BAYES BUSINESS SCHOOL
CITY, UNIVERSITY OF LONDON

SMM750 Group Assignment

ONLINE WINE MARKET

COMPETITOR ANALYSIS

Group 5

Linh Nguyen

Soumya Ogoti

Wenxu Tian

Aparna Viswanathan

Fan Xia

Table of Contents

Research scope

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

To establish a preliminary understanding of the wine market, the analytics team investigated online wine consumer purchasing behaviour through the operations of current competitors in the market. The reasons for online wine shopping can be categorised into three aspects: **cheaper prices**, **detailed product descriptions**, and **more options available** in online shops (Bonn et al., 2016; From The Vine, 2022). In this case, we have determined that our analysis of the wine market would be based on information surrounding **wine features** (including wine type, country, year, and ABV), **price** (for a bottle of 75cL), and **reviews** (including the number of reviews and scores) which would 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 ranges 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, we have utilised the Python libraries BeautifulSoup and Selenium to collect data from the competitor 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 who provide wines in a variety of selections with detailed product descriptions which can be utilised for the analysis. The four companies are **Laithwaites**, **Virgin Wines**, **Decantalo** and the wine section page on **Morrisons** website. To be more specific, our team selected Morrisons wine as a good representative of mass-market wine retailers as they mainly target consumers with a limited budget. In addition, Virgin Wines is also considered because 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. Finally, apart from all three British brands mentioned above, other markets outside of the UK are also considered. A retailer named Decantalo which is based in another significant wine market, Spain is analysed on the ground that it offers products to an international market.

1. Data acquisition

After identifying the research purpose and important variables for our analysis, as mentioned above, the following 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 chosen 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. They also, to some extent, have a consistent structure for wine listings, which may facilitate scrapping. In this project, only web scraping is considered due to the concentration on the price ranges and product catalogues of the competitors.

Our main method is to use Selenium to access the web pages and collect the URLs to each product listing. Then, we used Python's "requests" library to get the content from the web page and extract relevant information.

Most of the data cleaning was also done at this stage since the context surrounding the data is already available which made debugging much easier. The challenges and solutions during the web scraping process will be discussed further in this report.

2. Data cleaning

In this part, we:

- Analysed the missing values to find out the underlying causes. We have found that values could be missing for different reasons, for example:
    - Prices: The product is out of stock
    - ABV: The product is a collection instead of a single bottle
    - Year: The product is non-vintage
    - Size: The size needed to be inferred, e.g., Magnum is short for 1.5L
    - Rating: The product did not receive any reviews
    - The information is simply not available on the product page

- Analysed the values' distributions by plotting histograms and the set of values available. This process identified extreme outliers (a >£26,000 bottle) and incorrect information (a product with an ABV of >1000%).
- Changed the price column type from object to float64.
- For the column country, we:
    - corrected and standardised the country names (e.g., ": Portugal" → "Portugal", "U.K" → "UK"
    - Inferred the country names: "Cahors" → "France", "Produce of the EU" → "EU"

    - Filled the missing values with “unknown”

- For ABV values, since ABV takes values between 0 and 100% the rows with values 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 from the analysis.
- For missing values in the year column, it is observed that the percentage was 5%, hence, we decided to keep these rows to use the other column values for further analysis of the data.
- Standardised wine types, e.g., “Red wine”, “Red”, “Red Wine”, etc., are replaced with “Red”.

After this stage, we obtained datasets with the variables mentioned in the next part. The details of our relevant features engineering are also discussed. The codes for the scraping and cleaning part were carefully reviewed and updated to be the most efficient before all datasets were merged. The last step is visualizing the key features of our datasets to achieve meaningful insights for our business.

3. Variable description

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

Table 1. Variable Description

Variable	Description	Unit
name	Name of the wine	
country	Country of the wine's origin	
country_code	Country code of the wine's origin	
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 was not sufficient to justify. In this case, all the wines with less than 5 reviews have a rating score of 0.

Furthermore, since there are extreme outliers in the wine prices, the log of prices is taken to visualise the price distribution of different wines to address the skewness in price data.

As the bottle sizes vary among wines, to compare prices of wines on a fair basis, the team scaled the prices 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 were 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 was added for cleaner visualisation.

4. Summary statistics of all and each retailer

Table 2. General information of 4 retailers

	size(cL)	price	num_review	rating	scaled_price	logprice	price_fixed	age	score
count	9562.000000	7893.000000	9204.000000	9204.000000	7893.000000	7893.000000	7893.000000	8816.000000	9562.000000
mean	75.138465	39.471793	47.390374	1.803781	39.607413	1.295246	39.471793	3.124433	0.736561
std	10.319106	315.391753	300.767534	2.185503	315.395653	0.354276	315.391753	2.365533	1.605461
min	12.500000	2.000000	0.000000	0.000000	2.587500	0.412880	2.000000	0.000000	0.000000
25%	75.000000	11.800000	0.000000	0.000000	11.730000	1.069298	11.800000	2.000000	0.000000
50%	75.000000	16.730000	0.000000	0.000000	16.690000	1.222456	16.730000	3.000000	0.000000
75%	75.000000	26.430000	2.250000	4.200000	26.510000	1.423410	26.430000	4.000000	0.000000
max	300.000000	26384.190000	7538.000000	5.000000	26384.190000	4.421344	26384.190000	67.000000	5.000000

Table 3. Laithwaites

	size(cL)	price	num_review	rating	scaled_price	logprice	price_fixed	age	score
count	1109.000000	864.000000	751.000000	751.000000	864.000000	864.000000	864.000000	1027.000000	1109.00000
mean	77.235798	18.647373	199.600533	3.994274	18.212548	1.195228	18.647373	2.296008	2.15257
std	24.459814	17.551781	592.869375	0.524780	15.484475	0.205678	17.551781	2.412608	1.99632
min	18.700000	2.790000	1.000000	1.000000	4.990000	0.698101	2.790000	0.000000	0.00000
25%	75.000000	11.990000	8.000000	3.700000	11.990000	1.078819	11.990000	1.000000	0.00000
50%	75.000000	13.990000	43.000000	4.000000	13.990000	1.145818	13.990000	2.000000	3.40000
75%	75.000000	18.990000	158.500000	4.300000	18.990000	1.278525	18.990000	2.000000	4.00000
max	300.000000	225.000000	7475.000000	5.000000	225.000000	2.352183	225.000000	27.000000	5.00000

Table 4. Morrisons

	size(cL)	price	num_review	rating	scaled_price	logprice	price_fixed	age	score
count	509.000000	509.000000	509.000000	509.000000	509.000000	509.000000	509.000000	309.000000	509.000000
mean	76.624754	9.227996	5.345776	3.075049	9.245197	0.912216	9.227996	3.249191	1.183497
std	22.210107	6.512815	14.100116	1.824167	6.421140	0.191624	6.512815	2.313015	1.880223
min	12.500000	2.000000	0.000000	0.000000	2.587500	0.412880	2.000000	0.000000	0.000000
25%	75.000000	6.000000	1.000000	2.000000	6.000000	0.778151	6.000000	2.000000	0.000000
50%	75.000000	7.750000	3.000000	3.800000	7.750000	0.889302	7.750000	3.000000	0.000000
75%	75.000000	10.000000	6.000000	4.500000	10.000000	1.000000	10.000000	4.000000	3.400000
max	225.000000	56.000000	208.000000	5.000000	56.000000	1.748188	56.000000	10.000000	5.000000

Table 5. Virgin Wines

	size(cL)	price	num_review	rating	scaled_price	logprice	price_fixed	age	score
count	701.000000	701.000000	701.000000	701.000000	701.000000	701.000000	701.000000	625.000000	701.000000
mean	74.768188	19.404708	392.948645	3.906847	19.737282	1.222381	19.404708	2.852800	3.542796
std	8.721368	17.084742	802.789750	1.048323	17.553324	0.219907	17.084742	2.082612	1.508185
min	37.500000	5.000000	0.000000	0.000000	5.000000	0.698970	5.000000	0.000000	0.000000
25%	75.000000	11.990000	15.000000	4.000000	11.990000	1.078819	11.990000	2.000000	3.900000
50%	75.000000	14.990000	60.000000	4.100000	14.990000	1.175802	14.990000	2.000000	4.100000
75%	75.000000	19.990000	292.000000	4.300000	19.990000	1.300813	19.990000	4.000000	4.300000
max	150.000000	255.000000	7538.000000	5.000000	255.000000	2.406540	255.000000	21.000000	5.000000

Table 6. Decantalo

	size(cL)	price	num_review	rating	scaled_price	logprice	price_fixed	age	score
count	7243.000000	5819.000000	7243.000000	7243.000000	5819.000000	5819.000000	5819.000000	6855.000000	7243.000000
mean	74.748723	47.626702	1.118735	1.283777	47.833653	1.352380	47.626702	3.267688	0.216747
std	2.469440	366.862011	4.530541	2.095685	366.871814	0.369769	366.862011	2.357828	0.985741
min	50.000000	6.650000	0.000000	0.000000	6.650000	0.822822	6.650000	0.000000	0.000000
25%	75.000000	12.780000	0.000000	0.000000	12.780000	1.106531	12.780000	2.000000	0.000000
50%	75.000000	18.530000	0.000000	0.000000	18.620000	1.269980	18.530000	3.000000	0.000000
75%	75.000000	30.310000	1.000000	4.000000	30.500000	1.484300	30.310000	4.000000	0.000000
max	75.000000	26384.190000	94.000000	5.000000	26384.190000	4.421344	26384.190000	67.000000	5.000000

5. Results and Analysis

5.1 Price range in wine types within the 4 retailers

Listed price	Log-transformed price

Figure 1 - Decantalo

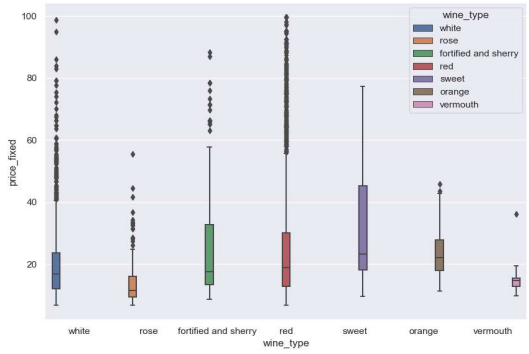


Figure 2 - Decantalo

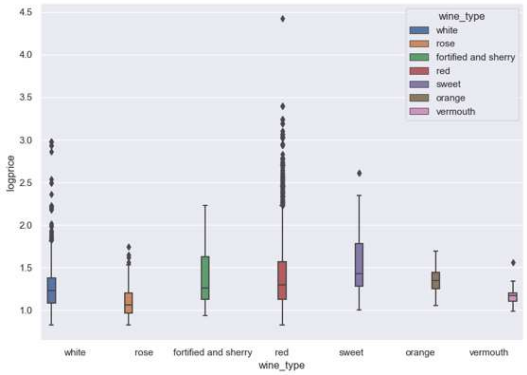


Figure 3 - Laithwaites

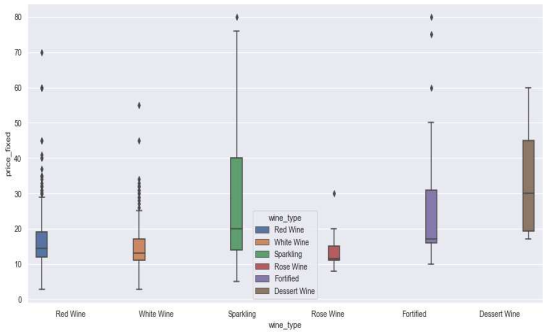


Figure 4 - Laithwaites

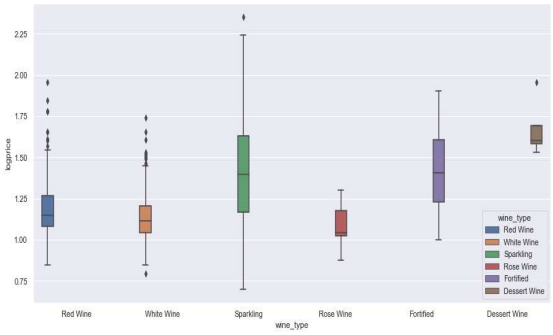


Figure 5 - Morrisons

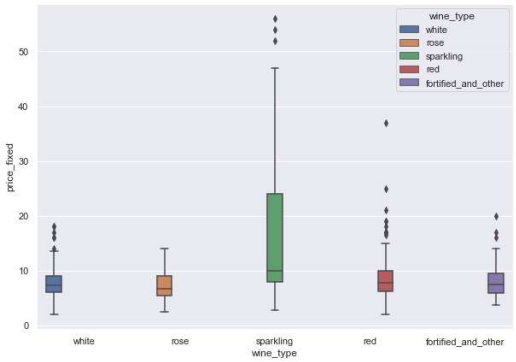
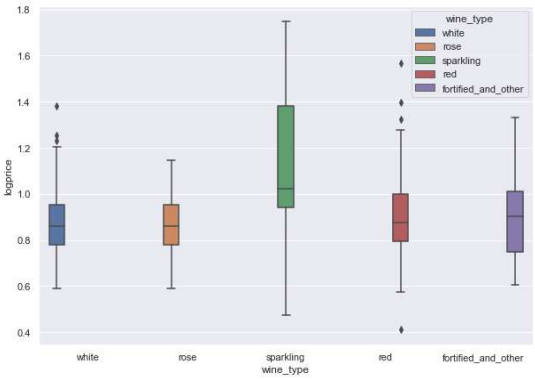


Figure 6 - Morrisons



Listed price	Log-transformed price

Figure 7 - Virgin Wines

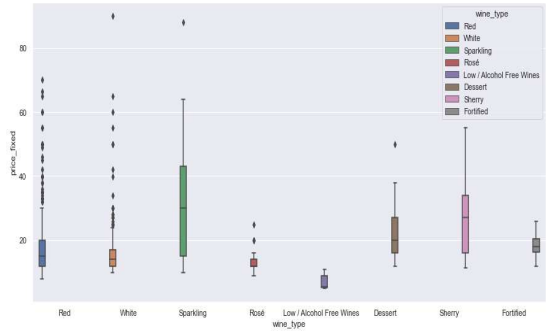
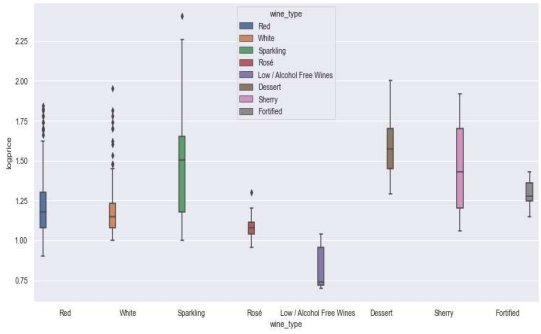


Figure 8 - Virgin Wines



The figures above show the listed and log-transformed prices by wine type of all retailers. In the visualisations for listed prices, only prices below £100 per bottle were included so the boxes can be visible. This is also the range that our company considers focusing on to target the public.

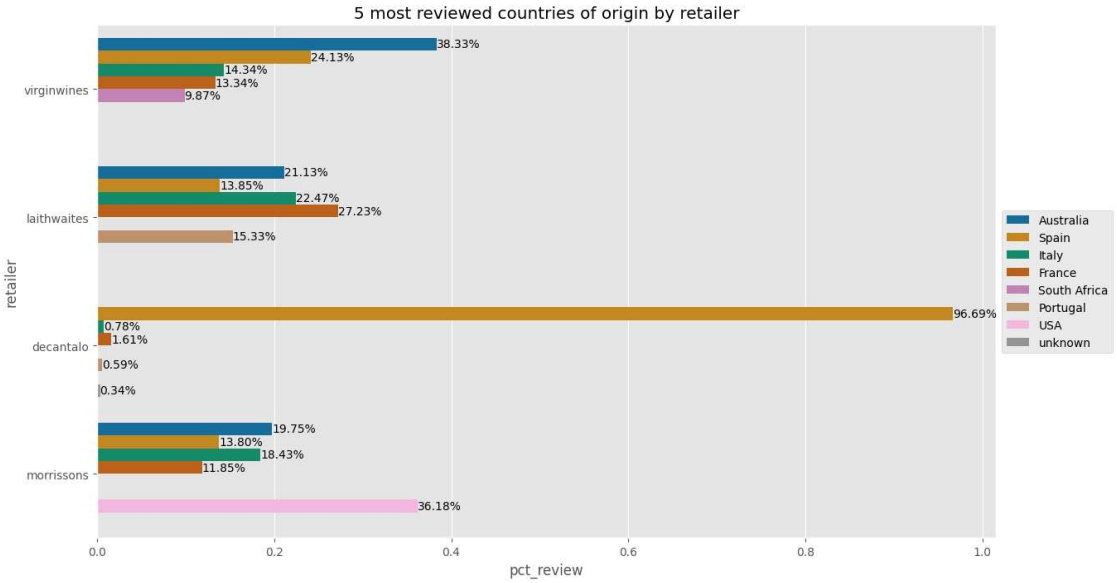
Among all the wine types, red wine has a wide price range in the case of Decantalo while it is sparkling wine in the case of the others. In particular, it is demonstrated in the log price plots that Decantalo has the widest price range with products' prices at an outstandingly higher level than others with the highest price per bottle being around £26,000.

Decantalo's average price per bottle is the highest being approximately £47. Virgin Wines and Laithwaites focuses on the middle price range with an average price of £19 and only sparkling type having some expensive products. Morrisons consistently stays in lower price range with the average price at £9, and only its sparkling type has some more expensive products, but the highest prices are still generally lower than in other retailers.

Therefore, we can conclude that although the price range in wine types varies from brand to brand, it is highly associated with the company's brand positioning and marketing target. The business scope and focus of the online wine shop should be carefully considered before determining the price range for each type of wine.

To identify the main product ranges for each retailer, it is natural to consider the volume sold or revenue of each product range. However, such information is not readily available. Therefore, we have decided to use the number of reviews as a proxy for a product's popularity. By this, we assume that the more reviews a product has, the more popular it is, and the more likely it is one of the main products of a retailer.

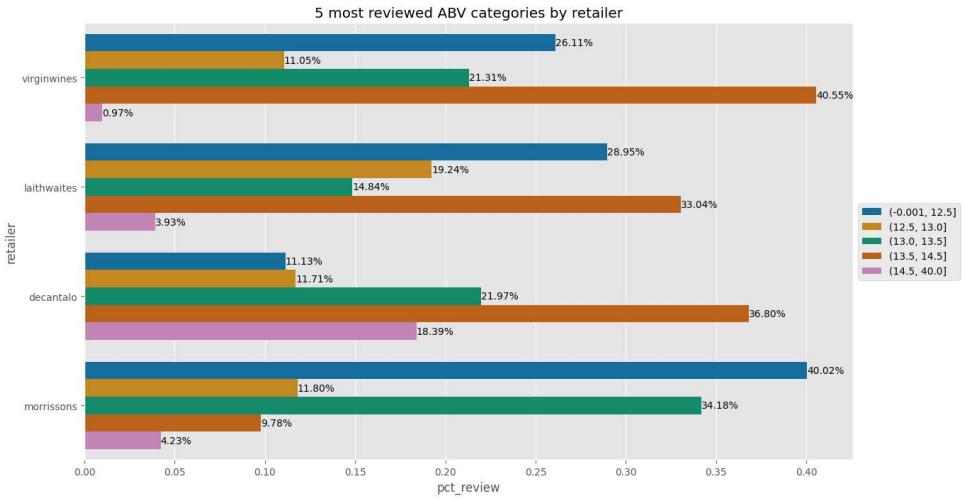
5.2 Top 5 most reviewed countries of origin by retailer



This bar chart shows the top 5 countries with wines being mostly reviewed on 4 retailers' websites by percentage. The chart suggests that Australia, Spain, Italy, and France are consistently the most popular wine origins. Interestingly, the USA is significantly more popular in the case of Morrisons compared to the other retailers.

Although the aim of plotting this chart is to identify which countries produce the most popular wine from the 4 retailers' data, the results can be significantly biased by the company's own product basis and only reflects the retailer's choice of 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. This makes sense because Decantalo is a Spanish retailer.

5.3 Top 5 most reviewed ABV categories by retailer

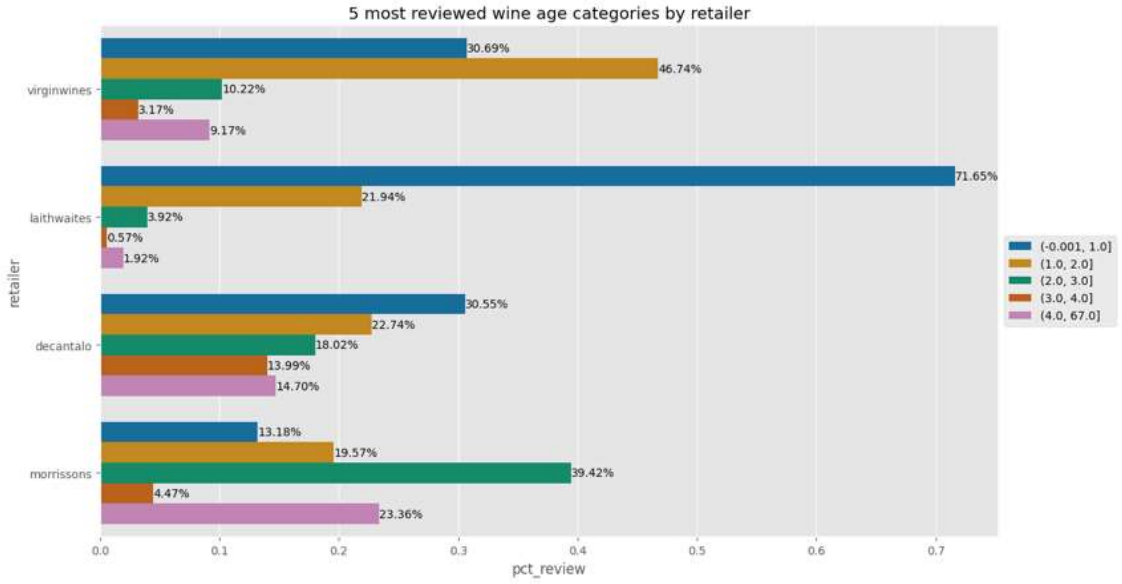


This graph shows the top 5 most reviewed ABV categories by retailer. The ABV values are binned into 10 deciles for ease of analysis.

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. However, the distribution is not the same across retailers. For example, Morrisons' most reviewed wine range has less than 12.5% ABV while the same range only ranked 5th in the case of Decantalo. It would be interesting to investigate further into this: Is it because of the difference in pricing, or country of origin?

Also, further investigation can be conducted by the marketing team to learn the market needs, for example, are wines with no more than 12.5% ABV more preferred by female consumers? After obtaining more information, a detailed promotion strategy can be implemented to boost sales.

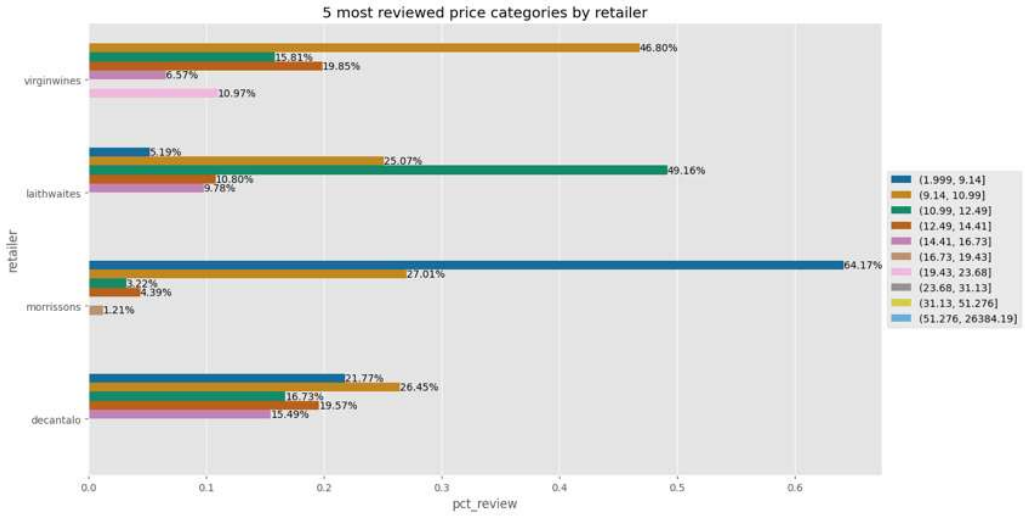
5.4 Top 5 most reviewed wine age categories by retailer



This bar chart indicates the top 5 most reviewed wine age by retailer. Although the age categories vary from retailer to retailer, most of the wines are under 2 years old. This can be explained by the fact that young wines can usually be non-expensive, which makes them more popular in the mass market. In this scenario, it can be implied that if our shop concentrates on mass-market wine sales, then young wines could be paid more attention.

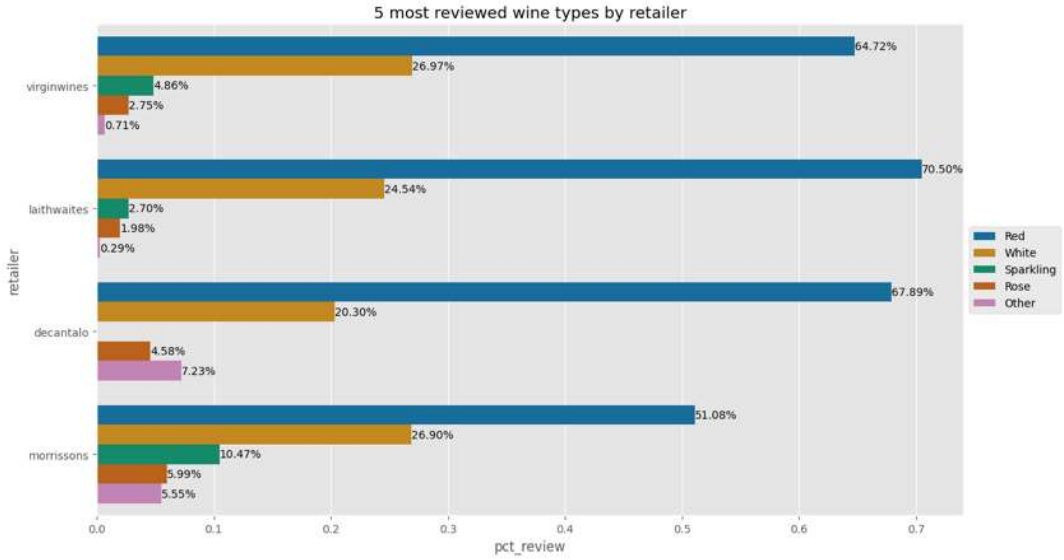
Interestingly, Morrisons' older wines received more reviews than the other categories, which is not true for the remaining 3 competitors. This might be explained by the fact that they generally price their wines lower than the other competitors, which makes their older wines more accessible to the mass market.

5.5 Top 5 most reviewed price categories by retailer



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 retailers, while Virgin Wines and Laithwaites focus more on the middle market. Finally, the price ranges at Decantalo are distributed on a more equal basis.

5.6 Top 5 most reviewed wine types by retailer



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

5.7. Conclusion

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 purchasing the major types of wine such as red wine, and white wine to be the main product offering.

Additionally, wines with ABV below 12.5% and between 13.5% and 14.5% are more popular. 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.

6. Web scraping challenges and solutions

During the data collection process, we encountered four main problems: Anti-scrapping, varying web page structures, sudden crashes, and poor data quality.

6.1. Anti-scrapping

There are some retailers who implemented various anti-scrapping methods on their websites. In our case, they were Waitrose Cellar and Ocado.

Waitrose can identify if requests were made from a scrapping engine such as Selenium and actively block them. We have attempted to use a macro tool to simulate clicks on the browser to get the URLs, but Waitrose also blocks GET requests to their product page listings. We have managed to get around this by using the headers sent to Waitrose from a manual browsing session. However, considering getting the URLs is an extremely slow and error-prone process and the wide range of products that Waitrose has, we have decided not to proceed with this retailer.

Like Waitrose, Ocado also blocks scrapping engines, although the method is not as sophisticated as Waitrose. In addition, they intentionally include special characters in the names of the class elements to prevent scrapping. Finally, they state clearly in their terms of use that web scraping is only allowed if there is a data exchange agreement.

This issue was identified at the initial research stage and as a result, Morrisons was chosen as an alternative.

6.2. Varying web page structures

Another challenge we faced is that different websites have different web page structures.

For Morrisons's website, the product listings are only loaded if the product containers are scrolled to in the browser. Therefore, we needed to use Selenium to maximize the window, scroll down, wait for the listings to appear, and click on "Show more" to show the next batch of products. Meanwhile, for other websites, only clicking the "Next arrow" button was required. In addition, Morrisons' product listing pages have poor consistency. For example, the country of origin sometimes has a dedicated section, but in other cases, it is specified in the product description with different spellings, e.g., "Product of X", "Wine of Italy", etc. In this case, we searched for those keywords from the bottom of the page instead from the top.

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 available in the product description. Therefore, for Laithwaites, information regarding *year* was subtracted from *name* of the wine in the heading. Also, the product attributes are organised into lists instead of classes, and the lists vary by product. As such, we need to text-match the list description to get the correct product attribute.

For Virgin Wines, the “Next” button is poorly defined which makes simply clicking on the button unreliable. However, we noticed that the page number is encoded in the URL. We have then decided to use a loop to iterate through the listing pages by modifying the page number in the URL.

For Decantalo, the number of reviews is queried with JavaScript from the webpage. As a result, it was not possible to get the information with the ‘requests’ library. We had to use Selenium to render the product page so that the reviews are displayed, which significantly slowed down the scrapping process. This, combined with the fact that Decantalo has 10-15 times more products on their page compared to the other retailers, and their website is slow, one scrapping run alone took more than 5 hours to complete.

6.3. Website crashes

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 that the correct data had been collected. Furthermore, the scraping process was not linear as it involves going back and forth to fix edge cases and debugging.

There were two websites that had sudden crashes which were 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 required significant extra time and effort. In this project, the crash of Laithwaites lasted for a few hours while it took more than 2 days for Virgin Wines. Hence, although there was a delay, the scraping code could still be reviewed and updated as planned.

6.4. Poor data quality

Some data quality issues were identified as follows:

- Different spelling of product information: For example, bottle sizes are specified in ‘mL’, ‘ml’, ‘cl.’, ‘Magnum’ (which is 1.5L), … For this issue, we had to handle the unit text and convert them to ‘cl’ and converted Magnum to 150cl.

- Wine type not consistent or not listed in the product page. As a result, we started a browsing session for each wine type, and relied on the product classification being correct.

- Incorrect product classification: There were instances of wine cases, and even candles being classified as wine bottles. We keyword-matched the product name to identify such products and excluded them from the analysis.

- Missing information: Where possible, we tried to look in other parts of the websites for the information (in the case of Morrisons). However, this is not always possible. For instance, if a product on Laithwaites or Decantalo is sold out or out of stock then the price value is missing. This method was not reliable

due to the unstructured nature of text data and hence does not guarantee the correct information being retrieved.

- Special characters: There were unnecessary spaces or special characters in the product attributes. We simply applied string processing methods to address them.

Dealing with the above issues required a significant number of trials and errors due to the sheer number of edge cases. Overall, we have found that saving the URL with the listing helped us quickly identify and analyse the anomalies, and we could avoid rerunning the scrapping from the beginning by setting up checkpoints during the scrapping run.

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.