**1. Introduction**

The purpose of this project is to analyze a wine dataset to uncover insights about wine prices, ratings, and reviewers. We aim to answer key questions such as:

- What is the distribution of wine ratings across different countries?

- How does wine price vary by country, region, or variety?

- Which wine varieties have the highest average ratings?

- Is there a correlation between price and points (higher price = better ratings)?

- Which tasters or wineries consistently rate wines higher?

**2.Tools and Technologies Used**

- Python (pandas, seaborn, matplotlib, numpy)

- Google Docs(for documentation)

- VS Code (for scripting)

- Organized folder structure (data,notebooks, outputs, scripts.)

**3. Dataset Description**

The dataset includes information about wines from various regions, with the following key columns:

- Country: The country of wine origin.

- Price: The price of the wine.

- Points: The rating given to the wine.

- Variety: The grape variety used.

- Taster Name: The reviewer’s name.

- Additional columns such as winery, designation, taster\_twitter\_handle and region.

**4.Data Preprocessing**

- Missing values in columns like `price` and `country` were handled by imputation.

- Cleaned data saved in the `data` folder as `cleaned\_wine\_data.csv`.

**5. Methodology**

The project followed these steps:

1.Data Cleaning:

- Dropped the unnecessary column `Unnamed: 0`.

- Handled missing values and inconsistent formats.

- Verified the integrity of numerical columns such as `points` and `price`.

2. Exploratory Data Analysis (EDA):

- Used Python’s pandas and seaborn libraries to visualize trends and patterns.

- Plotted box plots, histograms, scatter plots, and count plots.

3. Key Questions Addressed:

- Correlation analysis between price and points.

- Identifying top-rated wine varieties.

- Understanding rating and pricing diversity by country.

**6.Insights and Findings**

Distribution of Wine Ratings Across Different Countries:

- France, Italy, and the USA dominate in terms of the diversity and range of wine ratings.

- Countries like Portugal and Germany also show consistent high ratings.

Price by Country of Wine Origin:

- Significant variation in price ranges across countries.

- Italy and Spain have a wide range of prices, while Romania and Bulgaria have lower price ranges overall.

Wine Varieties with Highest Average Ratings:

| Variety | Average Points |

|----------------------|----------------|

| Terrantez | 95.00 |

| Tinta del Pais | 95.00 |

| Gelber Traminer | 95.00 |

| Bual | 94.14 |

| Sercial | 94.00 |

| Riesling-Chardonnay | 94.00 |

| Ramisco | 93.00 |

| Garnacha-Cariñena | 93.00 |

| Blauburgunder | 93.00 |

| Muscadelle | 92.50 |

Correlation Between Price and Points:

- Higher-rated wines tend to be more expensive, with some clusters and outliers indicating other influencing factors.

- Only 1,089 wines have ratings below 82 points, and 881 wines have ratings above 95 points.

Reviewer and Winery Analysis:

- Roger Voss shows a strong affinity for wines like Salon and Château Pétrus.

- Michael Schachner has a preference for Italian wineries like Tenuta dell'Ornellaia and Petra.

**7. Visualizations**

The following visualizations were created and saved in the `outputs` folder:

1.Top 20 Tasters: Count plot showing the most frequent reviewers.

2.Price by Country: Box plot showing price variation by country.

3. Points Distribution by Country: Box plot showing the spread of wine ratings across different countries.

4.Top Wine Varieties: Bar plot highlighting varieties with the highest average ratings.

5. Correlation Analysis : Scatter plot of price vs. points.

6. Points Categories: Count plot categorizing wines by their ratings.

**8. Code Details**

Scripts Folder

- Scripts are stored in the `scripts` folder.

- Key scripts include:

- `data\_cleaning.py`: For handling missing values and formatting issues.

- `eda.py`: Contains code for visualizations and insights.

- `correlation\_analysis.py`: Code for analyzing relationships between variables.

**9. Conclusion**

The analysis highlights:

- The strong correlation between price and ratings, with some exceptions.

- The diverse range of wine quality in countries like France and Italy.

- Reviewer preferences influencing the ratings of certain wineries.

**10. References**

- Dataset source:Kaggle

- Libraries: pandas, matplotlib, seaborn