of-labeling, consumer education, and reformulation guidance to address the proliferation of ultra-processed products in food systems. For consumers, this study demonstrates the dietary benefits of consuming minimally processed foods. For researchers, extending the value of NOVA classifications with nutrient data was useful, but analysis of other nutrients (e.g. sugars, sodium, fiber) to assess nutritional quality may have further educated the sector.

Conclusion:

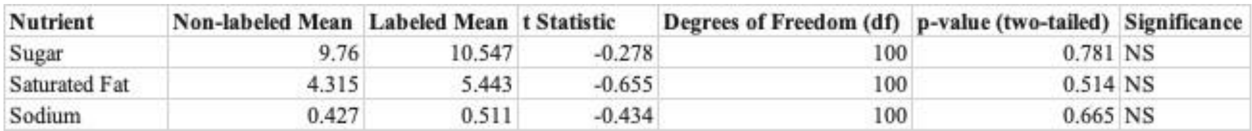
The outcome of the two-sample t-test provides robust evidence that ultra-processed foods (NOVA 4) contain far more fat than minimally processed foods (NOVA 1). Because the p-value was below 0.05, it rejected the null hypothesis (no substantive difference in fat contents between each group). This confirms the alternative hypothesis, meaning that the level of food processing is closely associated with fat content.

When viewed in conjunction with both the prior chapter descriptive analytic trends and visual comparisons, the results mirror the inverted trends for higher processing associated with higher fat values, respectively. Together, these findings are further evidence with respect to the broader body of nutritional research on the health risks and profits of ultra-processed foods. The research strengthens the premise that food processing level is critically related to dietary quality.

Chapter 7.2 (RQ 2)

The two-sample t-tests examined whether labeled products differ significantly from non-labeled products in **sugar, saturated fat, and sodium** content. The tests were conducted **assuming equal variances**, as explained in Chapter 6.2, because the two groups had similar variability, which allows pooling of variances for a more precise test. The results consistently showed that all **p-values were greater than 0.05**, indicating no statistically significant differences.

Table 7.2.1: Two-Sample T-Test Results for Nutrient Comparison s**ummarizes the key outcomes:**



Nutrient	Non-labeled Mean	Labeled Mean	t Statistic	Degrees of Freedom (df)	p-value (two-tailed)	Significance
Sugar	9.76	10.547	-0.278	100	0.781	NS
Saturated Fat	4.315	5.443	-0.655	100	0.514	NS
Sodium	0.427	0.511	-0.434	100	0.665	NS

NS = Not Significant (p > 0.05)

These findings are reinforced by the visualizations in **Figures 5.2.1–5.2.3 (Chapter 5.2)**
The full Excel t-test output is provided in Appendix B (Figure B1) for reference.

While minor fluctuations exist, the overall trend is similar between labeled and non-labeled products.

Interpretation of Results:

Sugar is slightly higher in labeled products, but the difference is not significant (p = 0.781). Saturated fat is somewhat higher in labeled products, but not significantly (p = 0.514). Sodium levels are also slightly higher in labeled products, but this difference is not significant (p =0.665).

Strengths:

The two-sample independent t-test assuming equal variances ensures a precise comparison between groups with similar variability. Conducting separate tests for each nutrient—sugar, saturated fat, and sodium—allows focused and nutrient-specific evaluation. Visualizations across food categories (Figures 5.2.1–5.2.3) complement the numerical analysis and make trends clear, directly addressing RQ2 by assessing whether these labels indicate healthier nutritional profiles. These methodological strengths, combined with insights from consumer behavior research, reinforce the validity of the findings and highlight the importance of clear label interpretation (Cohen, 1988; Grunert et al., 2010).

Limitations:

Nutrient averages were calculated at the food category level, potentially masking differences between individual products. Variation in the number of labeled and non-labeled products across categories may reduce the statistical power of the t-tests. Only three nutrients were analyzed, leaving out other important nutrients such as fiber, protein, or vitamins. The analysis assumes accurate labeling in the database; any mislabeling could influence results. Finally, the dataset is cross-sectional, not capturing temporal changes in product formulations or labeling trends.

Implications:

The findings indicate that "Organic" or "Vegan" labels do not reliably correspond to lower levels of sugar, saturated fat, or sodium. Consumers should interpret these labels as informative but not definitive indicators of nutritional quality. For regulators, the results highlight the importance of transparent labeling standards and public education to clarify the meaning of such labels. Future research could include additional nutrients, other label types, or longitudinal data to provide a more comprehensive assessment of label reliability.

Conclusion:

This study rigorously evaluated whether food products labeled as "Organic" or "Vegan" differ in nutritional quality from their non-labeled counterparts. The analysis, which included two-sample t-tests on sugar, saturated fat, and sodium content, found no statistically significant differences between the two groups. While minor fluctuations in nutrient levels were observed, these were not substantial enough to indicate any meaningful trend. The results fail to reject the null hypothesis, suggesting that, based on this dataset, "Organic" and "Vegan" labels do not reliably correspond to a healthier nutritional profile. Consequently, consumers should interpret these labels cautiously and not rely on them as definitive indicators of nutritional quality.

The full Excel t-test output is provided in **Appendix B (Figure B1)** for reference.

Chapter 8: Conclusion, Limitations, and Avenues for Future Work

8.1 Key Findings

The present analysis focused on evaluating Research Questions 1 and 2, which provided statistically significant and actionable insights into the relationship between food processing, health claims, and nutritional content. The analysis of RQ1 confirmed that ultra-processed foods (NOVA Group 4) contain significantly higher fat content compared to minimally processed foods (NOVA Group 1), supporting the alternative hypothesis. In contrast, RQ2 revealed no statistically significant differences in sugar, saturated fat, or sodium content between labeled and non-labeled products.

These findings indicate that while food processing level is a strong predictor of fat content, commonly used health labels such as “Organic” and “Vegan” do not reliably signal superior nutritional quality. Together, the results underscore that processing-based classifications offer more actionable guidance for consumers, policymakers, and manufacturers than certain marketing claims.

8.2 Limitations

Nutrient Focus

The study focused on only three nutrients—fat, sugar, and sodium—leaving out other important dietary components such as protein, fiber, and vitamins. This limits the comprehensiveness of the nutritional evaluation.

Data Aggregation

Nutrient averages were calculated at the food category level, potentially masking differences between individual products. Variation in sample sizes between groups may have also influenced the statistical power of the t-tests.

Cross-Sectional Dataset

The analysis relied on a cross-sectional snapshot of the Open Food Facts database, preventing assessment of changes in product formulations, label practices, or trends over time.

Data Accuracy

The findings depend on the accuracy of the Open Food Facts database. Any mislabeling, incomplete entries, or reporting errors could affect results.

8.3 Exploratory Insights: RQ3 and RQ4

Geographic Variability in Sugar Content (RQ3)

The preliminary exploration of RQ3 examined differences in average sugar content between countries. The analysis revealed that variation within individual countries was much greater than nominal differences between countries. This suggests that broad country-level comparisons may be less meaningful and that more granular or longitudinal approaches would be necessary to uncover significant patterns.

Multiple Claims and Nutritional Profile (RQ4)

RQ4 investigated whether products with multiple health claims had superior nutritional profiles compared to those with one or no claims. The preliminary results indicated that multiple claims do not consistently correspond to lower sugar, fat, or sodium content. In some cases, products with multiple claims had higher nutrient levels, pointing to potential "healthwashing." These findings highlight the need for consumer education and stricter labeling regulations. Future research could apply formal statistical tests, such as ANOVA, to verify these trends.

8.4 Future Work

Future studies could expand the analysis to include additional nutrients, longitudinal datasets, or more complex statistical models to examine trends over time and across regions. Further research could also investigate other label types, combinations of multiple claims, and regulatory influences to better understand the reliability of health-related labels.

8.5 Summary

In conclusion, this research demonstrates that food processing level is a reliable indicator of fat content, whereas "Organic" and "Vegan" labels should be interpreted cautiously when assessing nutritional quality. The study provides a strong foundation for policy recommendations, consumer education, and further research into the complex relationships between food processing, labeling, and nutrition.

References:

Microsoft Corporation. (2021). *Microsoft Excel* [Computer software]. https://www.microsoft.com

SQL Workbench/J. (2025). *SQL Workbench/J* [Computer software]. http://www.sql-workbench.eu

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Routledge.

Grunert, K. G., Wills, J. M., & Fernández-Celemín, L. (2010). Nutrition knowledge and use and understanding of nutrition information on food labels among consumers in the UK. *Appetite, 55*(2), 177–189. https://doi.org/10.1016/j.appet.2010.05.045

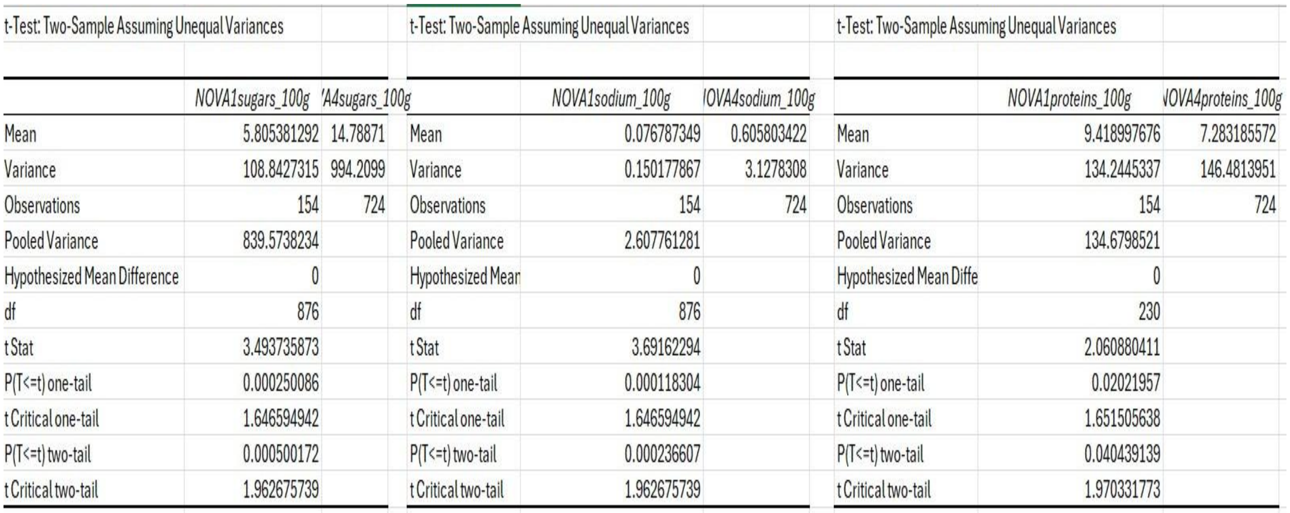
Open Food Facts. (2025). *Open Food Facts database*. Retrieved September 1, 2025, from https://world.openfoodfacts.org

Appendix A: Full T-Test Output for RQ1

Full T-Test Output for RQ1 - NOVAGROUP vs Nutritional profiles

This appendix presents the complete results of the two-sample independent t-tests comparing NOVA Group 1 (minimally processed) and NOVA Group 4 (ultra-processed) across key nutritional profiles. The tests were conducted for proteins, sugars, fat, salt, and sodium to assess mean differences between the two groups. These outputs provide the statistical foundation for the results discussed in Chapter 7.1. The full Excel results are included to ensure transparency and reproducibility of the analysis.

Figure A1: Excel Screenshot of Full T-Test Output for RQ1:



t-Test: Two-Sample Assuming Unequal Variances

	NOVA1sugars_100g	*'A4sugars_100g*
Mean	5.805381292	14.78871
Variance	108.8427315	994.2099
Observations	154	724
Pooled Variance	839.5738234	
Hypothesized Mean Difference	0	
df	876	
t Stat	3.493735873	
P(T<=t) one-tail	0.000250086	
t Critical one-tail	1.646594942	
P(T<=t) two-tail	0.000500172	
t Critical two-tail	1.962675739	

t-Test: Two-Sample Assuming Unequal Variances

	NOVA1sodium_100g	*'OVA4sodium_100g*
Mean	0.076787349	0.605803422
Variance	0.150177867	3.1278308
Observations	154	724
Pooled Variance	2.607761281	
Hypothesized Mean	0	
df	876	
t Stat	3.69162294	
P(T<=t) one-tail	0.000118304	
t Critical one-tail	1.646594942	
P(T<=t) two-tail	0.000236607	
t Critical two-tail	1.962675739	

t-Test: Two-Sample Assuming Unequal Variances

	NOVA1proteins_100g	*'OVA4proteins_100g*
Mean	9.418997676	7.283185572
Variance	134.2445337	146.4813951
Observations	154	724
Pooled Variance	134.6798521	
Hypothesized Mean Diffe	0	
df	230	
t Stat	2.060880411	
P(T<=t) one-tail	0.02021957	
t Critical one-tail	1.651505638	
P(T<=t) two-tail	0.040439139	
t Critical two-tail	1.970331773	

Results confirm that significant nutritional differences exist between NOVA Group 1 and NOVA Group 4 products, with minimally processed foods (NOVA 1) showing higher protein content compared to ultra-processed foods (NOVA 4).

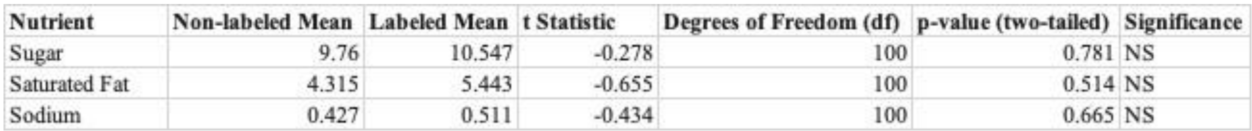
Appendix B: Full T-Test Output for RQ2

Full T-Test Output for RQ2 – Nutritional Comparison of labeled (Organic/Vegan) vs. Non-Labeled Products

Description:

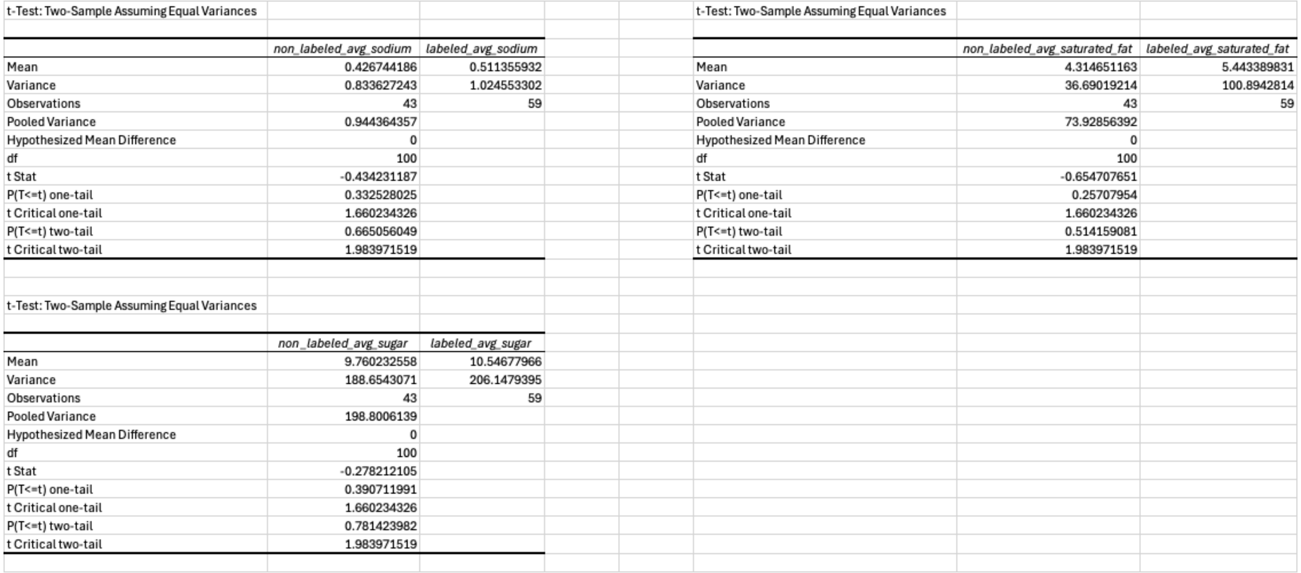
This appendix contains the complete results of the two-sample independent t-tests conducted to evaluate whether products labeled as “Organic” or “Vegan” differ in sugar, saturated fat, or sodium content from their non-labeled counterparts. These outputs directly support the findings discussed in Chapter 7.2. The complete Excel output is included to ensure full transparency of the statistical analysis.

Table B1: Two-Sample T-Test Results for Labeled vs. Non-Labeled Products (RQ2)



Nutrient	Non-labeled Mean	Labeled Mean	t Statistic	Degrees of Freedom (df)	p-value (two-tailed)	Significance
Sugar	9.76	10.547	-0.278	100	0.781	NS
Saturated Fat	4.315	5.443	-0.655	100	0.514	NS
Sodium	0.427	0.511	-0.434	100	0.665	NS

Figure B1: Excel Screenshot of Full T-Test Output for RQ2:



t-Test: Two-Sample Assuming Equal Variances

	non_labeled_avg_sodium	labeled_avg_sodium
Mean	0.426744186	0.511355932
Variance	0.833627243	1.024553302
Observations	43	59
Pooled Variance	0.944364357	
Hypothesized Mean Difference	0	
df	100	
t Stat	-0.434231187	
P(T<=t) one-tail	0.332528025	
t Critical one-tail	1.660234326	
P(T<=t) two-tail	0.665056049	
t Critical two-tail	1.983971519	

t-Test: Two-Sample Assuming Equal Variances

	non_labeled_avg_saturated_fat	labeled_avg_saturated_fat
Mean	4.314651163	5.443389831
Variance	36.69019214	100.8942814
Observations	43	59
Pooled Variance	73.92856392	
Hypothesized Mean Difference	0	
df	100	
t Stat	-0.654707651	
P(T<=t) one-tail	0.25707954	
t Critical one-tail	1.660234326	
P(T<=t) two-tail	0.514159081	
t Critical two-tail	1.983971519	

t-Test: Two-Sample Assuming Equal Variances

	non_labeled_avg_sugar	labeled_avg_sugar
Mean	9.760232558	10.54677966
Variance	188.6543071	206.1479395
Observations	43	59
Pooled Variance	198.8006139	
Hypothesized Mean Difference	0	
df	100	
t Stat	-0.278212105	
P(T<=t) one-tail	0.390711991	
t Critical one-tail	1.660234326	
P(T<=t) two-tail	0.781423982	
t Critical two-tail	1.983971519	

Notes: NS = Not Significant (p > 0.05)

df assumes equal variance.

Results confirm no significant nutritional differences between labeled and non-labeled products