Data Science Challenge:

Uncovering the Hidden Profiles in Wine Data

The dataset

These data are the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The analysis determined the quantities of 13 constituents found in each of the three types of wines.

The attributes are:

- Alcohol
- Malic acid
- Ash
- Alcalinity of ash
- Magnesium
- Total phenols
- Flavanoids
- Nonflavanoid phenols
- Proanthocyanins
- Color intensity
- Hue
- OD280/OD315 of diluted wines
- Proline

The dataset can be downloaded from this link (*Extra point if downloaded via api call or within the code)*

Objective:

Explore the provided wine dataset to uncover interesting patterns, profiles, or correlations among the different chemical constituents of the wines. The challenge is open-ended, encouraging creative and exploratory analysis.

Task Description:

Data Exploration:
Perform an initial exploration of the dataset to understand the distribution and relationship of various attributes like Alcohol, Malic Acid, Flavanoids, etc.
Use visual tools and statistical techniques for a comprehensive analysis.

You can use a notebook to show and document your findings

Clustering Analysis:
Apply clustering techniques (e.g., K-Means, Hierarchical Clustering, DBSCAN) to group the wines based on their chemical properties. Determine the optimal number of clusters and justify your choice.

Analyze the characteristics of each cluster. What does each cluster represent in terms of wine properties?

Creative Insights and Storytelling:
Beyond the technical analysis, craft a narrative around your findings. What interesting stories do the data tell about the wines?
Are there any unexpected correlations or findings that could be of interest to wine makers or enthusiasts?

Presentation:
The task must be presented in a Github repository where all the code and documentation should be. *Extra points for the good use of this tool*

Optional Task:

Containerization and Data Retrieval:

- Containerize the Analysis: Dockerize the entire analysis pipeline. This includes creating a Dockerfile and necessary configurations to build a Docker image that can run the analysis. *Separately from the python notebook where you documented your findings, you could create a dockerfile that executes the code fluidly and print some findings*
- API Data Retrieval: Enhance the task by adding a component where the wine dataset is retrieved via an API call from the link provided. This could involve setting up a simple API (using Flask or FastAPI) that serves the dataset, which the Docker container can then access to perform the analysis.

Provide documentation on how to set up and run the Docker container, and how to access the data through the API. *(A readme will suffice)*

Evaluation Criteria:

- Depth and Creativity in data exploration and clustering analysis.
- Technical Proficiency in applying data analysis and clustering techniques.
- Quality of Insights: Uniqueness and depth of the insights and stories derived from the data.
- Optional Task: Success in containerizing the analysis and setting up data retrieval via an API, along with clear documentation.

Dataset link: https://storage.googleapis.com/the_public_bucket/wine-clustering.csv