Dataset on Wine Characteristics and Ratings - Vivino

Data documentation template adapted by Hannes Datta.
Originally based on Gebru, Morgenstern, Vecchione, Vaughan, Wallach, Daumeé, and Crawford. (2018). Datasheets for Datasets.

Created by: Group 10
Tara Gijsbers: 2080556
Claudia van Hoof: 2059425
Ashley Saarloos: 2151534
Iris Verzijl: 2127933
Fleur de Wolf: 2040813

Course: Online Data Collection & Management

Tilburg University

10-10-2024

1. Motivation

1.1 *What primary research question, business problem, or knowledge gap motivated the creation of this dataset ("data context")? How does the dataset offer insights into new phenomena, contribute to developing new models, or streamline gathering essential information? Why is this dataset valuable to the broader research community or industry stakeholders?*

Rating information is found to have a significant influence on consumers, which is demonstrated by their effect on market behaviors. Higher scoring wines experience an increase in popularity and demand (Babin & Bushardt, 2019). It is therefore important for wine manufacturers to understand the effect of price on wine ratings of these wine ratings. By investigating this relationship, it can contribute to the broader literature on consumer behavior and pricing strategies on online platforms. As the dataset consists of user-generated ratings, it provides insights into real-world consumer behavior. Additionally, this study can broaden knowledge about how price affects consumer perceptions in a competitive and diverse marketplace. Furthermore, understanding the determinants of wine ratings can offer insights into online review systems and their influence on consumer decision making.

To fully explore how price affects ratings, it is essential to work with enriched datasets that provide clear insights into this relationship. For this reason, the dataset created in this study has been scraped from Vivino, a global platform that features over 3 million wines which allows users to buy and rate wines. If price is found to be an important determinant of wine ratings, it can be used as a marketing tool to affect how the wine is perceived and rated by consumers. Moreover, this dataset includes additional wine attributes, such as the brand, the year of vinting, and the review count. This variety allows for a more comprehensive analysis of the factors influencing ratings. Therefore, using this dataset, prediction models can be developed to forecast how new wines with certain attributes will be rated, thereby providing valuable insights for both academics and for the industry.

1.2 *Please compare the various websites and APIs you assessed relevant to your data context. A table may help to compare data sources. Why did you choose your specific data source? Discuss the research fit, extraction method (e.g., web scraping vs. APIs), efficiency of resource use, and any other factors that made it emerge as the best choice. For tips, see challenges 1.1 and 1.2 in Boegershausen et al. 2022.*

There are several websites that include wines alongside their characteristics and ratings that were considered for the development of our dataset. Table 1 shows the websites and APIs that we have taken into consideration, with the observations of the website and the explanation of usability for our research question.

[1*] https://arxiv.org/abs/1803.09010

Name Website	**Observations**	***Explanation usability for project***
Wine Enthusiast https://www.wineenthusiast.com/	Wine Enthusiast is a website that contains articles about wines and lists of wines including the following characteristics: rating score, price, vintage, grape variety and region.	*The website only displays ranking lists including wines with scores higher than 80. Therefore, the sample of wines on this website is not representative of the entire wine market. Furthermore, the wines are blind-tasted by experts (therefore, they do not know the price when giving the rating), which eliminates price bias.*
Vivino https://www.vivino.com/NL/nl/	Vivino features a diverse selection of wines and displays several wine characteristics, including the price of the wine and the average star rating the wine received. The ratings are provided by consumers.	*This enables us to learn more about the relationship between rating and price since we can observe the ratings for different price levels.*
Winespectator https://www.winespectator.com/	Winespectator contains a database including more than 400,000 ratings and also provides several wine characteristics.	*The wines on this website are also blind-tasted by expert raters, which eliminates price bias. Furthermore, logging in is required to get access to all wine listings. Therefore, this website is not suitable for web scraping.*
Gall & Gall https://www.gall.nl/	Gall & Gall displays a wide variety of wines, their characteristics, and consumer ratings.	*However, the wines on this website have very few ratings, which limits the generalizability and distorts results of later analyses.*
Wine Market Data https://www.winemarketdata.com/	Wine Market Data offers APIs to track wine prices and extensive wine market data.	*Wine Market Data only tracks the UK market, thereby limiting its generalizability to other markets. Additionally, rating scores are not included in this data, which makes it unsuitable for this project. Another disadvantage is that these APIs are a paid service.*

Wine-Searcher API https://www.wine-searcher.com/trade/api?srsltid=AfmBOoqmlkhfLAnry_aDL8_Ztu1_pxQ49JcHiU9vNLvn6kx8VyfSlMgR	The API of Wine-Searcher offers information about wines and their characteristics, including prices and a rating.	*However, the ratings of the Wine-Searcher API are also expert-ratings instead of consumer ratings. This limits the usability for our project. Furthermore, access to the API is paid.*

Table 1: Considered Data Sources

After observing these websites and contemplating their usability for our project, we came to the conclusion that Vivino will give us the best insights to properly analyze the relationship between wine pricing and their ratings. Vivino offers a wide range of wines including various price classes and diverse ratings. Therefore, our dataset includes a representative sample of all wines in the wine industry. The ratings on Vivino are user-generated, and therefore reflect the experience of actual wine consumers in a natural setting. Furthermore, it ensures that price is included in the assessment of the wine, which would not be the case in blind-tasted expert ratings. This price bias is crucial for the research question of this project.

In order to find a comprehensive conclusion for our research question we opted to use web scraping instead of using an API. Due to the limited availability of suitable APIs for our project and taking into account cost considerations, web scraping was the best extraction method for this project. Furthermore, using web scraping will allow us to gain a more flexible and customizable access to the data needed for our research.

1.3 *Please identify potentially relevant contextual information (Boegershausen et al. 2022, challenge 1.3). Provide an overview here and save any relevant information in the digital submission of your project.*

Vivino.com is an online platform and mobile app that offers an extensive database of wines, consumer reviews, and consumer ratings. Along these reviews and ratings, it provides information about characteristics of wines and its current price. From the ‘wines’ page it is possible to filter on specific wine types, price, ratings, grapes, regions, countries, wine styles and food pairings. This site uses an infinity scroll, loading more wines as you reach the end of the page.

1.4 *Who created this dataset (e.g., which team). Mention you are students of the Marketing Analytics program at Tilburg University.*

This dataset was created by five students of team 10 of the course Online Data Collection & Management within the Master’s program Marketing Analytics at Tilburg University.

2. Data Extraction Plan

2.1 *Please describe your data extraction plan in such a way that another researcher or team could replicate your data collection process. Which information to extract from which pages, how to sample, at which frequency to extract the data, and how to process the data during the collection. In your description, explain how you tackled the various validity, legal/ethical, and technical challenges. See Boegershausen et al. (2022, challenges 2.1-2.4) for tips.*

First of all, to check whether it is legally and ethically possible to scrape data from Vivino.com, an inspection via robots.txt was conducted before selecting the website for data scraping. The robots.txt file, which outlines the site's preferences regarding automated access, disallows many areas of the site, but these are recommendations rather than legal restrictions. Therefore, it was determined that scraping the site is permissible under these guidelines.

To address our research question it is necessary to collect the prices and the average rating. To collect this information we opted to scrape the wine page. The wine page on Vivino provides a large assortment of different wines with different filtering and sorting options. However, Vivino provides a default filtering and sorting option where not all the wines are displayed. The wine page makes use of an infinity scroll, so there is no limit to iterating through pages or endpoints. As the URL changes when filtering and sorting is applied, these dynamic filter options can not be faced by selenium. Each filter or sorting option creates a unique url, and all these unique urls are therefore used for scraping the data. The required information is publicly available, so logging in is not needed. While personal and potentially sensitive information is displayed on the website, information like username or reviews will not be scraped to limit legal and ethical risks.

The platform may deploy algorithms that affect how data is presented to users as there are several sorting methods available on the website. To limit algorithmic biases on page, we opt for the most neutral sorting method, and sort on ‘Lowest price first’, which is an object criterion that is unlikely to be influenced by factors such as algorithms or commercial interests of Vivino. When changing the filter to “Lowest price first”, it can immediately be seen that a higher number of wines is displayed than when using other sorting methods, such as “Best Picks”, or “Popular”. Screenshots of all filters and sorting methods can be found in the additional documents.

While it would be possible to retrieve information on all wines on the Vivino website, it is technically infeasible to extract this data within the time frame of this project. Therefore, we decided to extract a sample of these wines through filtering, and base our research on the Spanish wines. While focusing on Spanish wines, we still explore a broad variety of wines by including all wine types, all prices, and all rating levels to ensure a wide range. This will leave us with a sample of 8,041 wines. This balances validity and technical feasibility by ensuring the sample is rich enough to perform analyses, but also technically feasible to extract. According to Fernández-Olmos & Florine (2023), Spanish wines are generalizable to other wines belonging to the Old World Wine Regions, including countries such as France and Italy. These regions are the originators of the winemaking industry, and within these regions

wines are highly differentiated. Therefore we can infer that our chosen sample will likely be a highly representative sample of the whole population of wines.

Every wine category on Vivino is included in the dataset: red, white, rosé, sparkling, dessert and fortified. The wine selection for the rosé, sparkling, dessert and fortified wines are limited and therefore, data on every wine presented on the website in those categories can be collected with no issues. The red and white wines have a broader selection. This larger selection could break the code. Section 3.2 addresses this potential issue, including our way of solving it.

From each wine on the page, we extracted the hyperlink pointing to that specific wine, the brand, the name of the wine, the average rating score, the number of reviews, and a timestamp to increase validity.

As mentioned in Section 1.3, Vivino is a dynamic website that is constantly changing. However, archival info regarding prices are not available on the Vivino website. As such, it is not possible to know what the price of a wine was at the time the rating was given. To account for this, it would be necessary to collect the data on price close to the date of each given rating over a longer period of time. However, given the scope and timespan of this project, and to limit the burden on the Vivino server, we extract the data only once. With the purpose of not overloading the server, we will make one request per two seconds, using a timer. Using this extraction frequency, it took approximately 5 hours to complete the data extraction.

Since we are working with smaller extraction batches to mitigate the chance of the code breaking, we opted to process and save the data only after each batch is extracted. The data will be stored in a CSV file. A CSV file is chosen over a JSON file due to the fact that the data collected is tabular data which can easily be stored in rows and columns. The data collected will be assigned to appropriate columns to ensure consistency and facilitate later analysis. As the data does not include any sensitive or personal information, ethical risks are limited. To further validate the scraped data, a sample will be cross-checked against the website

2.2 *Does the dataset contain data that might be considered confidential or sensitive (e.g., personal data such as usernames or IP addresses, data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, financial or health data, biometric or genetic data, social security numbers, passwords, etc.)? If so, please provide a description.*

The dataset does not contain any information that could be considered as confidential or sensitive information. Although the rating data is generated by individual users, only the average rating is collected, and no personal information about the users who provided the ratings is included. Since the data will be collected at one point in time, we do not work with high scraping frequencies that could possibly make the data more sensitive

2.3 *If the dataset relates to people, is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.*

As mentioned above, the dataset does not relate directly to individuals. Only the average rating of all individuals that have rated the wine will be investigated, not including any personal information of the individual raters. Therefore, it is not possible to identify individuals directly or indirectly from the dataset.

3. Data Extraction Process

3.1 *When was the data collected?*

The data was collected on the 5th of October, 2024.

3.2 *Please describe any technical challenges you encountered while scaling your data collection. How did you resolve them? Please provide a clear explanation of the debugging process (see Boegershausen et al. 2022, challenge 3.1).*

As described in Section 2.1, some pages we aimed to extract data from presented a large number of wines. Due to this, the web scraper broke down and the data collection stopped. We faced these difficulties for the red and white wine categories, which were the larger categories. To deal with this, we decided to capture the desired information in smaller extraction batches, thereby reducing the likelihood of technical issues and allowing us to collect all necessary data without overwhelming the web scraper. We extracted the red wine page in 4 different batches and the white wines in two smaller batcher to ensure that the web scraper did not break. To ensure that all the data from a category was stored in one data file, a for-loop was developed that looped through these different pages.

In order to overcome this issue we selected a sample of the red and white wines in multiple extraction batches. For red wines, we collected the data in four parts: wines in the price range of €0-€10, of €10-20, of €20-40 and of €40-€500+. For white wines, we will collect the data in two parts: wines in the price range of €0-€20, and of €20-€500+. We thus sampled directly from the source, and extracted data from 10 different pages. When all data of all the 10 pages is scraped, it leaves us with a sample of 8,079 wines as of 05-10-2024 (this number can slightly differ when re-extracting the data caused by updates on the site like listing and delisting wines, as seen in the screenshot in Section 1.3).

An important note is that, because filtering on the website is only possible in whole euros, the filtering of the red and white wines per price range will leave duplicates in the data collected. The wines that are exactly €10, €20 or €40 euro will be collected double for red wines, for white wines these are only the wines that are exactly €20. To resolve this issue, in the cleaning process of the collected data, we will remove these duplicate observations.

To visualize and understand the exact filters and pages used for this project, please see the additional document named ‘Vivino Filter Overview’,

3.3 *What measures or monitoring systems were in place to ensure and validate the quality of the extracted data? Can you describe how these monitoring systems functioned? (see Boegershausen et al. 2022, challenge 3.2).*

To ensure and validate the quality of the extracted data, some measures were implemented. First, in order to verify whether the raw data was correctly parsed, the raw data and parsed data were compared for a sample of the extracted information. This was done for each extraction batch to monitor the quality of the data. Additionally, a timestamp is added to the code. File sizes and the number of observations were also regularly checked to detect any abnormalities or missing data, which could indicate parsing errors and double or incomplete extraction.

3.4 *Can you specify the infrastructure you used for the deployment and execution of your data collection?*

The data was collected using a manually programmed web scraper, using both BeautifulSoup and Selenium in Python. Selenium is used to scroll through the infinity pages and make sure all wines are loaded completely so the data can be scraped. BeautifulSoup helps structure and parse the HTML content, making it easier to look for the classes and tags that contain the information we are interested in scraping. Additionally, loops are used to collect the data in small pieces and by using a for-loop we managed to put the small pieces together in one dataset. The data collection was deployed from a local computer.

4. Preprocessing, cleaning, labeling

4.1 *After collecting the data, did you perform any data processing? If yes, please provide specific examples and explain the reasoning behind each step.*

Several steps were taken to process the data after collection to improve interpretability and enhance later analysis. All these steps were performed using R and RStudio. First, we added a column including the category of each wine, meaning red wines would be assigned ‘Red’, white wines ‘White’, and so on for each of the categories (red, white, rose, sparkling, dessert, fortified). This ensures that, when the separate datasets are merged, the category of each wine could still be recalled.

After the categories were added, we merged the different datasets that were scraped of the different wine categories into one, all-encompassing dataset including all wine categories. This dataset contained the complete sample of wines scraped from the Vivino website, thus 8,079 wines.

As a follow-up, we extracted the unique wine ID and the year for the different wines that we scraped. The wine ID was retrieved from the hyperlink, as this identifier is not directly displayed on the website. Since this wine ID identifies the type of wine, but not the specific year the wine was harvested, this alone is not a unique identifier. In order to uniquely identify each wine, the year is also needed. Therefore, we also extracted the year from the hyperlink. The wine ID and the year of the wine uniquely identifies each wine on the Vivino website, so this combination can be used to trace back each wine. The

pages of each wine can be found back on the website using the following URL format: 'Vivino.com/NL/nl/{Brand-sperated-with-dash}/w/{WineID}?year={Year}'

For the variable price, the Dutch numeral notation was used. As such, a period was used as the thousands separator and a comma was used to initiate decimals. Therefore, to operationalize the variables and make sure the data is usable for analyses within R, the thousands separator is removed and the decimal separator is substituted for a period. For the same reasons of operationalization, the currency notation (€) is removed from the price column. Lastly, the price variable and rating variable are converted as numeric variables to make them suitable for analysis.

Furthermore, for some reasons that are still unclear for us, the Vivino website offers some wine multiple times on their website. That means, within one and the same page, a wine could be listed more than once. As previously stated, some wines also got extracted multiple times as a result of the subsampling in the red and white wine category (as discussed in Section 3.2). Therefore, it was necessary to remove all duplicate rows within the dataset, so only the unique rows would be retained, preventing possible biased calculations in further research using this dataset. We removed all duplicate rows from the dataset, leaving us with a final sample of 7,585 wines that is ready to use for analysis.

Finally, having made use of the timer function in Python, the timestamps were presented as the Unix time. To make the database more readable and understandable, we converted these timestamps into readable dates of extraction. After the timestamp was transformed to readable data, we transformed the numeric value to a date value.

Ultimately, we rearranged the variables in a logical order to create a dataset that is easy to understand and interpret for new users. The final data structure, including the variables and the order of these variables, can be found in Section 4.4.

4.2 *Were any measures implemented to ensure privacy, such as anonymizing user data? Please describe the methods used.*

Prior to the web scraping process, we thoroughly discussed and identified the specific data that we intend to scrape for this research project. During the coding process, our aim was to write the code in such a way that the final dataset would be excluded from any privacy-sensitive data. Meaning, any data that could form any legal or ethical issue is not included in the data set. Consequently, no additional privacy protection measures were necessary throughout the execution of our project.

4.3 *How did you address and clean out any implausible or erroneous observations in the dataset?*

During the pre-processing phase of the data, several erroneous observations were identified, including duplicate entries for certain wines. Although this issue was anticipated prior to the web scraping process, it was not feasible to eliminate duplicates during the scraping without risking the loss of some observations. To address this, duplicates were subsequently removed using the code that retained only unique values.

4.4 *Did you modify the data structure for long-term storage, like rearranging the dataset or renaming columns for clarity? If so, provide details on these changes and their rationale.*

During the scraping process, the database was already given a clear structure and the columns were named clearly. However, to ensure the database is suitable for long-term use, we made several adjustments to variables. For instance, the timestamp variable was originally in an unreadable format consisting of numbers. We converted this into a more user-friendly, clear date format.

Additionally, we encountered a limitation where the scraping code could not directly capture the wine ID. However, the hyperlink pointing to the individual wine did include the wine ID. These hyperlinks could be found in the HTML code of the pages we scraped from. Thus, we also scraped these hyperlinks and later isolated the wine ID.

We have organized the dataset in a way that makes it easy to read and navigate for users. The structure of the dataset is visible in Table 2.

Timestamp	hyperlink	WineID	Brand	Wine	Year	Category	Price	Rating	ReviewCount

Table 2: Overview Variables in Dataset

4.5 *What potential threats or biases could arise from your pre-processing steps? Please elaborate on any risks associated with the modifications made to the data and how they might impact the dataset's integrity or utility.*

One potential threat to the dataset's integrity arises from the fact that some of the wines of which we have extracted data were not directly sold on the Vivino site. Vivino provides hyperlinks to other websites where these specific wines are available and sold. Typically, these specific wines are advertised at the lowest price found online ("Available from x price"). However, the use of the lowest price can introduce bias, as it may disproportionately influence customer ratings, potentially skewing perceptions of the value and quality. This in turn affects the dataset's overall consistency and introduces a potential bias into any analysis based on pricing.

Additionally, relying on external websites for price data carries risks. Since Vivino has no direct control over the data on these other sites, there may be inconsistencies in pricing across different channels, which could compromise the accuracy and integrity of our pre-processed dataset. These inconsistencies could result in misleading conclusions about market conditions or customer behavior, reducing the overall utility of the dataset for further analysis.

Another significant threat stems from how Vivino assigns the same wine ID to all wines of the same type, regardless of the vintage year. This can lead to multiple entries sharing the same ID, even though they may have different prices based on the vintage. If not properly addressed during pre-processing, this could cause confusion for anyone analyzing the dataset, as they may inadvertently compare aggregate data from different vintages under the assumption that it refers to the same wine. This poses a risk to the integrity of the dataset, as vintage-specific variation in price and quality could be

overlooked, resulting in biased or inaccurate conclusions. Section 4.1 describes how we resolved this issue, by both extracting the wine ID and year from the hyperlink. It also explains how the URL should be formed for each unique wine by including the wine ID and year.

5. Data Inspection

5.1 *Please provide a variety of meaningful summary statistics and plots. For example, consider means/SDs for continuous variables, frequency distributions for categorical variables or – in the case of plots – bar charts, line plots, or histograms. This part of the documentation is intended to illustrate the richness of the collected data.*

The final dataset consists of 7,585 wines from 1,592 unique brands. Table 3 provides the summary statistics of rating, review count, and price. The average rating of these wines is 3.9, which is based on an average of 640.07 reviews per wine. The average price of a wine in the dataset is €36.08.

Summary Statistics for Continuous Variables

Mean_Rating	SD_Rating	Mean_ReviewCount	SD_ReviewCount	Mean_Price	SD_Price
3.9	0.31	640.07	2054.22	36.08	93.24

Table 3: Summary Statistics for Continuous Variables

Figure 1 and 2 show the distribution of harvest years of the wines and the distribution of the wine ratings respectively. The wine harvest years range from 1926 to 2023, showing a rich distribution of the data (see Figure 1). The wine ratings are approximately normally distributed (see Figure 2).

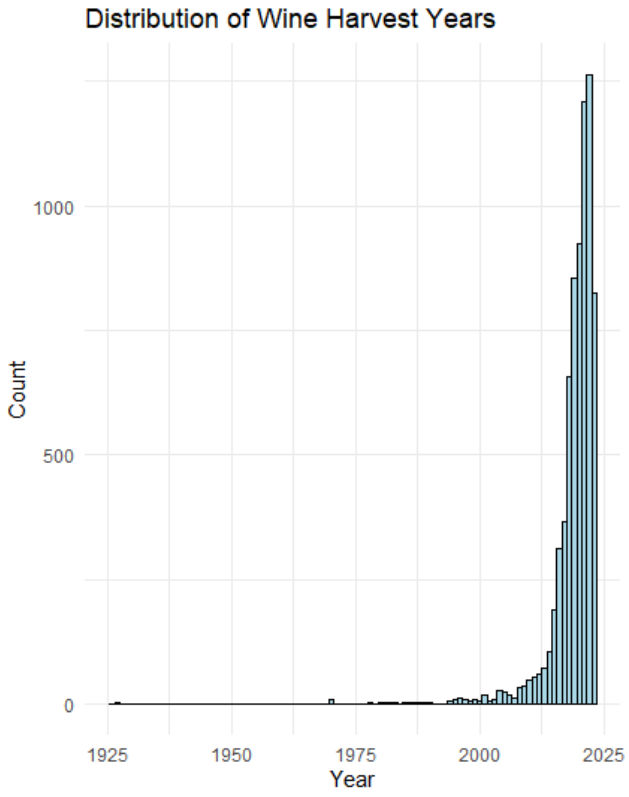

Figure 1: Distribution of Wine Harvest Years

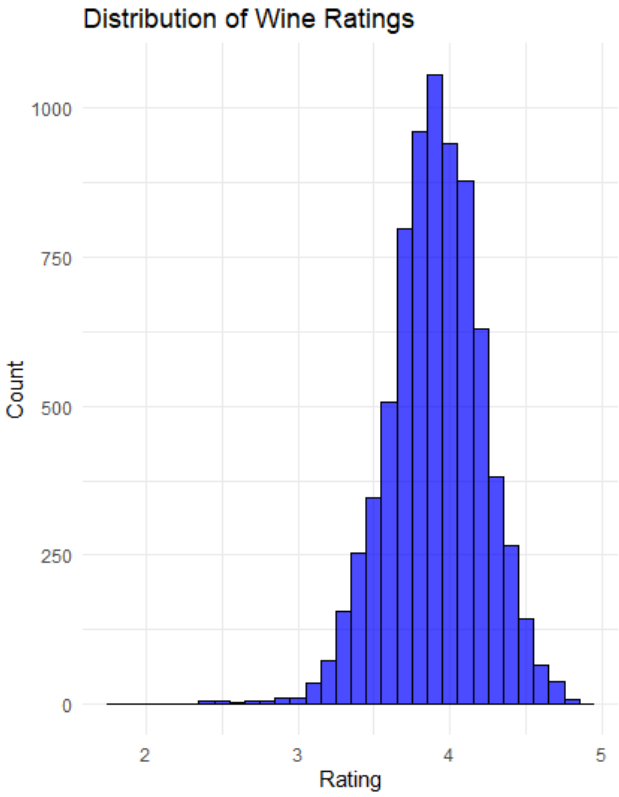

Figure 2: Distribution of Ratings

Table 4 and Figure 3 display the frequency distribution of the wine categories. Red wines are most prevalent in the dataset, accounting for 60.6% of the total wines, followed by white wines, accounting for 24.7% of all wines.

Frequency Distribution of Wine Categories

Category	Frequency	Relative Frequency
Red	4595	60.6%
White	1870	24.7%
Sparkling	522	6.9%
Rosé	328	4.3%
Fortified	217	2.9%
Dessert	53	0.7%

Table 4: Frequency Distribution of Wine Categories Table

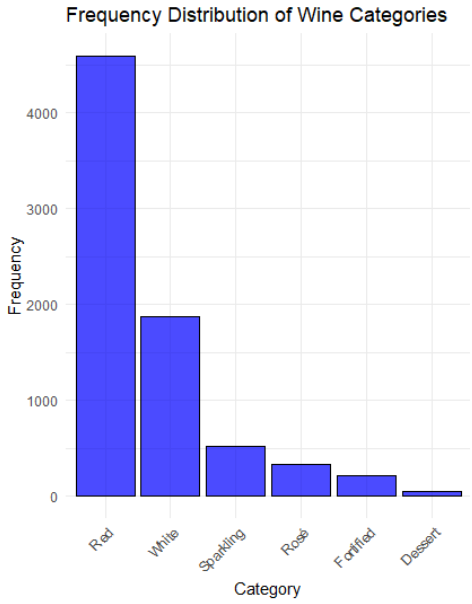

Figure 3: Frequency Distribution of Wine Categories Plot

Figure 4 displays the scatter plot of ratings and price. A logarithmic transformation was applied to the price variable to reduce the skewness and better highlight the distribution of lower prices, as these are more prevalent in the dataset.

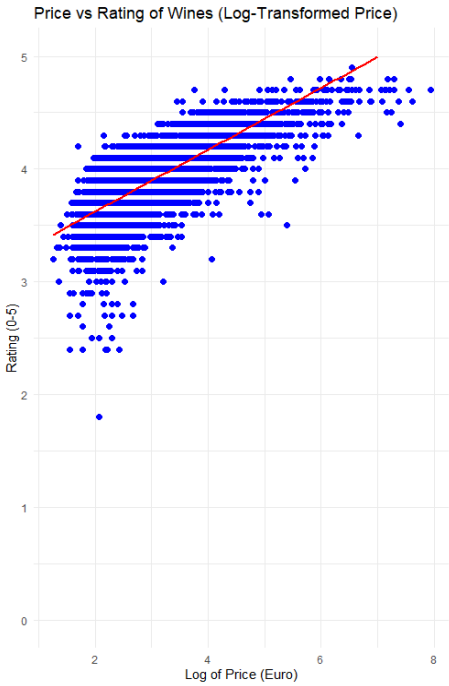

Figure 4: Scatter Plot of (Log-Transformed) Price vs Ratings

5.2 *Is any information missing from individual instances? If so, please describe why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information but might include, e.g., redacted text.*

The year column has 234 missing entries, which accounts for 3.09% of the total data entries. These values are missing as these wines are all non-vintage wines, which are wines that are produced from grapes harvested in multiple years. As such, there is no single year that represents the harvest year of the wine. There is no other information missing from the dataset.

6. Uses

6.1 *Has the dataset been used for any tasks already? If so, please provide a description.*

In the short period of time after concluding the dataset and writing this document the dataset has not been used for any tasks yet. However, we do not rule out employing the dataset in the near future for analysis.

6.2 *What (other) tasks / research projects could the dataset be used for? Provide a set of potential research questions or ideas for research projects.*

The dataset provides different opportunities for future research and analysis. One potential research opportunity could be to create a forecast tool on how new wines will be rated. By analyzing the patterns within this dataset a prediction can be made for the ratings based on variables such as price, year, and wine category.

Secondly, the dataset could potentially be used to study price differences across various categories of wines (e.g., red, white, rosé) or across different wine brands. Insights could provide wine producers and marketers to understand how pricing trends differ by wine category or brand and could guide them to better pricing strategies. The dataset could provide information that certain categories or brands can command a premium price due to quality perception, brand loyalty, or market demand.

A possible research project with this dataset could be to compare the number of reviews with the highest-rated products. Are the wines with the most reviews also the highest rated, or is there a discrepancy between popularity and quality? Studying the relationship between the number of reviews and the rating of wines can provide insights into consumer behavior and how popularity interacts with perceived quality. From a marketing perspective, if there is a large discrepancy between the most-reviewed wines and the best-rated wines, it may point to an opportunity for lesser-known, high-quality wines to gain more exposure. It could also indicate that certain wines achieve popularity due to factors unrelated to quality, such as effective marketing, brand recognition, or pricing strategies.

Furthermore, future research could provide a more in-depth analysis of the relationship between a wine's age and its rating. Researchers might investigate whether more vintage wines consistently receive higher consumer ratings, as well as examine how the perceived quality of wine develops or changes as it matures over time.

6.3 *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?*

The dataset's composition introduces limitations that should be taken into account for future uses. The ratings reflect the preferences and experiences of Vivino users. This could introduce biases based on the demographics of the users (e.g., predominantly wealthier, urban consumers) or their wine knowledge. Thereby, as the dataset lacks personal or demographic information of reviewers, no conclusions can be made about why certain wine categories, regions or brands receive biased ratings. Also, because the dataset lacks personal or demographic information of reviewers, it is unsure whether the consumer preferences are generalizable to the whole wine-drinking population.

Additionally, the data is collected at one point in time and therefore does not capture price changes or trends over time. This could lead to misleading conclusions which should be taken into account in future analyses.

6.4 *Are there tasks for which the dataset should not be used? If so, please provide a description.*

The dataset is not suitable for research that requires demographic data of reviewers, as the dataset only includes ratings, prices, brands, and wine categories, but lacks any individual information of the reviewers. Any insights into consumer behavior would be limited to aggregated review data without the ability to segment users by demographics. This can lead to biased or incomplete conclusions if the preferences of different demographic groups play a significant role in wine ratings or purchasing behaviors.

Additionally, the dataset is limited to Spanish wines listed on Vivino, meaning it only covers a subset of the overall wine market, which could vary significantly from other sources or platforms. This limitation makes the dataset unsuitable for comprehensive market analyses or competitive studies that require data from multiple platforms, wine producers, or regions. Since it does not represent the entire wine industry, studies aiming to compare market share, distribution, or pricing strategies across multiple platforms or countries may consider this dataset incomplete and potentially misleading.

Finally, as the date when each review or price was not recorded, this dataset is not suitable for studying trends over time, such as how wine ratings or prices evolve in response to factors like seasonal demand, marketing campaigns, or economic conditions.

References

Babin, B. J., & Bushardt, C. (2019). Third-party ratings and the US wine market. *International Journal of Wine Business Research*, *31*(2), 151–162. https://doi.org/10.1108/ijwbr-08-2017-0052

Fernández-Olmos, M., Ma, W., & Florine, P. (2023). Linking Spanish wine farmers to international markets: Is direct export better than indirect export in improving farm performance? *Economic Analysis And Policy*, *81*, 153–163. https://doi.org/10.1016/j.eap.2023.11.027