

SOC 2070 Final Report Proposal

1. Project Summary

1.1 Problem statement

According to the Global Wine Market Industry Overview, the global wine market is valued at \$333 billion in 2023 and is expected to grow at a compound annual growth rate of 5.2% through 2027. In such a fast-growing and competitive market environment, online consumer reviews have become an important factor in purchasing decisions. Consumer-generated reviews detail product features and experiences, and they also reflect emotional feedback. These reviews can reveal the strengths and weaknesses of a product, promote the brand, and reflect broader market trends and consumer preferences. However, most traditional wine review analysis is based primarily on numerical ratings. Although these ratings are relatively simple and intuitive to collect, they need to reveal the deeper semantic information in the review text, thus making it difficult to reflect the whole story behind consumer opinions.

1.2 Methods and Analysis

This study aims to overcome the limitations of traditional analytics methods by applying Natural Language Processing (NLP) techniques to delve into key themes in consumer sentiment and wine reviews. This study used data from the Wine Researcher platform, aggregating a large amount of high-quality wine consumer review data, including detailed information on review text, ratings, and prices. Firstly, the study normalized the raw reviews, including work splitting, deletion of stop words, word shape reduction, and cleaning of extraneous characters to ensure data consistency and analysis efficiency. Additionally, the analysis applied sentiment analysis tools to categorize consumer sentiment and capture sentiment polarity and semantic nuances in the reviews. To identify the concerns that consumers repeatedly mention in both positive and negative reviews, the study extracted the core themes in the reviews through theme modeling techniques. Moreover, the study combined statistical methods to explore the relationship between sentiment, ratings, and price. The quantitative analysis also revealed patterns of sentiment distribution across price points and categories.

1.3 Impacts and Contributions

The study contributes to advancing the application of Natural Language Processing techniques in consumer behavior research by integrating sentiment analysis and topic modeling to analyze wine reviews. This study provides wine producers with specific suggestions to help optimize product characteristics, pricing strategies, and market positioning. Also, the study reveals the key factors that drive consumer satisfaction to help brands better match their products with consumer preferences and market trends.

2. Motivation

2.1 Background

The global wine market has grown rapidly in recent years and to become a massive industry. According to the Global Wine Market Industry Overview, the global wine market was valued at \$333 billion in 2023 and is expected to grow at a compound annual growth rate of 5.2% through 2027 (Anderson and Pinilla, 2018). In such a competitive and diverse market, consumer opinion has become an important influence on brand success, especially in online reviews, where consumers describe their strengths, weaknesses, features, and feelings about a product through authentic feedback. These reviews have a positive impact on brand communication and reveal broader market trends and consumer preferences. However, the review data is under-explored by previous studies.

2.2 Problem Statement

Existing methods for analyzing wine reviews rely on numerical ratings (e.g., 1-5 star ratings or 100-point scoring systems). While these ratings are simple, intuitive, and easy to compare, they do not reveal specific points of satisfaction or dissatisfaction with a product. For example, a highly rated wine may be criticized for being overpriced, while consumers may love a lower-rated wine for its value for money (Cao, Duan, and Gan, 2011). The details and consumer sentiment behind numerical ratings are often hidden in the review text, and this semantic information must be effectively mined in traditional analyses.

In addition, consumers often refer to key characteristics of wines in their reviews, such as taste, price, and origin. However, the relationship between these attributes and consumer satisfaction has been explored less by existing analytical methods, limited to simple word frequency statistics or keyword extraction, and lacking systematic topic modeling and sentiment association analysis (Blei, Ng, and Jordan, 2003). At the same time, there are obvious differences in the concerns of different consumer groups. For example, price-sensitive consumers may be more concerned with value for money, while high-end consumers may be more concerned with the complexity of flavor or uniqueness of origin. Simple ratings cannot reflect these details.

3. Research Objectives

This study aims to explore the association between consumer sentiment and wine attributes (e.g., category, price, rating, etc.) by applying Natural Language Processing (NLP) techniques to deeply mine sentiment and thematic information in consumer reviews and ultimately to guide with practical applications for the wine industry. The specific objectives of this study include the following:

1. Reveal the distribution characteristics of consumers' emotions towards different wine categories and price ranges

Consumers show different affective tendencies towards different wine categories and different price ranges. To quantify such differences, this study will utilize high-quality review data from the Wine Researcher platform to categorize and analyze the sentiment polarity of the reviews and systematically quantify the sentiment distribution characteristics across categories and price ranges.

The analysis aims to reveal the different emotional responses of consumers to premium and budget wines. Also, this study can fill a research gap. Existing studies have primarily focused on simple statistical analysis of ratings and sales data, while quantitative analysis of the relationship between consumer sentiment distribution and wine attributes is still insufficient. This study will fill this gap by providing a new analytical perspective on the wine market.

2. Identifying key themes driving consumer satisfaction

Consumer satisfaction is often closely linked to their interest in specific product characteristics, such as flavor descriptions (e.g., "fruity" or "full-bodied"), value-for-money (e.g., "value for money" or "price"), or the quality of the wine. "or "overpriced"), and origin (e.g., "Napa Valley" or "Bordeaux").

This study will utilize topic modeling techniques to extract key themes that drive consumer sentiment from review texts, focusing on three main areas: flavor descriptions, price sensitivity, and origin preferences. By analyzing consumer preferences for taste, aroma, and aftertaste, the study aims to identify how specific flavor profiles influence sentiment. It will also explore price-related themes to understand consumer tendencies regarding value-for-money perceptions. Additionally, the study will examine consumer concerns about wine origin, distinguishing between traditional regions (e.g., France, Italy) and emerging regions (e.g., Chile, Australia). These findings will reveal the core factors shaping consumer sentiment and provide valuable insights for producers and brands to understand the drivers of consumer satisfaction better.

3. Quantitatively modeling the association between consumer sentiment and wine attributes

Sentiment, rating, and price are three key correlates of consumer reviews, but their relationships are still underexplored. This study aims to address this gap by analyzing the sentiment-rating relationship to determine whether positive sentiment is significantly associated with high ratings, and how neutral or negative sentiment is distributed among low-rated reviews. In addition, the sentiment-price relationship will be explored by examining the sentiment distribution of wines at different price points, such as whether high-end wines are more likely to elicit positive sentiments and whether low-priced wines are more strongly associated with perceptions of value for money.

4. Applying the model to the real world to provide insights to guide industry practice

The ultimate goal of this study is to apply the constructed sentiment analysis and quantitative model to provide specific guidance insights for wine producers, marketers, and consumers:

- (1) For producers: help optimize product characteristics and market positioning by analyzing the core drivers of consumer satisfaction.
- (2) For marketers: Utilize the sentiment-price-score model to develop precise pricing strategies and marketing promotions for different markets and consumer segments. For example, for high-end consumers, highlight high ratings and taste descriptions; for price-sensitive consumers, emphasize value-for-money and origin characteristics.
- (3) For consumers: Through quantitative analysis, we can help consumers better understand the characteristics of wines in different categories and price ranges to make more informed choices based on their personal preferences at the time of purchase.

4. Hypotheses

To achieve the research objectives, the study proposes the following hypotheses:

Hypothesis 1 (H1): Positive emotions are more likely to be associated with high ratings.

While consumers' ratings are usually a direct expression of their satisfaction, the emotions behind the ratings provide a deeper level of perception. This hypothesis suggests that positive emotions expressed in reviews are usually more likely to be associated with high ratings, while negative emotions are often seen in low-rated reviews.

Hypothesis 2 (H2): There is a significant difference in the distribution of emotions elicited by wines in different price ranges.

Consumers show significant differences in emotions towards wines in different price ranges. For example, high-priced wines may receive more positive emotions due to their quality but may also be questioned for their value for money due to their high price. In contrast, low-priced wines may receive more neutral or positive emotions. This hypothesis proposes a significant regular difference in sentiment distribution across price ranges.

Hypothesis 3 (H3): Themes in consumer reviews have a significant effect on sentiment polarity.

Themes frequently mentioned in consumer reviews, such as flavor descriptions of wines ("round," "tart"), price ratings ("value for money" or "too expensive"), and origin characteristics (e.g., 'Bordeaux,' 'Napa Valley') had a significant effect on affective expression. This hypothesis suggests that flavor-related reviews are more likely to be positive, while price-related reviews are more likely to be neutral or negative.

Hypothesis 4 (H4): Positive consumer sentiment towards high-priced wines may diminish with increasing price.

Although consumers usually favor high-priced wines for their superior quality, further price increases may weaken consumers' recognition of their value for money, leading to a marginal decrease in positive effect. This hypothesis proposes that consumers' positive sentiments towards high-priced wines may peak within a specific price range and diminish beyond this price range.

Hypothesis 5 (H5): There is a significant difference in the distribution of emotions triggered by different wine categories.

Significant differences in wine categories' use, cultural background, and flavor profiles may elicit different affective responses. This hypothesis suggests that red wines may be more likely to elicit positive emotions due to their complexity and cultural identity. In contrast, white wines and sparkling wines may exhibit more neutral or negative emotions due to their limited use.

5. Literature Review

The analysis of consumer reviews has become a growing interest in understanding consumer behavior in industries such as e-commerce, hospitality, and food and beverage. Previous research has relied heavily on numerical ratings or basic textual analysis, which often misses the true meaning and emotional content of consumer reviews (Cao, Duan, and Gan, 2011). However, traditional methods often fail to capture the nuances of consumer feedback, limiting the feasibility of findings and their applicability for organizations. Textual data can provide information on consumer emotional sentiment and specific product attributes that are key indicators of consumer satisfaction with a product.

Advances in Natural Language Processing (NLP) have dramatically changed the ways of analyzing textual data, especially when data scientists can go deeper with techniques such as sentiment analysis and topic modeling. Sentiment analysis tools (e.g., VADER and BERT) have performed well in classifying sentiment polarity in textual data (Hutto and Gilbert, 2014; Devlin et al., 2019), while topic modeling approaches such as LDA can extract recurring themes from large amounts of unstructured text (Blei, Ng, and Jordan 2003). These methods have been widely used to analyze consumer feedback in product reviews, hotel ratings, and food delivery services to gain deeper insights into customer preferences and decision-making processes (Hu and Liu, 2004). However, their application in the wine industry is still limited.

The wine industry has a significant comparative advantage and could benefit from these techniques due to the highly subjective nature of wine reviews. Existing research in the wine domain has focused on structured data, such as expert ratings, chemical composition, and sensory evaluations, while consumer-generated reviews have been less explored. However, consumer reviews are a rich source of information about how non-specialist audiences perceive a wine's flavor, aroma, value, and origin, among other things. While recent research has begun to apply NLP techniques to food and beverage research, few studies have explored the relationship between consumer sentiment, numerical ratings, and wine review pricing. Furthermore, combining sentiment analysis with topic modeling to identify key drivers of consumer satisfaction is still an underdeveloped area of research.

6. Data Overview

6.1 Data Sources and Collection

The dataset for this study is sourced from the Wine Searcher API, a comprehensive wine information platform that aggregates wine data from retailers and critics on a global scale. The dataset comprises over 150,000 wine reviews, including detailed information about each wine's characteristics, evaluation scores, and prices. The data collection process involved systematic API querying to gather information about wines across different regions, varieties, and price points. Each entry in the dataset represents a unique wine review and includes standardized scores, price information, geographical indicators, and detailed tasting notes.

6.2 Variables and Measurement

The dataset includes ten key variables: country of origin, detailed tasting descriptions, wine designation (classification level), numerical evaluation scores (80-100 point scale), retail price, geographical indicators (province and two levels of regional classification), grape variety, and winery. The scoring system (points) follows the industry-standard 100-point scale, providing a standardized measure of wine quality. Price data represents retail values in US dollars, offering a market-based assessment of each wine's positioning. The geographical hierarchy (country → province → region_1 → region_2) allows for multi-level spatial analysis of wine production and evaluation patterns.

6.3 Limitations and Considerations

While the Wine Searcher API provides a comprehensive dataset, several limitations should be noted. First, the data represents wines that are actively marketed and reviewed, potentially underrepresenting smaller producers or regions with limited market presence. Second, the scoring system reflecting subjective evaluations by wine critics. Third, price data may fluctuate based on market conditions and regional variations in wine distribution systems.

Despite these limitations, the dataset offers several advantages for this research. The large sample size ($n > 150,000$) provides robust statistical power for analysis. The standardized scoring system enables comparative analysis across regions and varieties. The inclusion of detailed tasting descriptions allows for both quantitative and qualitative analysis of wine evaluation language. Additionally, the hierarchical geographic data structure creates possibility for multi-scale spatial analysis of wine production and evaluation patterns.

The systematic organization of variables and comprehensive coverage of global wine production makes this dataset particularly suitable for analyzing relationships between objective characteristics (price, variety, origin) and subjective evaluations (scores, descriptions). The standardized nature of the data collection and evaluation processes provides a reliable foundation for investigating patterns in wine evaluation and market positioning.

6.4 Exploratory Data Analysis

Initial exploratory analysis reveals several key patterns in wine evaluation and pricing structures. Figure 1 presents a box plot distribution of wine scores across major wine-producing countries. This visualization highlights the variation in scoring patterns across regions, with traditional wine-producing countries such as France and Italy showing notably different score distributions compared to emerging wine regions. The plot includes median scores, quartile ranges, and outliers, providing insights into the evaluation standards across different wine-producing regions.

A closer examination of Figure 1 reveals distinct patterns in wine scoring across different countries. Traditional wine-producing nations like France, Germany, and Austria consistently show higher median scores (around 89-90 points) and larger interquartile ranges. This suggests both higher overall quality and greater variation in their wine productions. In contrast, emerging wine regions such as Romania and Greece display lower median scores (around 85-86 points) with more compressed distributions. Notably, countries like Australia and the United States occupy a middle ground, with median scores around 87-88 points, which reflects their established but relatively younger wine-making traditions. These patterns suggest that historical wine-making heritage may influence not only the quality but also the consistency of wine production across different regions.

Figure 2 provides an analysis of price-score relationships across ten major wine varieties, revealing distinct patterns in how different varietals are priced and scored. The visualization employs density-based plotting with trend lines, showing both the distribution of wines and the overall price-quality relationship for each variety. Cabernet Sauvignon and Pinot Noir demonstrate the steepest price increases with higher scores, suggesting that quality improvements in these varieties command significant price premiums. By contrast, varieties like Zinfandel and Riesling show more moderate price-score relationships.

The density patterns, indicated by color gradients from purple to yellow, reveal that most wines cluster in the 85-90 point range across all varieties, but the price dispersion varies considerably by varietal. Bordeaux-style Red Blends and Cabernet Sauvignon show particularly wide price distributions at higher scores, indicating greater price variability in their premium segments. This pattern suggests that factors beyond quality scores, such as brand reputation or regional prestige, may play a significant role in price determination for these varieties. Notably, white wine varieties like Riesling and Sauvignon Blanc show more concentrated pricing patterns, suggesting more standardized pricing in these categories.

These visualizations collectively provide evidence of complex relationships between geographic origin, price positioning, and quality evaluation in the wine market. They suggest that wine evaluation patterns are influenced by both objective characteristics (variety or origin) and market factors (reputation or pricing strategies), supporting the need for further detailed analysis of these relationships.

7. Method

This project will employ natural language processing techniques to analyze a corpus of wine review texts, with the goal of identifying keywords, sentiment patterns, and cultural preferences

that provide insights into how wines are perceived and described. The methodology is specifically designed to address the key research questions around what linguistic features and patterns characterize wine reviews, and how these vary by region, price point, and wine characteristics.

First, wine-specific stopword elimination will be performed to remove common generic terms like 'wine', 'drink', and 'bottle' that occur frequently but provide little informational value. Removing these terms allows the analysis to focus on the more unique and expressive vocabulary that captures the sensory properties and qualities of each wine. A wine-specific stopword list will be developed by analyzing word frequencies across the corpus.

Next, a color-coded wine keyword categorization system will be implemented to classify descriptive terms into major categories such as flavor (fruit, berry, vanilla), structure (acidity, tannins, dry), and character (ripe, fresh, rich). Organizing the keywords taxonomically and visually via color-coding enables identification of the dominant sensory themes, comparison of term usage frequency, and analysis of relationships between categories. This novel categorization scheme is an important element for uncovering the core linguistic patterns in wine descriptions.

A key goal of this keyword frequency analysis is to identify and categorize the most common descriptive terms used in wine reviews, revealing patterns in how wines are portrayed across different countries and price tiers. Understanding which wine attributes are most frequently highlighted provides insight into what characteristics consumers and reviewers focus on when evaluating wines. Comparing keyword usage between countries and price levels uncovers variations in consumer preferences and helps determine which factors are most influential in driving wine perceptions and purchasing decisions in each market segment.

To quantify sentiment, both a points-based wine quality rating and sentiment analysis of review text will be conducted. The scores provide a direct numeric measure of perceived quality, while sentiment analysis algorithms will determine the overall positive or negative language and emotional tone of the unstructured text. Integrating both sentiment mining approaches allows the project to assess alignment between the quantitative ratings and qualitative descriptions.

The sentiment analysis aims to evaluate the relationship between language use and wine ratings/prices, determining how specific descriptors correlate with wine quality scores and market positioning across different countries. Positive and negative sentiment keywords will be mapped against the corresponding rating and price data to measure the strength of association. This helps understand how the choice of evocative and emotional language reflects the reviewer's overall assessment of the wine's quality and value proposition. Comparing these sentiment-score relationships cross-culturally reveals differences in how countries assign value to wines and the role language plays in justifying that valuation.

We also hope to include cross-reference correlation of wine quality ratings, price segments, and regional variations to uncover relationships between these variables and the language used. Opinions of a wine's quality and value may be shaped by price expectations and regional reputation. Analyzing how scores and keywords differ by price and provenance offers insight into underlying biases and preferences.

Finally, the project integrates analysis of country-specific sentiment patterns and preferences. Expectations, perceptions and descriptions of wine are shaped by cultural context. Examining the frequency, clustering and correlation of terms within each major wine producing country can reveal distinctive national attitudes and values associated with wine. This cross-cultural comparative lens is key to understanding the diversity of wine language across the world.

Our comprehensive approach combining natural language processing, wine-specific keyword categorization, sentiment analysis, and cross-reference data correlation aims to thoroughly examine the textual and numeric data available in wine reviews. The keyword frequency analysis will determine the primary linguistic features and uncover consumer wine preferences. The sentiment analysis will evaluate how language use aligns with price and rating perceptions in different markets. By integrating these complementary techniques, the project aims to generate new empirical insights into the relationship between wine language, quality assessment, and consumer valuation across cultural contexts. The methodology leverages established text mining practices while introducing innovative domain-specific elements to yield a meaningful and novel contribution to understanding the language of wine appreciation.

8. Preliminary Results

8.1 Keywords Analysis

The keyword frequency analysis provides valuable insights into the linguistic patterns and prevalent terminology used in wine reviews. The results underscore the centrality of flavor descriptions, particularly fruit-related terms, in the language reviewers employ to characterize wines. Generic "fruit" is the most common keyword overall, appearing over 60,000 times in the corpus. More specific fruit descriptors like "cherry," "blackberry," "plum," and "apple" also occur with high frequency, suggesting that reviewers often identify particular fruits in a wine's flavor profile. The strong emphasis on fruit indicates that it is the predominant sensory aspect reviewers highlight in their tasting notes.

Beyond flavors, the analysis reveals frequent use of technical terms related to wine structure and body. Words associated with acidity, tannins, and dryness, such as "acidity," "tannins," "dry," and "crisp," comprise the second most prevalent category after fruit descriptors. The common presence of these terms points to a relatively professional and knowledgeable tenor in the reviews, as these words reflect industry-specific concepts central to the winemaking process and how the wine feels in the mouth. At the same time, abstract subjective descriptors like "ripe," "rich," and "good" also appear regularly, signaling that reviewers seek to convey overall quality impressions in tandem with concrete tasting details.

Notably, a mix of objective structural terms and subjective evaluative language emerges in the top keyword list, revealing a duality in how wines are discussed. Reviewers balance precise technical details with more impressionistic qualitative judgments, combining an analytical approach with an appreciation of the wine's hedonistic qualities. This marriage of the concrete and abstract is a distinctive feature of wine writing.

Color descriptors, especially "black" and "red," are consistently used in the reviews, though they are not as dominant as fruit or structural terms. These color keywords contribute to depicting the

wine's visual appearance, which is an important component in professional sensory evaluation. Aging-related terms have a more modest but still significant presence, with "oak" and "blend" appearing most frequently, reflecting the widespread use of oak barrels in winemaking and the common practice of blending multiple grape varieties.

The keyword frequency results also highlight the role of texture descriptors in constructing a multisensory profile of the wine. Terms like "soft" and "crisp" add a tactile dimension to the tasting notes, conveying how the wine feels on the palate. Though not as prevalent as the fruit or structural categories, texture keywords help paint a more vivid and complete picture of the tasting experience.

8.2 Sentiment Analysis

The sentiment analysis reveals a generally positive emotional tone in the majority of wine reviews. The frequent use of affirmative descriptors like "good," "ripe," "rich," and "fresh" indicates that most reviewers highlight the enjoyable qualities and desirable characteristics of the wines they assess. These words carry an inherently positive sentiment, and their prevalence confirms an overall favorable impression conveyed in the tasting notes.

However, the sentiment is nuanced and varied rather than uniformly positive. Wine reviews often involve detailed sensory descriptions that encompass a range of flavors, textures, and structural elements, some of which are depicted in neutral or even slightly negative terms. For example, words like "dry," "crisp," "acidic," or "tannic," while not necessarily hedonic, are more technical descriptors that don't always align with positive sentiment. The presence of these terms shows that reviewers provide a balanced and objective assessment of the wine's attributes, rather than just focusing on praising its qualities.

In terms of the relationship between sentiment and price, the analysis shows a positive correlation, with higher-priced wines receiving more favorable sentiment scores on average. This finding suggests that price often serves as a proxy for quality in the wine market, and reviewers may be influenced by price information when forming their evaluations. Additionally, pricier wines are often crafted with greater care and attention, using higher-quality grapes and production methods, which can contribute to a more positive sensory experience.

The sentiment analysis also reveals some differences in emotional tone across wine varieties and regions. For instance, reviews of bold red wines like Cabernet Sauvignon or Syrah exhibit a higher proportion of powerful, intense descriptors that convey a sense of richness and depth. In contrast, reviews of lighter white wines like Pinot Grigio or Sauvignon Blanc feature more delicate, refreshing language that evokes a brighter, crisper sentiment. Similarly, wines from certain renowned regions like Napa Valley or Bordeaux are associated with a more reverential or prestigious sentiment, reflecting their established reputation and perceived quality.

Examining the relationship between sentiment scores and wine ratings, the analysis finds a strong positive correlation. Wines with higher numerical ratings are characterized by more effusive praise and superlative descriptions in the reviews, translating to higher sentiment scores. This alignment between sentiment and rating suggests that the overall emotional tone of the review is a good indicator of the critic's quality assessment.

These sentiment analysis results provide valuable insights into the emotional dimensions of wine evaluation and appreciation. They highlight the predominance of positive sentiment in wine reviews, the nuanced and balanced nature of the language used, and the relationships between sentiment and factors like price, variety, region, and rating. The findings enhance our understanding of how critics communicate their sensory experiences and quality judgments, and how sentiment varies across different segments of the wine market. This information can be useful for wine producers, marketers, and consumers alike, as it sheds light on the affective aspects of wine discourse and decision-making.

9. Study Significance

This proposed study has the potential to make significant contributions to the field of wine marketing and consumer research. By applying advanced text analysis techniques to a large dataset of wine descriptions, this research aims to provide valuable insights into the prevalent keywords, sentiment patterns, and linguistic conventions that shape how wines are perceived and evaluated. The keyword frequency analysis will shed light on the central role of sensory descriptors in differentiating and characterizing wines, while the sentiment analysis will explore the emotional dimensions of wine appreciation. The findings from this study could have practical implications for wine producers, marketers, and retailers seeking to craft effective product descriptions, target specific consumer segments, and align their offerings with prevailing taste preferences and quality expectations. Moreover, the study's innovative methodology, combining natural language processing techniques with statistical analysis, has the potential to offer a valuable template for future research in this domain. Funding this project will support the development of data-driven approaches to understanding consumer perceptions and behavior in the wine market.

10. Conclusion

Through this proposed research, we anticipate generating a comprehensive and nuanced understanding of the language and sentiment employed in wine reviews. The keyword frequency analysis will likely reveal the dominant sensory characteristics, such as flavors, aromas, and structural components, that reviewers prioritize when assessing wine quality. The sentiment analysis is expected to show that wine reviews generally convey a positive emotional tone, while also highlighting the nuanced interplay between objective and subjective terminology used to provide balanced assessments.

We also expect to uncover significant relationships between sentiment and key variables such as price, rating, and wine variety. These findings will offer actionable insights for wine industry professionals seeking to optimize their product offerings, marketing strategies, and consumer communication. By understanding the language and sentiment that resonates with different consumer segments and aligns with quality expectations, wine producers and marketers can more effectively position their brands and craft compelling narratives around their wines. The outcomes of this study have the potential to inform and enhance decision-making processes across the wine value chain, from production to branding and retail.

In conclusion, this proposed research project aims to leverage advanced text analysis techniques to provide a data-driven understanding of the language and sentiment in wine reviews. The expected findings will contribute to the growing body of knowledge on consumer perceptions and preferences in the wine market, while also offering practical implications for industry stakeholders. By funding this study, you will support innovative research that bridges the gap between academic insights and real-world applications, ultimately fostering a more informed and competitive wine sector.

Bibliography

Anderson, Kym, and Vicente Pinilla. *Wine's Evolving Globalization: A New Comparative History*. Princeton: Princeton University Press, 2018.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3 (2003): 993–1022.

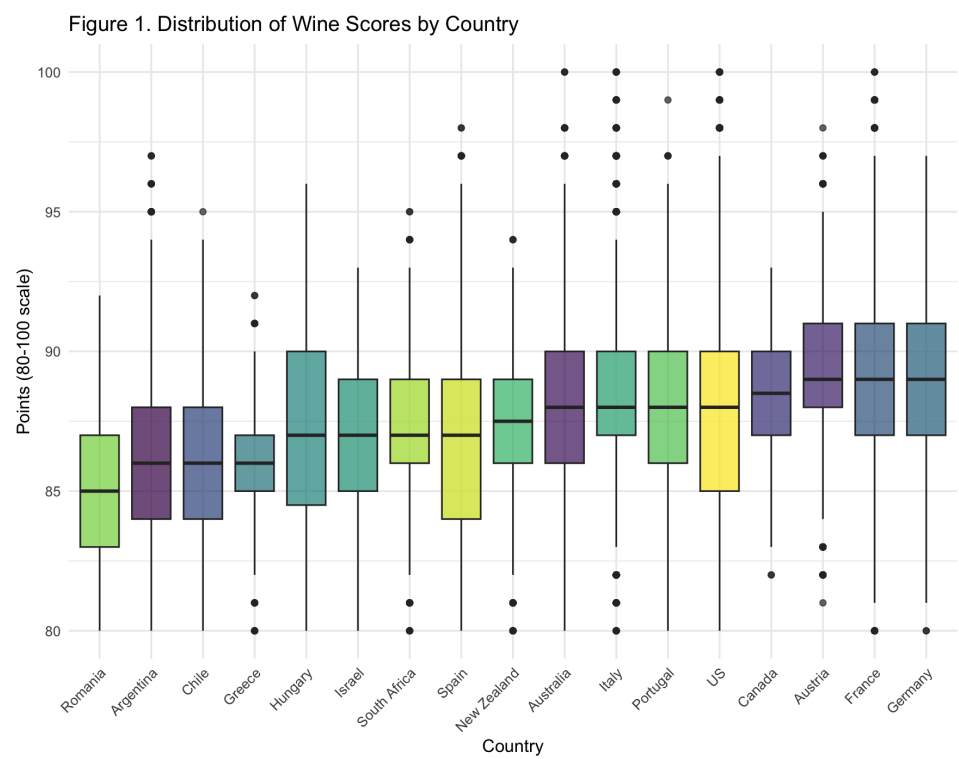
Cao, Qiang, Wenjing Duan, and Qi Gan. "Exploring Determinants of Voting for the 'Helpfulness' of Online User Reviews: A Text Mining Approach." *Decision Support Systems* 50, no. 2 (2011): 511–521.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171–4186. Minneapolis: Association for Computational Linguistics, 2019.

Hu, Minqing, and Bing Liu. "Mining and Summarizing Customer Reviews." In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 168–177. New York: ACM, 2004.

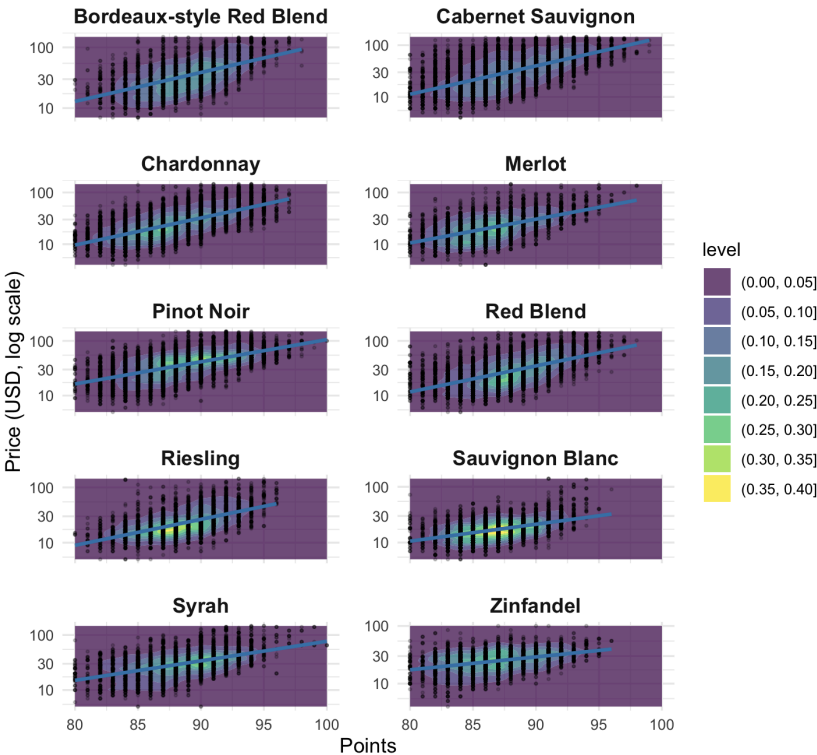
Hutto, C.J., and Eric Gilbert. "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text." In *Proceedings of the 8th International Conference on Weblogs and Social Media*, 216–225. Ann Arbor: AAAI, 2014.

Graphs and Tables



[Figure 1: Distribution of Wine Scores by Country] A box plot visualization comparing wine score distributions across major wine-producing countries, highlighting potential regional quality variations and evaluation patterns.

Figure 2. Wine Scores vs. Price Relationships by Variety



[Figure 2: Price vs. Points Scatter Plot with Variety Overlay] A scatter plot examining the relationship between wine scores and prices, with different grape varieties represented by distinct colors, revealing potential patterns in how different varieties are evaluated and priced.

Table: Table 1. Sentiment Analysis Summary

Sentiment	Distribution	Average.Price	Average.Points
Positive	127617	33.97	88.23
Neutral	13202	28.33	86.46
Negative	10111	29.09	85.47

[Table 1. Sentiment Analysis Summary] displays the overall results from our sentiment analysis in wine reviews.

