Cloud-Hosted Notebook Data Manipulation

1. Project Overview

The CloudHosted Notebook Data Manipulation project demonstrates data manipulation techniques on a dataset of various cereal brands (`cereal.csv`). This project involves tasks such as data cleaning, feature engineering, and aggregation to extract insights. The notebook is hosted on Google Colab for ease of access, and the project includes a CI/CD pipeline for automated code quality checks and testing using GitHub Actions.

Goals:

- Perform data manipulation tasks on a real-world dataset.
- Implement automated testing and linting using a CI/CD pipeline.
- Host the notebook on a cloud platform for easy sharing and reproducibility.

2. Notebook Access and Setup

Google Colab Access
The notebook is hosted on Google Colab. You can access it directly using this [Colab Link](link_to_your_colab_notebook).

Repository Access
All project files, including the notebook, are available in the GitHub repository:
[GitHub Repository Link](link_to_your_github_repository)

Running the Notebook Locally
1. Clone the repository:
   ```bash
   git clone https://github.com/yourusername/yourrepo.git
   ```
2. Install the required dependencies:
   ```bash
   pip install r requirements.txt
   ```
3. Open the notebook using Jupyter Notebook or JupyterLab.

3. Data Manipulation Tasks

Data Loading
The dataset (`cereal.csv`) is loaded directly from GitHub into the notebook using the following code:

```python
import pandas as pd
url = 'https://raw.githubusercontent.com/yourusername/yourrepo/main/cereal.csv'
df = pd.read_csv(url)
```

Data Cleaning
Remove Duplicates: Ensures unique entries by dropping duplicate rows.
Handle Missing Values: Missing values are either filled with default values or removed if necessary.

Feature Engineering
Calorie Category: The dataset is enriched by adding a new column, `calorie_category`, to categorize cereals based on calorie content:
Low (<100 calories), Medium (100150 calories), High (>150 calories).

```python
import numpy as np
conditions = [
    (df['calories'] < 100),
    (df['calories'] >= 100) & (df['calories'] < 150),
    (df['calories'] >= 150)
]
choices = ['Low', 'Medium', 'High']
df['calorie_category'] = np.select(conditions, choices, default='Unknown')
```

SugartoFiber Ratio: A `sugar_fiber_ratio` column is created to represent the balance of sugar and fiber in each cereal, helping identify healthier options.

Aggregation and Filtering
Manufacturer Analysis: The dataset is grouped by `manufacturer` to calculate each brand's average calories, protein, and fiber.
HighProtein Filter: Cereals with more than 4 grams of protein per serving are filtered to highlight options with higher protein content.

Summary of Data Manipulation Results
Summarized insights, such as calorie distribution across different manufacturers and the proportion of high protein cereals, are provided at the end of the notebook.

4. CI/CD Pipeline Explanation

The CI/CD pipeline is set up using GitHub Actions and includes the following steps:

1. Install Dependencies: Installs all necessary libraries from `requirements.txt`.
2. Linting with flake8: Checks code style and syntax to ensure consistency and readability.
3. Testing with pytest: Run unit tests to validate the correctness of scripts.
4. Optional Notebook Execution: Converts the notebook to a Python script using `jupytext` and runs it to ensure all cells execute without errors.

Triggering the Pipeline:
The pipeline is triggered on each `push` or `pull_request` event, automating testing and linting to maintain code quality.

5. Key Insights and Findings
From the data manipulation tasks, the following insights were gained:

- Calorie Distribution: The majority of cereals fall into the `Low` calorie category, indicating a focus on lower calorie options within the dataset.
- HighProtein Cereals: Only a small selection of cereals has more than 4 grams of protein per serving, which might be valuable for consumers seeking high protein options.
- SugartoFiber Balance: Certain cereals exhibit a favorable sugar to fiber ratio, providing a potential measure for identifying healthier choices.

These insights can help consumers or nutritionists understand trends in cereal composition and guide better dietary choices.

6. Troubleshooting and Known Issues
Common Issues:

- ModuleNotFoundError: If a module is not found, ensure that all dependencies are listed in `requirements.txt` and are correctly installed with:
  ```bash
  pip install r requirements.txt
  ```
- CI/CD Pipeline Errors: If the pipeline fails, check the GitHub Actions logs to identify the specific step where the error occurred. Common issues include unmet dependencies or syntax errors detected by `flake8`.
- Notebook Execution Errors: Ensure that all file paths in the notebook match the repository structure, especially if you're loading the dataset from GitHub.

7. Conclusion

The Cloud-Hosted Notebook Data Manipulation project successfully demonstrates data manipulation, feature engineering, and analysis on a real-world dataset. The CI/CD pipeline ensures that the project remains error-free and follows best practices in coding standards. Hosting the notebook on Google Colab and setting up a pipeline with GitHub Actions enhances the reproducibility and maintainability of the project, making it easier for collaborators and reviewers to verify results.