Hyperiondev

TASK

Exploratory Data Analysis on the Wine Dataset

Visit our website

Introduction

The objective of this report is to conduct an in-depth exploratory data analysis (EDA) of the 'wine' dataset, which contains detailed information on 1103 wines. By employing rigorous data exploration techniques and visualization methods, our goal is to uncover significant trends and key factors influencing the characteristics and quality of wines.

Utilizing the extensive array of attributes provided in the dataset (country, points, price, province, variety, regions) we aim to reveal the underlying narratives within the world of wines. From exploring correlations between grape variety and wine quality to identifying prevalent regions and characteristics associated with high-rated wines, this report seeks to offer a comprehensive insight into the diverse landscape of winemaking.

Through our data-driven approach, we aspire to shed light on evolving consumer preferences, regional wine production dynamics, and overarching trends shaping the global wine industry.

DATA CLEANING

To clean the ‘wine’ dataset, the following methods were employed to enhance its relevance and accuracy for analysis:

1. Removing Unnecessary Columns:

Initially, unnecessary or redundant columns that were not utilized in the data analysis were identified and removed. This process ensured that only relevant information remained, making it easier to derive meaningful insights.

2. Standardization of Textual Data:

Certain columns containing textual data, including 'winery', 'province', 'variety', 'region_1', and 'region_2', exhibited inconsistencies due to the presence of special characters. To ensure uniformity and accuracy, special characters were substituted with their corresponding accented letters. A rigorous validation process, involving Google searches for each data point, was undertaken to verify the correct spelling and integrity of textual entries.

3. Feature Engineering:

To facilitate deeper exploratory analysis, the attribute 'colour' was introduced to the dataset. This categorical attribute provides insights into the colour spectrum of wines, enabling researchers to discern patterns and preferences associated with different hues. This addition enriches the dataset and offers valuable avenues for nuanced investigation into colour-specific trends within the wine dataset.

MISSING DATA

The missing data in the 'price', 'region_1' and 'region_2' columns were identified. These missing values can be classified as MCAR (Missing Completely At Random), indicating that their absence is consistent across all observations and unrelated to any observed or unobserved factors within the dataset (HyperionDev, 2021).

To address the missing data:

1. Handling missing 'region_1' and 'region_2' data:

Initially, the two columns were consolidated into a single 'regions' column. This consolidation aimed to minimize data gaps, as missing information in one column could potentially be supplemented by data from the other. However, despite this merging, some data remained missing. Deleting these observations entirely was deemed impractical, as it would compromise the dataset's analytical depth and potentially introduce bias into subsequent analyses, particularly since the missing data pertained to specific regions that would be entirely omitted from the dataset if removed. Consequently, a new category labelled 'Unknown' was created to accommodate these missing values within the 'regions' column.

2. Handling missing ‘price’ data:

To address missing prices, a strategy was employed to replace them with the mean price of wines from the same country. This approach was chosen as it serves as a statistically sound method to mitigate biases in the dataset. By substituting missing prices with the average price of wines from the corresponding country, we aim to maintain data integrity while ensuring a representative replacement value.

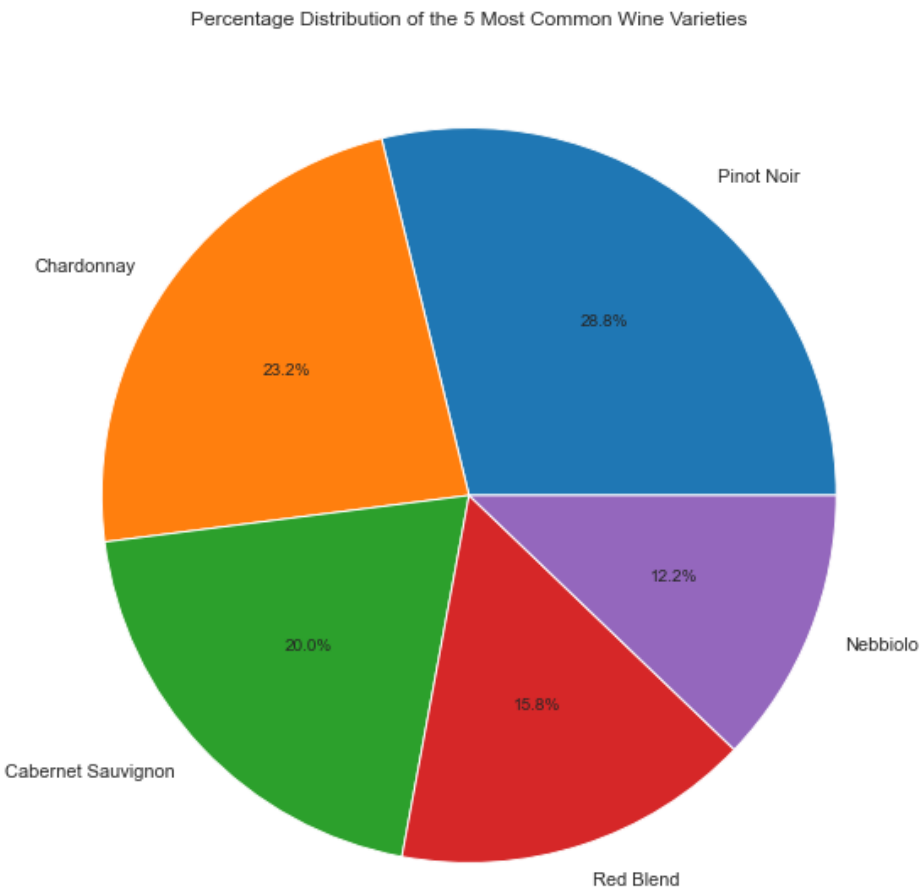
DATA STORIES AND VISUALISATIONS

Valuable insights were gained from the data by creating and analysing visual representations. Below are the various data stories along with their visualizations:

Top 5 Most Common Wine Varieties

Figure 1 illustrates the percentage distribution of the 5 most common wine varieties. Pinot Noir emerges as the most prevalent variety at 28.8%, succeeded by Chardonnay at 23.2%, Cabernet Sauvignon at 20.0%, Red Blend at 15.8%, and Nebbiolio at 12.2%. The above data distribution provides insights into the popularity and prevalence of different wine varieties within the dataset. It offers valuable information about consumer preferences and market trends, indicating which wine varieties are most commonly consumed or produced. Understanding the distribution of these varieties can inform decisions related to product offerings, marketing strategies, and production planning within the wine industry.

Figure 1



Points Ratings for the top 5 most Common Wine Varieties

In **Figure 2,** the wine variety with the most normally distributed points ratings is Pinot Noir, followed by Cabernet Sauvignon, Chardonnay, Red Blend, and Nebbiolo. This finding contrasts with the percentage distribution, which was unexpected.

The insights from this observation suggest that while certain wine varieties may be more prevalent in terms of overall percentage distribution, the distribution of points ratings across wines of different varieties can vary significantly. Specifically:

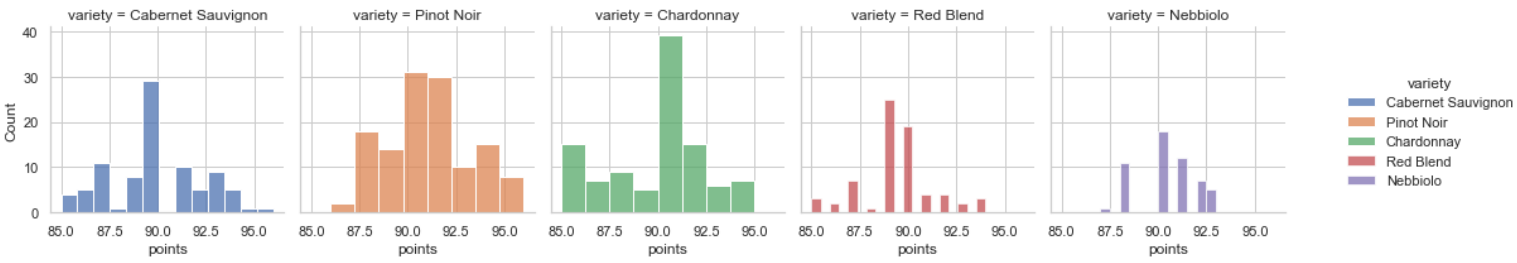
Pinot Noir appears to have a more normally distributed range of points ratings compared to other varieties. This could indicate a more consistent quality perception among Pinot Noir wines.

Cabernet Sauvignon, Chardonnay, Red Blend, and Nebbiolo exhibit deviations from normal distribution in their points ratings. This implies that there may be more variability in quality perceptions among wines within these varieties.

The discrepancy between percentage distribution and the distribution of points ratings underscores the importance of considering both factors when evaluating the popularity and quality of wine varieties. It suggests that while certain varieties may be widely consumed, the perception of quality among consumers may not always align with their prevalence.

Overall, these insights highlight the complexity of consumer preferences and perceptions within the wine industry, indicating that factors beyond mere popularity play a role in shaping how wines are perceived and appreciated.

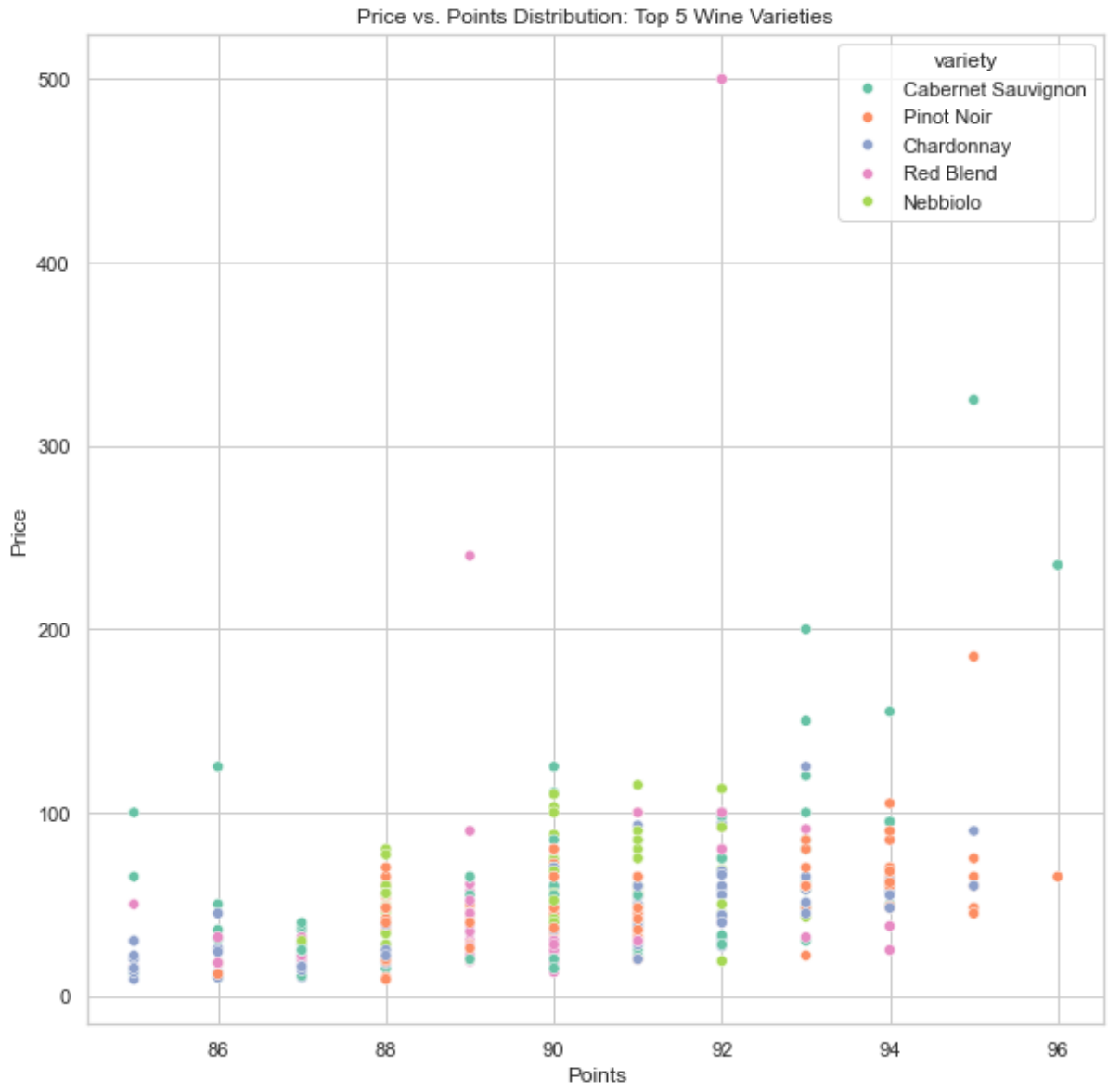
Figure 2: Histogram Grid Plot: Distribution of Points Ratings for the top 5 most common wine varieties



Price and Points Distribution for the 5 Most Common Wine Varieties

In **Figure 3**, a positive correlation is observed between wine price and point rating across the top 5 wine varieties. As the point rating increases, so does the price. This suggests that higher-priced wines tend to receive higher point ratings, reflecting a perception of higher quality among consumers. Conversely, lower-priced wines typically receive lower point ratings. This correlation underscores the influence of price on perceived quality in the wine market and highlights the importance of price as a factor in consumer decision-making.

Figure 3



Distribution of Wines by Country

Figure 4 illustrates the distribution of wines by country in descending order, beginning with the country with the highest frequency of wines (the US) at the top and descending to the country with the lowest frequency of wines (Slovenia) at the bottom. Overall, these insights provide a snapshot of the distribution of wines by country, offering valuable information about the geographical diversity of wines in the dataset.

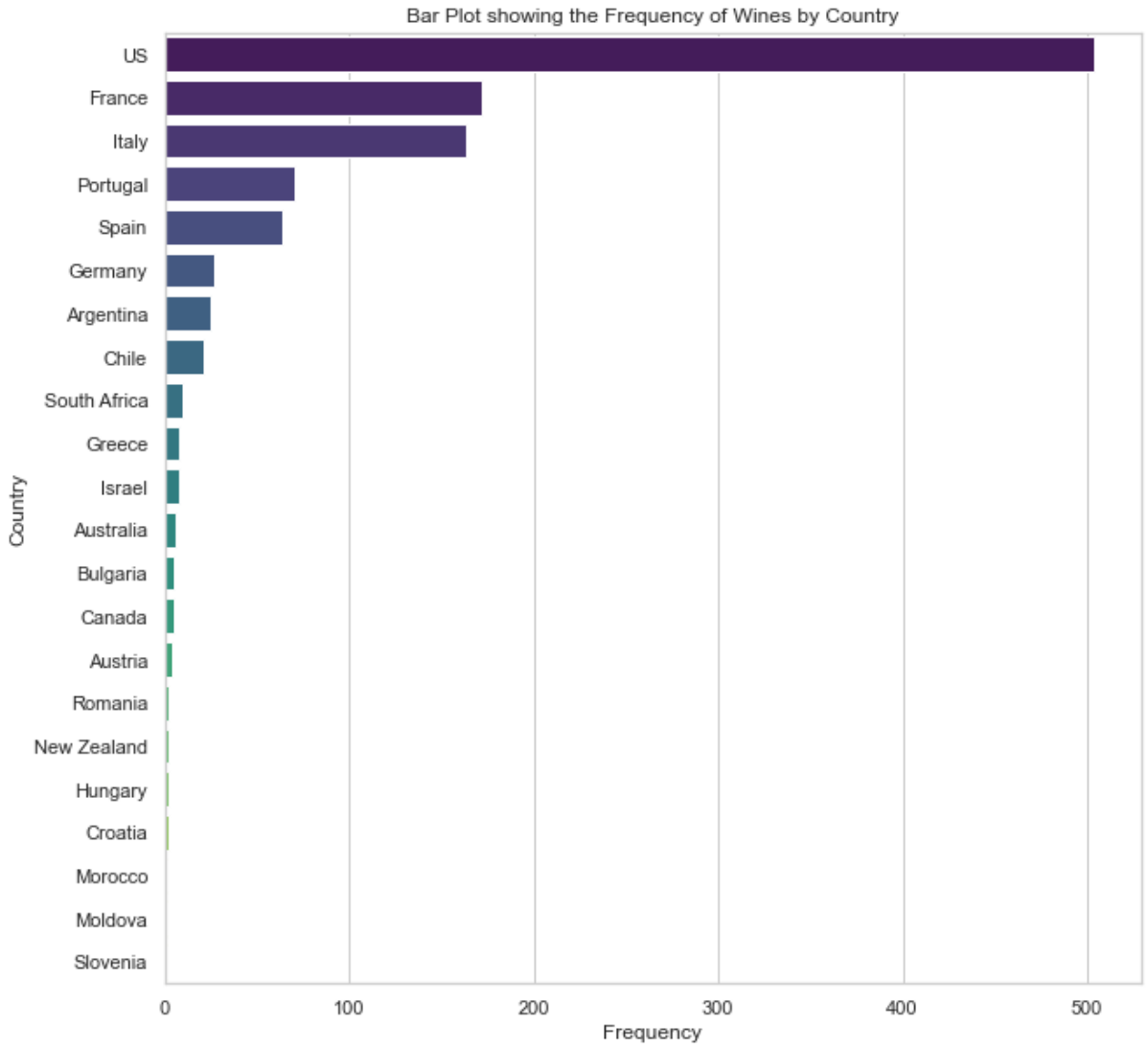
Higher frequencies of wines from certain countries, such as the United States, could reflect strong wine production industries, consumer demand, and established wine cultures within those regions.

Conversely, lower frequencies of wines from other countries, like Slovenia, may indicate smaller or emerging wine industries, limited consumer demand, or economic factors influencing production and distribution.

Additionally, variations in the frequency of wines from different countries could be influenced by factors such as historical winemaking traditions, government policies, market dynamics, and international trade agreements, all of which contribute to the unique wine cultures and landscapes across various regions.

The distribution of wines by country not only provides insights into the composition of the dataset but also offers a glimpse into the broader cultural, economic, and social contexts shaping the global wine industry.

Figure 4

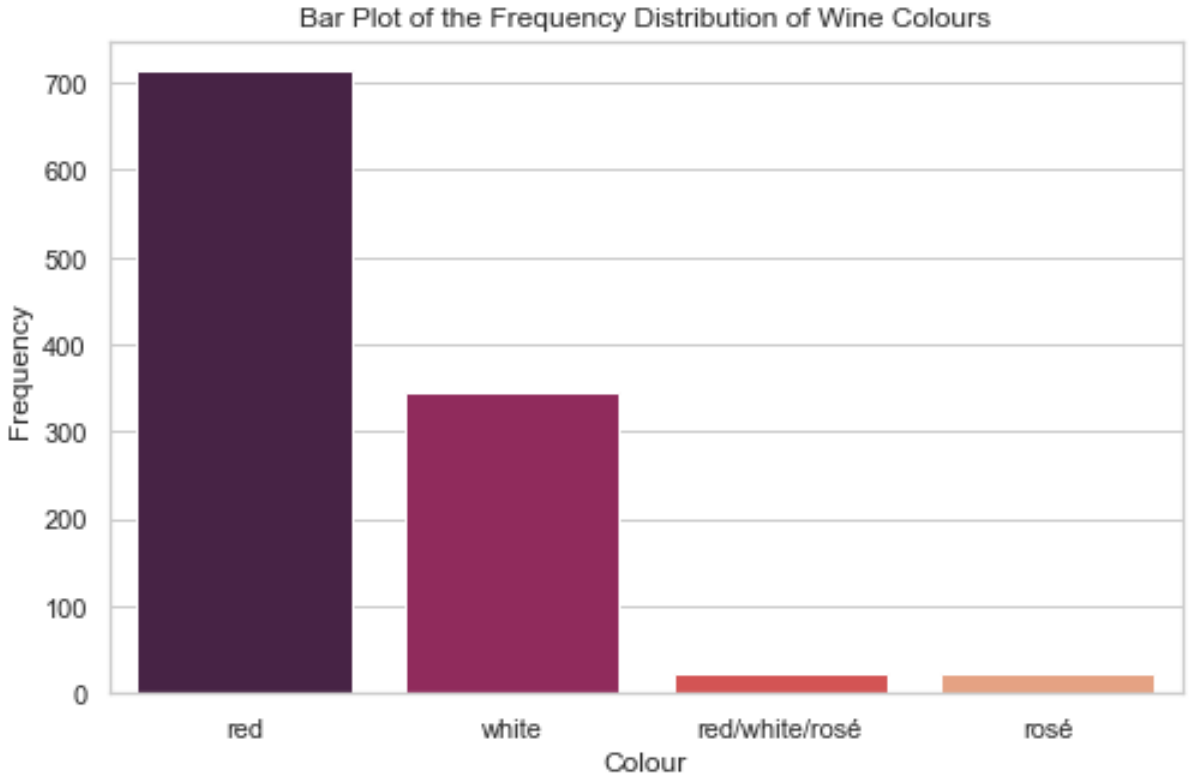


Frequency Distribution of Wine Colours

In Figure 5, it is evident that red wine has the highest frequency, followed by white wine, and then red/white/rosé, with rosé wines having the lowest frequency. The highest frequency of wines in the dataset is attributed to red wine. This suggests that red wine is the most commonly encountered type among the observations , and that white wine emerges as the second most prevalent type. This indicates that white wine is also popular and frequently represented in the dataset, albeit to a slightly lesser extent than red wine. While the wines that fall into the re/white/rosé and rosé categories are still present in the dataset, they are less commonly encountered than red and white wines. Overall, these insights offer valuable information about the distribution and prevalence of different wine

colours within the dataset, providing context for understanding consumer preferences and market trends in the wine industry.

Figure 5



Red Wine Price Variations Across Countries

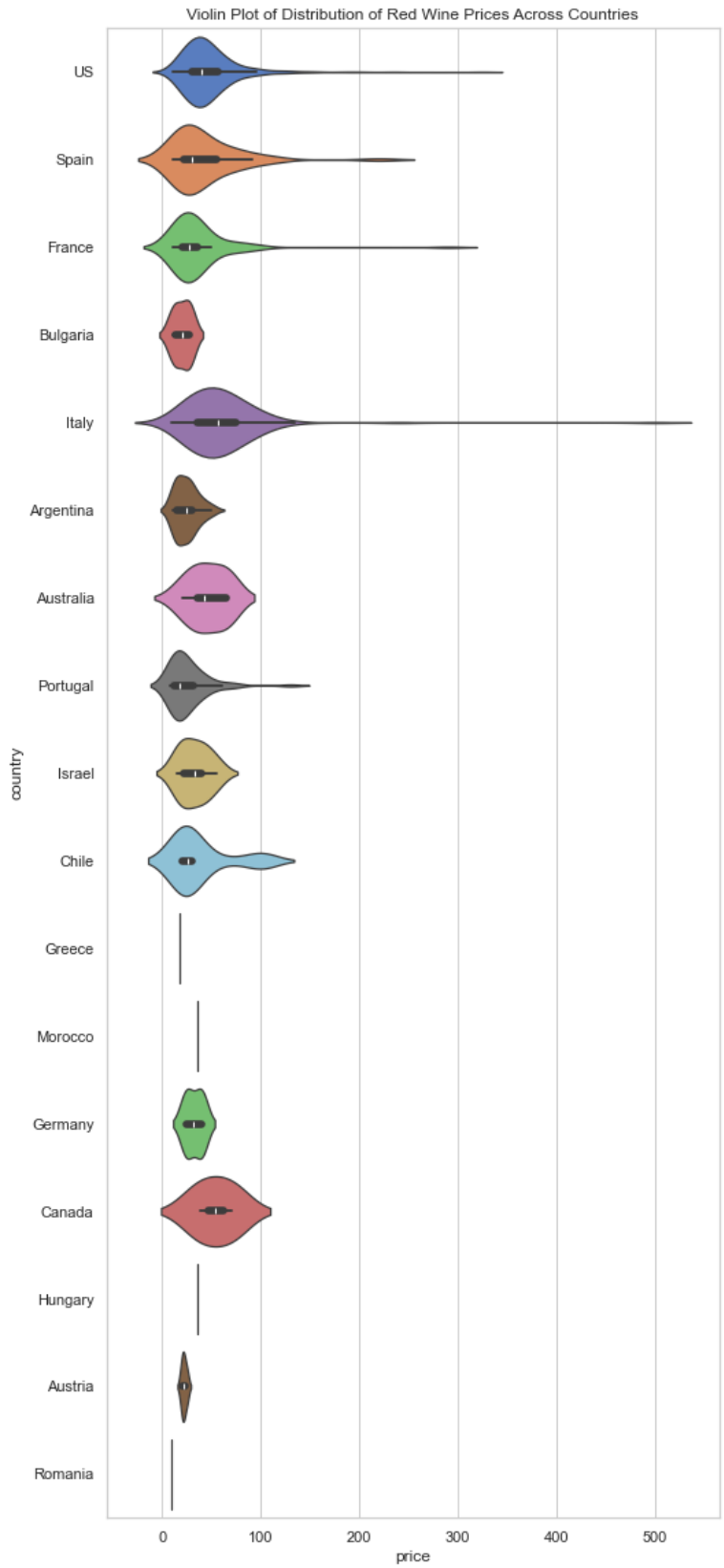
Figure 6 illustrates the distribution of red wine prices across countries. It is evident that prices in Canada, France, Germany, Austria, Israel, and Bulgaria exhibit a relatively normal distribution. Conversely, prices in the US, Spain, Australia, and Portugal are skewed to the right, while in other countries, the distribution is skewed to the left.

The observation of normally distributed prices in countries such as Canada, France, Germany, Austria, Israel, and Bulgaria suggests a balanced market where prices are distributed evenly around the central tendency.

In contrast, countries like the US, Spain, Australia, and Portugal display a right-skewed distribution of red wine prices. This indicates that a larger proportion of wines in these countries are priced higher than the average, with fewer wines falling in the lower price range.

Conversely, the distribution of red wine prices in other countries is left-skewed, implying that a majority of wines in these regions are priced lower than the average, with fewer wines being priced higher.

Figure 6



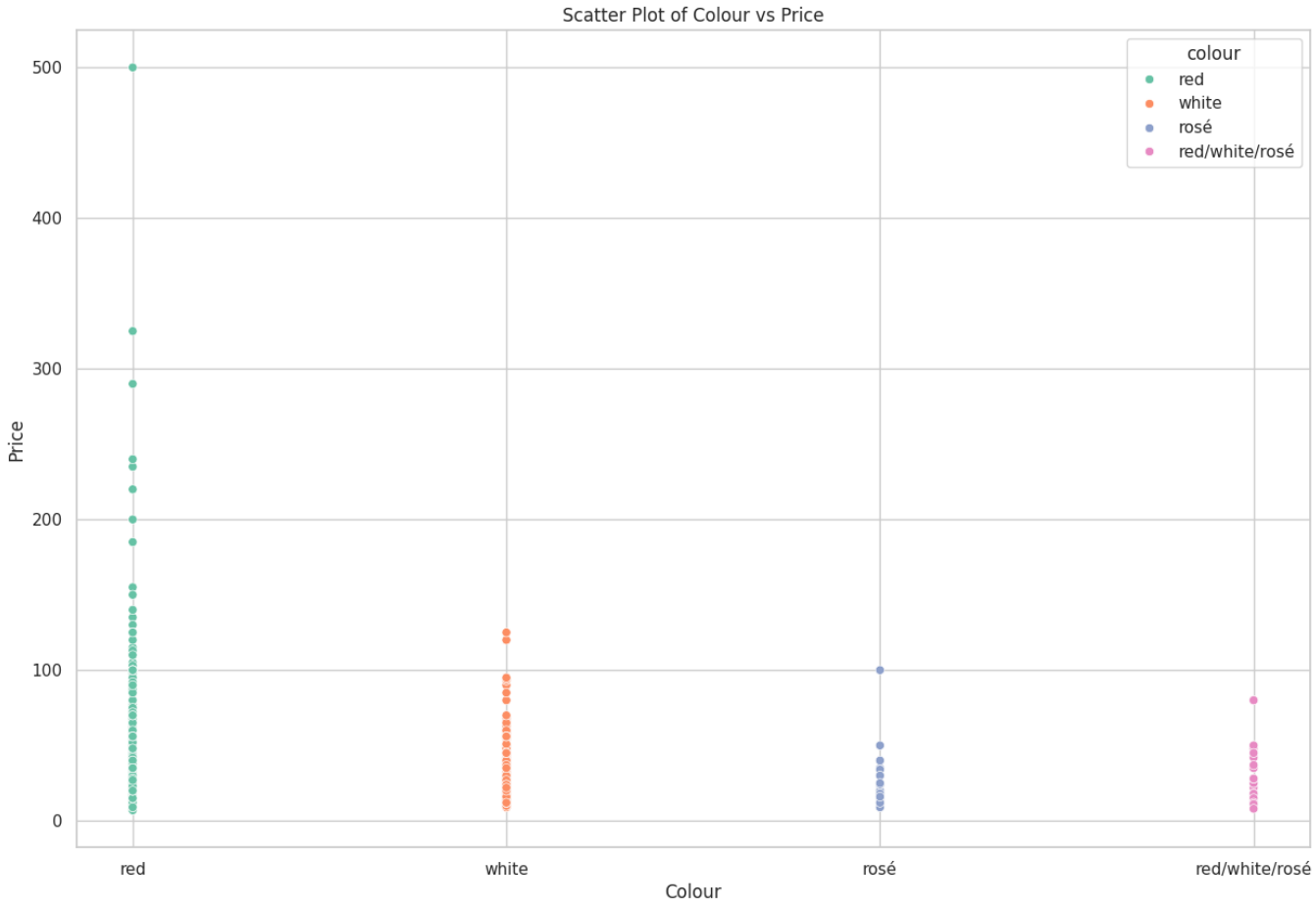
The Relationship Between Wine Colour and Price

In Figure 7, the relationship between wine colour and price is depicted, revealing a consistent trend where red wines are generally the most expensive, followed by white wines, rosé, and finally wines that can be categorized as red/white/rosé blends.

This suggests that there may be factors unique to red wines, such as production costs, aging processes, or market demand, contributing to their premium pricing. This hierarchy in pricing aligns with common perceptions of the market, where red wines are often associated with higher quality and higher price points compared to white and rosé wines (Sommelier Business, 2024).

Overall, these insights shed light on the relationship between wine colour and price, providing valuable information for producers, distributors, and consumers within the wine industry.

Figure 7



References

HyperionDev. (2021). Data Analysis - Preprocessing. [Educational notes]. Retrieved from Dropbox-NK23110009394.

Sommelier Business. (2024). Wine Pricing Strategy: Profitability and Adjustments. Retrieved from https://sommelierbusiness.com/en/articles/menu-intel-1/wine-pricing-strategy-profitability-and-adjustments-14.htm

THIS REPORT WAS WRITTEN BY : Nagitta Kasirye-Koikanyang