Exploratory Data Analysis on the Wine Dataset

Introduction

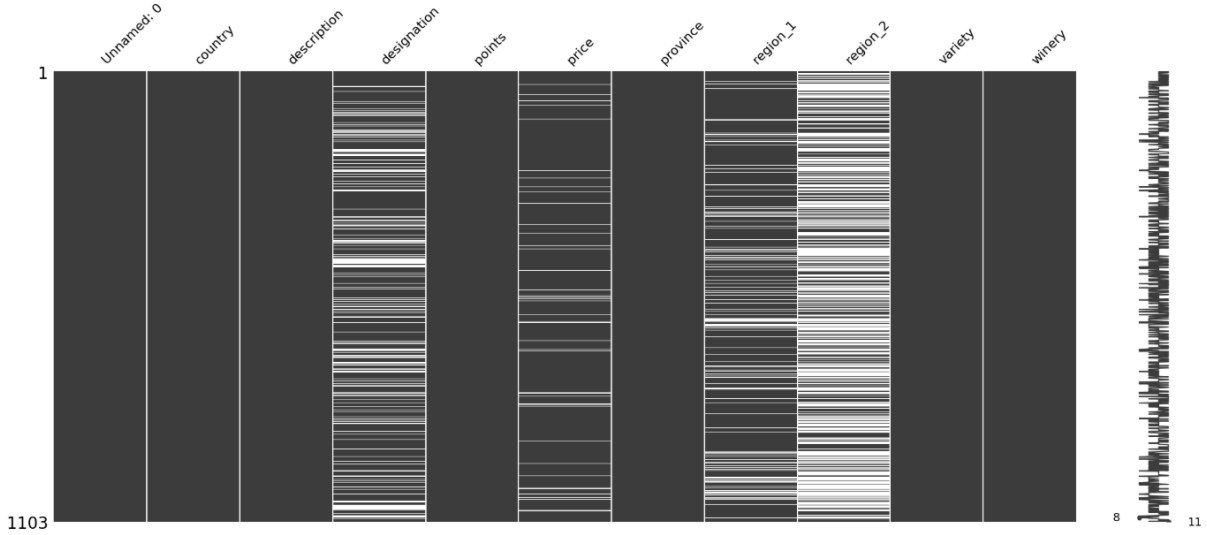
The wine dataset contains information on a range of wines. Each wine in the dataset includes a unique identifier (index), a description of the wine, its variety, price in dollars and a quality rating measured on a 100-point scale. The dataset also captures geographical information about where each wine was produced. This includes the country, province, region 1 and region 2. Additionally, the designation field holds a specific name or label assigned to the wine (such as a vineyard or special reserve), while the winery column identifies the wine producer.

Data Cleaning

A preliminary inspection of the dataset was conducted using the head() function to view the first five records and gain an initial understanding of the data structure. A duplicate index column was identified and removed to avoid redundancy. Encoding issues like special characters were corrected and trailing whitespace was stripped from string-based fields. The unique values of country, province, region 1, region 2 and variety were analysed to ensure that the data was free of encoding errors and duplicate categories. These cleaning steps improved the quality and readability of the dataset, forming the basis for visual analysis.

Missing Data

A count of missing data found missing data in the designation, price, region 1 and region 2 fields. A matrix of this missing data was created using missingno. All missing values were replaced with the string ‘Unknown’ for consistency and to preserve the row count.



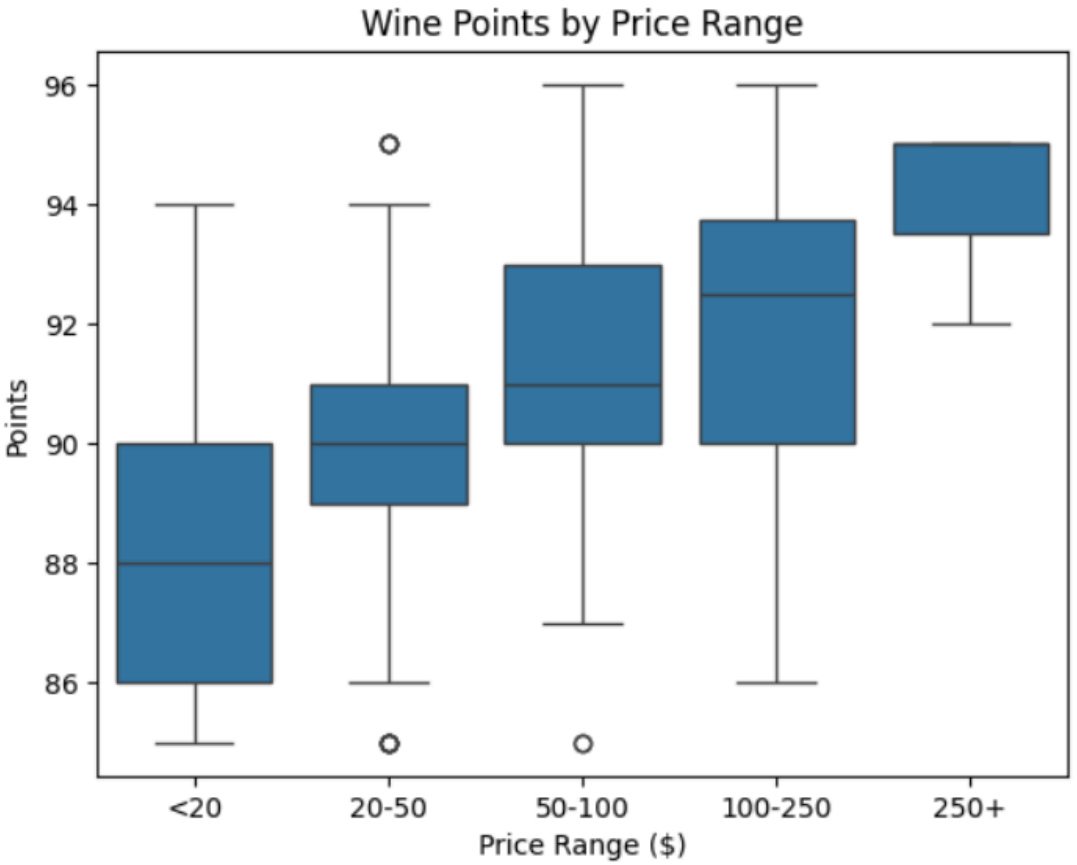
Data Stories and Visualisations

Wine Scores and Price

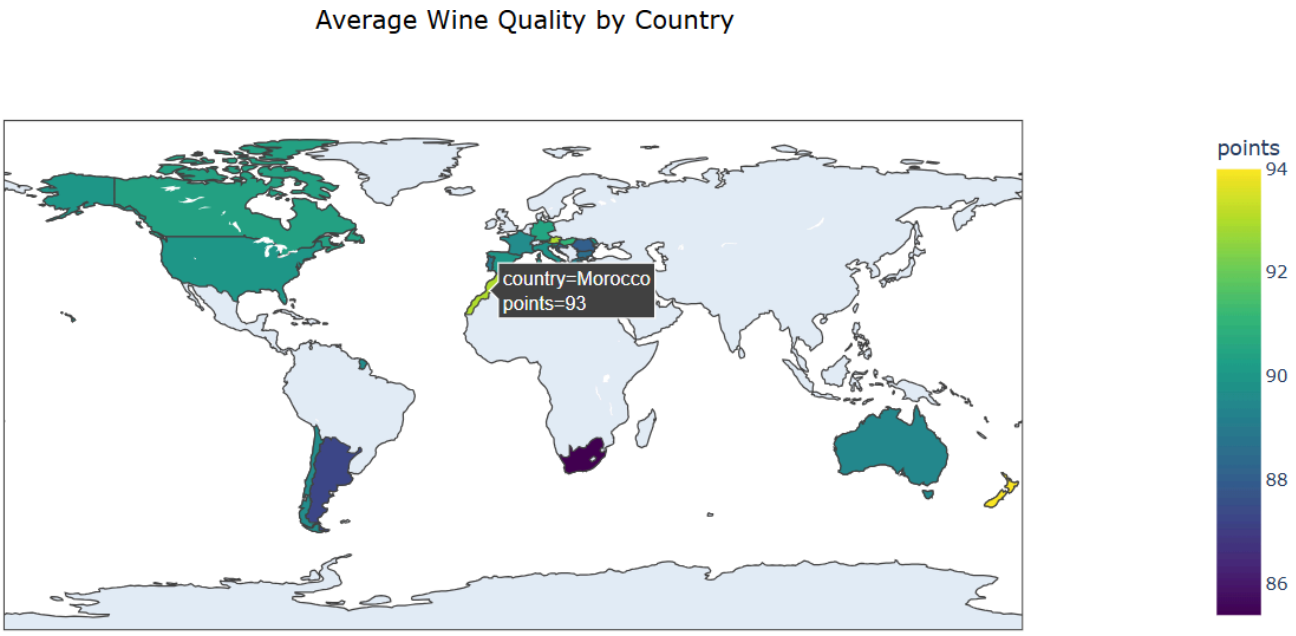
The table below provides information on a few of the top wines. This information includes the variety, winery, country, province, points and price. The top 4 wines all have a score of 96 and the rest have a score of 95. However, the table doesn't show that there are 21 wines with the score of 95, which means that this table provides a limited idea on the top scoring wines. The table does show that there is a wide range of prices and indicates that Spain, United States (US) and France are countries where there are high scoring wines.

Variety	Winery	Country	Province	Points	Price
Tinta De Toro	Bodega Carmen Rodríguez	Spain	Northern Spain	96	110
Pinot Noir	Ponzi	US	Oregon	96	65
Sauvignon Blanc	Macauley	US	California	96	90
Cabernet Sauvignon	Heitz	US	California	96	235
Chardonnay	Bergström	US	Oregon	95	90
Pinot Noir	Patricia Green Cellars	US	Oregon	95	48
Tannat	Vignobles Brumont	France	Southwest France	95	90
Chardonnay	Center of Effort	US	California	95	60
Tinta De Toro	Numanthia	Spain	Northern Spain	95	220
Tempranillo Blend	Muga	Spain	Northern Spain	95	79

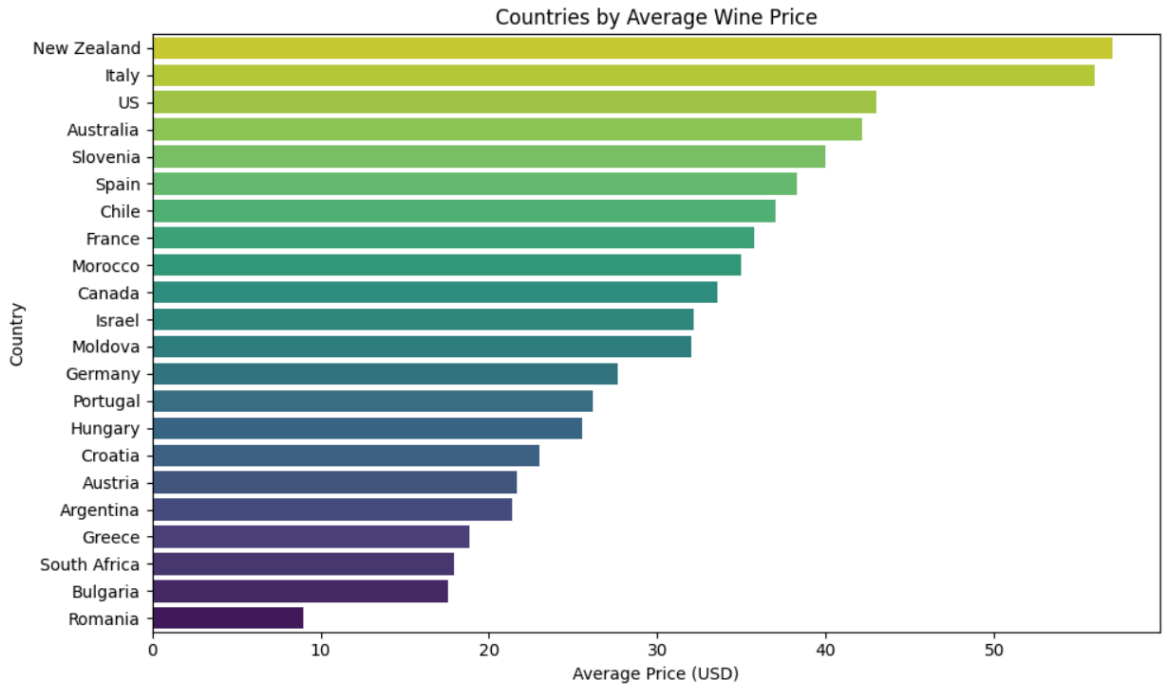
The boxplot below shows the relationship between wine points and price. It shows that the price ranges **under 20**, **50-100** and **100-250** have more variability in wine points. This indicates that high scoring wines can be found in cheaper price ranges. However, the general trend shows that the average wine score (per price range) increases with price. The **20-50** and **50-100** range have outliers. Potential reasons for outliers and variability in the relationship between wine scores and prices could be related to the cost of making the wines and scoring being subjective. This is the first visualisation that shows that the wine points range is between 85 and 96, with the average wine score being 90. This indicates that the majority of the wines in the dataset are relatively high scoring.



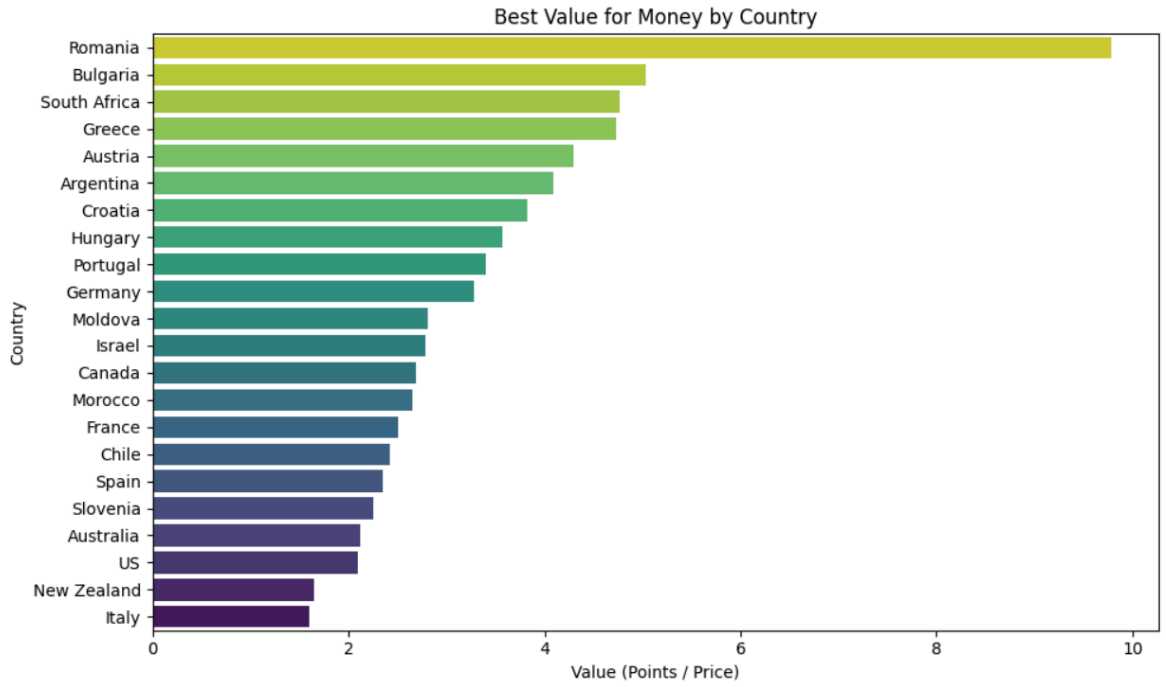
The interactive map below shows the average wine points per country. It shows New Zealand has the highest average points (94) followed by Austria and Morocco who both have an average score of 93. South Africa has the lowest average score of 85.4. There are no extreme differences in countries' average points. The key shows that the average points are all between 85 and 94.



The horizontal bar graph below shows the average wine price per country. Wines from New Zealand and Italy are significantly more expensive than wines from other countries. Whereas Romanian wines are significantly cheaper than wines from other countries. Most of the countries in the dataset have an average wine price between $20 and $40.

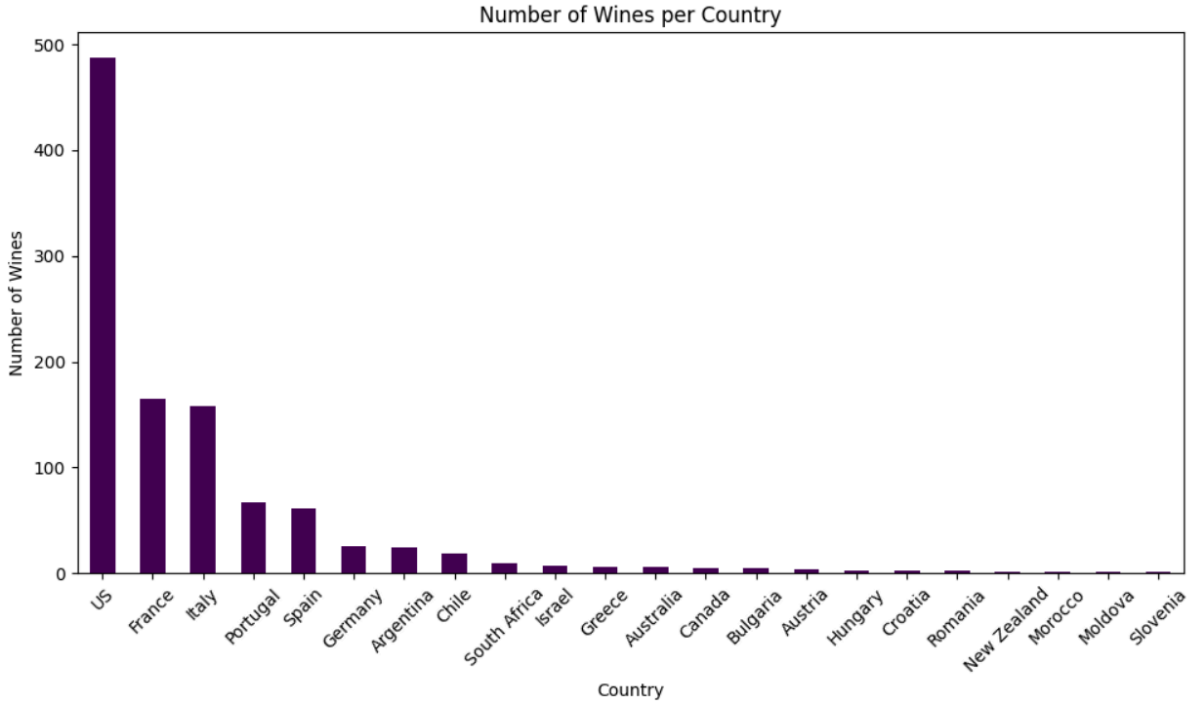


The horizontal bar graph below shows the average value for money (points / price) in each country. Romania has a significantly higher value for money compared to other countries. Romania doesn't have the highest average points but they have significantly cheaper wine prices. Whereas, New Zealand and Italy have the lowest value for money. Both countries had the highest average wine prices but New Zealand had the highest average score whereas Italy had a more average wine score. This indicates that the small difference in points is not necessarily worth the extra price from a cost perspective. However, expensive wines likely have a reason for their cost and many consumers may purchase wines regardless of their cost.

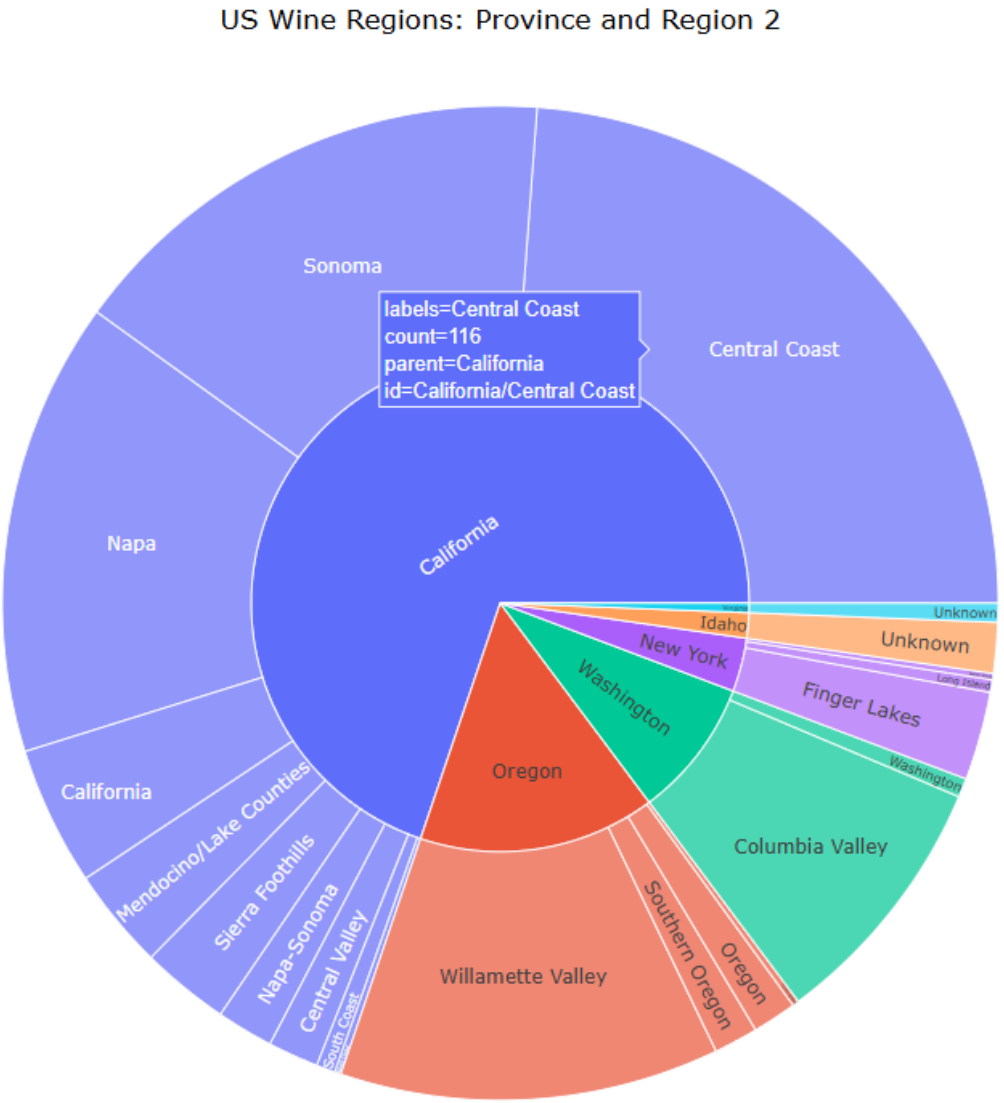


Geography of Dataset

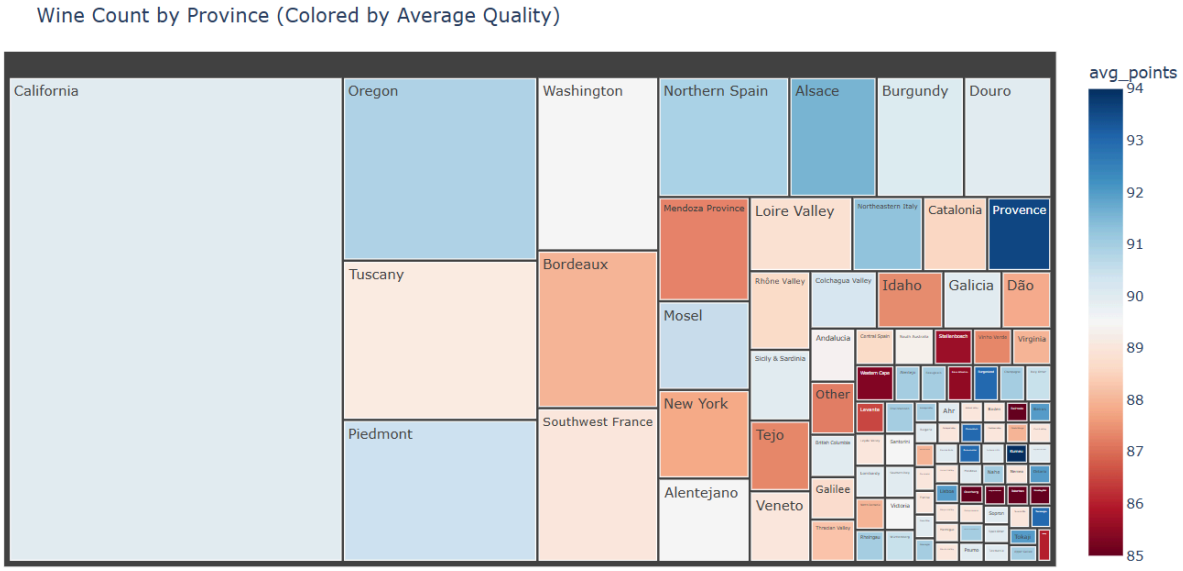
The bar chart below shows how many wines in the dataset belong to each country. The majority of the wines in the dataset are from the United States (US), followed by France and Italy. Previously we spoke about New Zealand, Austria and Morocco having the highest average wine score, but since they don't have a lot of wines in this dataset it's not necessarily an accurate assumption on all wines from those countries.



The multi-layer pie chart below explores the geography of the data within the US because it is the country with the most wines in the dataset. It shows the province in the centre and region 2 on the outer side. Region 2 is unique to the US and is a more general category compared to region 1 (which can be very specific). California has the most wines (340), followed by Oregon (75) and then Washington (44). The top three areas in California with the most wines are Central Coast (116), Sonoman (79) and Napa (72).

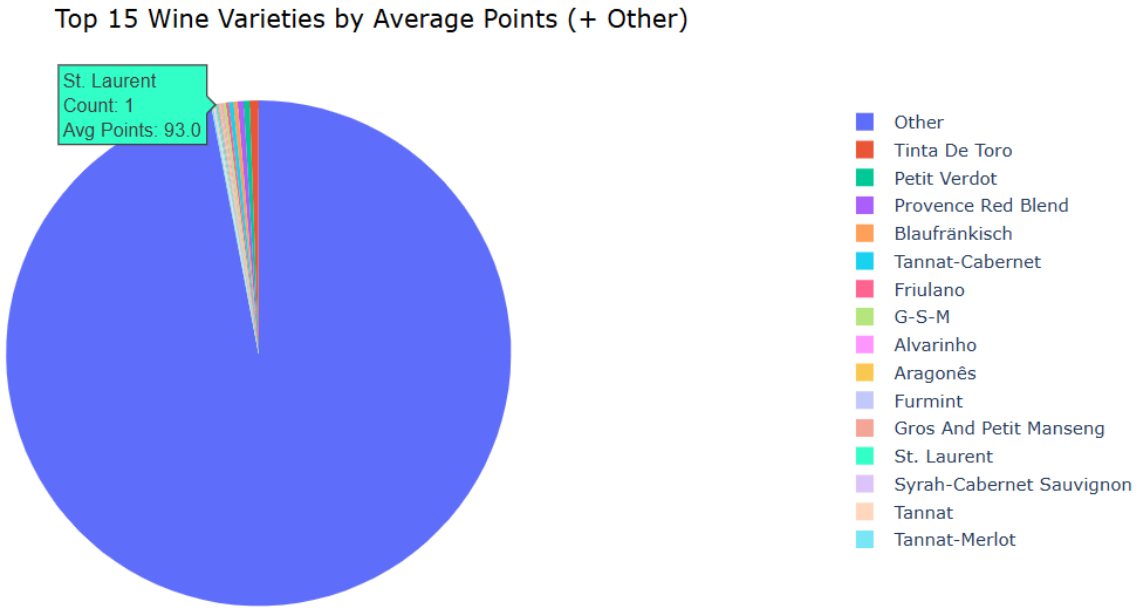


The treemap below looks at all provinces based on how many wines each province has but uses colour to indicate the average quality of the wines in that province. Dark blue indicates a high average score and dark red indicates a lower average score. The province with the highest average rating (94) is Kumeu which is in New Zealand but it only has 1 wine. The second highest province is Provence (93.6) which is in France and it has 10 wines. There are four provinces which have 93 as an average rating but they only have 1 or 2 wines each. Out of the top 3 US provinces the highest average is in Oregon with a 90.88 average.

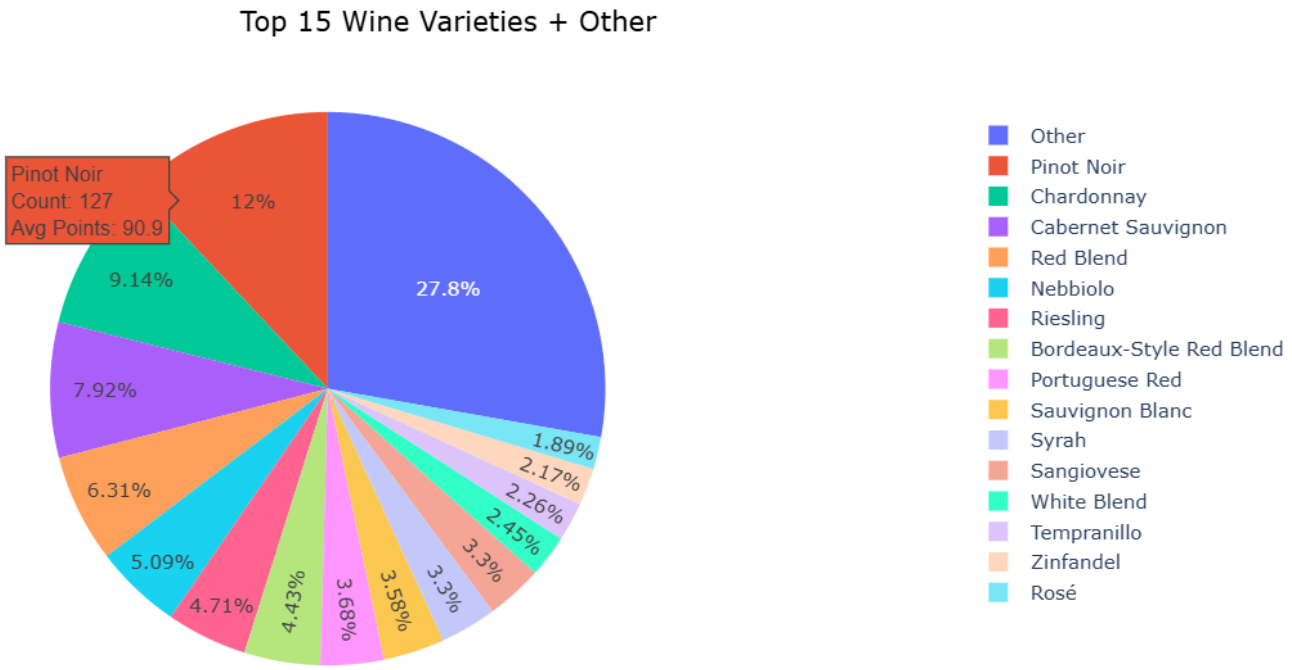


Wine Varieties

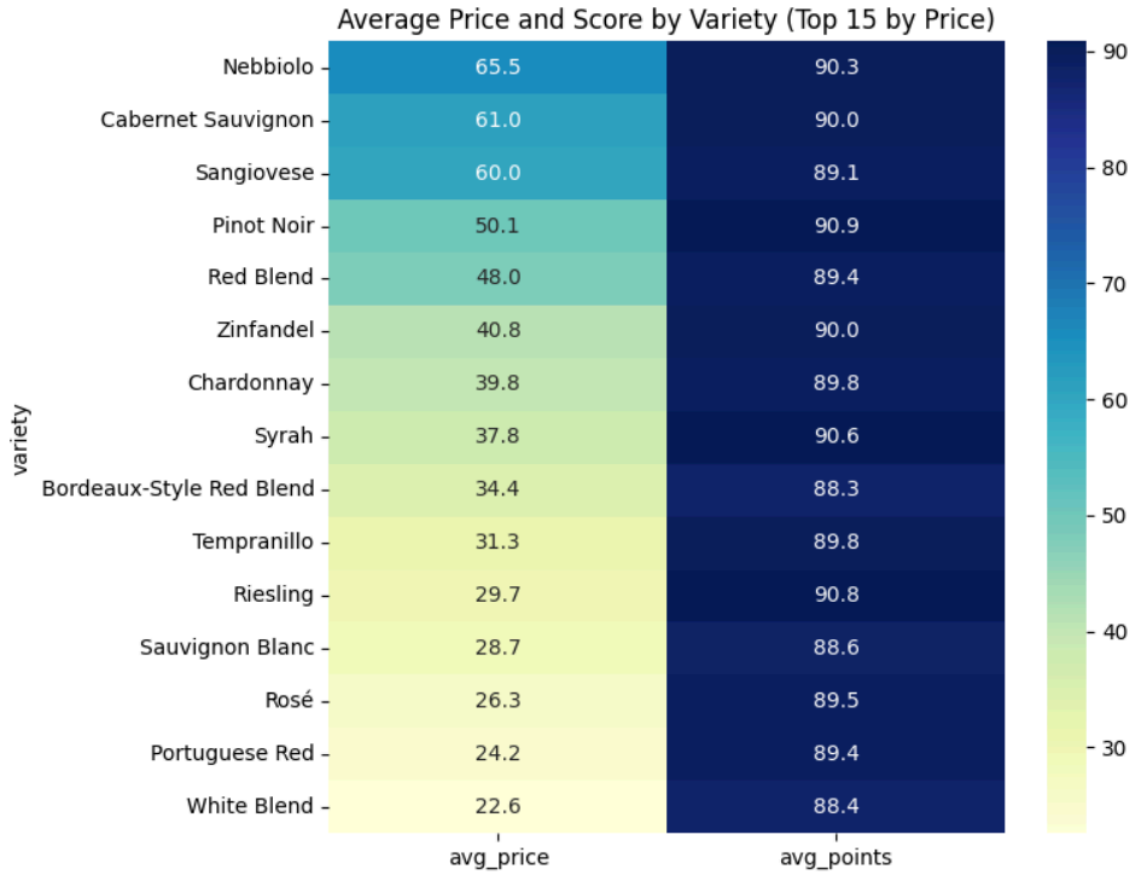
The pie chart below shows the top 15 wine varieties based on the average score. These varieties have between 1 and 6 wines each, with an average score between 91 and 94. This potentially indicates that more niche varieties likely have higher average scores. This could be because only the best wines were selected for that variety because they are less popular and the standard is higher for that variety. However, this pie chart doesn't give us a general understanding of the overall wine varieties in the dataset.



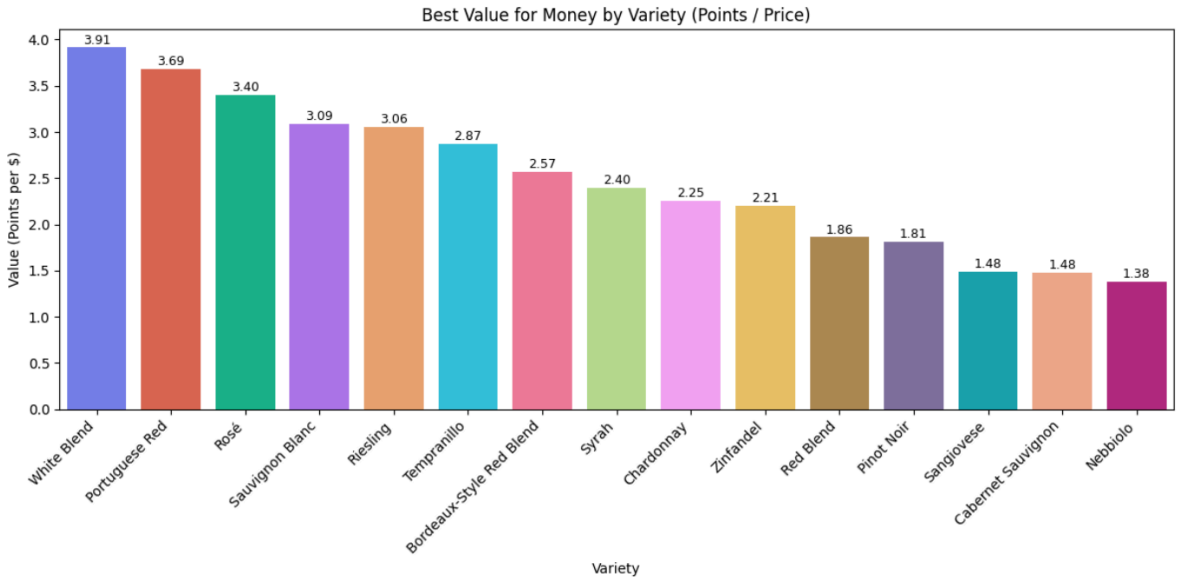
There are 114 unique wine varieties in the dataset. For better readability, the pie chart shows the top 15 wine varieties and groups the remaining varieties into the category ‘other’. The top 3 wine varieties based on count are Pinot Noir (12%), Chardonnay (9.14%) and Cabernet Sauvignon (7.92%). The interactive pie chart shows the count for each wine variety and average points. The average score for the top 15 wine varieties ranges between 88 and 91. Whereas the count varies from 127 to 20.



The heatmap below shows the average price and points for the top 15 wine varieties which have 20 or more wines. The most expensive wine varieties are Nebbiolo ($65.5), Cabernet Sauvignon ($61) and Sangiovese ($60). With most of the wine varieties average price falling between $20 and $40.



The bar graph below shows the value for money score for each of the top 15 wine varieties which have 20 or more wines. This shows the wine varieties in the reverse order to the previous heatmap. This means that the cheaper the wine variety, the better the value for money. This strengthens the idea that since all the wines have relatively high scores, a huge price increase doesn't make sense for just a few points especially for the average consumer.



Descriptive words

The description field holds text that describes the wine. This word cloud shows the popular words used to describe wines. The most prominent words used are flavor, wine and aroma. The word cloud includes specific flavours (like licorice and cherry) as well as general descriptive words (like rich and vintage).

This report was written by: Isabella Gloor