

**TASK**

**Exploratory Data Analysis on the wine Dataset**

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

**Introduction**

The data set is compiled of data about wine, this includes the price, points/ratings, variety of grape, province, country and winery. The data can be used to show which wines or areas have specific points or prices, varieties. This is useful for consumers, investors, companies and distributers.

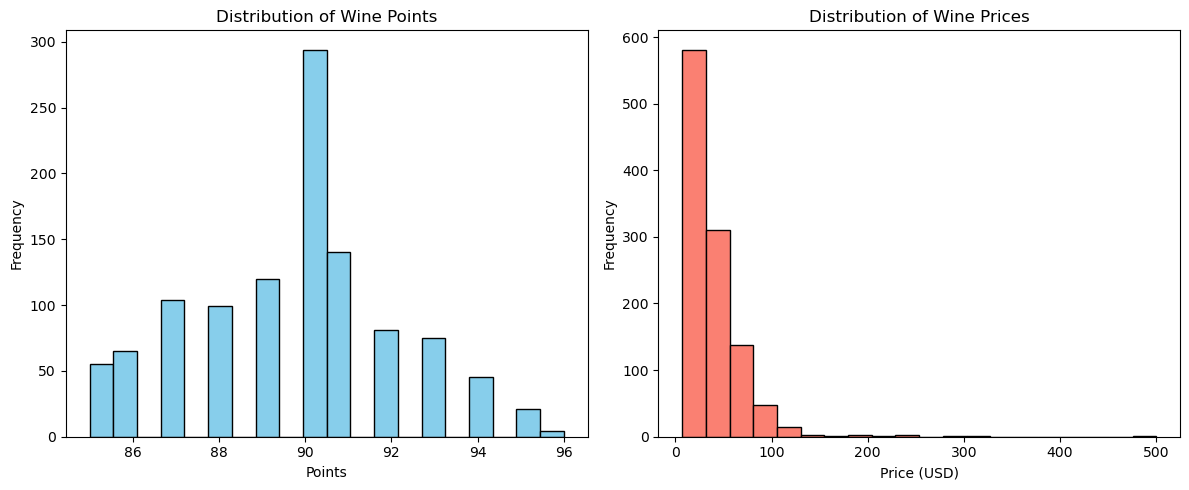
**DATA CLEANING**

* **Removing the duplicate rows from the Data Frame**
  + This helps to make the data more concise without duplicates getting in the way.
* **list of columns to remove**
  + I then printed a list of all column headings to see which ones I could use for processing and which ones I should drop.
  + I decided to drop ‘Unnamed: 0’, ‘description', 'region\_2'
  + I then checked the new dataset to see if the columns are dropped by using the ‘. head()’ function
* **Checking data types for categories**
  + I printed a list of each column and their corresponding datatype; this will show me which columns may need to be converted to numerical datatypes for calculations. All data was already fine

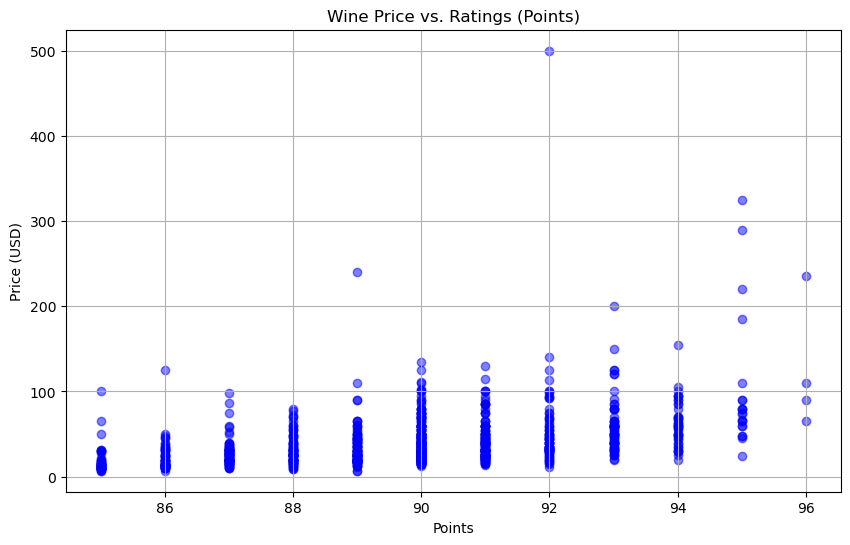
**MISSING DATA**

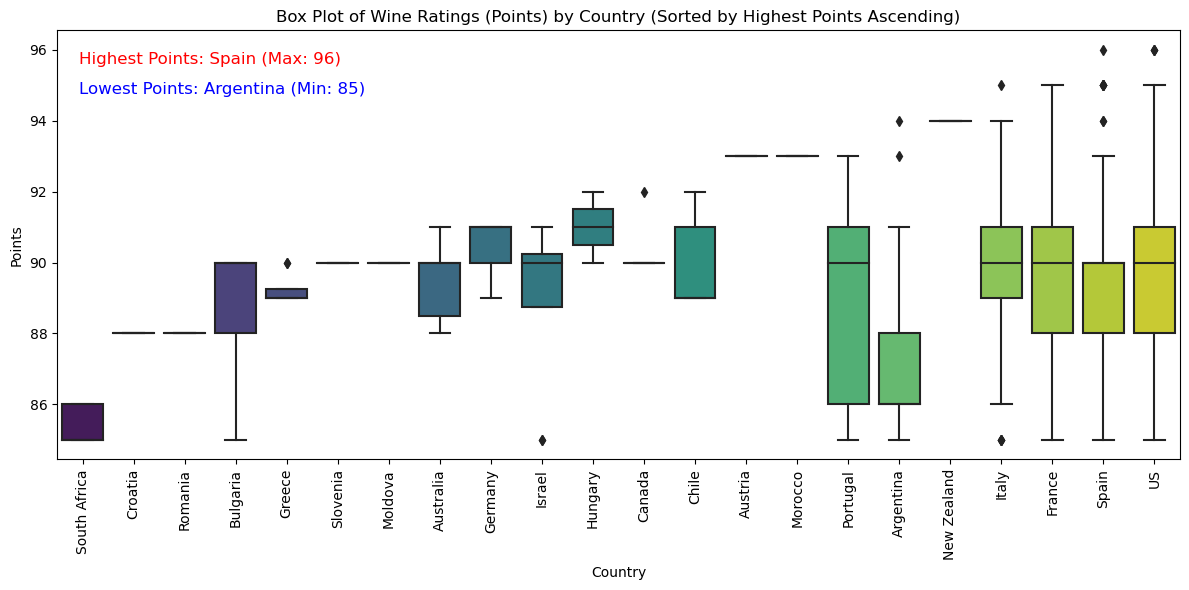
* **Checking ‘Nan’ values**
  + I printed a list of the total Nan values in each column, this way I can decide which columns/group data needs to be changed or dropped.
  + region\_2 had more than half total Nan values, so I dropped it,
  + because price had only 57, I used the median price
  + designation I thought I may need so I created a new group for them called ‘Unknown’ that way it I can still use some of the data for visuals
  + Lastly, I printed the Nan counts again to see if all is on 0

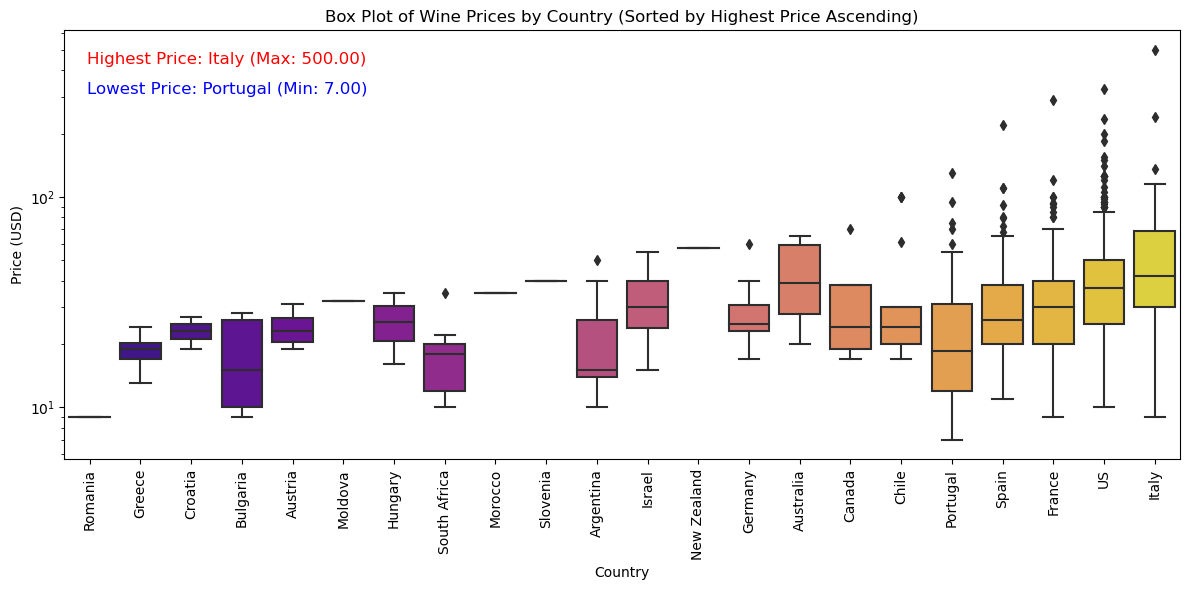
**DATA STORIES AND VISUALISATIONS**

****

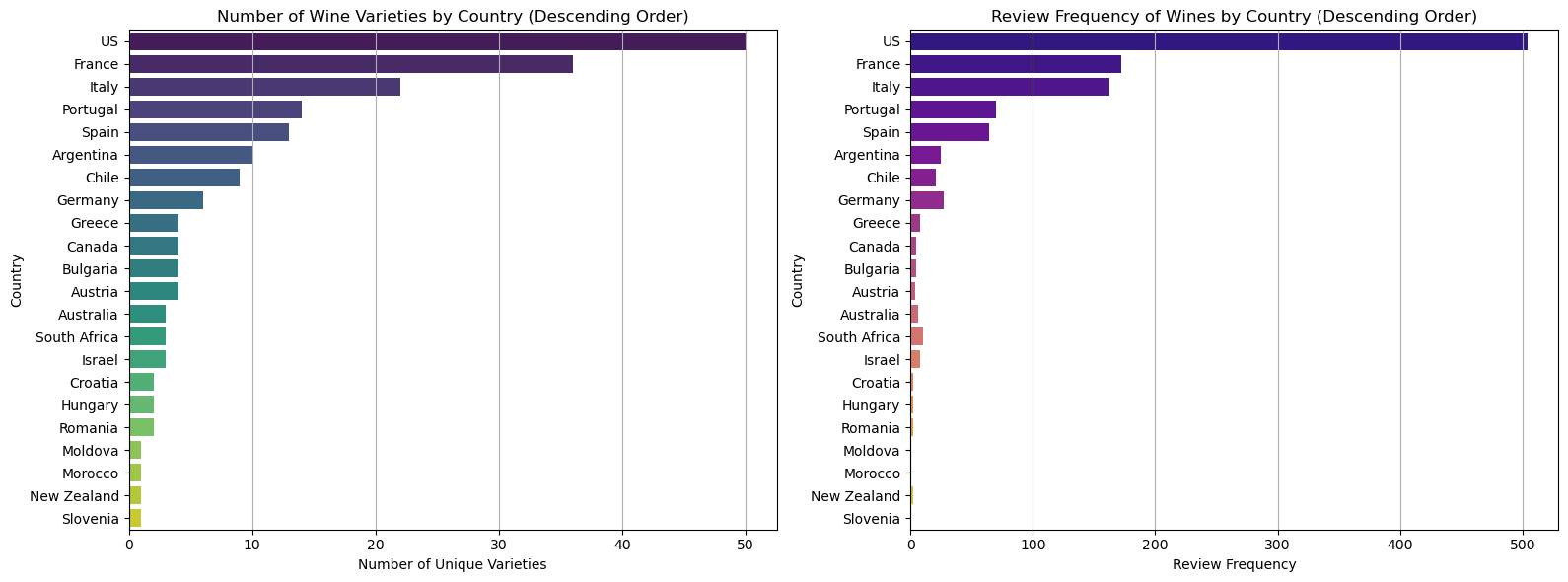
* **The left graph shows the distribution of wines according to what points they are.**
  + The graph shows that as the points/ ratings of the wine increases so does the distribution up until a high of 90 points which has the highest frequency then declines as the ratings go higher.
  + This can be useful to find the right market and aim at the points you would like to reach for the right distribution
* **The right graph shows the distribution of wines according to price**
  + The graph shows that as the price increases the distribution becomes less.
  + This can be useful to find the right market price for distributing your wines, if you want to sell a lot of wines at a lower price or fewer at a larger price.



* **The graph shows the price of wines to the points it is rated by**
  + The graph shows an increase in price the more points the wine has, most of the wines stay within 100 USD however there are a few higher point wines that go beyond 150 USD
  + This is useful to price your wine accordingly when gaining ratings from wines, a safe price for 86 - 88 points would be 10 - 80 USD or 90 - 92 points would be 10 - 140 USD
* **The graph shows a box plot of wine points by country**
  + The graph shows that Spain has the highest rated wines with a max of 96 Points
  + This is beneficial for knowing which countries make the best rated wine for consumers and distributers and their market.

****

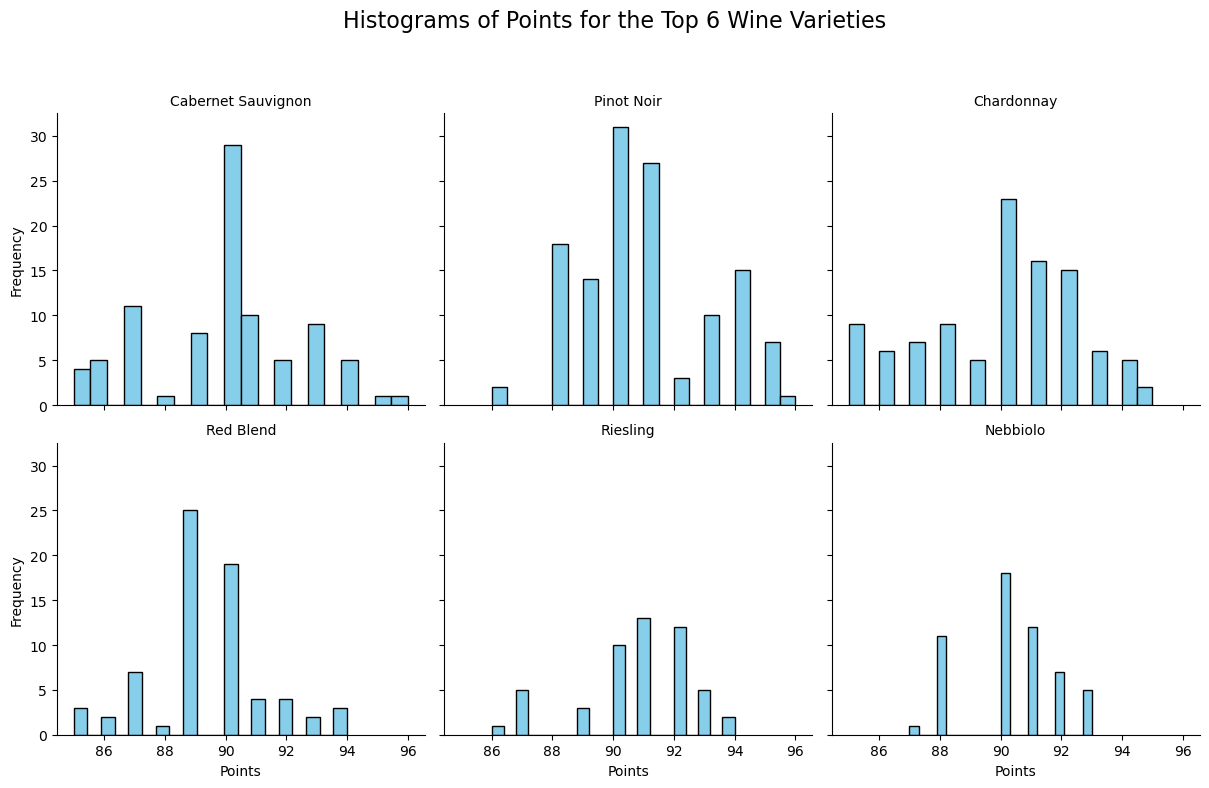
* **The graph shows a box plot of wine price by country**
  + The graph shows that Italy has the highest priced wines with a max of 500 USD
  + This is beneficial for knowing which countries make the highest priced wines



* **The graphs show a bar plot of the number of varieties in each country and the number of frequent wines in each country.**
  + The graphs shows that US has the most variety and also the most frequent wines, there is a clear indication that the more variety in a country the more frequency of wines.
  + This is beneficial for Wineries and wine distributors as they can use this data to make strategic decisions about which countries to target for exports or marketing campaigns and can help wineries assess their competition and make informed decisions about pricing, sales & marketing, and distribution.

#### 

* **The graphs show a bar plot of the highest top 10 average price wineries and the average rating of each as well**
  + The graphs shows that Gaja has the highest average price with a high average rating but not the highest
  + This is beneficial for investors or customers when buying and investing in wineries.
* **The graphs show a bar plot of the average wine price per country**
  + The graphs shows that New Zealand has the highest average price and Romania the lowest
  + Understanding regional variations in wine prices can help wineries better understand their target consumers.

****

* **The graphs show a multi-plot grid of the frequency of the top 6 varieties by points**
  + The graphs shows that there is more frequency in pinot Noir by 90 – 92 points and overall has a high frequency.
  + This is useful for investors and business owners to choose which varieties to make their wine from and which ones make the best rated wines.

**THIS REPORT WAS WRITTEN BY : RYAN COX**

