

**TASK**

**Exploratory Data Analysis on the Wine Dataset**

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

**Introduction**

The dataset to be discussed below includes data about wines and their corresponding attributes which uncleaned range from country, description, designation, points, price, province, variety and winery.

Through analysing this data and presenting this report it is hoped that persons of interest, such as wine consumers, gain important insights. The specific insights which are observed are: Does a wines price indicate its quality according to the points it has scored; Correlation between wine price and its points score; Wine variety by country; and the distribution of wine varieties.

**DATA CLEANING**

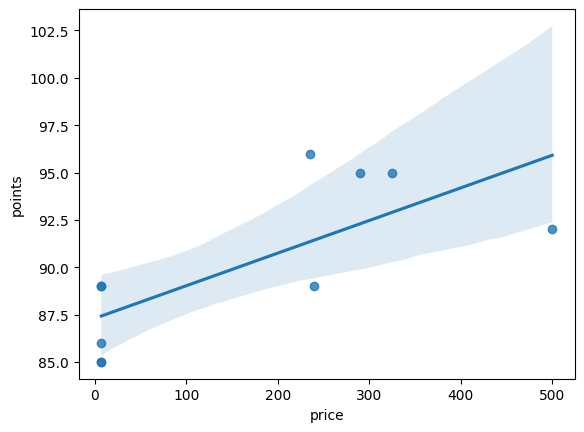
To clean the data, duplicate rows were removed and columns containing irrelevant and incomplete data, specifically columns region 1, region 2, description and designation were removed so they did not skew the data or make it harder to read.

**MISSING DATA**

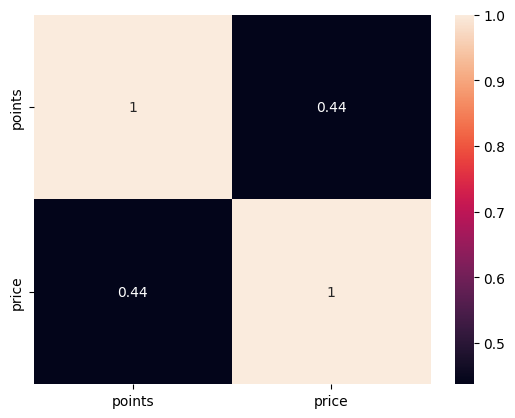
All the columns and rows with missing data were removed so that they would not skew the data analysis or make it harder to read.

**DATA STORIES AND VISUALISATIONS**

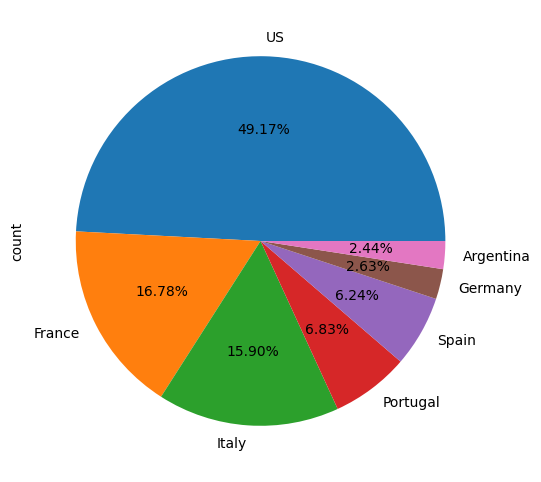
The first piece of data that was looked at is the points the wines were given as against their price. This is interesting because price should in theory correlate strongly towards the scores a wine receives. It appears to be the case that whilst there is a general correlation between price going up and the points going up there is a high degree of variance between each point. So you can by a wine that is very expensive but is barely better than a much cheaper wine. Price should therefore not be considered as a sufficient proxy for wine quality.



To investigate the thesis that price should not be considered as a sufficient proxy for quality a heatmap was generated and can be seen below. The map holds the correlation as 0.44 between 1 and -1 which means that there is a relationship, but it is a rather weak one. This further strengthens the thesis.



People interested in wine would also be interested to know where they should visit to get the widest varieties of wines. The pie chart below shows the variety of wine per country. An assumption that can be made is that those countries with the highest variety receive the highest demand for their wines as their economies are large enough and have sufficient capital to purchase the wines they want.



Wine resellers should be interested in which wine varieties have the greatest number of brands. The graph below indicates which variety has the most wine brands. We can again assume that the wine with the most variety is the wine that is in the most demand because that variety would not be produced if the demand for it was small.

. A bar chart with different colored bars

Description automatically generated with medium confidence

**Summary**

The outcomes of this analysis have led to the following insights: price is not a good proxy for determining wine quality; wine varieties are made mostly in the United States followed by Western European countries; and the wines of the top four varieties are Pinot Noir, Chardonnay, Cabernet Sauvignon, and Red Blend.

Further studies regarding wine should look at better proxies for wine quality than price by checking ratings against say sales or popularity. Another thing that should be investigated further is the which countries produce the wine with the best quality and which countries sell the most wine to other countries.

That would require more data collection than used in the scope of this study, but the insights generated already are valuable to businesses and consumers and further studies may derive even more valuable insights.

**THIS REPORT WAS WRITTEN BY: ANDREWN MARK MACDONALD HELLENS**