From Bytes to Bites: Harnessing ChatGPT to Investigate Global Food Trends

Kasra Lekan, Rachel Holman, Brendan Puglisi, Matthew Bonner
University of Virginia
kl5sq, dnw9qk, btp6ht, mkb2dxw@virginia.edu

Abstract

We investigate the efficacy of ChatGPT at performing analysis on various aspects of a database using Ray - a distributed Python framework for parallel processing. We used the Open Food Facts Data which is a crowd-sourced database created by Stéphane Gigandet in 2012, it currently has over 2 million products listed from over 182 countries. First, we completed our analysis of the dataset without LLM assistance. Then, we performed an end-to-end analysis by providing ChatGPT with prompts to generate code relating to exploratory data analysis, summary statistics, and data visualizations. Finally, we compared ChatGPT's results to the results of our analysis. In doing so, we hope to highlight both the benefits and limitations of using ChatGPT for end-to-end data analyses using Ray.

1 Introduction

Large Language Model tools like ChatGPT are increasingly being applied to workflows for creating software prototypes or performing data analysis. These tools can complete simple coding tasks and generate natural language explanations to questions as quickly and efficiently. With only a handful of prompts, ChatGPT is capable of brainstorming data analysis ideas, generating code, refining existing code, and tailoring responses with user feedback, effectively allowing users to engage in pair programming. In this work, we seek to explore the capabilities and limitations of ChatGPT (version 3.5) concerning data analysis. To do so, we analyzed a dataset without using ChatGPT (human-only data analysis) and subsequently used ChatGPT for end-to-end data analysis. We logged errors made by ChatGPT and corrected them when necessary to proceed with the analysis process.

Using the publicly available data from OpenFoodFacts.org [2], our research team along with ChatGPT will examine trends relating to food. Food is a principal influence on both health and environmental impact. By analyzing food data, we will elucidate key trends at the intersection of food, health, and environment across contexts.

2 Human Analysis

2.1 Approach

Our general approach was (1) Exploratory Data Analysis (EDA), (2) Brainstorming potential directions, (3) Creating a small version of the dataset for rapid testing, (4) Writing analysis code using small-scale packages like Pandas, (5) Applying Ray to scale our analysis code, and (6) Compiling results.

2.2 Findings

Food is a ubiquitous and vital aspect of life. Accessing quality food is not equitably distributed in all contexts, but the differences in the types of foods we eat and how they are packaged can make a significant impact on our health and our environment. The varying data analysis methods outlined in the previous sections resulted in useful findings that shed light on impact of food and how changes to our diets can affect the world.

2.2.1 Weights

A few features we were initially interested in were ecoscore, nutriscore, and quantity. The scatterplots between these variables showed very little relation between them. We looked further into the relationships serving size may hold with other variables such as country. We found that the countries with the highest average quantity were Ecuador (3313 grams), Peru (1099 grams), and Bulgaria (935 grams). For reference the United States had on average 413 grams, France had 391 grams, China had 366 grams, and Italy had 526 grams. Most nations hovered around the mid 300s for their average grams per item.

The last quantity connection we investigated was how it related to food category. We found that the food categories with the largest quantities were fresh eggs (38,133 grams), 'preparations-de-boulangerie' (25,000 grams), 'preparations-pour-pains-aux-graines' (25,000 grams), 'riz-long-parfume'

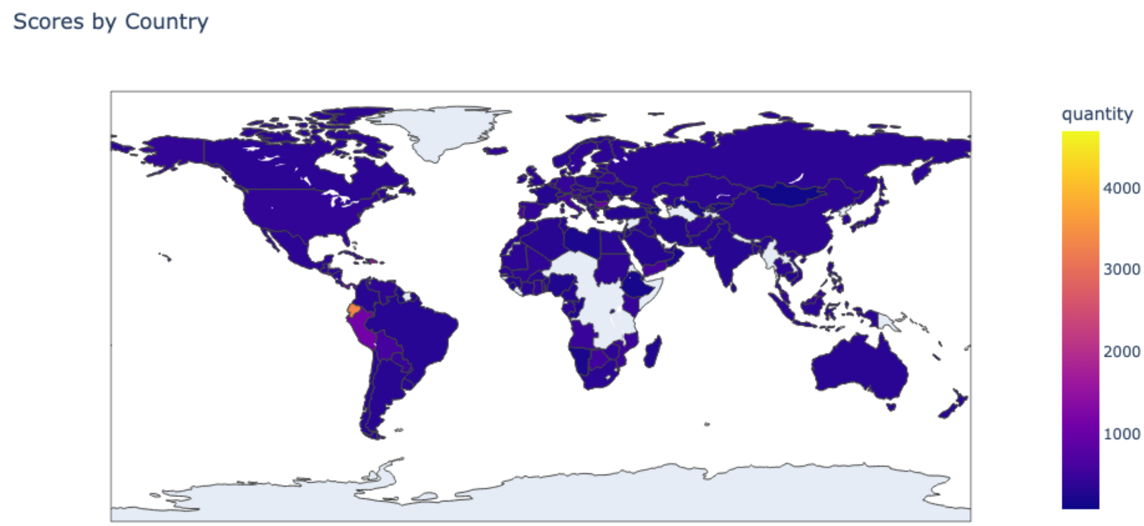

Figure 1: Average Weight of Items per Country (Grams)

(20,000 grams), and fresh zucchini (17,036 grams). Some of the food categories with the lowest average quantity were lamb liver (.54 grams), vanilla soy based drinks (1 gram), and fish-snack (4 grams). This presents an interesting look at the data as not all of these numbers seem accurate. This is an open source dataset and so there could be issues here ranging from data entry to misunderstanding. Some may have interpreted quantity as the serving size - as is likely the case with the fish snack. While others may have seen it as the size of a bulk delivery, this could possibly explain the large quantity of fresh eggs.

2.2.2 Food Groups

We wanted to further explore the variables of ecoscore and nutriscore, so we decided to look at the respective scores for each food group. The data contained food groups that were too specific to derive actionable insights from, so we decided to aggregate them into 10 main groups. These were 'Fruits and Vegetables', 'Cereals and Potatoes', 'Composite Foods', 'Beverages', 'Salty Snacks', 'Fish, Meat, and Eggs', 'Milk and Dairy', 'Fat and Sauces', and 'Sugary Snacks' and 'Unknown'. First we looked at nutriscore which is a metric relating to the nutritional value of products. Based on how it is calculated, a smaller score means it has the most nutritional value. We produced two visuals comparing food groups and nutriscore: a bar-plot and a spider graph. To no surprise, fruit and vegetables had the lowest score of -1.05, while sugary snacks had the highest of 17.03. Looking more in depth at this group, chocolate products actually have the worst nutritional value based on this data. Another interesting insight in our analysis was that protein based items including the groups 'Fish, Meat, and Eggs' and 'Milk and Dairy' had very similar nutritional value to 'Salty Snacks' and almost twice as bad as carbohydrates in the 'Cereals and Potatoes' group.

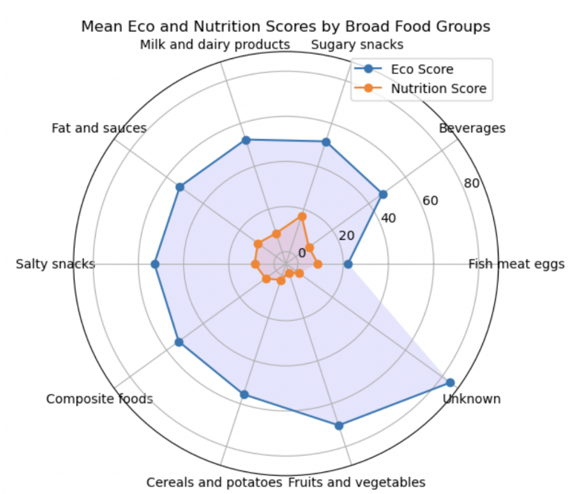

Figure 2: Nutrition vs. Economic Scores by Food Group

We then created a similar aggregate analysis along with visualizations for ecoscore, this metric is calculated where the higher the score the more eco-friendly the product is. An interesting insight we found from this was that most groups had very similar ecoscore in the range of 45-50. A standout group was ‘Fish, Meat, and Eggs’ which had a far worse score of 22.14, looking deeper into this group we discovered that the subgroup ‘lean fish’ is the most detrimental to the environment. This is interesting as we would have thought red meat was the worst due to the large carbon footprint of livestock farms however, it is important to keep in mind that this is solely based on the products contained in the database. Another standout group was ‘Fruits and Vegetables’ which had the highest score of 70.08, proving they are the most eco-friendly.

While in general, we found that there was not much correlation between a food’s nutriscore and its ecoscore, some small correlations did emerge when breaking the food down into the aforementioned groups. Sugary snacks and dairy products had the the strongest negative correlations whereas composite foods and salty snacks had slight positive correlations. This was a good example of how dividing data into groups can reveal relationships that get washed out when dealing with the data as a whole. Furthermore we looked at box-plots with these groups and discovered the greatest spread of ecoscores were ‘Fish, meat, and eggs’ and ‘Milk and dairy products’ whereas ’Beverages’ and ‘Composite foods’ had the smallest spreads. As for nutriscore, the greatest spreads were found with ‘Milk and dairy products’ and ‘Fruits and vegetables’ and the lowest were with ‘Beverages’ and ‘Unknown’.

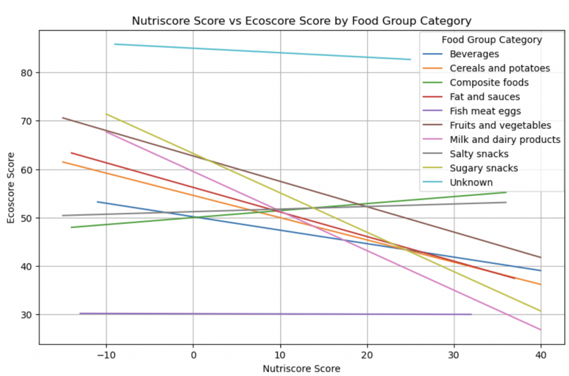

Figure 3: Correlation Between Nutrition and Economic Scores by Food Group

With this information, we were interested in further exploring where our data was coming from. We found that 37.65% of our data came from France, 20.50% came from the US, and in a distant third 7.54% came from Spain. The most common product in our dataset was ‘Filet de poulet’ making up 0.14% of our data. The second most common was ‘Miel’ with 0.11%. The 3 most common brands in our data were Carrefour with 1.36%, Auchan with 1.06%, and U with 0.48% of our data.

2.2.3 Analysis Conclusions

Through this analysis of ecoscore and nutriscore, we discovered that based on this dataset chocolate seems to be one of the worse items for both environmental and nutritional health. More data needs to be collected and further analysis needs to be completed to determine the geographical areas responsible for the most environmental degradation in regards to food products. Along with understanding the nutritional value of products made in certain countries.

3 ChatGPT as a Data Analysis Tool

We attempted to leverage ChatGPT to perform an end-to-end analysis that follows the general approach we followed when performing our analysis as described in Section 2.1. As we progressed through the analysis process, we attempted to use a pair-programming approach in instances when ChatGPT generated code that did not accomplish our analysis standards.

3.1 Brainstorming Plans

Since the OpenFoodFacts dataset has been public for many years, we opted to test ChatGPT’s brainstorming capability without providing any metadata. By prompting “*Analyze and summarize the OpenFoodFacts.org data to extract key insights and trends. Develop visualization methods to represent this data effectively and provide natural language explanations of the visualizations.*”, ChatGPT returned 5 unique ideas in chronological order from cleaning the data to writing conclusions about the results. The system specifically proposed various types of plots and analysis techniques for the user to create and apply. Box plots, histograms, word clouds, and bar charts were suggested to visualize characteristics of ingredients, brands, and nutrition information to name a few.

3.2 Solo-Programming

ChatGPT is capable of generating error-free code in various programming languages, however, in our analysis we focused on its understanding of Python and its use of Ray [5]. Without providing any code to build from, we prompted ChatGPT to implement the plans it had devised in the previous section.

When asked to “*develop code in Python to analyze and summarize the OpenFoodFacts.org data*”, ChatGPT effectively produced Python code using Pandas for data manipulation and Matplotlib/Seaborn for visualization. The response was thorough, including importing packages, providing in-line comments, and handling potential missing values in the data. Although the system noted the importance of cleaning the data first, it did not implement that section of code likely because we did not express how we had downloaded the data or what columns we were most interested in. Apart from loading and cleaning the data ourselves first, the only change necessary was ensuring the suggested column names in the generation of visuals were correct. We also subsequently provided the columns to ChatGPT. We found that this approach did not address the preprocessing gaps but did implement proper data loading.

Figure 4: Word Cloud of the Most Commonly Used Ingredients Generated by ChatGPT

We were particularly intrigued by ChatGPT's word cloud visual because that is not a figure typically used in data analysis. Pictured above, this visual was the only one our human team had not considered making prior to our AI consultation.

3.3 Refining Specific Code

Having established that ChatGPT can create code from scratch, we evaluated how it responds to code provided in a prompt. This was tested by asking ChatGPT to make sections of code more efficient and to fix code that caused an error.

During the data cleaning stage of our analysis, we developed code to ensure all text in a column was lowercase, devoid of excess white spaces and special characters, and split by commas to extract a list of unique values. This cleaned list of values took our research team several lines of code to create, but when fed into ChatGPT with the task "*make this more efficient*", we received output that achieved the same functionality using a third of the lines of code.

When asked to "*rewrite to fix this error: LinAlgError: SVD did not converge in Linear Least Squares*", ChatGPT expertly assessed what could have caused the error and returned a working line of code with detailed explanation. This error-correcting capability proved effective with issues as simple as typos to those as complex as regularizing regression models to reduce collinearity between variables. Using AI to assess errors was significantly faster and more readily applicable than finding and applying solutions found via online search.

3.4 Paired-Programming

While ChatGPT is capable of completing an entire analysis with very little human intervention, we found that with a limited model like ChatGPT, there is great value and opportunity for learning by using the system as a partner for data-intensive projects. ChatGPT is capable of remembering previous prompts and responses which it can then tailor using user feedback. The ability to communicate back and forth conversationally allows users to experience a similar sensation to talking virtually with another programmer.

We partnered with ChatGPT to develop and refine visualizations to best understand trends in the data. After providing a small subset of our data, we worked alongside the system to code analytical solutions more apt to the problem than either of us could have written alone. Taking lines of code for one visual, we would ask ChatGPT to "*stagger the labels for readability*", "*make points slightly transparent and add trend lines*", and so on until we achieved the optimal outcome.

Because ChatGPT was given a look at the dataset we were working with, it was able to develop plots comparing specific columns by name with the understanding of which datatype each variable was. This allowed our team to use suggested code as it was without needing to alter variable names. Additionally, because ChatGPT provides both code and natural language explanations of the code, we were able to gain a deeper understanding of how our thought process differed from the AI system's.

4 Comparing ChatGPT and Human Approaches

Both the ChatGPT and the human methodologies exhibited distinct strengths and weaknesses. Our comparative analysis primarily focused on three fundamental criteria. The first was an application in which creativity was prevalent in both approaches, but ChatGPT could not make actionable insights without further guidance. ChatGPT had good ideas, but could not implement them without assistance from us. we still had to lead it in analysis along with data preprocessing. In addition, ChatGPT needed assistance in implementing ray batching when prompted with Python code. This could be due to us attempting to use pandas functions as we were most familiar with them instead of using built-in ray functions. Additionally, humans were able to pick apart the analysis and address issues with data, whereas chat missed these completely unless directed to the issue explicitly. The next criterion was code efficiency. ChatGPT could write code that was significantly more concise than humans. For example, when provided with our original code, ChatGPT could almost always return code with the same functionality in fewer lines. Sometimes up to 3 times fewer lines of code, as mentioned previously. Our third criterion was bugs, in which both ChatGPT and human written code had errors but assessing the errors with AI was significantly faster and more readily available than human debugging. This was especially prevalent when implementing visuals in which ChatGPT worked much faster than going to several online forums to manually answer our specific problem. In general, we found pair programming with the LLM to have the greatest potential currently. Especially when working with large datasets, as ChatGPT produced visuals that were too cluttered since it did not attempt to divide the data into groups, even though such groups (e.g. country, food group) were inherent to the dataset. This grouping was a natural first step for human analysis. In light of this, analysis of small-scale datasets that can be fully appended (at least in terms of metadata and a few rows) to the LLM's context window is very effective, however, big data projects still cause ChatGPT to falter. Additionally, we find that brainstorming with ChatGPT can be a powerful tool even for a data analysis that does not use an LLM for programming.

5 Related Work

The OpenFoodFacts dataset [2] has been leveraged for food data analysis in past research [1] which we explored while forming our methodology.

Many of our findings relating to ChatGPT's capabilities and limitations are corroborated by Guo et al. [3] when they compared AI to human ability. Their research found that ChatGPT is objective, focused only on the given topic, formal, and unemotional while humans approach work subjectively colloquially and emotionally with the ability to shift between topics and ideas organically.

Hassani et al. [4] specifically explores the use of ChatGPT in data analysis, mostly from a qualitative perspective, arguing that while promising, ChatGPT's outputs can be difficult to interpret and are insufficiently reliable for replacing a full-time data scientist.

6 Conclusion

While ChatGPT is impressively capable of end-to-end data analysis, especially for small-scale data, ultimately it lacks robust support for big data code production. We also found evidence that the large number of brainstorming ideas that ChatGPT created needed to be filtered by a human expert to perform effective analysis.

Additionally, we discovered the importance of dedicated distributed processing frameworks like Spark. Translating our small data analysis code (using Pandas, MatplotLib, etc.) to leverage the parallel speeds provided by Ray required a significant engineering effort that was prone to errors.

7 Metadata

The presentation of the project can be found at:

file link

The code/data of the project can be found at:

`https://github.com/anrath/BigDataProject`

References

[1] Eloi Chazelas, Mélanie Deschasaux, Bernard Srour, Emmanuelle Kesse-Guyot, Chantal Julia, Benjamin Alles, Nathalie Druesne-Pecollo, Pilar Galan, Serge Hercberg, Paule Latino-Martel, Younes Esseddik, Fabien Szabo, Pierre Slamich, Stephane Gigandet, and Mathilde Touvier. Food additives: distribution and co-occurrence in 126,000 food products of the French market. *Scientific Reports*, 10(1):3980, March 2020. Publisher: Nature Publishing Group.

[2] Open Food Facts. https://world.openfoodfacts.org/data.

[3] Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection, January 2023. arXiv:2301.07597 [cs].

[4] Hossein Hassani and Emmanuel Sirmal Silva. The Role of ChatGPT in Data Science: How AI-Assisted Conversational Interfaces Are Revolutionizing the Field. *Big Data and Cognitive Computing*, 7(2):62, June 2023. Number: 2 Publisher: Multidisciplinary Digital Publishing Institute.

[5] Philipp Moritz, Robert Nishihara, Stephanie Wang, Alexey Tumanov, Richard Liaw, Eric Liang, Melih Elibol, Zongheng Yang, William Paul, Michael I. Jordan, and Ion Stoica. Ray: A Distributed Framework for Emerging AI Applications, September 2018. arXiv:1712.05889 [cs, stat].

Appendix

7.1 Human Findings

7.1.1 Food Groups

7.1.2 Geographical

After obtaining a better understanding of ecoscore and nutriscore based on food groups, we wanted to understand these metrics on a more detailed geographical level. Starting with nutriscore, we observed that Tajikistan has the worst nutriscore by far of 23, the average being 9.29 and the median being 9.19. The United States had a slightly better score than the average of 8.56, While Africa had several countries with the best scores close to 0. It is important to note that this score is not necessarily representative of the nutritional health of a country, but rather the nutritional value of products in this dataset a country produces. For example, a quick query reveals that Chad, a country with a nutriscore of 0, only had a few products in this dataset and most consisted of beverages such as flavored water, which evidently do not represent the cuisine of Chad or the general nutritional health of the country.

Figure 5: The Food Categories with Heaviest Average Items

Figure 6: The Food Categories with Lightest Average Items

We applied a similar form of cautionary analysis to observing ecoscore across the geographical landscape. On average, countries had an ecoscore of 51.25 and a median of 50.54. Most countries were consistent with this average value, but a few stood out. Including Sudan which had an average ecoscore of 109, the highest by far. Upon further investigation, it was determined that there were only two foods in this database relating to this country which were both syrup. Upon further research, we discovered that syrup actually has a net positive carbon footprint and is more sustainable than other sugars and sweeteners. However, as previously stated this is by no means representative of the country's food products

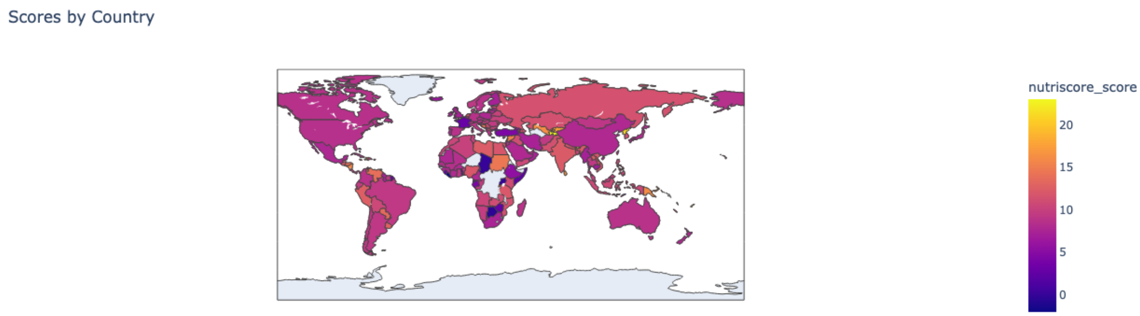

Figure 7: Nutriscore Geovisual

ecoscore as a whole.

7.1.3 Extreme Products

To better understand the metrics of ecoscore and nutriscore, we decided to observe the components that make up the products with the worse respective ratings. Concurring with our previous analysis of nutriscore and food groups, we found that the products with the worse nutriscore for the most part contained excessive sugar. More specifically most of these products were candies and chocolate. We used spider plots to visualize the main components of these unhealthy foods focusing on three main ones: Walnut Bonbons, Coffee Creamer, and Snickers. We determined that these products were mostly comprised of carbs, sugars, and fat with little to no protein except for the coffee creamer which obviously includes dairy.

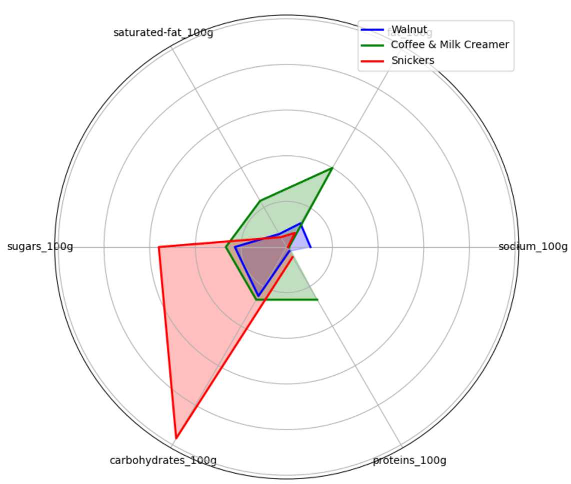

Figure 9: Nutriscore Spider Plot

Completing a similar analysis for ecoscore, we discover that our previous observation that meat items have the worse ecoscores only holds true in the aggregate, but not for the extremes. The items found to have the worse ecoscores are typically sweet treats and coffee. The items we decided to take a closer look at were Nesquik, Choco Coffee Mix and Mini Donuts. We thought it was interesting that chocolate and coffee items have a worse ecoscore than meat products, however upon further research it is justified since cocoa and coffee bean farming is well known to be associated with deforestation and require significantly more water than most other crops.

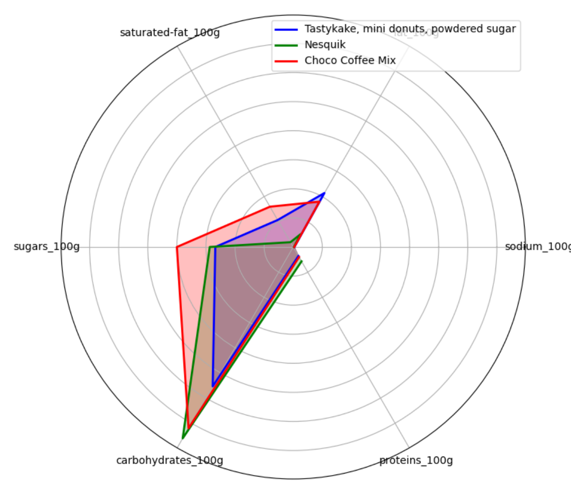

Figure 10: Ecoscore Spider Plot

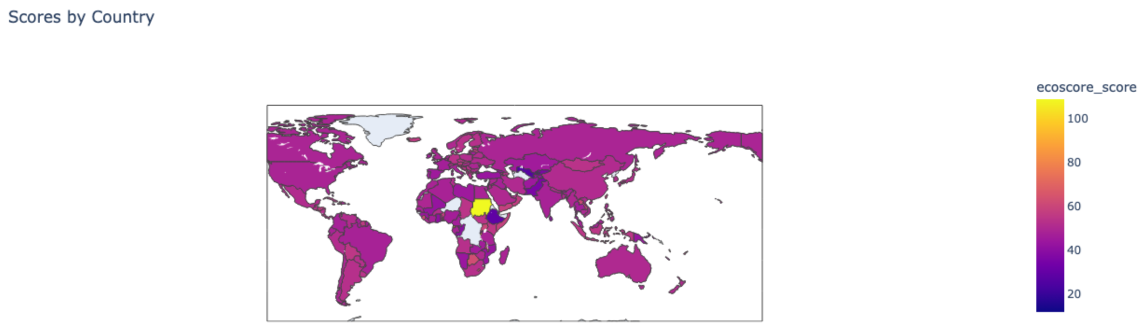

Figure 8: Ecoscore Geovisual

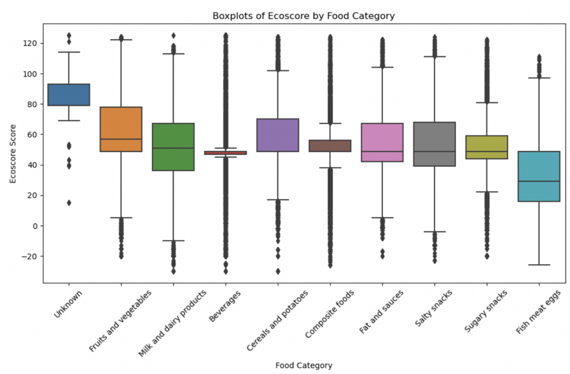

Figure 11: Economic Score by Food Group Boxplot

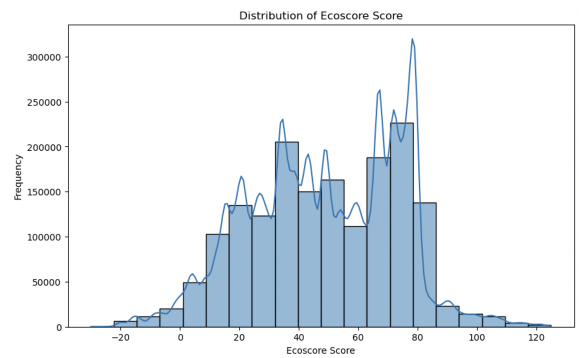

Figure 13: Distribution of Economic Scores Generated by ChatGPT

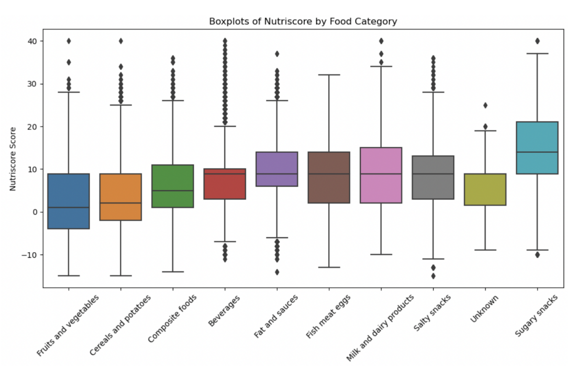

Figure 12: Nutrition Score by Food Group Boxplot

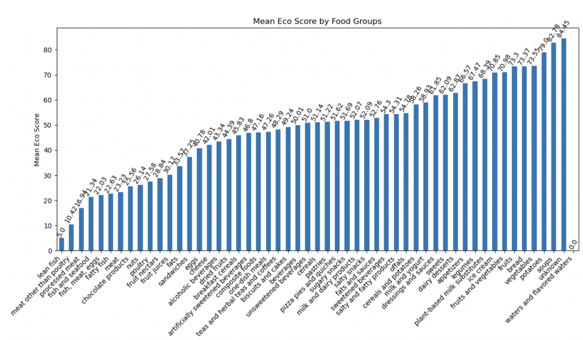

Figure 14: Economic Score by Food

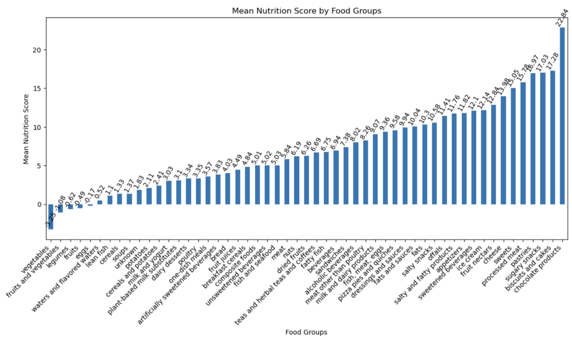

Figure 15: Nutrition Score by Food

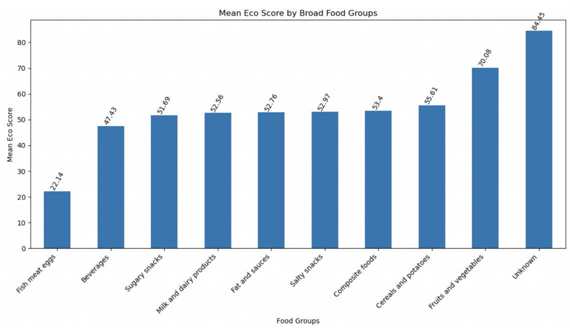

Figure 16: Economic Score by Food Group

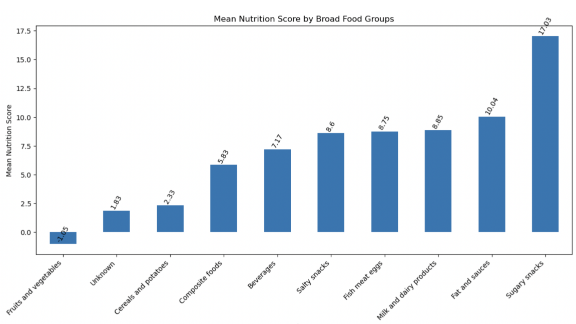

Figure 17: Nutrition Score by Food Group