|  |  |
| --- | --- |
| Virgin Wines UK | Life's too short for boring wine | Virgin  Logo, icon  Description automatically generated  Icon  Description automatically generated  Decántalo  **Online wine market**  **COMPETITIVE ANALYSIS** | SMM750 Group Assignment  **Group 5**  Linh Nguyen (170037344)  Soumya Ogoti (220045527)  Wenxu Tian (210059418)  Aparna Viswanathan (220004767)  Fan Xia (200039922) |

Table of Contents

[Research scope 2](#_Toc118323740)

[1. The data sources and working process 3](#_Toc118323741)

[2. Variable description 4](#_Toc118323742)

[3. Summary statistics of all and each retailer 6](#_Toc118323743)

[4. Results and Analysis 8](#_Toc118323744)

[4.1 Price range in wine types within the 4 retailers (log price and listed price) 8](#_Toc118323745)

[4.2 Top 5 most reviewed countries of origin by retailer 10](#_Toc118323746)

[4.3 Top 5 most reviewed ABV categories by retailer 10](#_Toc118323747)

[4.4 Top 5 most reviewed wine age categories by retailer 11](#_Toc118323748)

[4.5 Top 5 most reviewed price categories by retailer 12](#_Toc118323749)

[4.6 Top 5 most reviewed wine types by retailer 12](#_Toc118323750)

[5. Web scraping challenges and solutions 14](#_Toc118323751)

# Research scope

During the pandemic period, global beverage market has been under a negative impact due to closure of restaurants, bars, and clubs, as well as a sharp decrease in tourism sector. However, the boost in e-commerce sales has partly offset the adverse trend in wine industry (Wittwer & Anderson, 2021).

To establish the preliminary understanding of the wine market, the team investigated online wine consumer purchasing behaviour. The reasons for online wine shopping can be categorised into three aspects: **cheaper price**, **detailed product description**, and **more options available** in online shops (Bonn et al., 2016; From The Vine, 2022). In this case, we determine that our analysis on wine market will be based on information surrounding **wine features** (including wine type, country, year, and ABV), **price** (for a bottle of 75cL), and **reviews** (including number of reviews and scores) which is going to be collected from competitors’ websites for analysis.

To be clearer, the key objectives are to: (1) identify the most popular wines sold across the companies and the key features of best-selling wines; (2) identify the most frequent price range for wines in order to learn product portfolio and business focus (e.g. how much should be invested on mass-market wines and niche-market wines respectively in our e-commerce shop). Therefore, web scraping is utilised to access data and BeautifulSoup and Selenium are applied as tools for scraping data from websites.

Given the fact that this online wine shop plans to target a broad market scope with various wine options, the team carefully selected four online retailers which provide wines in a variety of selections with detailed product descriptions which can be utilised for analysis. The four companies are **Laithwaites**, **Virgin Wines**, **Decantalo** and wine section page on **Morrisons** website. To be more specific, the team selected Morrisons wine as a good representative of mass-market wine retailer as the wines it targets consumers with a limited budget. What is more, Virgin Wines is also taken into account due to the fact that unlike Morrisons, Virgin Wine not only has cheap wines but also sells middle-end wines which are priced over £100. Laithwaites, which provides wines from £2 to £1,400 is also considered to be valuable in this case as the fine wines in its product offering will improve the analysis accuracy. Last but not least, apart from all three British brands mentioned above, other markets outside of the UK is also considered. A website called Decantalo which is based on another significant wine market, Spain, is analysed on the ground that it offers products to an international market.

## The data sources and working process

After identifying the research purpose and valuable variables that are important for our analysis, as mentioned above, these websites were selected as our data sources:

* <https://www.laithwaites.co.uk/wines>
* <https://www.virginwines.co.uk/>
* <https://www.decantalo.com/uk/en/wine/>
* <https://groceries.morrisons.com/browse/beer-wines-spirits-103120/wine-champagne-176432>

Those websites were selected not only because of their companies’ positions in the online wine retail market but also because their websites contain sufficient information that we require for our analysis. In this project, only web scraping is considered since we focus on the price ranges and product catalogues of the competitors.

Regarding our prioritization, first, we started with the web scraping. We selected the suitable websites and collected information from them to produce clean and usable datasets. The challenges and solutions during web scraping process will be discussed further in this report.

The next stage is to prepare the datasets usable for the analysis. The missing values were checked, and histograms were plotted to observe the values, unique value to detect if there is any incorrect or out of context information. The column type of price was change to the appropriate type which is from object to float64. For column country, the missing values were filled with “unknown”. For ABV values, since ABV takes values between 0 and 100m the rows which have value larger than 100 were dropped. We also checked the total number of missing values in ABV and found that it was not significant (<1%), so those rows were also dropped. For missing values in year, it is observed that the % was 5%, hence, we decided to keep these rows to use the other column values for further analysis of the data. This can be removed when fitting a model with year.

After this stage, we obtained datasets with variables mentioned in the next part. The details of our relevant features engineering are also discussed. The last step is visualizing our datasets to achieve meaningful insights for our business.

## Variable description

In this section, a table of variable description is presented to explain the variables in more details.

Table 1. Variable Description

|  |  |  |
| --- | --- | --- |
| Variable | Description | Unit |
| name | Name of the wine |  |
| country | Country of the wine where it was produced |  |
| country\_code | Country code of the wine where it was produced |  |
| wine\_type | Type of the wine, e.g. red wine, white wine, sparkling wine, rosé wine etc. |  |
| year | Year when the grapes for the wine were harvested | Years |
| price\_fixed | Listed price of the wine per bottle | Price in GBP (£) |
| logprice | Log price of the wine per bottle | Price in GBP (£) |
| scaled\_price | Scaled price of the wine to 75 cL | Price in GBP (£) |
| score | Rating score of the wine | Lowest (0) to highest (5) rating |
| num\_review | Number of reviews of the wine | No. of reviews |
| abv | Alcohol by volume, which measures alcohol content of wine | Percentage (%) |
| age | Age of the wine which is calculated by deducing *year* from current year (2022) | Years |
| size(cL) | Volume of the wine | cL |

In terms of ratings, for wines with less than 5 reviews, the ratings are not considered in our analysis as the number of reviews are not sufficient to justify. In this case, all the wines with less than 5 reviews have a rating score of 0. Furthermore, since the dataset for our analysis is quite large, log of price is taken to visualise the price distribution of different wines so as to address the skewness in price data. As the bottle sizes vary among wines, in order to compare price of wines on a fair basis, the team scales price of the wine to 75cL which is the most common size of wine bottles. Besides, the age of wines is calculated by deducting year when grapes was harvested from the current year (2022). For the wine type, apart from the main types such as red, white, sparkling and rose, there are some other wine types such as orange and sherry which account for a relatively small percentage. Therefore, other wine types except red, white, sparkling and rose are categorised as “others” for further analysis. Lastly, since there are a number of unique country values, a new column for country codes is added for cleaner visualisation.

## Summary statistics of all and each retailer

Table 2. General information of 4 retailers

Table

Description automatically generated

Table 3. Laithwaites

Table

Description automatically generated

Table 4. Morrisons

Table

Description automatically generated

Table 5. Virgin Wines

Table

Description automatically generated

Table 6. Decantalo

Table

Description automatically generated

## Results and Analysis

### 4.1 Price range in wine types within the 4 retailers (listed price and log price)

|  |  |
| --- | --- |
| Figure 1 Decantalo\_listed price | Figure 2 Morrisons\_listed price |
| Chart, box and whisker chart  Description automatically generated | Chart, box and whisker chart  Description automatically generated |
| Figure 3 Laithwaites\_listed price | Figure 4 Virgin Wines\_listed price |
| Chart, box and whisker chart  Description automatically generated | Chart, box and whisker chart  Description automatically generated |
| Figure 5 Decantalo\_logprice | Figure 6 Morrisons\_logprice |
|  | Chart, box and whisker chart  Description automatically generated |
| Figure 7 Laithwaites\_logprice | Figure 8 Virgin Wines\_logprice |
| Chart, box and whisker chart  Description automatically generated | Chart, box and whisker chart  Description automatically generated |

By plotting the relationship between price (listed price and logprice) and wine type across four companies, it can be seen that among all the wine types, sparking wine has a wide price range and it has a relatively high average price.

* Decantalo and Laithwaites: have wider price range, some products are very expensive (red for Decantalo and Sparkling for Laithwaites)
* Virgin Wines focus on middle price range, only Sparkling type has some expensive products. Morrisons: consistently stay in lower price range, only Sparkling type has some more expensive products but the highest prices are still lower than other websites.

### 4.2 Top 5 most reviewed countries of origin by retailer

A picture containing graphical user interface

Description automatically generated

As shown above, this bar chart shows the top 5 countries with wines being mostly reviewed on 4 retailers’ websites by percentage. The aim of plotting this chart is to identify which countries produce the most popular wine from the 4 retailers’ data. However, the results can be significantly biased by the company’s own product basis and only reflects the retailer’s choice in wine instead of consumers. Taking Decantalo as an example, it is observed that the majority of wines sold on this platform were produced in Spain.

### 4.3 Top 5 most reviewed ABV categories by retailer

Chart, bar chart

Description automatically generated

This graph is designed to investigate top 5 most reviewed ABV categories by retailer and it is clear to see that for most of the retailers, ABV with a range below 12.5% and between 13.5% to 14.5% is generally preferred by consumers as those wines get most reviews. Further investigation can be conducted by the marketing team to learn the market needs, for example, is wine with no more than 12.5% ABV more preferred by female consumers? After obtaining more information, detailed promotion strategy can be implemented to boost sales.

### 4.4 Top 5 most reviewed wine age categories by retailer

Timeline

Description automatically generated

Similar as what we presented above, this bar chart indicates the top 5 most reviewed wine age by retailer. However, the results appear to be diverse from company to company. In a real business case, wine age can be considered along with the wine type and country and region. For example, 2009 can be one of the best year for Bordeaux wine.

* Most of the wines are less than 2 years from this chart?

### 4.5 Top 5 most reviewed price categories by retailer

Chart, timeline, bar chart

Description automatically generated

The findings in this bar chart clearly demonstrate the market positioning of these four selected retailers, as Morrisons’ most popular wines are priced at the lowest among the four companies, followed by Virgin Wines and Laithwaites. While the price ranges at Decantalo are distributed on a more equal basis.

### 4.6 Top 5 most reviewed wine types by retailer

Chart, bar chart

Description automatically generated

From the chart shown above, it is undeniable that amongst most of the wine retailers, red wine is ranked to be the most popular type, followed by white wine and sparkling wine. Rosé wine the fourth favourable wine by majority.

Overall, it can be concluded that in our business scenario, while considering the product portfolio of the e-commerce wine shop, it is recommended to consider purchase the major types of wine such as red wine, white wine, sparkling wine and rosé wine to be the main product offering. Additionally, wines with ABV below 12.5% and between 13.5% and 14.5% is more preferrable by consumers. To maximise the profit, while considering the pricing strategy, the results of the most reviewed price categories can be considered along with the cost factors to set a reasonable price range for our products.

## Web scraping challenges and solutions

During the data collection process, there are four main problems we have encountered: IP blocking, slow and unstable load speed, different webpage structures, sudden crashes, and poor quality of data.

In terms of IP blocking, one example is that for some websites such as Waitrose Cellar and Ocado, they are highly secured which do not allow web scraping. Specifically, Waitrose Cellar puts many security layers to prevent web scraping despite different scraping tools utilized. Ocado only allows web scraping if there is an exchange for permission to scrape another website. This was identified at the initial research stage and as a result, Morrisons was chosen as an alternative.

During the scraping process, after the CSV files were saved successfully, a code review was conducted to ensure that our code would run efficiently, and the correct data had been collected. However, during this review process, there were two websites that had a sudden crash which are Virgin Wines and Laithwaites. The only solution in this situation is to either wait for the websites to operate normally or to choose another website, which both requires a lot of extra time and effort. In this project, the crash of Laithwaites lasted for few hours while it took more than a day for Virgin Wines. Hence, although there was a delay, the scraping code could still be reviewed and updated as planned.

To deal with the slow and unstable load speed, Python time sleep function is utilised so Selenium can wait while the web page is loading. It is important to choose a suitable time for Selenium to wait, otherwise, the website cannot load, and all information will not be collected. A typical example in this project is Decantalo’s website. A time sleep of 3 seconds had been utilized in this case so the function could be run smoothly.

Another challenge we faced is that different websites have different web page structures. For example, for Laithwaites website, *year* of wine is included at the end of the *name* of a wine (e.g. Cabalié Cuvée Vieilles Vignes 2021) which can only be found on the heading of the page, while for the other three websites, it is only available in product description. Therefore, the code was customized so for Laithwaites, information regarding *year* was subtracted from *name* of the wine in the heading. Another example is how information was collected from page to page of Morrisons’s website. We needed to use Python function to ask Selenium to scroll down and click “Show more” then all pages would appear. Meanwhile, for other websites, only the click of the “Next arrow” button was required.

Regarding the problem of poor data quality, there exists some formatting issue such as unnecessary spaces or special characters between words. The spaces and special characters were removed to obtain the exact value for data. In addition, it is important to get the right values and convert them into the correct data types so our datasets would be clean and usable for the analysis. Additionally, some information is missing on the websites. For instance, if a product on Laithwaites or Decantalo is sold out or out of stock then the price value is missing. To address this issue and to ensure the efficiency in the following cleaning process, “if not” statement is used when check the information under the same class to determine which item does not have corresponding information and “Not found” is typed to replace the missing value.References

Bonn, M.A., Kim, W.G., Kang, S. and Cho, M., 2016. Purchasing wine online: The effects of social influence, perceived usefulness, perceived ease of use, and wine involvement. *Journal of Hospitality Marketing & Management*, *25*(7), pp.841-869.

From The Vine, 2022. Shop Smart: Reasons To Buy Your Wine Online. [online] Available at: https://www.wtso.com/blog/shop-smart-reasons-to-buy-your-wine-online/ [Accessed 31 Oct. 2022].

Tanford, J.A., 2006. E-commerce in Wine. *JL Econ. & Pol'y*, *3*, p.275.

Wittwer, G. and Anderson, K., 2021. COVID-19 and global beverage markets: Implications for wine. *Journal of Wine Economics*, *16*(2), pp.117-130.