University of New Haven

TAGLIATELA COLLEGE OF ENGINEERING

Electrical & Computer Engineering and Computer Science

Tagliatela College of Engineering

TECHNICAL REPORT

SEMESTER

CONTENTS

Executive Summary

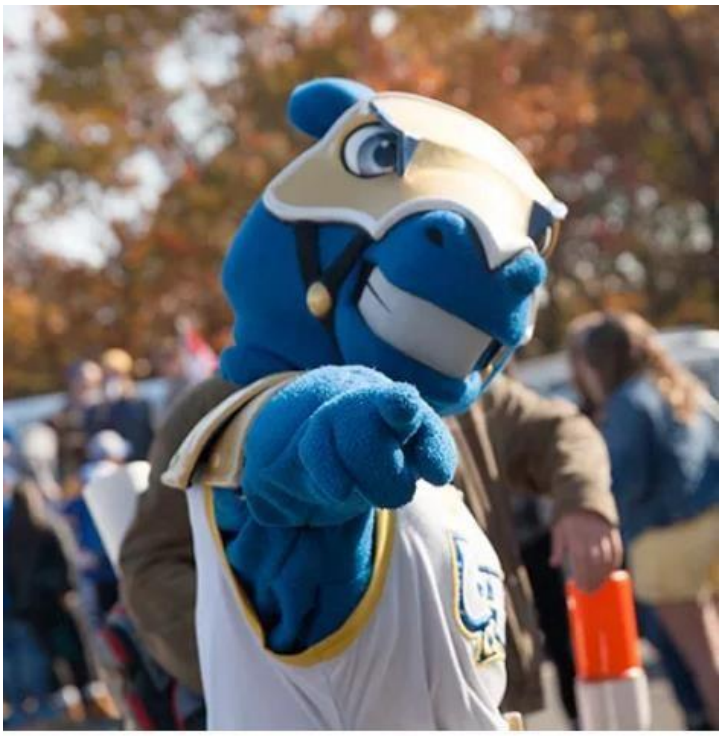
This project dives into the analysis of food and nutrition data using Python, uncovering meaningful insights into the nutritional content of various food categories. Through a detailed process of data cleaning, exploratory data analysis (EDA), and visualization, it highlights critical metrics such as protein, fat, fiber, and calories. The findings shed light on the top nutrient-rich foods and major caloric contributors, providing practical guidance for making healthier dietary choices.

Team Members:
SAI KIRAN NELLURU
MOHAMMED TAQI AHMED IYAAZ
DESHIK SASTRY YARLAGADDA

Questions?
Contact :
snell10@unh.newhaven.edu
tmoha19@unh.newhaven.edu
dyarl1@unh.newhaven.edu

Technical Report

Exploration of Food and Nutrition Data Using Python



Highlights of Project

Highlights of Project

- **Tools Used:** Python (Pandas, Matplotlib, Seaborn, Plotly).
- **Data Cleaning:** Addressed missing values, inconsistencies, and duplicates.
- **EDA:** Visualized distributions of macronutrients across food categories.
- **Category Analysis:** Grouped data by food type to identify nutrient-rich categories.
- **Specific Insights:** Identified top protein, calorie, and fat contributors.

Submitted On:
December 9, 2024

Abstract

The project uses Python's robust data processing and visualization tools to analyze a comprehensive food and nutrition dataset. By identifying nutrient-dense foods and revealing category-level nutritional trends, it provides actionable insights into dietaryplanning. Advanced visualizations, like 3D scatter plots and heatmaps, complement the analysis to make the data more accessible and actionable.

Here you can find a few useful tips on coming up with a great pitch

Cover Page

PREPARED BY
Sai Kiran Nelluru
Mohammed Taqi Ahmed Iyaaz
Deshik Sastry Yarlagadda

Affiliation
University of New haven

Course or Department
Data Science

Instructor
Dr. Ardiana Sula

Date of Submission
12/08/2024

Contact :
snell10@unh.newhaven.edu
tmoha19@unh.newhaven.edu
dyarl1@unh.newhaven.edu

Introductory Section

The relationship between food and health is as old as time, but modern tools allow us to understand it on a deeper level. This project employs Python's analytical prowess to explore a rich dataset of nutritional information, cleaning and visualizing it to discover trends and insights that can guide dietary choices. The findings demonstrate the importance of various food categories in delivering essential nutrients.

Review of available research

In recent years, there has been growing attention to how nutrient tracking can influence health outcomes. Studies often focus on single nutrients or specific food categories, leaving room for a broader approach. This project fills that gap by combining comprehensive data analysis with advanced visualization techniques. Heatmaps, 3D scatter plots, and category-level summaries bring a new perspective, revealing how dietary choices impact overall nutrition.

Methodology

1. **Data Cleaning:**
    - Addressed inconsistencies and missing values by leveraging techniques such as imputation and outlier removal.
    - Standardized units of measurement to ensure seamless comparison across data points.
    - Employed statistical summaries and visualization checks to validate the integrity of the data before proceeding to analysis.
    - Identified and eliminated duplicate entries to maintain dataset uniqueness.
2. **EDA:**
    - Created histograms, bar charts, and boxplots to explore distributions and identify key trends in nutritional data.
    - Grouped foods by category to understand the variance in nutrient density within and across categories.
    - Conducted correlation analyses between macronutrients such as calories, fat, and protein to uncover interdependencies.
    - Detected anomalies and flagged potential data inconsistencies for further investigation.
3. **Category Analysis:**
    - Utilized pie charts and stacked bar charts to visualize the proportional contributions of macronutrients by food category.
    - Highlighted categories with the highest and lowest contributions to key nutritional metrics, such as fiber and protein.
    - Identified potential gaps in nutrient offerings within specific food groups.

4. **Advanced Visualizations:**
    - Designed funnel charts to analyze calorie density in desserts and drinks, identifying the most calorically dense options.
    - Used 3D scatter plots to visualize multi-dimensional relationships between protein, fat, and caloric content across food categories.
    - Created heatmaps to represent nutrient density and detect patterns across various food types, providing a comprehensive view of the data.
5. **Additional Statistical Techniques:**
    - Conducted principal component analysis (PCA) to reduce dimensionality and identify dominant trends in the dataset.
    - Performed clustering analysis to group foods with similar nutritional profiles, aiding in targeted recommendations for balanced diets.
    - Implemented hypothesis testing to confirm significant differences in nutrient distributions across categories.

Results Section

pipeline

1. Data Acquisition: Collected a raw dataset of food and nutrition information from publicly available sources and verified its credibility.
2. Data Cleaning: Removed errors, handled missing values through imputation, and standardized units for consistency.
3. Exploratory Data Analysis (EDA): Conducted detailed analysis using visual tools such as histograms, scatter plots, and correlation matrices to identify patterns and relationships.

4. Visualization: Created compelling visualizations, including bar charts, pie charts, funnel charts, and 3D scatter plots, to represent findings effectively.
5. Insights Extraction: Highlighted key observations from the visualizations to guide actionable conclusions.
6. Data Validation: Ensured all calculations were accurate and reproducible through cross-validation and consistency checks.
7. Data Engineering pipeline: Your data engineering pipeline for your project (provide your designed schema):
    - Data Ingestion: Tools and applications
    - Data Storage: Tools and applications
    - Data Processing: Tools and applications
    - Data Consumption: Your App (Tools and applications)
        - Model Deployment: Describe how you've created a deployable environment for your model.
        - Data Visualization**:** Showcase the results through comprehensive visualizations.
    - Deployment

RESULTS

- Protein-Rich Foods: Identified top 20 protein-packed foods, including vegetables, grains, and seeds.
- Caloric Content: Pinpointed foods with the highest calorie density.
- Fat Analysis: Showed top 20 high-fat foods with visual aids.
- Category Contributions:
    - Breads and Cereals: High calorie and carbohydrate contributors.
    - Seafood: A standout category for protein.
- Fiber Insights: Vegetables and legumes emerged as leaders in fiber content.
- Correlation Analysis: Revealed strong links between fat levels and caloric content.

Discussion

The findings from this project highlight the diverse nutritional landscape of common food categories. Seafood stands out as a top protein source, while bread and cereals dominate caloric intake. These insights underscore the value of nutrient tracking for balanced diets. Addressing challenges like missing data has strengthened the reliability of the analysis, but future research could enhance it by integrating global consumption patterns and exploring temporal trends in food production.

Conclusion

In Conclusion, analyzing food and nutrition data using Python has proven to be a powerful way to uncover meaningful patterns. This project not only supports healthier dietary choices but also sets the stage for more advanced studies, including longitudinal analyses and international comparisons. By visualizing data in an accessible manner, it empowers individuals and organizations to prioritize nutrition in their decision-making.

References

https://pandas.pydata.org/docs/
https://seaborn.pydata.org/examples/index.html
https://matplotlib.org/stable/contents.html
https://aws.amazon.com/s3/
https://fdc.nal.usda.gov/
https://www.kaggle.com
https://plotly.com