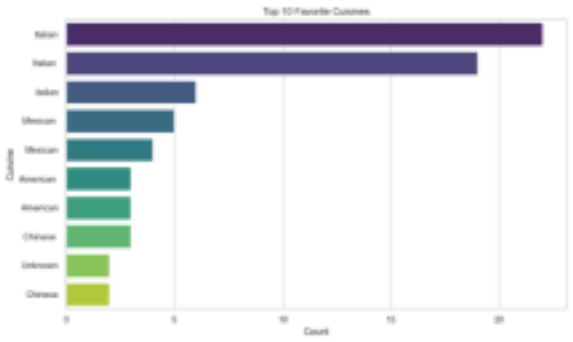
Summary of the Food Dataset Analysis Project

Overview

Here,we analyze the food_coded.csv dataset from kaggle, which contains data on college students' dietary habits, preferences, and related factors such as GPA, gender, and exercise. The project aims to clean the dataset, perform exploratory data analysis (EDA), and create visualizations to uncover insights into dietary patterns and their potential relationships with academic performance and lifestyle factors.

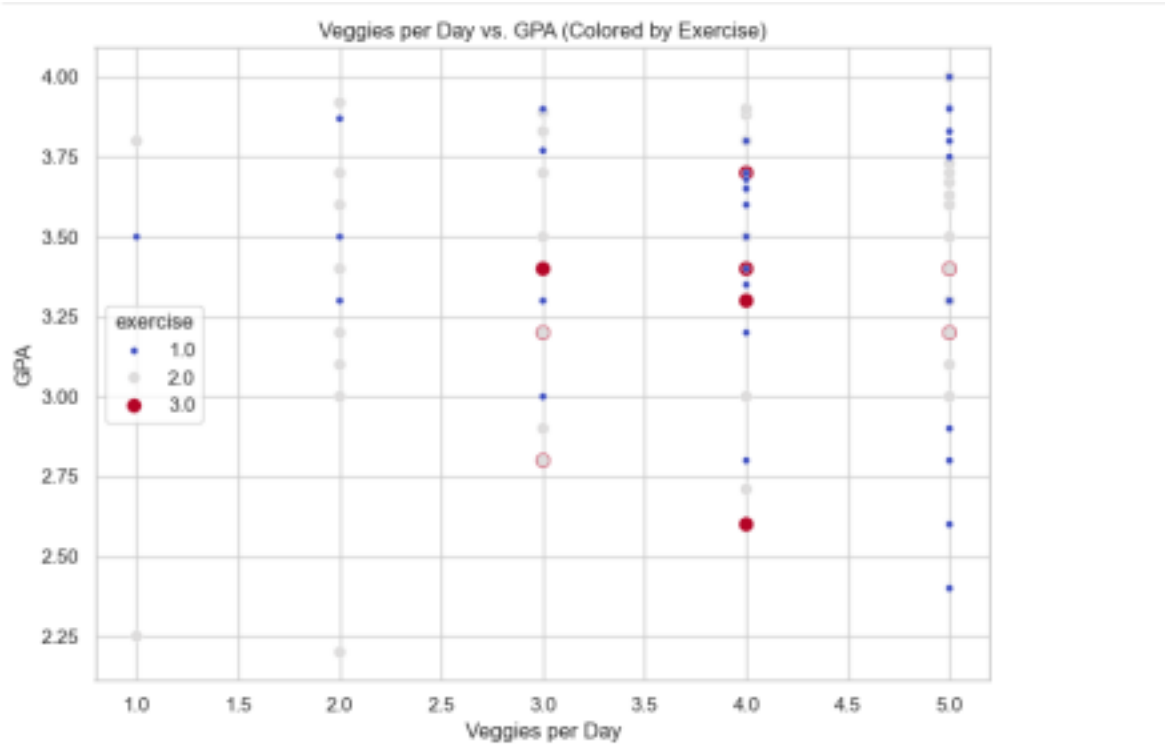
Key Components

1. **Data Loading and Cleaning**:
    - The dataset is loaded using pandas from food_coded.csv.
    - Cleaning steps include:
        - Converting 'nan' strings to actual NaN values.
        - Converting GPA to numeric, handling non-numeric entries.
        - Filling missing numerical values with medians and categorical values with 'Unknown'.
        - Removing duplicate rows.
    - This ensures a clean dataset for analysis.
2. **Exploratory Data Analysis (EDA)**:
    - **Dataset Exploration**: Displays the dataset's shape, info, and descriptive statistics to understand its structure.
    - **Analysis**:
        - Groups data by Gender to compute average GPA, calorie intake (calories_day), and exercise levels.
        - Filters students with high vegetable consumption (veggies_day >= 4). - Sorts by GPA to identify top performers.
        - Aggregates counts of favorite cuisines (fav_cuisine) to find the top 10.
3. **Data Visualization**:
    - Three visualizations are created using matplotlib and seaborn:
        - **Bar Plot**: Shows the top 10 favorite cuisines (e.g., Italian, Mexican) by count.



        - **Box Plot**: Displays GPA distribution by gender (1=Female, 2=Male), highlighting medians and outliers.

- **Scatter Plot**: Plots veggies_day vs. GPA, with points colored and sized by exercise level, to explore potential correlations.



- Plots are displayed inline in Jupyter and saved as PNG files (cuisine_barplot.png, gpa_boxplot.png, veggies_gpa_scatter.png) for reporting.

4. **Key Findings**:
   - **Cuisine Preferences**: Italian cuisine is the most popular, followed by Mexican

and Chinese, likely due to their availability in campus dining.

- **GPA and Gender**: Females (Gender=1) have a slightly higher median GPA than males (Gender=2), with similar variability and outliers.
- **Diet and Academic Performance**: A weak positive correlation exists between vegetable consumption (veggies_day) and GPA, especially for students who exercise regularly (exercise=1). Students with high veggie intake (4–5 per day) tend to have higher GPAs.
- **Surprising Insight**: Many students report high vegetable consumption (>=4 per day), challenging the stereotype of college students relying on junk food.

5. **Conclusion**:
   - The analysis suggests a potential link between healthy eating (high veggie intake) and better academic outcomes, though not definitive.
   - Italian cuisine's dominance reflects campus dining trends.
   - Future analyses could explore comfort foods or exercise's impact on health perceptions.
6. **Exporting**:
   - Instructions are provided to export the notebook to PDF using jupyter nbconvert or HTML-to-PDF methods, ensuring saved plots are included.
7. **Dataset Choice**:
   - The food_coded.csv dataset was chosen for its mix of numerical (e.g., GPA, calories) and categorical (e.g., cuisine, comfort food reasons) variables, ideal for demonstrating data cleaning, EDA, and visualization.
   - Its focus on college students makes it relatable for studying dietary and lifestyle patterns in a specific demographic.

This project demonstrates fundamental data science skills—cleaning, analyzing, and visualizing data—while providing insights into college students' dietary habits and their potential academic implications. Let me know if you need a mock dataset, further detai