CASE 01: From Grapes to Glass - the Story of Wine

This case study focuses on analysing data collected from various wine brands across different countries.

For the analysis and visualization, you will be provided a folder “Wine_Stats” containing 8 csv files. Each file corresponds to a country where the wine is produced. The files include several statistics related to each wine brand in a uniform format. The table below outlines the key attributes included in the dataset:

Attribute	**Description**
Name	The name of the wine
Rating	The average rating of the wine
Number of Ratings	The number of ratings the wine has received
Price	The price of the wine in USD
Region	Country, Region, and the wine producing area Ex: Chile / Aconcagua / Casablanca Valley
Winery	The winery that produces the wine
Wine style	The style of the wine (Ex: Chilean Carménère)
Alcohol content	The percentage of alcohol in the wine
Grapes	The type of grape used in the wine
Food pairings	Suggested food pairings for the wine
Bold	A score representing the boldness of the wine
Tannin	A score representing the tannin level in the wine
Sweet	A score representing the sweetness of the wine
Acidic	A score representing the acidity of the wine

Apart from the above, you will also receive an additional file “wine_reviews.csv”. This contains 500 customer reviews received for the wine Merry Edwards Sauvignon Blanc 2023. This file should be used to complete Task 03.

TASK 01: Maintain a GitHub Repository

- From the beginning, create and maintain a GitHub repository for the project.
- Follow proper version control practices and GitHub etiquettes (e.x: meaningful commits).
- We will limit our evaluation to the Python scripts and Jupyter notebooks present in the repository. Please ensure all your code is pushed promptly!
- Refer to the marking grid to ensure all necessary components are addressed for evaluation.

TASK 02: Data Preparation

To achieve the passing mark, the following tasks are mandatory. Implementing advanced techniques will earn extra credit. Carry out the below tasks in a Jupyter Notebook.

1. **Reading and combining data**
   - Load all 8 CSV files into a list.
   - Concatenate the files into a single DataFrame, named wine_df.

2. **Initial data exploration and cleaning**
   - Examine the DataFrame structure, including its features and data types.
   - Remove any duplicate records.
   - Remove null records if they exist.

3. **Handle outliers and missing values**
   - Perform outlier removal and missing value imputation only if necessary.
   - State the reason for any such actions (you can state the reasons within the notebook).

4. **Adding new columns to the Dataframe**
   1. Country
      - A string column indicating the country where the wine is produced

      Hint: *Extract this information from the region column using appropriate processing steps.*
   2. Country_region
      - A string column indicating the region of the country where the wine is produced.

      Hint: *Extract this information from the region column using appropriate processing steps. Country region is indicated after the country.*
   3. The column “Food pairings” contains values in a list format. You are required to create new variables to store each list element.

- Ex: If a particular wine has food pairings as [‘Beef’, ‘Pasta’, ‘Lamb’, ‘Poultry’] you will create 4 new columns with the values as follows:

Beef	Lamb	Poultry	Pasta
TRUE	TRUE	TRUE	TRUE

Hint*: At the end of this step, you will introduce 21 new columns each representing a food.*

5. Column Removal

- Drop irrelevant columns and provide reasons.

TASK 03: Deploying a HuggingFace Model

Complete this task in a separate Jupyter notebook. Treat it as an independent task, and there’s no need to consider it in relation to the rest of the tasks.

- Read the data from the wine_reviews.csv file. It has 500 customer reviews received for a particular wine brand.
- Select a suitable zero-shot classification model from HuggingFace and provide the rationale behind selecting the model.
- Using the selected model, classify the reviews into one of the below classes;
    1. talks about food combinations
    2. talks about taste
    3. talks about value for money
    4. other
- Add the predicted labels to the dataset as a new column (name the column “talks_about”).
- Visualize the spread of the above categories using a suitable chart.
- Ensure to push both the updated dataset and the notebook to the GitHub repo.

TASK 04: Dashboard Creation

- Design a dashboard using Plotly Dash that tells an insightful story with the data.
- Be SMART!!! There are many different charts you can use to visualize data. Refer to Plotly documentation to decide the best and most interactive charts to tell your story.
- Refer to the marking grid to cover all required aspects.

SUBMISSION GUIDELINES

- **Python scripts and notebooks:** Push to a public GitHub repository
- **Dashboard:** Screen record and submit as a video
- **Presentation:** A maximum of 5 slides explaining what you did in the analysis
- Upload the below items to the Google Form:
    1. GitHub repository link (public)
    2. Video clip of the dashboard
    3. PowerPoint Presentation

MARKING GRID

	Task	Weight	Evaluation Criteria - minimum requirements	
Git	**Maintain a git repo for the project**	10%	1.1	All the team members should be added to the project
			1.2	Maintain branches for each component/member
			1.3	At least two commits per member
			1.4	At least one completed pull request
			1.5	Make commits on-the-spot (not at the end)
			1.6	Maintain proper branch naming conventions
			1.7	Maintain meaningful commits
			1.8	Main branch should be free of conflicts
Pandas	**Data preparation**	30%	2.1	Read data files
			2.2	Merge files
			2.3	Remove duplicate/null records
			2.4	Impute missing values (only if required)
			2.5	Outlier removal (only if required)
			2.6	Pivoting / Grouping
NLP	**Deploying a Hugging Face model**	20%	3.1	Pick a suitable model
			3.2	Reliability of the model
Visualization	**Dash Dashboard**	40%	4.1	Use correct charts to represent data
			4.2	Include at least 5 different types of charts
			4.3	Call the charts to a dashboard
			4.4	Use interactive features on the dashboard (ex: filters)
			4.5	Clarity of the dashboard
			4.6	Story-telling

To pass, you must score at least 65% of the allocated marks in each section.

If you have any queries reach out to us via:
uvinir@uom.lk
samujitha.senaratne@acuitykp.com