Is Vegan Food the Most Cost-Effective Source of Protein?

Abstract

This study investigates whether vegan food is the cheapest source of protein compared to non-vegan options. Utilizing nutritional data scraped from Tesco Ireland's website, we analyzed a comprehensive dataset of over 12,000 products, including their prices, protein content, and dietary classifications. Each product was evaluated based on its cost per gram of protein to identify cost-effectiveness. Our findings reveal significant trends in protein affordability across dietary categories, shedding light on the economic viability of vegan diets for protein intake. By comparing the distributions and averages of protein costs, this analysis provides actionable insights for budget-conscious consumers seeking affordable dietary choices. The results aim to inform dietary decisions while contributing to the ongoing discourse on the accessibility and sustainability of plant-based diets.

Dataset

Data Collection

The dataset was compiled by scraping nutritional information for all products available on the Tesco Ireland website. The initial raw data contained 11,841 rows and 3,392 columns, representing a wide variety of products and attributes. This unexpected volume of columns arose due to inconsistencies in how nutritional attributes were recorded, with each product often having its unique column names, even for identical attributes. Additionally, the dataset lacked explicit classification for vegan and non-vegan products.

Challenges in Raw Data

- **Column Redundancy and Errors**: Many columns represented the same attribute with variations in naming conventions, such as "salt," "sale," "slat," "sal," "salt(g)," and "salt g." These inconsistencies arose from brand-specific naming practices and occasional spelling errors.
- **Data Size**: Although the dataset initially appeared large due to its 39 MB size and high dimensionality, it did not meet the "volume" criterion of big data. However, the dataset exhibited significant "variety" due to the diversity of product attributes.

Data Cleaning

To prepare the data for analysis, extensive cleaning was performed using Python:

1. **Column Standardization**: Columns were renamed and standardized by identifying and merging those representing the same attributes. For example, variations of “salt” were consolidated into a single column.
2. **Attribute Selection**: Unnecessary columns were removed, reducing the dataset to key nutritional attributes such as protein, fat, calories, carbohydrates, and salt.
3. **Error Handling**: Spelling errors and inconsistencies were addressed during the cleaning process.
4. **Data Reduction**: After cleaning, the dataset was reduced to 1 MB, retaining only relevant and structured information.

Vegan Classification

The original product pages did not explicitly classify products as vegan or non-vegan. However, the website provided a filter in the product listing for vegan products. By leveraging this filter, an additional dataset of vegan products was scraped. This dataset was integrated with the cleaned nutritional dataset, introducing a new binary column, **“vegan,”** with True indicating vegan products and False otherwise.

Final Dataset

The cleaned and integrated dataset included:

- **Rows**: 11,841 (representing unique products).
- **Columns**: The cleaned dataset included only the most relevant columns: url, ProductName, Unit, price, unit_price, NetContents, unit_carbs, unit_energy, unit_fat, unit_protein, and vegan.

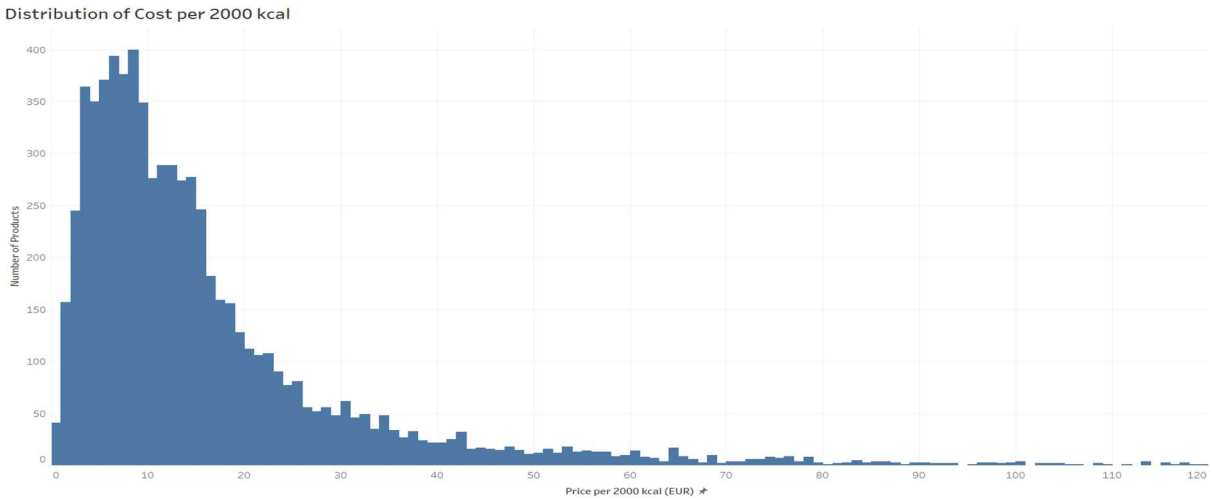
By addressing the challenges of inconsistent data representation and merging the datasets effectively, the final dataset was prepared for analyzing the cost-effectiveness of vegan and non-vegan protein sources.

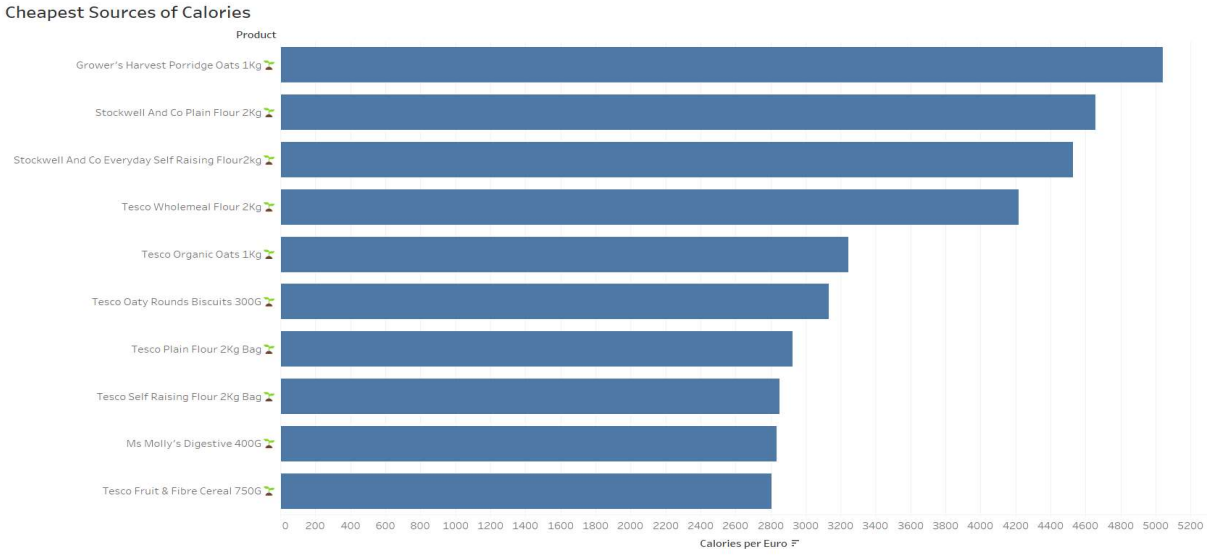
Data Exploration

Before diving into the analysis of protein cost-effectiveness, we explored the affordability of food products in terms of energy and protein content. The goal of this exploration was to identify trends and insights into the cost of fulfilling basic dietary requirements, regardless of nutrient balance.

Affordability of Calories

To determine the cheapest ways to meet the average human's daily energy requirement of **2000 kcal**, we calculated the cost of fulfilling this caloric requirement for all products in the dataset. The distribution of these costs is visualized in the attached histogram, which plots the **number of products** on the y-axis against the **price per 2000 kcal** in euros on the x-axis.





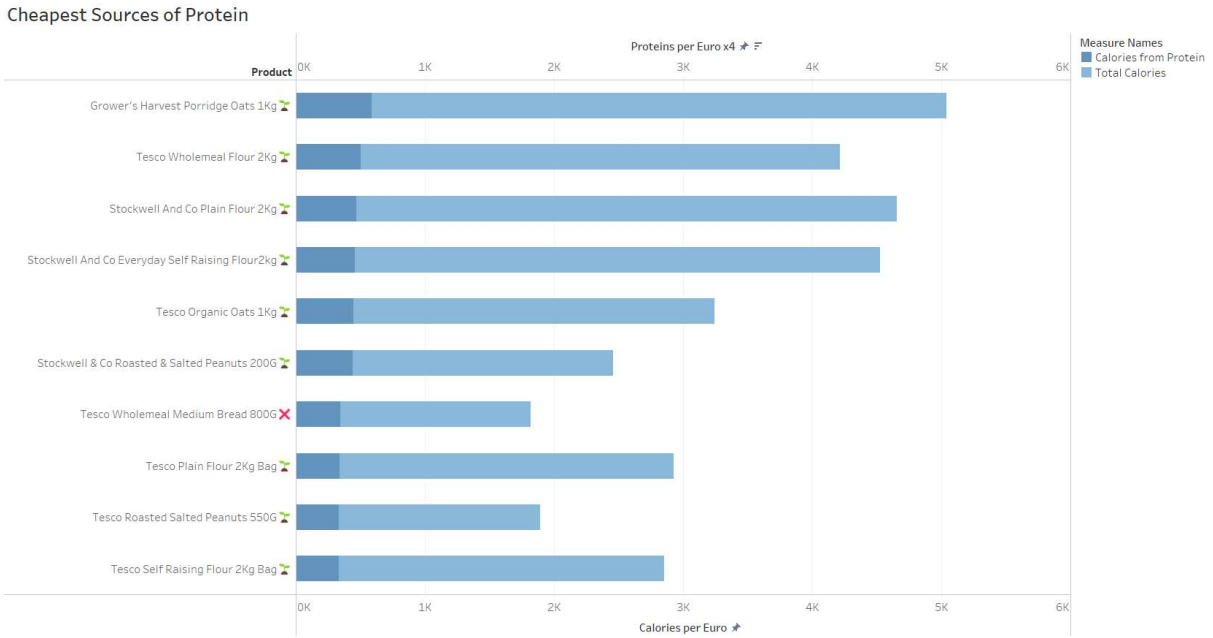
Findings:

1. Most products fall within the **€2 to €15 range** for 2000 kcal.
2. Among the **top 10 cheapest products** for fulfilling this caloric requirement:
    a. **Every product in the top 10 is vegan.**
    b. The cheapest product overall is oats, capable of meeting the 2000 kcal requirement for **€0.40**.

Note that this analysis focuses solely on the cost of calories and does not consider macronutrient composition or other essential nutrients like vitamins and minerals.

Affordability of Protein

Next, we analyzed the cost of obtaining **100 grams of protein** from the available products. The results are visualized in a graph of the **top 10 cheapest sources of protein**.



Findings:

- Among the **top 10 cheapest sources of protein**, **all but one are vegan**.
- The **cheapest source of protein** remains oats, which can provide 100 grams of protein for **€0.68 per day**.
- However, while oats are cost-effective for protein, they are predominantly a carbohydrate-rich food, with only **11% of their calories coming from protein**. Consuming sufficient protein solely from oats would lead to an excessive caloric intake of over **3400 kcal daily**.

Protein and Caloric Balance

To align with dietary recommendations, we refined the analysis by filtering products based on the percentage of calories derived from protein. According to nutritional guidelines, **10-35% of daily caloric intake** should come from protein.

1. **Minimum 10% Calories from Protein**:
    - Filtering out products that provide less than 10% of calories from protein, we identified the **top 10 cheapest sources of protein**.
    - **Most of the top 10 products are vegan**, demonstrating that plant-based foods remain cost-effective when considering a balanced diet.
2. **Maximum 35% Calories from Protein**:

- For individuals with higher protein requirements, such as athletes, we focused on products where **at least 35% of calories come from protein**.
- In this category, **all of the top 10 cheapest products are non-vegan**, indicating the reliance on animal-based products for high-protein diets.

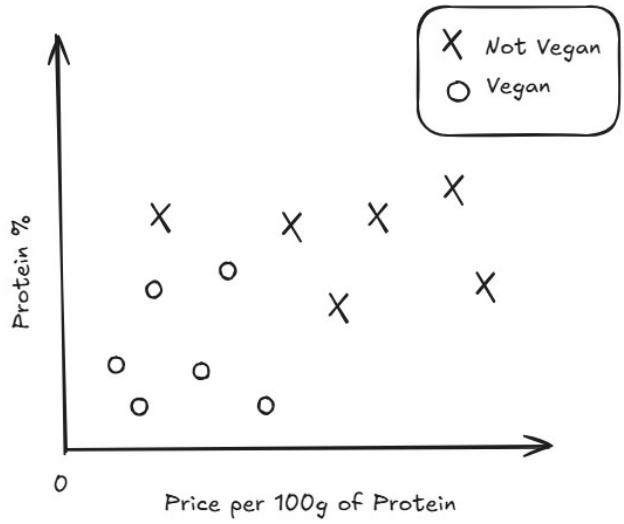
3. **Optimal 20% Calories from Protein**:

- For a balanced dietary approach, we settled on **20% calories from protein** as a benchmark. The graph of the **top 10 cheapest sources of protein** under this criterion highlights the shift in cost-effectiveness:
- Only **2 out of the top 10 products are vegan**, suggesting that while vegan products can be cost-effective at lower protein percentages, they become less competitive as protein density requirements increase.

This exploratory phase reveals that while vegan foods are highly cost-effective sources of calories and protein, their macronutrient composition often skews towards carbohydrates. For high-protein diets or more balanced macronutrient needs, non-vegan products, particularly animal-based, dominate the cost-effectiveness rankings. These insights set the stage for deeper analysis and visualization of the comparative affordability of protein across dietary categories.

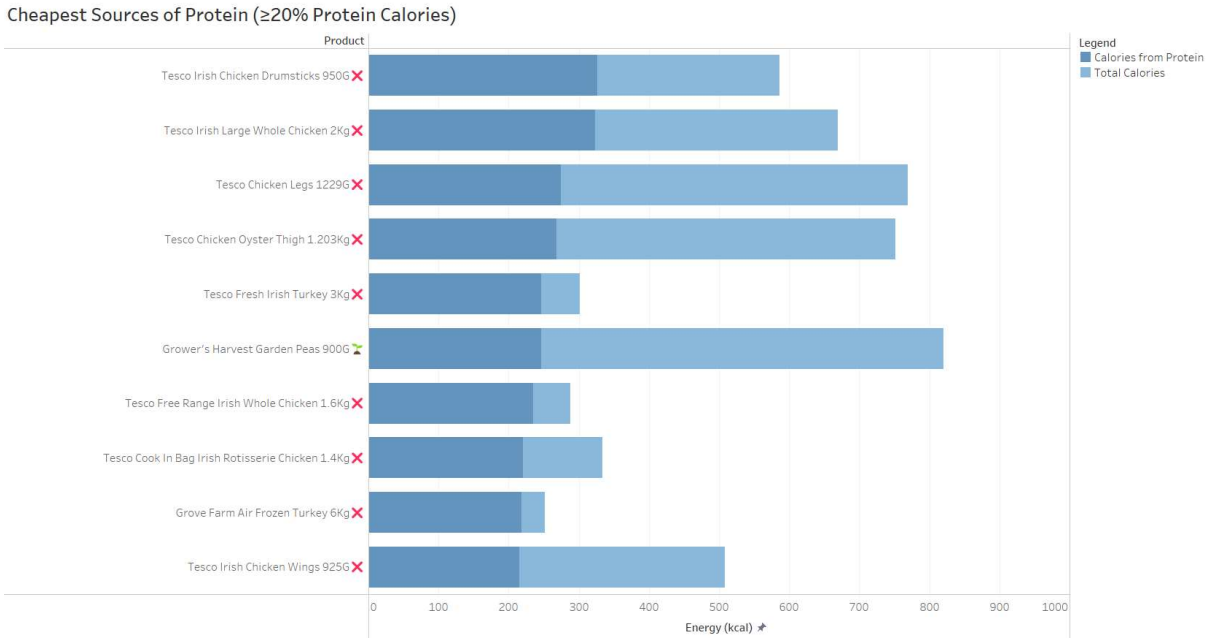
Visualization

In the design process, we also considered using a scatter plot to visualize the relationship between the cost of protein and the percentage of calories derived from protein. This approach would plot products with **Price per 100g of Protein** on the x-axis and **Protein Percentage of Calories** on the y-axis, using different colors or symbols to distinguish vegan and non-vegan products. While this method could highlight cost-protein trade-offs and outliers effectively, it lacked the clarity needed for non-technical audiences and did not explicitly show total calorie information, which is critical to the analysis. Ultimately, we opted for a horizontal bar chart for its straightforward presentation and ability to emphasize both protein contribution and caloric context.



Design Choice

For the final visualization, the selected graph is titled **"Cheapest Sources of Protein (≥20% Protein Calories)."** This visualization effectively communicates the affordability of protein sources while accounting for the percentage of calories derived from protein. The choice of a horizontal bar chart allows for a clear comparison of energy values (total calories and calories from protein) across different products.



Why This Chart?

This chart type was chosen because:

- **Comparative Clarity**: Horizontal bars facilitate an easy comparison of energy values across products.

- **Dual Information**: It simultaneously shows both the total calories and the subset of calories coming specifically from protein.
- **Highlighting Protein Contribution**: By filtering for products with at least 20% of calories from protein, the graph emphasizes the products that balance affordability with a protein-focused caloric profile.
-

Insights from the Visualization

1. **Non-Vegan Dominance**:
    - The top 10 cheapest sources of protein (≥20% protein calories) are overwhelmingly non-vegan, with only one vegan option (Grower's Harvest Garden Peas).
    - This suggests that while vegan protein sources may be cost-effective, they often lack the caloric density required for a balanced macronutrient profile at this protein threshold.
2. **Caloric Distribution**:
    - The dark blue segments represent calories from protein, while the light blue segments indicate total calories.
    - For most non-vegan products, a significant portion of total calories comes from protein, making them more suitable for individuals with high protein requirements without over-consuming calories.
3. **Trade-offs in Vegan Sources**:
    - Vegan products, while affordable, tend to derive most of their calories from carbohydrates or fats, as evidenced by their limited representation in the top 10 products at the 20% threshold.

Design and Styling

- **Colors**: The dark and light blue shades distinguish calories from protein versus total calories, ensuring the protein contribution is visually distinct.
- **Labels**: Product names are clearly labeled along with symbols indicating vegan (🍸) or non-vegan (❌) status, enhancing readability and context.
- **Focus on Energy Efficiency**: The x-axis, labeled in kcal, highlights energy density while keeping the focus on affordability.

Limitations

- **Lack of Micronutrient Data**: The chart does not account for essential micronutrients, which could impact the overall dietary quality.
- **Single Metric**: While the graph highlights the cost of protein at the 20% threshold, it does not provide insights into other thresholds or nutrient ratios.

Why This is the Final Choice

This visualization succinctly answers the research question by demonstrating that while vegan protein sources are often cheap, they may not provide an optimal balance of macronutrients at higher protein percentage thresholds. It highlights that non-vegan products dominate the cost-effective category for protein-focused diets, particularly at the 20% protein threshold, making it the ideal choice to convey the study's key insights.

Conclusion

This study analyzed the cost-effectiveness of vegan and non-vegan foods as protein sources using scraped nutritional data from Tesco Ireland. While vegan options like oats and peas are among the cheapest, they often derive most calories from carbohydrates, making them less optimal for higher protein requirements. Filtering for ≥20% of calories from protein revealed non-vegan products as the most cost-effective sources for balanced or protein-focused diets.

Tools Used

1. **Browser JavaScript**: To fetch Tesco product URLs.
2. **Python (Google Colab)**: For scraping, cleaning, and processing nutritional data.
3. **Tableau**: For creating the final visualization.

The horizontal bar chart effectively highlights the trade-offs between affordability and macronutrient composition, providing insights for consumers balancing cost with dietary needs. Future analyses could include micronutrient content and sustainability considerations.

This project was completed as a team effort, with most of the work shared evenly. Mervin fetched the product URLs using JavaScript and uploaded them as a JSON file to Google Drive. Both team members collaborated on scraping, cleaning, and processing the data using Python in Google Colab. Saurav created the final visualization in Tableau after jointly discussing and refining sketches with Mervin.

References

1. https://www.tesco.com
2. https://www.mayoclinichealthsystem.org/hometown-health/speaking-of-health/are-you-getting-too-much-protein