DOCUMENTATION

Data Cleansing

- Removed the 'username' column and 'designation' column from the dataset
- Combined 'region_1' and 'region_2' columns into one 'region' column
- Filled missing values in the 'price' column with the median value
- Removed rows with missing values in the 'variety' column
- Created a new 'year' column by extracting the year from the 'review_title' column

Data Analysis

- Count plot of the number of wines produced in the top 10 wine producing countries
- Box plot of the distribution of wine prices by variety
- Count plot of the top 10 wine varieties produced in the dataset
- Bar plot of the number of wines produced in the top 10 wine producing provinces
- Correlation matrix of the 'points' and 'price' columns of the DataFrame
- Filtering the DataFrame to identify high-quality wines priced under $20
- Implementation of a text classification model using logistic regression to predict the wine variety based on the review description
- Implementation of a Random Forest Classifier model to predict the wine variety based on various features

Top 5 Actionable Insights

- The majority of wines in the dataset come from the United States, followed by Italy and France.
- Some wine varieties, such as Pinot Noir, Cabernet Sauvignon, and Bordeaux-style blends, are more expensive on average than other varieties.
- The most produced wine variety in the dataset is Cabernet Sauvignon, followed by Pinot Noir, Chardonnay, and Red Blend.
- The province with the highest number of wines produced is California, followed by Washington and Tuscany.
- The text classification model achieved an accuracy of 71.5%, while the Random Forest Classifier model achieved an accuracy of 66.8% in predicting the wine variety based on various features.

This documentation summarizes my work and provides insights that can be actionable to the stakeholders.