



TASK

Exploratory Data Analysis on the wine Dataset

Visit our website

Introduction

The dataset comprises statistical data collected about wine. This dataset can be used to gain insight into the most popular wine, the most expensive wine, and which country produces the most popular wine. These are just a few of the deductions that can be made from the data.

Key attributes include:

- country : object
- points : int64
- price : float64
- province : object
- variety : object
- winery : object

points-group : category To conduct this analysis, we will utilize Python's data analysis and visualization libraries, including pandas, numpy, matplotlib, and seaborn. Techniques such as summary statistics, correlation analysis, and data visualization will be employed to uncover insights and highlight significant trends.

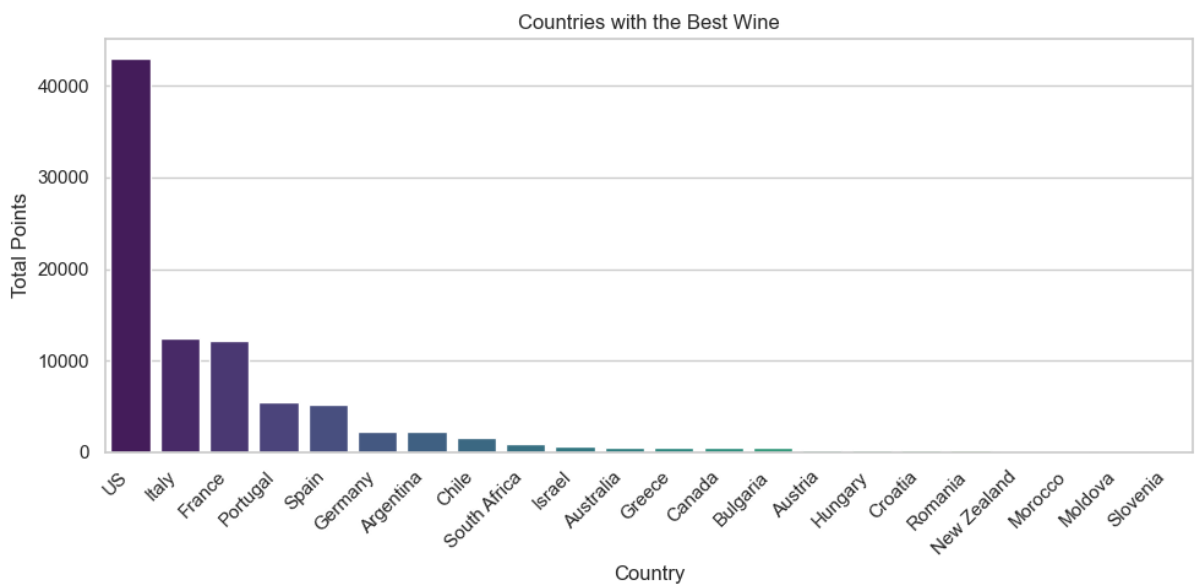
DATA CLEANING

1. Redundant and unnecessary data was removed from the dataset. The columns removed were identified as not relevant for the decision-making and evaluation of the data.
2. The removal of duplicate values from the dataset.
3. The removal of rows of missing data.
4. The removal of duplicate data was performed on the dataset.
5. The creation of points-group was done to have a better understanding of the wine quality and for easier data manipulation for the generation of graphs.

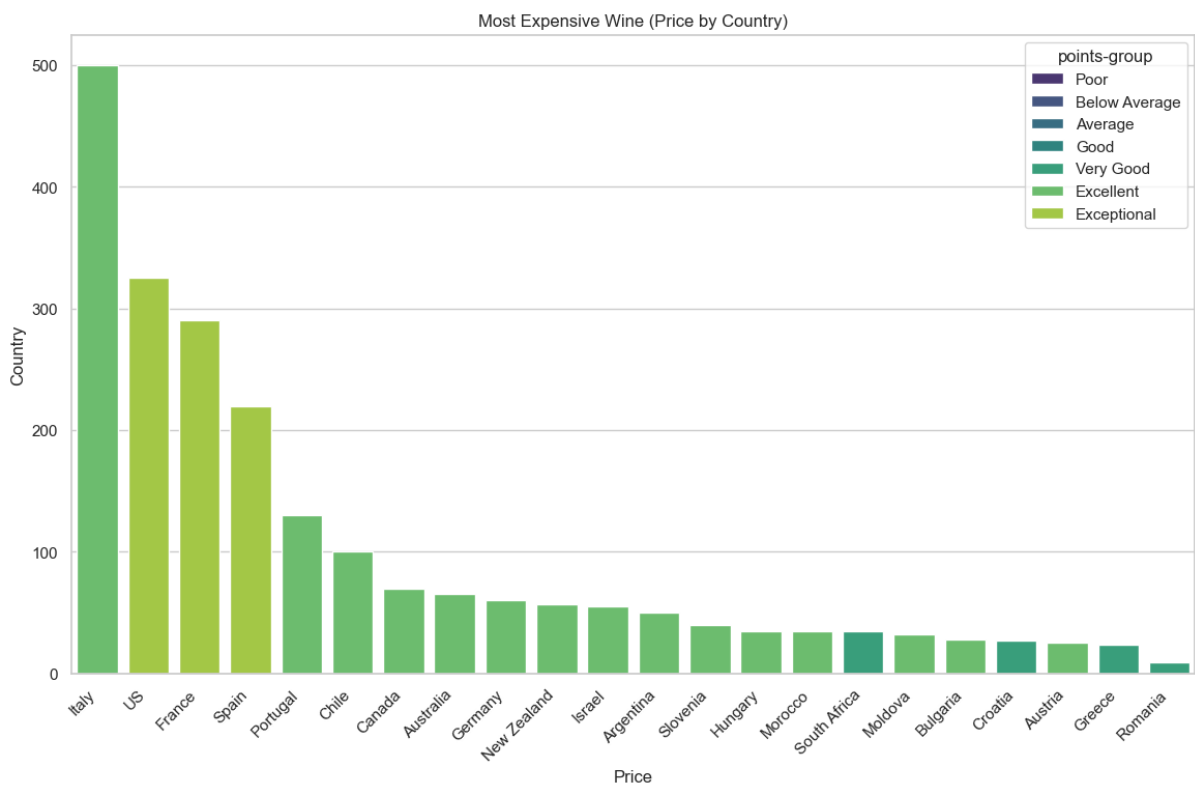
MISSING DATA

A check for missing data on the dataset was carried out. Upon inspection of the null check results the number of missing values was deemed small enough that the complete removal of the rows with the missing data would not affect the overall evaluation of the data.

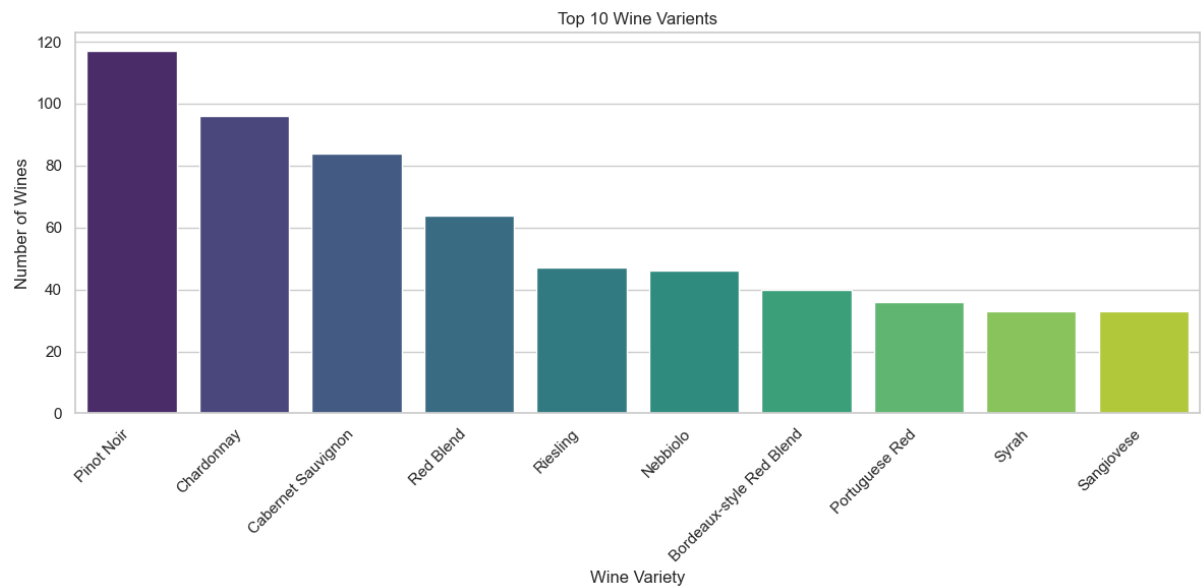
DATA STORIES AND VISUALISATIONS



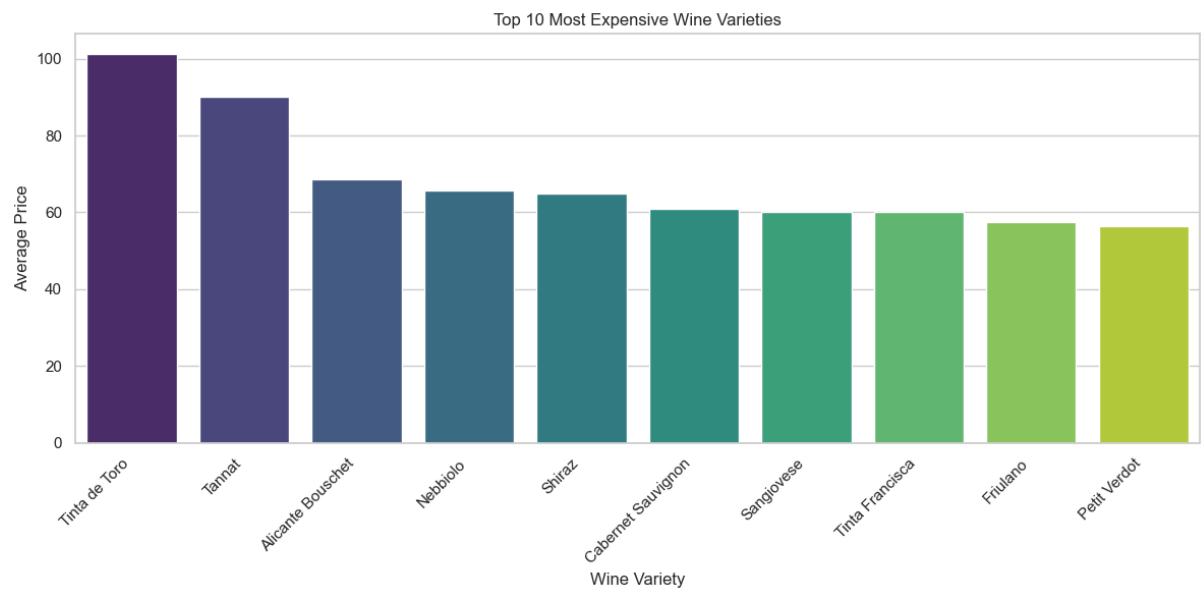
The graph shows us that the US is the favourite choice in wines it far exceeds the rest, while we can see that the next two countries after the US are more closely matched.



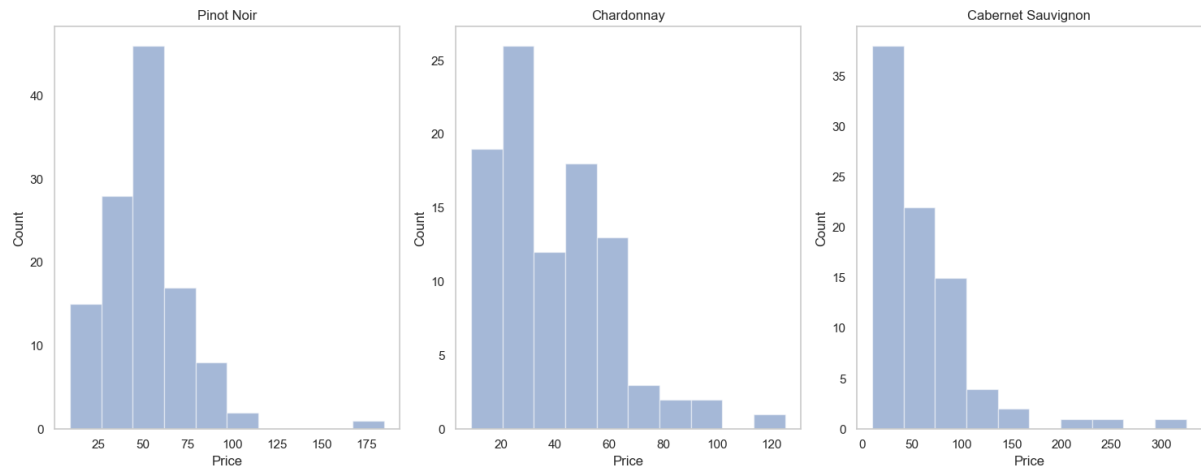
We can see that Italy produced the most expensive wines but is not deemed the best quality. The US, France, and Spain produce some of the best wine that is far cheaper than the wine from Italy. This however does not mean that Italy's wine is of poor quality it falls just below the best quality of the wine. There are a lot of other countries that also produce great wine at varying price points.



The graph shows the top 10 wine variants and we can see that while Pinot Noir is the favourite the other wines follow closely.



This graph gives insight into the most expensive wine and based on the previous graph the only wine that appears in both is Cabernet Sauvignon. This wine was in the top 3 favourite variants.



The graph above displays the relationship between price and popularity of the top 3 popular wines in the dataset. If we look at the most popular price point for each wine we can deduce the following:

- More People are willing to spend more on Pinot noir than the other two variants.
- While Cabernet Sauvignon might have the most expensive wines between the 3 variants the popular price point for this wine is the cheapest.
- Chardonnay has the widest range of popularity with regards to price point even though it can be noted that the \$30 price is the most popular.

THIS REPORT WAS WRITTEN BY : SHEENA WEBER
