

**TASK**

**Exploratory Data Analysis on the Wine Dataset**

[](http://www.hyperiondev.com/portal/)

**Introduction**

The Wine dataset contains information about wines, encompassing various attributes related to wine production, characteristics, and reviews. It includes details such as the country of origin, description, points awarded, price, province, region, wine variety, and winery.

The analysis provided insights into the distribution of wine prices and points, relationships between variables, and variations across countries and wine varieties. Factors influencing wine prices and points were explored, and common themes in wine descriptions were identified. The data cleaning process ensured a cleaner dataset for meaningful analysis. The visualizations and summaries offer a comprehensive understanding of the dataset, allowing for further exploration and insights into the world of wines.

**DATA CLEANING**

* When loading the data, the encoding of the file was detected and the file was read using this detected encoding in order to try and avoid errors when reading special letters and symbols.
* The following information was found and viewed in order to get a picture of the data:
  + The first five rows of the dataset.
  + The shape of the data frame.
  + The column information containing column name, non-null count and data type.
  + A description of the numerical columns including the count, mean, standard deviation, minimum and maximum values and the quartile values.
  + The number of unique values in each column.
  + The amount of missing data in each column.

**MISSING DATA**

* After viewing the data, the following was carried out:
  + The first column ‘Unnamed : 0’ was removed as it was an extra index column. Additionally the 'region\_1', 'region\_2' and 'designation' columns were removed due to large amounts of missing data.
  + Duplicate rows were removed.
  + The rows containing missing price values were removed as they amounted to around 5% of the data.

**DATA STORIES AND VISUALISATIONS**

Visualisations were made on the price and points of all the entries. Studies were then carried out on the different countries and the different varieties of wine, looking specifically at the count, average price and average points of counties and .wine varieties. Following this the top wine varieties in the top producing countries were studied. Lastly a Word cloud was done on the descriptions of all the wines.

* Price and Points:

In order to visualise the distribution of price and points amongst the wine data, a histogram and count plot of the price and points respectively were made.

A graph and chart with numbers

Description automatically generated with medium confidence

Most of the wine prices lie below $100 with a few outliers ranging up to $500.

As for the points: The wines in this data set are all fairly highly rated wines, with the most common points value given to be 90 points, with slightly more wines having points below 90 than above.

A pair plot was made to view the relationship between the price and points of wine. The price was plotted in the log scale since most of the data lies in the lower price range, the few outliers in the higher price range would make it difficult to see any patterns otherwise since the majority of the wine would be condensed to a small area in the lower end of the price range. This log scale allowed us to see a much clearer pattern between the price and the points, where there is a weak yet clear positive correlation between the two values.

A graph of a graph

Description automatically generated with medium confidence

* Countries:

Since the price data is skewed to the left with outliers in the higher price range, the average price for each country was found using the median. The average points was normalized to view discrepancies in the results better.

From the count and percentage of wines in each country, it can be seen that almost half the wines listed in the dataset are from the US. Following the US, Italy and France respectively have the next highest counts of wines. We see that there are many countries with very low wine counts, likely with only one or two entries. As is found to be the case when viewing a list of the actual counts per country.

A graph of different countries/regions

Description automatically generated

Two bar graph of the average price and average points for each country are presented below.

While it is interesting to see that New Zealand has the highest average price and points, it must be noted that there is only one entry for this country and so this may not be a good representative.

Of the top producing countries (those with the highest wine count) it is seen that Italy has the highest average price, followed by the US. These two countries rank right in the middle, however, for their average points.

On the other hand, Romania is the one of the top producing countries (7th highest) and has the second lowest average price of all the countries and 4th lowest average points. South Africa too is among the top producing countries (9th) and has one of the lowest average prices (4th lowest of all countries) as well as the lowest average points of all the countries.

There are a few aspects that may affect the price of wine in each country. It is seen from the price-points study above that there is a positive correlation between the points and the price, and we can see from the below two plots that the countries for the most part hold the same general position for both average price and average points. However, the prices may also be influenced by the country's economy or the quantity of wine being produced, amongst other factors. These factors could possibly be what effects the prices of countries like Australia where the average price is high, but the average points is low. Or countries like Austria, where the average price is rather low but the average points is very high.

A graph of different colored bars

Description automatically generated

A graph of average points

Description automatically generated

The results of the count, average price and average points have all been normalized and plotted together in the following bar graphs. These plots further verify the previous claims. It is also interesting to see that the countries that have a higher wine count appear to have slightly lower average prices and average points for the most part when compared to the countries producing fewer wines. However, there is insufficient data to be confident in this result and one would need far more entries to verify this.

A graph of different colored bars

Description automatically generated

A graph of different colored lines

Description automatically generated

* Wine Varieties:

Below is a table containing the wine varieties which have the highest and lowest average points and average prices.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Top 5 Varieties by Price | | | |  | Top 5 Varieties by Points | | | |
|  | Variety | Average Price ($) | Average Points |  | Variety | Average Price ($) | Average Points |
| 1 | Tinta de Toro | 91.5 | 94.17 | 1 | Tannat | 90.0 | 95.0 |
| 2 | Tannat | 90.0 | 95.0 | 2 | Friulano | 57.5 | 94.5 |
| 3 | Alicante Bouschet | 68.5 | 88.0 | 3 | Tinta de Toro | 91.5 | 94.17 |
| 4 | Nebbiolo | 68.0 | 90.36 | 4 | Provence red blend | 44.0 | 94.0 |
| 5 | Shiraz | 65.0 | 90.0 | 5 | Tannat-Cabernet | 40.0 | 93.67 |
|  | | | |  | | | |
| Bottom 5 Varieties by Price | | | | Bottom 5 Varieties by Points | | | |
|  | Variety | Average Price ($) | Average Points |  | Variety | Average Price ($) | Average Points |
| 1 | Chenin Blanc | 10.0 | 85.0 | 1 | Chenin Blanc | 10.0 | 85.0 |
| 2 | Ugni Blanc-Colombard | 10.0 | 86.0 | 2 | Pinot Noir-Gamay | 18.0 | 85.0 |
| 3 | Portuguese Sparkling | 11.5 | 85.0 | 3 | Portuguese Sparkling | 11.5 | 85.0 |
| 4 | Grenache-Syrah | 13.0 | 89.0 | 4 | Aglianico | 34.0 | 86.0 |
| 5 | Aragones | 14.0 | 91.0 | 5 | Glera | 22.0 | 86.0 |

A scatter Plot was also made of the average prices and average points of all wine varieties.

A graph with colored dots

Description automatically generated

It is evident that, of the top wines, Tinta de Toro and Tannat are the highest ranking in both price and points. Of the low ranking wines, Chenin Blanc holds the bottom spot in both points and price. The Portuguese Sparkling is also low in both price and points aspects.

Next, the same analysis is applied that was done for the countries analysis limited to the top 20 most occurring wine varieties in the dataset.

A comparison of different varieties of wine

Description automatically generated

We see that the variety with the highest count is that of Pinot Noir (making up ~12% of the dataset), followed by Chardonnay (~9.5%) and Cabernet Sauvignon (~8.5%) respectively. There are too many varieties the view them all, however, we can see from the above plot which are the most popular. It is interesting to note that of the top and bottom wines in average points and price, the only one featured in the top counts is the Nebbiolo. Otherwise it appears that the highest and lowest ranked wine varieties by price and points are not commonly listed.

A graph of different colored bars

Description automatically generated

A graph of different colors

Description automatically generated

When comparing the top 20 average price wine varieties and the top 20 average points wine varieties, we can notice that only six of the wine varieties appear in both top 20 lists. Those are Tannat, Tinta de Toro, Friulano, G-S-M, Provence Red Blend, and Tannat-Cabernet.

A graph of different colors

Description automatically generated

A graph of different colored bars

Description automatically generated

It is interesting to see that the average points appear to be higher for the 11th-20th most frequent wine varieties than the 1st-10th most frequent wine varieties.

* Countries and Their Top Wine Varieties:

The overall data for the countries different wine varieties has been presented. Now the data for top producing countries and the top wine varieties produced in those countries will be presented.

The below plot identifies the top 6 countries with the highest frequency of wine and the top 5 most frequent wine varieties within those countries. From these plots we can see the large amount of wine being produced in the US as we saw in the first visualisation of the countries section. We can again see the high ranking in price and points of the Tinta de Toro which we can see is a common wine in Spain. What is interesting to note is that of the top 6 countries, only Italy produces more of the most expensive wine variety (Nebbiolo) and for the most part, the frequency of wine in Italy decreases with price.

A group of bar charts

Description automatically generated with medium confidence

* Descriptions:

It may be interesting to see the most common words used to describe the various wines. For this a Word Cloud is used.

A word cloud of words

Description automatically generated

Amongst the descriptions, the most abundant are various fruit. Some examples include black cherry, pear, berry, peach, strawberry and plum to name a view. Also amongst the descriptions are various sweet flavours such as chocolate, vanilla and licorice. Somewhat popular are caffeine flavours such as coffee, mocha and espresso, as well dark and rich flavours like tobacco, oak, smoke and leather. The dataset overall appears to be made up of a large variety of wines.

**THIS REPORT WAS WRITTEN BY : Kathleen Sellick**

